# Population genomics of weak evolutionary forces: Base composition evolution in coding and non-coding gene regions in *Drosophila*

Yamashita, Haruka

Doctor of Philosophy

Department of Genetics
School of Life Science
The Graduate University for Advanced Studies, SOKENDAI
September 2022

# ACKNOWLEDGEMENT

I would like to express my appreciation to my advisor, Hiroshi Akashi for mentoring me. I would like to thank you for giving me the opportunity to pursue the interesting area of evolutionary genetics and for teaching me how to think critically as a scientist. I would like to thank my collaborators, Ziheng Yang, Tomotaka Matsumoto, and Kent Kawashima, for providing me with a concrete basis of data analysis. I acknowledge Tomotaka Matsumoto for teaching me fundamentals of population genetics in the reading club and for giving me critical comments in my analysis and writing. I would like to thank my thesis evaluation committee, Naoki Osada, Jun Kitano, Kuniaki Saito, Kazuho Ikeo, and Naoko Fujito, for the time you spent reading and thinking about my works. I would like to thank previous progress report committee members, Tetsuji Kakutani, Yasukazu Nakamura, and Kayo Hibino, for giving me comments to improve my projects. I would like to thank Evolutionary Genetics lab members. I am grateful to Tomoko Ohta for providing me with insightful comments in journal clubs and lab meetings. I would like to thank lab administrative assistants, Risa Takeshima, Ayumi Ashworth, and Hiroko Mochizuki, for helping me with paperwork and for sharing an enjoyable lunch time. I also would like to thank Hassan Daanaa for his support. Finally, I would like to thank my family and friends for their support.

# Table of Contents

# SUMMARY

One of the central questions in evolutionary genetics is whether natural selection is a primary force acting on new mutations. However, our understanding is limited, especially when natural selection is weak, because of technical challenges. Detection of natural selection requires large sample sizes (numbers of DNA changes) and reliable ancestral inference methods. This study aims to address these limitations in order to understand the role of weak natural selection in molecular evolution.

Codon usage bias, non-random usage of synonymous codons, can be described in population genetic models that include fitness differences among codons. We can test model predictions given ancestral and derived states for observed changes. Because the model applies roughly across codon positions, synonymous changes from different genes can be pooled to provide statistical power in the analysis.

I assembled a data set of genome sequences from species in the *Drosophila melanogaster* subgroup from the public databases. The species are an excellent system for our analyses in several aspects. Since these species are closely related, the degree of sequence differences reflects the number of mutations in a lineage. This feature allows us to reduce errors in ancestral inference compared to lineages with long genetic distances. Genome sequences are available from natural population samples and outgroup species. I employed DNA sequences for strains from Rwanda and Madagascar populations of *Drosophila melanogaster* and *D. simulans*, respectively, and reference sequences of *D. yakuba* and *D. erecta* as outgroups.

I focused on distinguishing among directional forces, natural selection and GC-biased gene conversion (gBGC), acting on synonymous mutations and intronic mutations. Since the two forces are predicted to give overlapping signatures in polymorphisms and divergence, I refer to them as “fixation bias”. To distinguish fixation biases from mutation biases, I analyzed site frequency spectrum (SFS), distributions of allele frequencies among variants segregating in a population. Mutations that have positive fitness effects are expected to be segregating at higher frequencies in a population than mutations that have negative fitness effects. The expectation can be tested by comparing SFS between two classes of mutations that are physically interspersed within DNA (*e.g.*, T→C vs. C→T). To construct SFS for

each mutation class, I employed an ancestral inference method that our lab has developed previously. Our lab has confirmed the method's reliability in studying the base composition evolution in the Drosophila species. The combination of large-scale data and a reliable method is one of the advances from previous studies.

In *D. simulans*, synonymous and intronic mutations show strong support for fixation biases consistent with the direction to increase GC content. I employed the categorization of synonymous families, groups of synonymous codons that differ by a single nucleotide. I estimated fixation biases for mutations for each synonymous family using SFS data. I found that the absolute values of the estimates are heterogeneous among synonymous families and between synonymous and intronic mutations. Because gBGC should not distinguish functional classes (*e.g.*, coding vs. non-coding regions), it cannot explain the heterogeneity in fixation biases. Fixation bias estimates are greater at X-linked loci than at autosomal loci. This difference may reflect the differences in factors, such as heterogamety in males, dosage compensation, gene content, and population-scale recombination rates. These factors have been predicted to differentiate the magnitudes of fixation biases between X-linked and autosomal genes but this work does not address which mechanisms contribute to the autosome vs X chromosome difference in fixation biases.

In *D. melanogaster*, fixation bias is strongly supported for synonymous mutations, but is acting in the opposite direction from that in *D. simulans*; AT-increasing changes are segregating at higher frequencies compared to GC-increasing changes for synonymous mutations only within NAY codons. This pattern is not found for other mutations but is consistently found for NAY codons at autosomal and X-linked loci. Moreover, the absolute values of fixation bias estimates are greater at X-linked loci than at autosomal loci. The results cannot be explained by gBGC. The results provide strong evidence that synonymous mutations in *D. melanogaster* are under ongoing adaptive evolution toward fitness optimum different from the closely related species, *D. simulans*. NAY codons are the only ones translated by tRNA that can undergo post-transcriptional replacement of guanosine with its modified form at wobble positions, called queuosine modification. The levels of queuosine modification are known to depend on substrate abundance in diets. One possibility is that diet changes may underly change in directions of selection between *D. simulans* and *D. melanogaster*. This study shows that population genomics approaches can contribute to the

discovery of functional effects of mutations that were not detected in experimental or ecological approaches.

**CHAPTER 1**

# General introduction

## Development of theories of molecular evolution

Fundamental questions in evolutionary genetics are what are the factors that drive evolution of genetic composition. Population genetics gives us a systematic framework to pursue such questions. More specifically, population genetics mainly considers two processes: the origination of new variations and the fixation of the variations in a population.

### *Neutral and nearly neutral theories of molecular evolution*

Until molecular variation data became available, biologists have believed that evolution proceeds under Darwin's theory of natural selection (Darwin 1859). The mechanisms of evolution became controversial due to the extensively high levels of naturally occurring variations at molecular level. Harris (1966) and Lewontin and Hubby (1966) applied the gel electrophoresis to proteins and showed high levels of allozyme polymorphisms (Harris 1966; Hubby and Lewontin 1966; Lewontin and Hubby 1966). The initial discussion was how natural selection can maintain the extensive variations in a population. Some form of balancing selection was the initial candidate. An alternative model was proposed independently by Kimura (1968) and King and Jukes (1968) (Kimura 1968; King and Jukes 1969). They categorized new mutations into two classes: strongly selected and selectively neutral. Among strongly selected classes, they think strongly advantageous mutations are very rare and have negligible impact on molecular evolution. Since strongly deleterious mutations are quickly removed from a population by natural selection, most mutations that go

to the fixation should be selectively neutral. The fixation process of the neutral mutations is driven by genetic drift, fluctuation in the allele frequency by random sampling of gametes in a finite population (Kimura and Ohta 1971). This model is referred to as the neutral theory of molecular evolution and explains early observations of protein variations: the existence of high levels of variations and clock-like properties of protein evolution (Zuckerkandl and Pauling 1965; Harris 1966; Hubby and Lewontin 1966; Lewontin and Hubby 1966).

Subsequent studies reported DNA variations that are not consistent with predictions of the neutral theory. Under the neutral model, heterozygosity is predicted to be proportional to population size and mutation rate. Therefore, a greater heterozygosity is expected for a large population. However, species that supposedly have different population sizes (human, fly, and yeast) show similar levels of protein polymorphism. This issue is often referred to as "invariance of heterozygosity" (Lewontin 1974).

Another issue was generation time effect of DNA divergence; DNA divergence was greater in lineages experiencing more numbers of generations per year, while protein divergence was proportional to absolute time. If mutation rates are constant between synonymous and nonsynonymous changes, substitution rates for those changes are expected to be equal.

Ohta proposed that nearly neutral mutations may explain the invariance of heterozygosity and generation time effect (Ohta 1972). The nearly neutral theory assumes a continuous distribution of selection coefficient, $s$, for new mutations around range of $s = 0$ (Ohta 1972, 1973) and she called mutations under selective pressure, $s < |1/N_e|$, nearly neutral mutations.

# Base composition evolution

## *Early observations of codon usage bias*

One of the well-studied phenomena supporting the nearly neutral theory is codon usage bias, non-random usage of synonymous codons. Codon usage bias was initially considered the best candidate for the neutral evolution (Kimura 1968; King and Jukes 1969) because synonymous mutations do not affect a primary protein structure. Codon usage bias is observed across a wide range of taxa (Grantham *et al.* 1980a; b, 1981; Bernardi and Bernardi 1985; Ikemura 1985; Ogasawara 1985; Aota and Ikemura 1986; Vicario *et al.* 2007; Drummond and Wilke 2008) including *Drosophila* (Shields *et al.* 1988; Moriyama and Powell 1997). Under the neutral theory, mutation bias is considered a major factor shaping codon usage bias (Sueoka 1962; Freese 1962; Grantham *et al.* 1980a; Filipski 1987).

Several observations were inconsistent with the neutral model of codon usage bias. The first observation is that frequently used codons ("major codons", "minor codons" for less frequent codons) tend to be ones recognized by abundant tRNA isoacceptors and the isoacceptors tend to have anticodons that form the Watson-Crick base pairing with the major codons (Ikemura 1981, 1985; Bennetzen and Hall 1982). The second is that highly expressed genes showed higher levels of codon usage bias than lowly expressed genes (Gouy and Gautier 1982; Bennetzen and Hall 1982; Sharp *et al.* 1986; Shields *et al.* 1988; Andersson and Kurland 1990). The third observation is that synonymous substitution rates show negative correlations with the degree of codon usage bias and with gene expression levels (Sharp and Li 1987; Shields *et al.* 1988). These observations are not expected under the neutral model, assuming a uniform mutational spectrum across genes.

## *Test of directional force among synonymous codons*

Directional selection can be detected using population genetics methods. One approach is the comparison of site frequency spectra (SFS) between mutation classes that are physically interspersed within DNA. The null hypothesis in this test is that the mutation classes are segregating at the same frequency in a population, which come from the neutral prediction. This approach was employed to study protein evolution; Frequency distributions are compared among protein polymorphisms based on allozyme data (Bulmer 1971) and frequency distributions are compared between nonsynonymous and synonymous polymorphisms (Sawyer *et al.* 1987). Akashi and Schaeffer (1997) expanded the method to compare SFS between forward and reverse mutations (*e.g.*, TTT→TTC vs TTC→TTT) among synonymous codons (Akashi and Schaeffer 1997). The test has rejected the neutral evolution of synonymous codons in *Drosophila* (Akashi and Schaeffer 1997; Begun 2001; Kern and Begun 2005; Poh *et al.* 2012; Jackson *et al.* 2017). In *Drosophila melanogaster* and *D. simulans*, previous studies have found that GC-increasing mutations are segregating at higher frequency than AT-increasing mutations (Akashi and Schaeffer 1997). Polymorphisms and divergence comparison also showed that GC-increasing mutations are fixed within a population at a higher rate than AT-increasing mutations in *D. simulans* (Akashi 1995). The disagreement from the neutral expectation was initially interpreted as the action of natural selection. The selection intensity, $N_e s$, was estimated in a range of 0 to 1 in the two populations (Akashi 1995, 1996; Akashi and Schaeffer 1997; McVean and Vieira 2001; Jackson *et al.* 2017), consistent with the prediction under the Li-Bulmer model (Li 1987; Bulmer 1991).

There is experimental evidence that suggests differences in translation efficiency between major and minor codons. Previous studies suggested shorter waiting time for isoaccepting

tRNA and effective rejection of non-cognitive tRNA in bacteria and yeast (Robinson *et al.* 1984; Varenne *et al.* 1984; Sørensen *et al.* 1989; Curran and Yarus 1989; Gardin *et al.* 2014). The fast translation may be beneficial because it increases free ribosomes in a cell and decreases total energy consumed for rejecting non-cognitive tRNA. Major codons are also found to be translated more accurately than minor codons in bacteria (Precup and Parker 1987). The accurate translation of major codons may be because of lower rate of mis-aminoacylation of isoaccepting tRNA, the efficiency of initial rejection of non-cognitive tRNA, and subsequent proofreading.

Non-selective force may also explain the observed patterns. Since most of the major codons in *Drosophila* species end in G or C (Shields *et al.* 1988; Akashi 1995; Akashi and Schaeffer 1997; Vicario *et al.* 2007), GC-biased gene conversion may also underlie codon usage bias. GC-biased gene conversion is a non-random repair of GC/AT mismatches into GC alleles during meiotic recombination. Meiotic recombination at homologous locus carrying different sequences (heterozygous locus) forms mismatches within heteroduplex DNA, which is subject to a DNA repair. When the repair process has a systematic bias toward particular nucleotides, such bias is expected to affect base composition evolution. GC-biased repair was found by transfection of viral DNA containing mismatches to monkey kidney cells (Brown and Jiricny 1988) and was suggested to result in GC-biased gene conversion (Brown and Jiricny 1989). Later studies show experimental evidence for GC-biased gene conversion as a consequence of meiotic recombination in yeast (Birdsell 2002; Mancera *et al.* 2008) and in humans (Odenthal-Hesse *et al.* 2014; Williams *et al.* 2015). GC-biased gene conversion was proposed to contribute to elevation of GC content in DNA sequences (Brown and Jiricny 1989; Holmquist 1992; Eyre-Walker 1993)(Brown and Jiricny 1989; Holmquist 1992; Eyre-Walker 1993).

One prediction of the GC-biased gene conversion model is a positive correlation between meiotic recombination rate (more specifically GC-biased DNA repair rates during meiosis) and GC content. Such correlations have been reported for many taxa: human (Ikemura and Wada 1991; Eyre-Walker 1993; Meunier and Duret 2004; Duret and Arndt 2008), yeast (Gerton *et al.* 2000; Jeffreys and Neumann 2002), mouse (Clément and Arndt 2013), and other taxa (Pessia *et al.* 2012). GC-biased gene conversion may affect substitution rates. A positive correlation between recombination rates and GC-increasing substitution rates is found for mammals (Duret and Arndt 2008; Clément and Arndt 2011; Galtier *et al.* 2018; Pracana *et al.* 2020), reptiles (Figuet *et al.* 2014), avians (Nabholz *et al.* 2011; Mugal *et al.* 2013; Rousselle *et al.* 2019), plants (Clément *et al.* 2017), and bacteria (Lassalle *et al.* 2015; Long *et al.* 2018). Polymorphism data can be employed to test GC-favoring fixation biases; one prediction is that GC-increasing mutations segregate at higher frequencies than AT-increasing mutations in populations. This prediction is supported in mammals (Eyre-Walker 1999; Smith and Eyre-Walker 2001). However, these lines of evidence are also consistent with directional selection favoring GC and evidence for GC-biased gene conversion in *Drosophila* species is still controversial (see below).

The impact of biased gene conversion has been studied in a framework of population genetics theories (Nagylaki 1983; Walsh 1983; Bengtsson 1986). A model of mutation, drift, and biased gene conversion (but without natural selection) can give a prediction of evolutionary dynamics that are not distinguishable from those under a model with weak natural selection (Nagylaki 1983). Therefore, in this study, I refer to directional selection and biased gene conversion as fixation bias. More specifically, fixation biases that elevate GC content will be referred to as “GC-favoring fixation bias”.

### *Intron base composition*

Population genetics approaches often employ putatively neutral sites. In *Drosophila*, currently the best candidate of neutral evolution is sites within short introns. Previous studies have shown that the middle regions (specifically, 8-30 bp regions) of short introns have the highest divergence and polymorphisms compared to other regions in the *D. melanogaster* genome (Halligan and Keightley 2006; Parsch *et al.* 2010). There is evidence for GC-favoring fixation biases in short introns in *D. melanogaster* and *D. simulans* (Jackson *et al.* 2017; Jackson and Charlesworth 2021). However, fixation biases have not been tested for mutations that do not change GC content of short introns.

## Goal of the study

The goal of the dissertation is to reveal the evolutionary forces shaping codon usage bias and intron base composition in *Drosophila*. More specifically, I focused on fixation biases (directional selection and biased gene conversion). To estimate fixation biases, I employed population genetics methods and compared the estimates among various mutation classes: synonymous mutations within individual amino acid coding families (synonymous families) and intronic mutations.

**CHAPTER 2**

# Base composition evolution in *Drosophila*

## Introduction

Codon usage bias is a well-established system to study the nearly neutral theory of molecular evolution. The observed synonymous polymorphisms (Akashi and Schaeffer 1997) agreed with the action of weak directional force, as predicted by the quantitative model of codon usage bias (Li 1987; Bulmer 1991). The directional force is often interpreted as directional selection because 1) codons that are used at higher proportion ("major codons") than other synonymous codons ("minor codons") are ones that are recognized by the most abundant tRNA. Major codons tend to form the Watson-Crick base pairings with the tRNA anticodons (Ikemura 1981, 1985; Bennetzen and Hall 1982), 2) the degree of codon usage bias is greater for highly expressed genes than lowly expressed genes (Gouy and Gautier 1982; Bennetzen and Hall 1982; Sharp *et al.* 1986; Shields *et al.* 1988; Andersson and Kurland 1990), and 3) synonymous substitution rates are lower in highly biased and highly expressed genes (Sharp and Li 1987; Shields *et al.* 1988).

Support for directional selection from population genetic approaches can be confounded by a different directional force that does not reflect fitness effects, such as biased gene conversion. Because theory predicts that biased gene conversion has equivalent outcome of evolution as directional selection (Nagylaki 1983), biased gene conversion needs to be taken into account in tests of directional selection. Jackson and colleagues (2017) attempted to distinguish directional selection for codon usage bias from GC-biased gene conversion by comparing fixation bias estimates among mutation classes (Jackson *et al.* 2017). Fixation bias

estimates at 4-fold redundant sites are greater than those at short intron sites and show elevated magnitudes for genes with high GC content at 4-fold redundant sites (Jackson *et al.* 2017). However, they did not examine synonymous mutations at 2-fold redundant sites.

In this study, I aimed to distinguish effects of directional selection and biased gene conversion acting on synonymous mutations and to examine whether directional forces vary among 2-fold synonymous families. I employed genome-scale data for population samples; 14 lines from the Rwandan population of *D. melanogaster* (Pool *et al.* 2012) and 21 lines from the Madagascar population of *D. simulans* (Rogers *et al.* 2014; Jackson *et al.* 2017) and two closely related species as outgroups. I tested fixation biases by comparing SFS for forward and reverse mutations within each of the synonymous families (*e.g.,* AAA → AAG vs AAG → AAA). The categorization of mutation classes is based on ancestral and derived states and designations of "forward" and "reverse" are arbitrary; We use the term forward mutations for A or T (W)→ G or C (S) changes among GC-altering mutations and for T→A and C→G among GC-conservative mutations. Mutations in the opposite direction are termed reverse mutations. As many studies have suggested non-stationary and lineage-specific evolution of synonymous mutations in the *D. melanogaster* subgroup (Akashi *et al.* 2006; Nielsen *et al.* 2007; Singh *et al.* 2007, 2009; DuMont *et al.* 2009), I employed a refined ancestral inference method with a non-stationary model (Yang 2007; Matsumoto *et al.* 2015; Matsumoto and Akashi 2018). Furthermore, I employed a maximum likelihood method (Glémin *et al.* 2015) to estimate fixation biases and compared estimates among mutation classes.

# Results

## *Population genomics analysis*

### Level of DNA polymorphisms

The *D. simulans* population has a greater level of polymorphisms than the *D. melanogaster* population (Fig. 1). Nucleotide polymorphism was roughly four-fold greater in *D. simulans* than in *D. melanogaster* for autosomal short introns, those with lengths 100bp or less for the *D. melanogaster* and *D. simulans* sequences. Watterson's $\theta$ $\omega\varepsilon\rho\varepsilon$ 0.0495 and 0.0110 for *D. simulans* and *D. melanogaster*, respectively. The greater diversity in *D. simulans* than in *D. melanogaster* is consistent with previous reports (Moriyama and Powell 1996; Andolfatto 2001; Langley *et al.* 2012; Jackson *et al.* 2017; Jackson and Charlesworth 2021). I compared the levels of nucleotide polymorphisms between X-linked and autosomal short introns. X vs autosome ratio (X:A ratio) of neutral variations is expected to be 0.75 under assumptions of 1:1 sex ratio and equal mutation rates in the sexes. The X:A ratio of short intron diversity in *D. simulans* (0.727) was slightly smaller than but close to the expectation, while the estimate was greater than the expectation in *D. melanogaster* (1.25). Previous studies have also shown the similar patterns of X:A ratio of diversity (Begun and Whitley 2000; Andolfatto 2001; Jackson *et al.* 2017). The results confirm that our methods of ortholog identification do not produce a systematic bias at least in the diversity estimates.

### Fixation bias tests in short introns

I simply assumed the selective neutrality in DNA variations in short introns to obtain a rough estimate of the population diversity. To filter out intronic mutations that show evidence of no neutrality, I tested against the neutral assumption for each pair of forward and reverse

mutations. In *D. simulans*, GC-increasing mutations show higher values of allele frequency estimates than AT-increasing mutations in short introns (Table 1; Fig. 2). This pattern was found for both autosomal and X-linked short introns (Table 1; Fig. 2). Interestingly, the magnitude of SFS differences seem greater for X-linked, compared to, autosomal short introns in this species (Fig. 2). On the other hand, GC-increasing and AT-increasing mutations do not show significant differences in *D. melanogaster* (Table 1; Fig. 2).

GC-conservative mutations (G↔C and A↔T) show a contrasting pattern. There was no statistically significant difference in SFS between forward and reverse changes within GC-conservative mutations in autosomal and X-linked short introns in both species (Table 1). Previous studies have employed GC-conservative mutations as an assumed neutral reference for *Drosophila* species (Maside *et al.* 2004; Galtier *et al.* 2006; Haddrill and Charlesworth 2008; Jackson *et al.* 2017; Jackson and Charlesworth 2021) but had not explicitly tested this assumption. In the following sections, we employ GC-conservative mutations within short introns as a neutral reference to estimate the magnitude of fixation bias parameters. We pooled GC-conservative mutations into a single class.

**Evidence for GC-fixation biases in *D. simulans***

Among synonymous polymorphisms, GC-increasing mutations segregate at higher frequencies than AT-increasing mutations in *D. simulans* (Table 2). Such SFS differences are predicted under fixation bias (natural selection and/or biased gene conversion) that favors GC-ending codons. The patterns are strongly supported at both 2-fold and 4-fold redundant sites (Table 2) and expand on previous results that combined data from 2-fold and 4-fold sites for smaller numbers of genes (Akashi and Schaeffer 1997; Kliman 1999) and that employed

only 4-fold sites in larger data sets (Jackson *et al.* 2017). Genome-scale data allow us to further refine the analysis to individual synonymous families. Each of the 10 synonymous families shows SFS differences similar to the general pattern of elevated frequencies of GC-increasing compared with AT-increasing mutations in this species (Table 3; Fig. 3a and 3b).

We estimated fixation bias parameters, $\gamma_{W\leftrightarrow S}$, in order to compare SFS differences across mutation classes using the maximum likelihood method (Glémin *et al.* 2015). This approach requires polymorphism data for a putative neutrally-evolving class of mutations and we employed GC-conservative mutations within short introns. We confirmed the fixation bias estimates strongly correlate with the magnitudes of SFS differences (Fig. 4). The fixation bias parameters can be considered summary statistics of magnitudes of SFS differences. Fixation bias estimates are heterogeneous among synonymous families in *D. simulans* (Fig. 3c). $\gamma_{W\leftrightarrow S}$ for synonymous changes are uniformly greater than that for GC-altering mutations within short introns (Fig. 3c) and are correlated between autosomal and X-linked loci (Fig. 3c; Spearman's rank correlation $r_S = 0.955$, $p < 10^{-5}$). Interestingly, X-linked loci show larger fixation bias estimates than autosomal loci across mutation classes (Fig. 3c; Wilcoxon signed-rank test $p = 0.0051$).

**Evidence for both GC- and AT-fixation biases in *D. melanogaster***

In contrast to the patterns in *D. simulans,* evidence for fixation biases is heterogeneous between 2-fold and 4-fold redundant sites in *D. melanogaster*. In the pooled analyses, 4-fold redundant sites show GC-increasing mutations segregating at higher frequencies compared with AT-increasing mutations but 2-fold sites show little support for SFS differences (sample sizes are similar, Table 2). Many previous studies have suggested that natural selection

became less effective in the *D. melanogaster* lineage since its split from *D. simulans* (Akashi 1995; Akashi and Schaeffer 1997; Jackson *et al.* 2017). This may be because of population size reduction (Akashi 1995; McVean and Vieira 2001) and/or reduction in selection coefficient (Clemente and Vogl 2012). The pooled analysis is consistent with weaker fixation biases in *D. melanogaster* than in *D. simulans.* The weaker fixation bias in *D. melanogaster* may reflect smaller population size or selection coefficient but this study does not address which factors are contributing to the observation.

Individual 2-fold synonymous families show striking patterns in *D. melanogaster* polymorphism. For five out of the 10 families, SFS for GC-increasing and AT-increasing mutations are indistinguishable (Table 3). However, four synonymous families (Asp, Asn, His, and Tyr) show a distinct pattern of AT-increasing mutations segregating at higher frequencies than GC-increasing mutations at both autosomal and X-linked loci (Table 3 and Fig. 5). In each case, X-linked loci show larger point estimates of fixation bias in favor of AT (Fig. 5c). These four families (NAY codons) correspond perfectly with those for which cognate tRNAs undergo queuosine (Q) modification in the 3rd (wobble) position in almost all eukaryotes (Harada and Nishimura 1972; White *et al.* 1973; Kasai *et al.* 1975). Such modifications are not known to occur for cognate tRNAs for other synonymous families (Fergus *et al.* 2015).

We further examined the statistical support for X vs autosome differences in fixation bias estimates at NAY codons. Because sample sizes are limited for some of the codon families (especially for X-linked genes), we pooled data for the four families and increased the numbers of bootstrap replicates. Among autosomal loci, all 10,000 replicates supported GC-favoring forces ($\gamma_{W\leftrightarrow S} > 0$) in *D. simulans* and AT-favoring forces in *D. melanogaster* ($\gamma_{W\leftrightarrow S} < 0$; Fig. 6). X-linked loci show larger magnitudes (absolute values) of fixation biases than

autosomal loci in both *D. simulans* (all replicates) and *D. melanogaster* (all but one replicate; Fig. 6). Opposing directions of fixation biases in *D melanogaster* and *D. simulans* and greater efficacy of fixation biases on X-linked, compared to autosomal, loci are well-supported for NAY codons.

**Robustness of individual family analysis**

I confirmed the robustness of our findings based on a different approach of SFS construction. I inferred ancestral nucleotides and their probabilities using sequence alignments, where codon positions are pooled among 2-fold synonymous families: Asp, His, Asn, Tyr, Cys, Phe, Ser2, Lys, Gln, and Glu. I used the ancestral states to construct SFS for each synonymous family. This approach differs from the approach employed for analysis above in ancestral states for which sites from different classes were inferred under a single set of phylogenetic parameters. Overall patterns were consistent with the results above. In *D. simulans*, GC-increasing mutations are segregating at higher frequencies than AT-increasing mutations across all of the 2-fold synonymous families (Table 4). In *D. melanogaster*, Lys show GC-increasing mutations segregating at higher frequencies than AT-increasing mutations (Table 4). Asp, Asn, His and Tyr in *D. melanogaster* show patterns of AT-increasing mutations segregating at higher frequencies than GC-increasing mutations (Table 4). This analysis confirms that AT-favoring fixation bias does not reflect outlier in polymorphism data but does reflect global force in *D. melanogaster*.

**Fixation biases for 4-fold synonymous families**

Individual 4-fold synonymous families show GC-increasing mutations segregating at higher frequencies than AT-increasing mutations in *D. simulans*; the pattern is found at both autosomal and X-linked loci (Table 5; Table 6). One exception was Gly. Synonymous mutations to GGG codons (GGH→GGG) are segregating at lower frequencies than mutations in the opposite direction at autosomal loci in *D. simulans* (Table 5). This pattern of SFS differences is found also at X-linked loci for GGC↔GGG mutations but not for GGA↔GGG and GGT↔GGG mutations (Table 6). X-linked loci show larger magnitude of fixation bias parameters than autosomal loci in *D. simulans* (Fig. 7; Wilcoxon signed-rank test $p$ = 0.00063, site classes that include 200 polymorphisms for each forward and reverse mutations).

In *D. melanogaster*, SFS differences are found for fewer mutation classes compared with *D. simulans* (Table 5; Table 6). Interestingly, GGW→GGG mutations are found at lower frequencies compared with mutations in the opposite direction at autosomal loci in *D. melanogaster* (Table 5), consistent with the pattern in *D. simulans*. This suggests fixation biases that reduce GGG codon usage within the Gly synonymous family in both *D. simulans* and *D. melanogaster*. Overall, a simple model of fixation biases that favors GC-increasing mutations can not explain SFS patterns of 4-fold families in *D. simulans* and *D. melanogaster*.

# Discussion

I observed heterogeneity in evolutionary parameters among mutation classes within *D. simulans* and *D. melanogaster*. I tested fixation biases acting on mutations within individual synonymous families using genome-wide polymorphism data. The results strongly support the major codon preference model of codon usage evolution in *D. simulans*. In contrast, I found little support for fixation biases within most synonymous families in *D. melanogaster* but found strong support for AT-fixation biases in NAY codon families.

## *Support for the major codon preference model in D. simulans*

The mutation model does not seem a likely explanation. The mutation model considers mutation bias and drift as main drivers of codon usage evolution. If there is no fixation bias between two nucleotides states, SFS are expected to be not distinguishable between forward and reverse mutations because mutations affect the number of polymorphisms for each frequency class to the same extent. But if the degree of mutation bias has dramatically changed since the MRCA of segregating alleles, forward and reverse mutations can show differences in SFS (Akashi 1997; Eyre-Walker 1997). For example, a recent increase of AT-increasing mutation rate is expected to result in excess of rare AT-increasing changes segregating in a population. If there was no change in GC-increasing mutation rate, GC-increasing changes may appear to segregate at higher frequencies compared to AT-increasing changes in a population. To explain the observed heterogeneity in fixation biases (between synonymous and intronic mutations, between autosomal and X-linked loci, and among synonymous families; Fig. 3c) by mutation effects, it requires three conditions; the degree of mutation bias change is greater in coding regions than in non-coding regions, greater in X-

linked than autosomal loci, and heterogeneous among synonymous families (Fig. 3c). Although mutation processes may be context-dependent (Assaf *et al.* 2017), a recent change in context-dependent mutations has not been reported for *Drosophila* species. It is unlikely that mutation effects are the only factor underlying the heterogeneity in fixation bias estimates. Overall, the patterns of elevated segregating allele frequency of GC-increasing mutations compared to AT-increasing mutations suggests that GC-favoring fixation biases are prevalent across intronic and synonymous mutations in *D. simulans*. Previous studies have also supported GC-fixation biases acting on synonymous mutations in *D. simulans* (Akashi and Schaeffer 1997; Begun 2001; Nielsen *et al.* 2007; Jackson *et al.* 2017). We expanded analyses on 2-fold synonymous families and revealed heterogeneity in the fixation biases among synonymous families.

GC-favoring fixation bias may include effects of GC-biased gene conversion and directional selection. It is still not clear if GC-biased gene conversion is effective in the *Drosophila* genomes. If GC-biased gene conversion is effective, we can expect a positive correlation between gene conversion rate, which can be inferred from recombination rates, and GC content at putatively neutral sites. However, different studies using the *D. melanogaster* genome reported controversial results; some studies show positive correlation between recombination rate estimates and GC content (Kliman and Hey 2003; Marais *et al.* 2003), but some show the lack of correlation (Haddrill *et al.* 2007; Campos *et al.* 2012; Comeron *et al.* 2012; Jackson and Charlesworth 2021). These studies employ different putatively neutral sites. Marais and colleagues (2003) employed only long introns (Marais *et al.* 2003) because long introns may contain a smaller proportion of sequences required for splicing than short introns (Mount *et al.* 1992). Other studies employed short introns (Jackson and Charlesworth 2021), intergenic and intronic sequences (Comeron *et al.* 2012) and short

and long introns and the third codon positions (Haddrill *et al.* 2007; Campos *et al.* 2012). In addition, GC-biased gene conversion alone cannot explain the difference in intron GC content and synonymous codon usage between autosomal and X-linked loci (Campos *et al.* 2013). SFS analysis showed GC-increasing mutations segregating at higher frequency than AT-increasing mutations in short introns in *D. melanogaster* (Robinson *et al.* 2014; Jackson *et al.* 2017; Jackson and Charlesworth 2021). However, the magnitude of SFS differences and fixation bias estimates do not show correlation with recombination estimates (Robinson *et al.* 2014; Jackson and Charlesworth 2021). For other *Drosophila* species, some studies suggest that GC-biased gene conversion may be active on the X chromosome in *D. simulans* (Haddrill and Charlesworth 2008) and in *D. americana* (de Procé *et al.* 2012). Our analysis does not reject the possibility that GC-biased gene conversion is acting across sites in the D. simulans genome. The existence of GC-biased gene conversion in *Drosophila* species still remains a question. If biased gene conversion rates are roughly constant between intronic and coding regions. fixation biases should be equally effective at these regions under the absence of natural selection. The different magnitude of fixation biases between intronic and synonymous mutations (Fig. 3c; Fig. 7) suggests the action of directional selection. Assuming that SI sites are better examples of neutral evolution than synonymous sites, the result suggests that GC-ending codons are advantageous over AT-ending codons in this species.

## *Autosomal vs X-linked loci*

Both directional selection and biased gene conversion may contribute to the greater efficacy of fixation biases at X-linked, compared to autosomal loci. Some mechanisms may bias the efficacy of fixation biases systematically between autosomal and X-linked loci in *Drosophila*. The first is the recessivity (expression reduction in heterozygotes) of

advantageous alleles. Fitness effects of advantageous mutations may be "masked" in heterozygotes if the advantageous allele is recessive. The average rate of the masking effect is greater at autosomal loci than X-linked loci, because of heterogamety of the X chromosome in fly males, assuming equal numbers of males and females in a population. The net difference in the masking effects may be seen as a difference in the efficacy of natural selection (Avery 1984; Charlesworth *et al.* 1987).

The second is dosage compensation. In *Drosophila*, it is known that dosage compensation is done by transcriptional upregulation of X-linked genes in males (Marín *et al.* 2000; Baker *et al.* 2003). Under the major codon preference model, highly expressed genes are expected to be seen in natural selection more frequently compared to lowly expressed genes. Higher expression levels (on average) of X-linked loci can cause greater efficacy of natural selection. Recessivity and dosage compensation may not be independent factors but dosage compensation may induce recessivity of advantageous alleles (Mank *et al.* 2010). Mank and colleagues discussed that the Kacser-Burns model of dominance mechanism (Kacser and Burns 1981) may be applied from the flux of metabolic pathway to amount of gene expression. The fitness effect (contribution to the amount of final product of a metabolic pathway) of a mutation may be small if a large number of genes is involved in the pathway. However, fitness effects of non-additive (*i.e.*, dominant or recessive) mutations may be equivalent to those for additive mutations when fitness effects are very small. We still do not know the impact of recessivity of slightly adaptive alleles on molecular evolution.

Another possibility is the gene content difference between X-linked and autosomal loci. If X-linked loci harbor genes under stronger directional selection at a higher proportion than autosomal loci, the observed pattern may be consistent with this scenario. Previous studies

have shown that the X chromosome of *D. melanogaster* includes female-biased expression genes at a higher frequency than autosomes (Ranz *et al.* 2003; Hambuch and Parsch 2005; Mikhaylova and Nurminsky 2011). Mutations within X-linked genes with female-biased expression should be seen by selection at greater number of times compared with those within autosomal genes with female-biased expression. Gene content difference between autosomal and X-linked loci may be a plausible factor driving greater efficacy of major codon preference.

GC-biased gene conversion may also be more effective at X-linked loci. Because of recombination suppression in *Drosophila* males (Morgan 1914; Chovnick *et al.* 1970), the X chromosomes experience recombinations at roughly 4/3 times greater number of times than autosomes (Langley *et al.* 1988). Because biased gene conversion is thought to occur at heteroduplex DNA formed during recombination (Szostak *et al.* 1983), a higher recombination rate may result in a greater rate of GC-biased gene conversion at X-linked loci compared to autosomal loci. Recombination rates may also be different between autosomal and X-linked loci if these loci differ in the distributions of factors that regulate meiotic recombination. Fine-mapping of double strand breaks (DSB), which are the initial steps of meiotic recombination, in yeast revealed a significant enrichment of recombination hotspots in promoter regions (Petes *et al.* 1991) and in DNase I sensitive regions (Wu and Lichten 1994). This pattern suggests that chromatin accessibility may be related to recombination initiation. Some DSB hotspots in yeast are also known to require the binding of transcription factors (White *et al.* 1991, 1993; Fan *et al.* 1995; Gerton *et al.* 2000) and some do not require the binding of known transcription factors (Kirkpatrick *et al.* 1999). Recombination initiation may be associated with these factors (*i.e.*, chromatin accessibility, binding of transcription factors and others) in yeast. If there are complex interactions among various factors

determining recombination patterns also in *Drosophila* (Comeron *et al.* 2012), and if these factors differ between X chromosomes and autosomes, recombination rates, consequently GC-biased gene conversion rates, might be different.

The mechanism of greater fixation biases may be explained by directional selection or biased gene conversion. However, again, GC-biased gene conversion does not explain the heterogeneity among mutation classes within X-linked loci (Fig. 3c).

## *Support for AT-favoring selection in D. melanogaster*

SFS differences between GC-increasing and AT-increasing mutations were strongly supported within particular synonymous families in *D. melanogaster* (Fig. 5a; Table 3). Importantly, only the Tyr, Asn, Asp, and His showed SFS differences toward increasing AT content (Fig. 5; Table 3) and SI did not show SFS differences between GC-increasing and AT-increasing mutations (Table 1). SFS difference may be expected if there is a relatively large change in the degree of mutation biases. If there is such change within AY contexts, SFS difference is likely to be seen also in SI, given that other synonymous families but Lys did not show significant SFS differences. Mutation effects are not a likely explanation for the SFS differences within NAY synonymous families. Similarly, biased gene conversion that increases AT content does not explain this because such force should be observed in SFS for short introns sites and other synonymous families if it exists. Furthermore, X-linked loci show a greater degree of SFS differences between AT-increasing and GC-increasing mutations compared to autosomal loci (Fig. 5c; Fig. 6). To explain all the results without natural selection, we need to assume context-dependent mutations and greater degree of mutation bias in the female germ line. To my knowledge, there is no such report to date and

the scenario may be unrealistic. Therefore, we conclude that AT-favoring natural selection is acting on mutations within NAY synonymous families in the *D. melanogaster* population.

Many previous studies have investigated evolutionary forces shaping codon usage bias of the *D. melanogaster* genes. However, the conclusion has been mixed. Previous studies that have pooled synonymous families along the genome supported that natural selection on synonymous mutations has reduced on the *D. melanogaster* lineage (Akashi 1995, 1996; Jackson *et al.* 2017). On the other hand, AT-increasing changes are found at higher frequencies among *D. melanogaster* strains than GC-increasing mutations (Poh *et al.* 2012). However, Poh and colleagues employed strains from African (Malawi) and non-African (Raleigh) samples as alleles in the same population. This seems an unrealistic assumption; African and non-African populations show different levels of DNA variations and do not share many within-species variations (Begun and Aquadro 1993). The pattern found by Poh and colleagues may reflect the effect of pooling samples from different population histories but may not reflect the effects of fixation biases.

Although there has not been a strong support for global AT-favoring fixation biases in *D. melanogaster*, substitution analyses revealed strong support for AT-increasing forces at a single locus, *Notch* (Bauer DuMont *et al.* 2004; Nielsen *et al.* 2007; Singh *et al.* 2007; Holloway *et al.* 2008). To detect non-neutral forces in this locus, Bauer DuMont and colleagues developed a method that takes mutation bias into account for the estimation of substitution rates. The authors found significantly greater numbers of AT-increasing synonymous substitutions than that of GC-increasing synonymous substitutions (Bauer DuMont *et al.* 2004). The similar pattern of nucleotide substitution rates was not observed in closely located introns. This suggests that mutation bias does not explain the strikingly

greater rate of AT-increasing substitutions for synonymous mutations in the *Notch* locus. Later studies applied a more parameter-rich phylogenetic method (Nielsen *et al.* 2007) and expanded the analyses on genome-wide (Singh *et al.* 2007). These studies found that the significantly greater rates of AT-increasing synonymous substitutions than GC-increasing synonymous substitutions are specific to the *Notch* locus; other loci showed slightly greater rates for GC-increasing synonymous substitutions and almost no difference (Nielsen *et al.* 2007; Singh *et al.* 2007). Holloway and colleagues applied another approach that attempts to detect lineage-specific accelerated evolution to *D. melanogaster* and closely related species. This approach was initially developed to study human-specific accelerated evolution and acted as an independent confirmation of the substitution bias in the *Notch* locus in *Drosophila* (Holloway *et al.* 2008). Overall, AT-favoring fixation biases on synonymous mutations have not been established as global forces acting genome-wide in *D. melanogaster*.

Next question is which NAT or NAC codons are ancestrally preferred codons. According to previous studies, NAC codons seem to be preferred in lineages prior to the split between *D. simulans* and *D. melanogaster*. In the previous studies, almost all the synonymous families show positive correlations between degree of codon usage bias and a proportion of G- or C-ending codons in *D. melanogaster* (Shields *et al.* 1988; Akashi 1995) and similarly in *D. simulans* (Akashi and Schaeffer 1997). Although variation in mutation bias among genes can shape a similar pattern, multivariate analysis including intron GC content was not consistent with prediction under mutational explanation. While there was a positive correlation between GC content at the 3rd codon positions and that at intronic sites for genes showing low codon usage bias, there was no correlation for genes showing high codon usage bias (Kliman and Hey 1994). Because it is less likely that mutation bias varies between non-coding and coding regions within a gene, fixation biases should have acted to establish a greater proportion of

GC-ending codons in *D. melanogaster*. Later analyses confirmed that NAC codon usage increases from lowly to highly biased genes in seven other *Drosophila* species: *D. simulans, D. sechellia, D. yakuba, D. erecta, D. ananassae, D. pseudoobscura, D. persimilis* (Akashi and Schaeffer 1997; Vicario *et al.* 2007). Putatively neutral intronic sites are AT biased in these species, similarly to *D. melanogaster* (Moriyama and Hartl 1993; Vicario *et al.* 2007). Assuming that intron GC content more closely reflects mutational equilibrium, NAC codons have been ancestrally preferred over NAT codons.

NAC codon preference is also supported by a different approach. Deviation from equal codon usage between conserved or non-conserved positions has been compared across *Drosophila* species for each synonymous family (Zaborske *et al.* 2014). If a synonymous change has an influence on translation accuracy, codon positions that are sensitive to replacement change should show enrichment of preferred codons (Akashi 1994). *D. melanogaster* showed preference for NAC codons over NAT codons (Zaborske *et al.* 2014). This discrepancy may be because NAT codon preference is completely independent of translational accuracy selection and/or is a very recent force. In either (or both) cases, the identity of the preferred codon may have changed from NAC codons to NAT codons recently in the *D. melanogaster* lineage.

## *Queuosine modification of tRNA*

NAY-decoding tRNAs share a common feature of chemical modification, queuosine modification. Queuosine is a modified nucleoside and was first identified in Tyr-tRNA of E. coli and was designated as “nucleoside Q” as an unknown nucleoside (Goodman *et al.* 1968; RajBhandary *et al.* 1969; Doctor *et al.* 1969). Queuosine comprises a 7-deazaguanosine core,

where nitrogen at the seventh position in purine is replaced with a carbon, to which an aminomethyl chain and cyclopentanediol moiety are attached (Ohgi *et al.* 1979). Queuosine is found in Tyr-tRNA, Asp-tRNA, His-tRNA, and Asn-tRNA (Goodman *et al.* 1968; RajBhandary *et al.* 1969; Doctor *et al.* 1969; Harada and Nishimura 1972) in various organisms, including bacteria, flies, plants, worms, and mammals (Kasai *et al.* 1975; Katze *et al.* 1982). Although queuosine is found across a wide range of organisms, eukaryotes cannot synthesize queuosine but salvage a precursor base, queuine (Reyniers *et al.* 1981; Ott *et al.* 1982; Kirtland *et al.* 1988; Siard *et al.* 1991; Gaur *et al.* 2007). The precursor base is involved in a transglycosylation reaction, where the precursor base is replaced with guanine base at the wobble position. Since eukaryotes rely on diet for the source of queuine, the queuine is considered "micronutrient" for eukaryotes by recent studies (Zaborske *et al.* 2014; Zallot *et al.* 2014; Fergus *et al.* 2015; Müller *et al.* 2019; Hayes *et al.* 2020). Interestingly, NAY-decoding tRNAs are the only tRNAs that are found to be subject to queuosine modification to date (Okada and Nishimura 1979; Fergus *et al.* 2015). The overlap with codon preference reversal within NAY synonymous families may be consistent with the contribution of queuosine modification of tRNA to preference reversal.

Previous studies have examined whether the presence of queuosine at the wobble positions changes the efficiency of protein synthesis but conclusions are controversial (Harada and Nishimura 1972; Grosjean *et al.* 1978; McNamara and Smith 1978; Owenby *et al.* 1979; Yokoyama *et al.* 1979; Smith and McNamara 1982). Some other studies suggest its impact on the protein synthesis process. Meier and colleagues compared histidine incorporation rates between CAC and CAU codon positions within mRNA of a virus coat protein using an *in vivo* translation system (Meier *et al.* 1985). *Drosophila* two His-tRNA isoacceptors, which have the identical primary sequence except for the wobble position, were

injected into *Xenopus* oocytes with the target mRNA. His-tRNA$_{GUG}$ shows a higher rate of histidine incorporation at CAC compared to CAU codons, whereas His-tRNA$_{QUG}$ shows little difference between CAC and CAU codons (but show a slightly higher rate for CAU codons) (Meier *et al.* 1985). Computational modeling of tRNA structures predict that Asp-tRNA$_{GUA}$ forms more stable binding with GAC compared to GAU codons but Asp-tRNA$_{QUA}$ shows slight bias in binding stability between these codons (Morris *et al.* 1999). Under queuosine-deficient conditions, C-ending codons are likely to be preferred over T-ending codons for efficient translation. The preferred codon under queuosine-rich conditions is still unclear. In addition, studies have reported other possible functions of queuosine modification such as difference in aminoacylation efficiency (Noguchi *et al.* 1982; Singhal and Vakharia 1983), reduced misread of stop codons (Bienz and Kubli 1981), elevated tRNA stability (Tuorto *et al.* 2012, 2015, 2018), and efficient translation of mitochondrial genome coded genes (Suzuki *et al.* 2020). These aspects of the effects of queuosine modification may impact molecular evolution.

Although there is a potential influence of the tRNA$_{QUN}$ abundance on codon preference, we do not know if the tRNA$_{QUN}$ abundance is likely different between the African populations of *D. simulans* and *D. melanogaster*. queuosine is synthesized *de novo* in bacteria (Kersten and Kersten 1990) and taken by eukaryotes through diet (Ott *et al.* 1982; Kirtland *et al.* 1988; Siard *et al.* 1991; Gaur *et al.* 2007), and/or gut microbiota (Reyniers *et al.* 1981; Nishimura 1983; Katze *et al.* 1984). A recent study has shown that Marula fruit odor activates Or22a-expressing olfactory sensory neurons of *D. melanogaster* and that the *Or22a* locus showed high values of genetic differentiation statistics (FST) between African and European populations (Mansourian *et al.* 2018). Gut microbiota may be affected by diet (Staubach *et al.* 2013) but may also affect the foraging preference of *D. melanogaster* (Wong

*et al.* 2017). Such change may consequently alter the amount of queuine/queuosine consumption and may lead to a change in the ratio of $tRNA_{GUN}$ and $tRNA_{QUN}$. We do not know the biological mechanisms underlying AT-favoring fixation biases but our analysis suggests queuosine modification of tRNA may be an important factor. The fitness effects of fixation bias change may be related to protein synthesis but other phenotypes may be key contributors.

# Methods

## *Identifying orthologs*

### Genome sequences

We identified orthologs for protein-coding genes among four species, *D. melanogaster*, *D. simulans*, *D. yakuba*, and *D. erecta*, from the *D. melanogaster* subgroup. We obtained genome sequences and gene annotations from FlyBase (ftp://ftp.flybase.net/genomes/) for *D. melanogaster* (r6.24 FB2018_05; last downloaded on 12th July 2019), *D. simulans* (r2.02 FB2017_04; last downloaded on 9th March 2020), *D. yakuba* (r1.05 FB2016_05; last downloaded on 12th July 2019) and *D. erecta* (r1.05 FB2016_05; last downloaded on 12th July 2019). In the following analysis, we employed the longest CDS for genes with multiple protein isoforms and filtered genes for which CDS lengths were not multiples of three. We obtained 13,867, 13,996, 14,489, and 13,369 predicted CDS from the *D. melanogaster*, *D. simulans*, *D. yakuba* and *D. erecta* genomes, respectively. As outgroup species for ortholog identification, we employed genome sequences and gene annotations for *D. ananassae* (r1.06 FB2018_04; last downloaded on 16th October 2019) and *D. pseudoobscura* (r3.04 FB2018_05; last downloaded on 16th October 2019). We found 14,125 and 14,390 predicted CDS matching to the criteria above, respectively.

### Identifying putative ortholog groups

We combined two approaches to group the predicted CDS: one is based on published ortholog groups, and the other is protein sequence homology. The first approach employed the FlyBase ortholog annotations for *D. melanogaster* genes across 12 *Drosophila* species (ftp://ftp.flybase.net/releases/FB2018_05/precomputed_files/orthologs/D.

melanogaster_orthologs_in_drosophila_species_fb_2018_05.tsv.gz; last downloaded on 26th December 2019; [Thurmond et al. 2018]). We obtained 13,493 putative ortholog groups that include at least two representatives among *D. melanogaster*, *D. simulans*, *D. yakuba*, *D. erecta*, *D. ananassae* and *D. pseudoobscura*. The second approach employed protein sequence similarity searches using OrthoFinder ([Emms and Kelly 2015]; last downloaded on 13th September 2019). We obtained 12,789 putative ortholog groups that include at least two representatives among *D. melanogaster*, *D. simulans*, *D. yakuba*, *D. erecta*, *D. ananassae* and *D. pseudoobscura*. FlyBase and OrthoFinder groups were fused if groups shared one or more members. After fusing, we obtained a total of 16,193 groups. Among these, 10,320 groups included a single representative each from *D. melanogaster*, *D. simulans*, *D. yakuba* and *D. erecta*. 1,384 groups had at least one representative each from the four species but greater than one representative for at least one species. Other groups were missing representatives from one or more species of the *D. melanogaster* subgroup and were not included for further analyses.

We further processed the groups to obtain the four-species ortholog sets that are likely to be evolving independently of other sets. We will refer to such sets as "*msye* ortholog sets". *D. ananassae* and *D. pseudoobscura* genes were included in the groups for processing when available.

**Groups with single representatives from each *D. melanogaster* subgroup species**

We excluded groups that may include mis-assigned orthologs and show questionable alignments from the 10,320 groups that include single representative each from the four species. We employed phylogenetic and DNA distance approaches. We aligned predicted

protein sequences within ortholog groups using the E-INS-I method within the MAFFT software package (Katoh *et al.* 2002) and replaced amino acids with codons in the corresponding positions. We removed codons at which any of the aligned codons included gaps and/or non-ATGC characters. Nine groups were eliminated because no codons remained after this process. We estimated gene trees under a maximum parsimony assumption and conducted bootstrap resampling of nucleotide sites ($n$ = 1000). Since a maximum parsimony method does not require computational power, it allows bootstrap analyses for each candidate group. We determined supported clades among gene trees for the bootstrap replicates for each group using the "majority rule extended method" (implemented as a default setting of *consense* program of Phylip). In this method, clades that are observed more frequently among replicates are shown in an output gene tree as long as it does not contradict with more frequently occurring groups. Bootstrap resampling, parsimony tree estimation and consensus tree estimation were conducted using *seqboot*, *dnapars* and *consense* programs respectively in Phylip ((Felsenstein 2004); version 3.697; last downloaded on 25th September 2019). We filtered groups in which *D. ananassae* and *D. pseudoobscura* genes were placed within a *D. melanogaster* subgroup clade (bootstrap support ≥ 50%) because *D. ananassae* and *D. pseudoobscura* are established as distantly related to this subgroup (Kopp and True 2002; Ko *et al.* 2003; Akashi *et al.* 2006; Pollard *et al.* 2006). Among the 10,311 phylogenetic trees, 26 groups were rejected because of the placement of *D. ananassae*/*D. pseudoobscura*. We did not filter based on estimated topologies within a *D. melanogaster* subgroup clade. Because *D. melanogaster*, *D. simulans*, *D. yakuba* and *D. erecta* are closely related (Ko *et al.* 2003; Akashi *et al.* 2006; Heger and Ponting 2007; Wong *et al.* 2007), we do not expect relationships within this clade to be resolvable for most single genes.

To reduce the proportion of misaligned/misannotated data, we filtered groups that show extremely high sequence distance for the closely related species. We tested levels of synonymous divergence ($d_S$) for all CDS pairs within groups. We made CDS alignments for *D. melanogaster*, *D. simulans*, *D. yakuba*, and *D. erecta* genes as described above. For each alignment, we generated sliding windows of 50 codons with a step-size of one codon and calculated pairwise $d_S$ for each window based on the Nei-Gojobori method (Nei and Gojobori 1986) implemented in CODEML (Yang 2007). $d_S$ estimates for each window were tested against threshold values of 0.7 for *D. melanogaster/D. simulans* pairs and 1.0 for other species pairs. We excluded a group if more than 75% of codons were found in high $d_S$ (> threshold for at least one pair) windows or if less than 40 codons remained after removing codons in high $d_S$ windows. We excluded 37 groups according to these criteria. For this filtering, and for other steps described below, arbitrary thresholds were chosen to filter groups with extreme values (usually a few percent).

**Extracting *msye* ortholog sets from multiple candidate-containing groups**

To extract *msye* ortholog sets from the 1,384 FlyBase/OrthoFinder groups, we examined phylogenetic relationships, DNA distances, synteny, and alignment lengths. For the phylogenetic approach, we constructed gene trees and determined consensus trees among bootstrap replicates as described above. From inferred gene trees, we extracted clades (sets of gene members that are monophyletic on a phylogenetic tree) that can be explained by simple scenarios of gene duplications on lineages prior to, and within, the *D. melanogaster* subgroup (Fig. 9). The clade support requirements (Fig. 9) were designed to exclude cases of gene conversion among paralogs following gene duplications. We found a total of 1,788 clades that may contain one or more *msye* ortholog sets. 1,148 of these clades came from inferred

ancestral duplication cases. Among these, 76 included terminal duplications and 1,072 did not show duplications within the *D. melanogaster* subgroup clade. 4 of the candidate clades showed internal duplications (all on the *ms* lineage in Fig. 1b). 636 clades showed terminal duplications and those clades did not include internal duplications and not from ancestral duplications.

We considered all possible sets within the candidate clades as candidate *msye* ortholog sets. We filtered candidate sets based on synonymous divergence as described above. In addition, we tested synteny conservation (*i.e.*, sharing of neighboring orthologous genes). For each candidate *msye* ortholog set, we obtained 20 neighboring genes (10 genes 5′ and 10 genes 3′ of a given gene within a genome) from *msye* ortholog sets identified among the 10,320 FlyBase/OrthoFinder groups. If fewer than 10 genes were available, we listed as many genes as possible. In each species pair, we counted the numbers of the neighboring genes that belong to the same *msye* ortholog set and summed the counts across all species pairs. We retained candidate *msye* ortholog sets that had the highest overall counts among candidates that came from the same clade.

For remaining clades that contain multiple candidates of *msye* ortholog sets, we selected one set having the greatest number of aligned codons across the four species (and randomly chose one of remaining candidates). Overall, we obtained 1,774 *msye* ortholog sets from the 1,788 clades that contain multiple candidates. In total, we obtained 12,022 *msye* ortholog sets: 10,248 from the FlyBase/OrthoFinder groups that contain a single representative each from the *D. melanogaster* subgroup and 1,774 from FlyBase/OrthoFinder groups that contain more than one *msye* ortholog set candidates.

Among these, we employed *msye* ortholog sets of which both *D. melanogaster* and *D. simulans* genes show predicted chromosomal locations in the same class either autosomal or X-linked. We considered "2L", "2R", "3L", and "3R" scaffolds in the *D. melanogaster* genome, and "Scf_2L", "Scf_2R", "Scf_3L", and "Scf_3R" scaffolds in the *D. simulans* genome as autosomal class. "X" and "Scf_X" are considered as X-linked class for *D. melanogaster* and *D. simulans*, respectively. 10,202 and 1,746 *msye* ortholog sets were employed for the further analyses as autosomal and X-linked loci data sets, respectively.

**Identifying putative intron ortholog groups**

We examined gene structure consistency within *msye* ortholog sets to identify putative intron orthologs. We analyzed 36,400, 37,197, 36,516 and 36,537 introns predicted in *D. melanogaster*, *D. simulans*, *D. yakuba*, and *D. erecta* genes, respectively. We considered an intron pair as candidate orthologs as those that showed consistency in splice site positions (*i.e.*, predicted splice sites occur within aligned codons within CDS alignments). We obtained 34,585, 34,509 and 34,440 such pairs for *D. melanogaster*/*D. simulans*, *D. melanogaster*/*D. yakuba* and *D. melanogaster*/*D. erecta*, respectively. We also retained candidate orthologous introns among the remaining introns if an intron pair maintained consistency of intron order within ortholog, assuming that intron inversion events within a gene are rare among closely related species. We obtained 235, 311 and 380 such pairs for *D. melanogaster*/*D. simulans*, *D. melanogaster*/*D. yakuba* and *D. melanogaster*/*D. erecta* pairs, respectively. Among the candidate pairs, we employed pairs in which *D. melanogaster* introns have lengths less than or equal to 100 bp for the further analyses. From the intron pairs, we constructed groups that include a single representative each from *D. melanogaster*, *D. simulans*, *D. yakuba* and *D.*

*erecta* and obtained 22,209 groups. We considered these groups as candidates for *msye* intron ortholog sets.

We filtered a fraction of the candidates for *msye* intron ortholog sets based on sequence distances. We calculated the number of aligned nucleotides and the level of sequence identity for all the intron pairs within the candidate groups. In addition, we also examined the number of gaps and gap lengths for intron pairs that do not show consistent splice site positions but maintain the order of introns within orthologs. We tested the statistics for each of *D. melanogaster*/*D. simulans*, *D. melanogaster*/*D. yakuba* and *D. melanogaster*/*D. erecta* pairs within candidates. If one or more pairs within a candidate group did not pass any of the tests, the group was excluded from the intron data set. We made intronic sequence alignments for each of *msye* intron orthogroups by using the E-INS-I algorithm implemented in the MAFFT software (Katoh *et al.* 2002). We filtered groups that include pairs of which the number of aligned nucleotides is less than 40 bp.

We employed different threshold values for sequence identities depending on species pairs. Threshold values are 0.50 for *D. melanogaster*/*D. simulans* and 0.40 for *D. melanogaster*/*D. yakuba* and *D. melanogaster*/*D. erecta* species pairs for pairs that show consistent splice site positions. For the other type of pairs, the threshold values were set to 0.712, 0.552 or 0.554 for *D. melanogaster*/*D. simulans*, *D. melanogaster*/*D. yakuba* and *D. melanogaster*/*D. erecta* species pairs, respectively. The threshold values that include most of the intron pairs that show consistent splice site positions are chosen. This strategy of threshold value choice was applied to other filterings below.

We filtered some candidate ortholog groups if a pair included relatively large gaps. We calculated sequence length divided by an alignment size (total number of sites including gaps

in an alignment) for a given pair and employed a smaller value as a statistic. We filtered pairs that do not show consistent splice site positions but maintain intron order by the following cutoff values: 0.833, 0.738 or 0.730 for *D. melanogaster/D. simulans*, *D. melanogaster/D. yakuba* and *D. melanogaster/D. erecta* species pairs, respectively. We also filtered the class of intron pairs if a pair included many gaps between two sequences. We calculated the number of gaps in each sequence for a given pair scaled to the number of aligned nucleotides and employed a larger value as a statistic. Cutoff values were set to 0.0317, 0.0405 or 0.0408 for *D. melanogaster/D. simulans*, *D. melanogaster/D. yakuba* and *D. melanogaster/D. erecta* species pairs, respectively.

After filtering, we retained a total of 21,998 *msye* intron ortholog sets. We excluded four sets among them because we could not resolve intron orthology. We employed 19,113 and 2,759 intron ortholog sets as autosomal and X-linked loci data sets, respectively, for the further analyses.

## *DNA sequences for samples within species*

We employed available genome data for lines established from natural populations of *D. melanogaster* and *D. simulans*. We downloaded genome sequences for *D. melanogaster* lines established from a Rwanda population (http://www.dpgp.org/dpgp2/DPGP2.html; last downloaded on 22nd July 2014; [Pool *et al.* 2012]). Among 22 lines reported in the Pool *et al.* study, 14 lines were employed for this analysis (RG2, RG3, RG5, RG9, RG18N, RG19, RG22, RG24, RG25, RG28, RG32N, RG34, RG36 and RG38N). We excluded five lines that show evidence for a high proportion of admixture with European populations [RG10, RG11N, RG15, RG21N and RG35; (Pool *et al.* 2012)]. In addition, we excluded three lines

that contain relatively high frequencies of ambiguous nucleotides (RG4N, RG7 and RG33). We converted the r6.24 gene annotations to the r5.28 annotation using an annotation mapping table (https://github.com/FlyBase/bulkfile-scripts/blob/master/D. melanogaster_r5_to_r6/D. melanogaster_r5_to_r6_mapping.tsv; last downloaded on 30th July 2019) and extracted CDS and intronic sequences from the genomic sequences for the Rwandan population samples. Genes with different numbers of exons between r5.28 and r6.24 were excluded from the analysis. We obtained a total of 13,691 protein-coding genes.

We employed a reference sequence (Hu *et al.* 2013) and DNA variant information for *D. simulans* lines from a Madagascar population (Rogers *et al.* 2014; Jackson *et al.* 2017) to reconstruct genome sequences for within-species samples. These data were downloaded from https://drive.google.com/drive/folders/0B4O-acc8EJwheS1HZ1hnWkpOQlE?usp=sharing (last downloaded on 1st March 2017). We analyzed the DNA variant information for 21 lines; 10 lines (MD06, MD105, MD106, MD15, MD199, MD221, MD233, MD251, MD63, and MD73) were sequenced by (Rogers *et al.* 2014), and 11 lines (MD03, MD146, MD197, MD201, MD224, MD225, MD235, MD238, MD243, MD255, and MD72) were sequenced by (Jackson *et al.* 2017). These 10 and 11 lines were sampled from the same localities in Madagascar by (Rogers *et al.* 2014) and Wiliam Ballard, respectively, as described in (Jackson *et al.* 2017). To create an annotation matching table, we compared genome sequences of the FlyBase reference (r2.02) and a reference sequence (Hu *et al.* 2013) used for reconstruction of genome sequences for the Madagascar samples (Rogers *et al.* 2014; Jackson *et al.* 2017). We compared scaffold sequences from the reference genomes using MUMmer (version 3.23; last downloaded on 6th March 2020; [Kurtz *et al.* 2004]). We compared sequences of pairs of five main scaffolds: 2L and Scf_2L, 2R and Scf_2R, 3L and Scf_3L, 3R and Scf_3R, 4 and Scf_4, and X and Scf_X. We obtained coordinates of 1-to-1 matching

regions using the *delta-filter* command with *-1* option in MUMmer. 99.9% of r2.02 sequences for the five chromosomes was found in the Hu *et al.* (2013) reference genome with almost perfect sequence match (minimum sequence identity is 99.99%). Using the annotation matching table, we converted r2.02 gene coordinates corresponding to the Hu *et al.* (2013) genome for CDS and intron sequence extraction. We extracted predicted CDS and intron sequences from the genome sequences for a total of 13,831 protein-coding genes.

## *Sequence alignments*

We made sequence alignments using sequences from the reference genome and added sequences from the natural populations. For CDS, we aligned amino acid translations for the ortholog CDS's using the E-INS-i algorithm implemented in the MAFFT program (Katoh *et al.* 2002) and back-translated to nucleotides. For intron analysis, we focused on SI, those with length 100bp or less in both *D. melanogaster* and *D. simulans*, for the following analyses. We aligned intronic sequences using the E-INS-i method, similarly.

We aligned within-species data using the reference sequence alignments of orthologs as a "backbone". We inserted sequences for the within-species lines (14 for the *D. melanogaster* population and 21 for the *D. simulans* population) to the alignments by mapping codon or nucleotide positions to the corresponding positions of the reference sequences. The *D. melanogaster* and *D. simulans* reference sequences were removed from the sequence alignments prior to analyses. The data set includes 10,122 CDS and 18,719 intron alignments for autosomal loci, and 1,746 and 2,705 for X-linked loci.

We filtered data from heterochromatic/low crossover regions because such regions may experience different mutational spectra (Takano-Shimizu 2001; Marais *et al.* 2001, 2003; Singh *et al.* 2005) as well as reduced efficacy of natural selection (Fisher 1930; Muller 1932; Crow and Kimura 1965; Felsenstein 1974) compared to euchromatic regions. We employed the cytogenetic positions of heterochromatic and other chromosomal regions experiencing low crossing-over defined in (Kliman and Hey 1993) and employed chromosome map positions of *D. melanogaster* genes for filtering (ftp://ftp.flybase.net/releases/FB2018_05/precomputed_files/genes/gene_map_table_fb_2018_05.tsv.gz; last downloaded on 29th December 2020).

We also filtered some regions within the remaining CDS and intronic sequence alignments. To focus on single nucleotide variants, we filtered positions that align with gaps in the reference or in any of the population lines. We used the *D. melanogaster* genome annotation to determine coordinates of codons/sites that overlap with predicted transposable elements and/or of transcripts from other genes. In addition, we restricted the analysis to sites that are included in all predicted CDS isoforms for a given gene. Finally, we filtered putatively functionally constrained regions within introns (Halligan and Keightley 2006; Parsch *et al.* 2010): 10 bases at the 5′ splice junctions and 30 bases at the 3′ splice junctions in the sequence alignments. Table 7 shows filtering statistics.

## *Inference of polymorphisms and fixations*

We define "synonymous family" as a group of synonymous codons that can be interchanged in single nucleotide steps; serine coding codons were split into a 2-fold (AGY, referred to as $Ser_2$) and a 4-fold (TCN, referred to as $Ser_4$) family. We analyzed ten

synonymous families (Phe, Asp, Asn, His, Tyr, $Ser_2$, Cys, Gln, Glu, and Lys) at 2-fold redundant sites and six families (Ala, Gly, Val, Thr, Pro and $Ser_4$) at 4-fold redundant sites.

We estimated probabilities of nucleotides at ancestral nodes in the gene tree shown in Fig. 1. Although the sequences examined are relatively closely related, ancestral inference under simple substitution models such as maximum parsimony can be unreliable when character states are biased and/or changing on the gene tree (Collins *et al.* 1994; Matsumoto and Akashi 2018). In addition, our analyses require inference of ancestral and derived states at segregating sites in recombining regions where gene trees may differ among sites. In order to address these issues, we employed a likelihood-based approach that attempts to incorporate uncertainty in ancestral inference. We employed the Bifurcating Tree with Weighting (BTW) method (Matsumoto and Akashi 2018), which allows reconstruction of ancestral nucleotides for both within- and between-species variation. Ancestral nucleotides and their probabilities were estimated under a non-stationary model, GTR-$NH_b$ model (Tavaré 1986; Matsumoto *et al.* 2015) using BASEML (Yang 2007). We employed a newly implemented option (available on BASEML in PAMLver4.9) that allows user-defined branches to share transition parameters. Here, we set parameters to be shared within (but not between) collapse sequence pairs in *D. melanogaster* and *D. simulans*. Ancestral inference was conducted separately for data from autosomal and X-linked loci. We employed probabilities of changes as counts for the numbers of polymorphic and fixed (in the sample) mutations for each of 12 mutation classes.

## *Polymorphism analysis*

We analyzed SFS for forward and reverse mutations between pairs of nucleotides (*e.g.*, A→G vs G→A). Among the six possible pairs, four are "GC-altering" (*i.e.*, W↔S) and two are "GC-conservative" (*i.e.*, A↔T and G↔C). We employed Mann-Whitney *U* tests to compare differences in SFS for forward and reverse mutations. This is a non-parametric approach to test for differences in locations of frequency distributions of data and the statistical power to detect weak fixation biases was examined in (Akashi 1999). Because the counts in our SFS are not integers, all counts were scaled by a factor of 100 and the resulting test statistic was adjusted accordingly (scaled down by the same factor). Direct SFS comparisons between mutation classes that are physically interspersed within DNA (Bulmer 1971; Sawyer *et al.* 1987) attempt to control for effects of linked selection and demographic history in the inference of fixation biases. This approach can be employed to test weak selection models of synonymous codon usage bias (Akashi and Schaeffer 1997; Akashi 1999).

We estimated fixation bias acting on GC-altering mutations using a maximum likelihood method (Glémin *et al.* 2015) that fits observed SFS to theoretical expectations (Wright 1938). SFS of putatively neutral mutations (here, intronic GC-conservative mutations) were employed to adjust for possible departures from steady-state SFS caused by demographic history and linked selection (Eyre-Walker *et al.* 2006). We employed the M1 model in the *anavar* software package (Muyle *et al.* 2011; Glémin *et al.* 2015) to estimate $\gamma_{W\leftrightarrow S}$, the fixation bias parameter. Positive and negative values of $\gamma_{W\leftrightarrow S}$ indicate fixation biases that elevate and reduce GC content, respectively. This parameter is an estimate of the product of $4N_e$ and either the selection coefficient ($s$), or the intensity of the conversion bias ($b$) in selection and biased gene conversion models, respectively. In our analyses, estimates are

strongly correlated with MWU test statistics scaled to sample size (Fig. 3) and can be considered as summary statistics for the magnitude of difference between SFS that can be compared across mutation classes, X-linked and autosomal loci, and across species.

We conducted bootstrap resampling to estimate confidence intervals (CIs) of statistics. The units for resampling were CDS or introns and we obtained 300 replicates unless otherwise noted. For each replicate, we re-estimated ancestral nucleotides and their probabilities and calculated counts for polymorphic and fixed mutations.

## *Statistical methods*

### Bootstrap estimates of confidence intervals

I employed two approaches for CI estimation. The input data for BASEML ancestral reconstruction are terminal node nucleotide configurations (TNNC) and a tree topology. The inference method produces sets of ancestral node nucleotide configurations (ANNC) each with an associated probability (each such set is a "joint reconstruction"). Here, terminal nodes are $m_{c1}$, $m_{c2}$, $s_{c1}$, $s_{c2}$, *y* and *e* and ancestral nodes are *ms*, *ye*, *m*' and *s*' (Fig. 1b). Each bootstrap replicate resamples TNNC (within units of CDS or introns). For ancestral inference of most results, I employed an approach in which joint reconstructions are determined for each bootstrap sample. For analyses to confirm "robustness of individual family" analysis and for SFS constructions for sliding windows, I resampled joint reconstructions for each bootstrap replicate. 95% CI was estimated as the range from 2.5th- to 97.5th-percentile of observed statistics among replicates (Efron 1979).

**CHAPTER 3**

# Conclusion

I observed heterogeneity in evolutionary parameters among mutation classes within *D. simulans* and *D. melanogaster*. I tested fixation biases acting on synonymous mutations within individual synonymous families using genome-wide polymorphism data. The results strongly support the major codon preference model of codon usage evolution in *D. simulans*. In contrast, I found little support for fixation biases within most synonymous families in *D. melanogaster* but found strong support for AT-fixation biases in NAY codon families.

I conclude that evolutionary parameters are strongly heterogeneous among sites in the genome and between species. Most of my findings are consistent with the nearly neutral prediction that most mutations are weakly selected. Under the nearly neutral model, the general assumption is that among weakly selected mutations, the vast majority is deleterious and that slightly advantageous mutations are to compensate for the fitness reduction by slightly deleterious mutations. However, my analysis, in collaboration with lab members, suggests that slightly advantageous mutations may also play a role in ecological adaptation in nature.

# Figures

**Figure 1. Gene tree of *Drosophila melanogaster* subgroup species.**

Data from *D. melanogaster* (*Dmel*), *D. simulans* (*Dsim*), and two outgroups, *D. yakuba* (*Dyak*) and *D. erecta* (*Dere*), were employed for population genetic analyses and/or ancestral inference. (a) Genetic distances among DNA sequences from the four species. Branch lengths are numbers of intronic nucleotide changes per 100 sites at autosomal loci. The within-species gene trees are rough depictions based on the strong and weaker excesses of rare polymorphisms (star-like trees) in *Dsim* and *Dmel*, respectively, compared to neutral equilibrium expectations. Gray circles indicate approximate positions of the MRCA within each population sample. (b) Tree topology employed for ancestral inference. Within-species variation was collapsed to two sequences ($m_{c1}$ and $m_{c2}$ for *Dmel*, $s_{c1}$ and $s_{c2}$ for *Dsim*). Ancestral nodes for within-species sequences are indicated as *m'* and *s'* for *Dsim* and *Dmel*, respectively. Ancestral nodes for between-species sequences are labeled *ms* and *ye* for the pairs *Dmel* / *Dsim* and *Dyak* / *Dere*, respectively. Note that only the topology is employed as input for ancestral inference (branch lengths are among the estimated parameters).

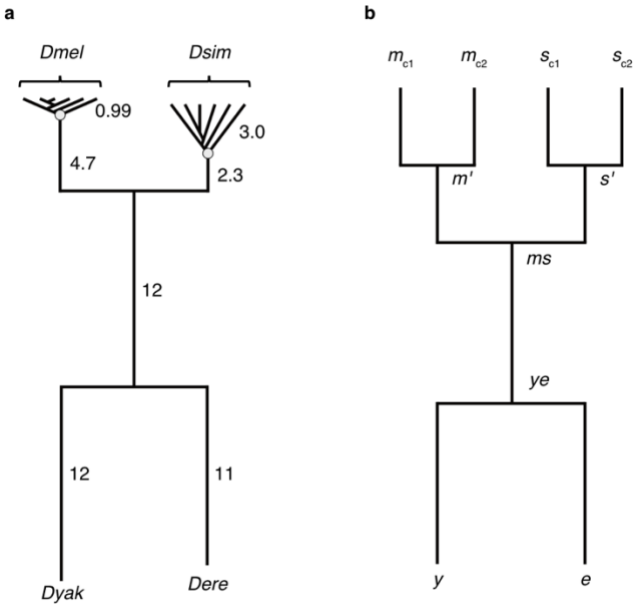


**Figure 2. Estimates of fixation biases acting on intronic mutations based on scaled Mann-Whitney *U* test statistic.**

SFS at SI sites are analyzed. The directions of SFS differences are indicated by colors for each pair of forward and reverse mutations. The magnitudes of SFS differences are indicated by color saturation which reflects MWU *z* statistics scaled to a square root of sample sizes (*sum of* forward and reverse polymorphisms) to give MWU *z'* values. Positive MWU *z'* indicates SFS of forward mutations skewed toward higher frequencies compared with reverse mutations. The top four rows are GC-altering mutations and the bottom two rows are GC-conservative. Gray shading indicates sample size < 200. *, ** and *** indicate $p < 0.05$, < 0.01 and < 0.001, respectively. The sequential Bonferroni method (Holm 1979) was applied within each species and chromosome class to account for multiple tests.

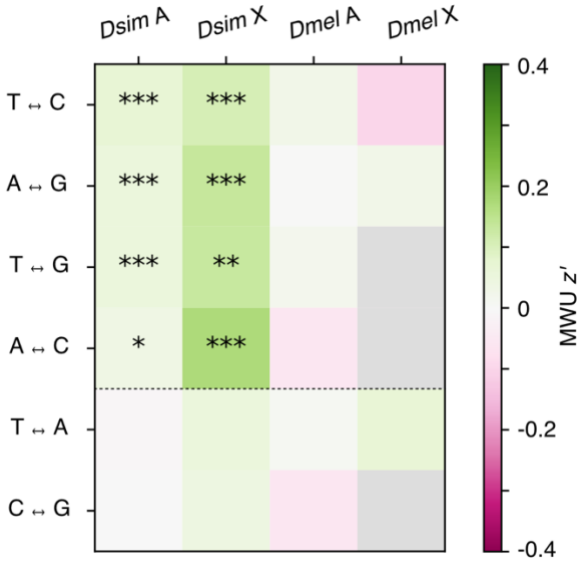


**Figure 3. Fixation biases on synonymous mutations in *D. simulans*.**

SFS comparisons for mutations at autosomal loci for Asp (a) and Lys (b) synonymous families (shown as examples). Data for some intermediate and high frequency classes were pooled as indicated in the x-axis labels. (c) Fixation bias estimates for GC-altering mutations at 2-fold redundant sites and at intronic sites. "intron" indicates data for SI. Autosomal (A) and X-linked (X) loci were analyzed separately. Data are plotted in order of magnitude of γ for autosomes. Error bars indicate 95% bootstrap CIs.

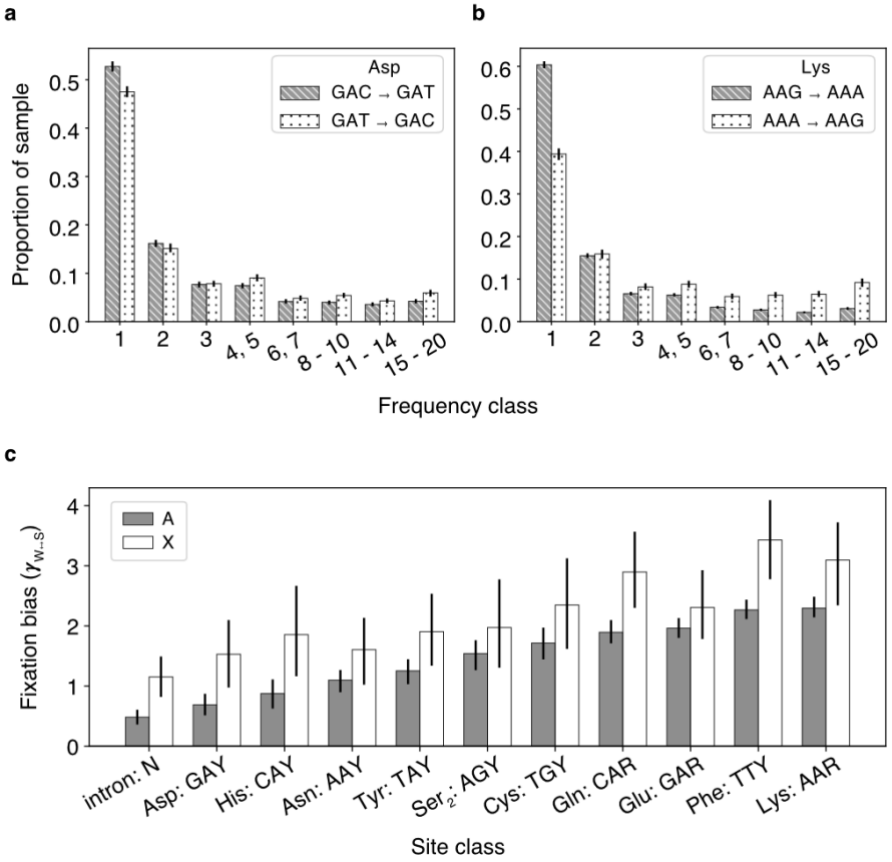


**Figure 4. Relationship between fixation bias parameter and scaled Mann-Whitney *U* test statistic.**

$\gamma_{W\leftrightarrow S}$ and MWU *z* statistics are compared for mutations for 2-fold synonymous families and intronic mutations (mutations at sites within short introns). Data for autosomal (A) and X chromosomal loci are plotted separately. MWU *z* for SFS comparisons of W→S vs S→W polymorphisms are scaled to a square root of the sum of sample sizes (*i.e.*, # W→S polymorphisms + # S→W polymorphisms) to give *z'* values. Error bars indicate 95% bootstrap CIs.

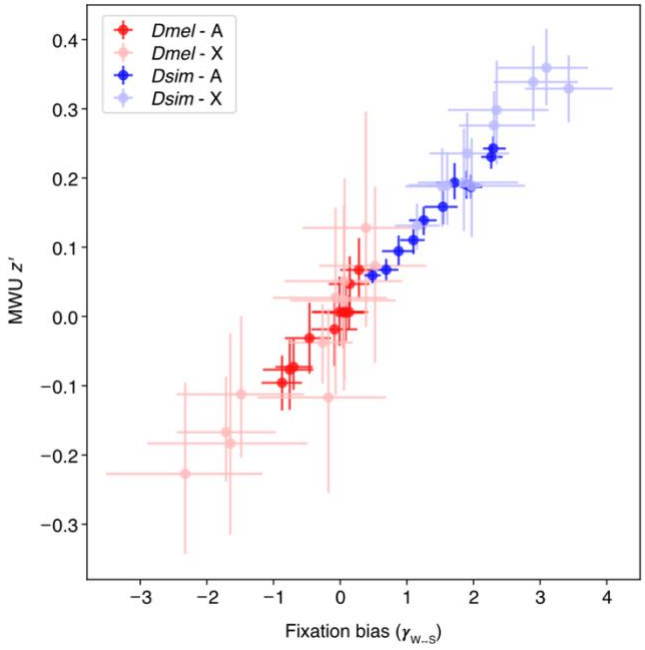


**Figure 5. Fixation biases on synonymous mutations in *D. melanogaster*.**

SFS comparisons for mutations at autosomal loci for Asp (a) and Lys (b) synonymous families (shown as examples). Data for some intermediate and high frequencies were pooled as indicated in the x-axis labels. (c) Fixation bias estimates for GC-altering mutations are plotted similarly to Fig. 2c (including order of site classes). X-axis labels are red for NAY families (see also Result section). Error bars indicate 95% bootstrap CIs.

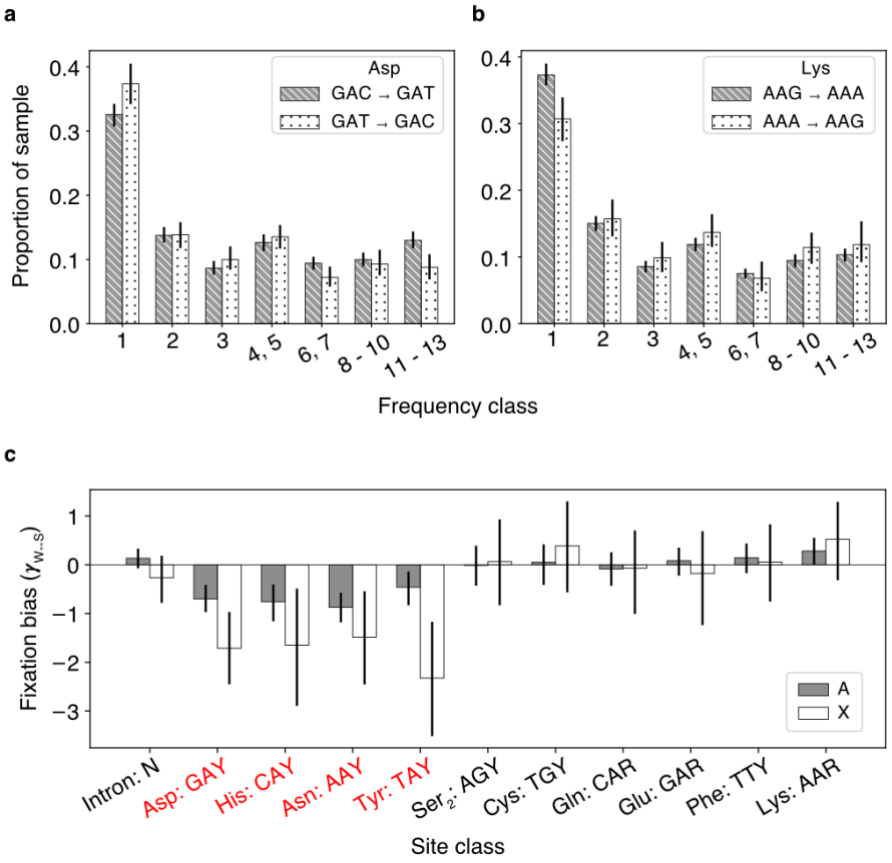


**Figure 6. Fixation bias for synonymous changes at NAY codons between species and chromosome classes.**

Bootstrap distributions of $\gamma_{W \leftrightarrow S}$ estimates for synonymous mutations pooled among NAY families. Data are plotted for autosomal (A) and X-linked loci within each species. 10,000 bootstrap replicates were conducted with resampling of both CDS and introns. Vertical dotted line indicates $\gamma = 0$.

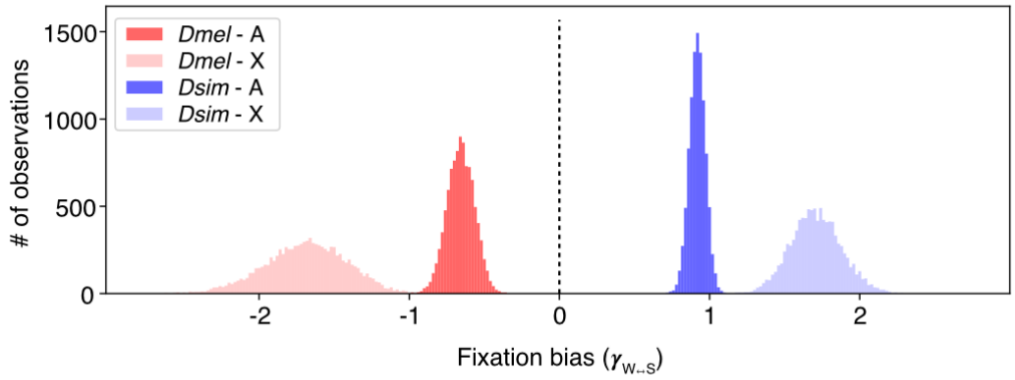


**Figure 7. Fixation biases for synonymous mutations at 4-fold redundant sites.**

γ estimates for each pair of forward and reverse mutations in *D. simulans* autosome (a) *D. melanogaster* autosome (b), *D. simulans* X chromosome (c) and *D. melanogaster* X chromosome (d). GC-conservative mutations at SI were employed as a neutral reference for estimation within each table (*i.e.*, within each chromosome class within species). The top four rows within each table are GC-altering mutations and the bottom two rows are GC-conservative. Gray shading indicates that sample size is less than 200. *, ** and *** indicate $p < 0.05$, $< 0.01$ and $< 0.001$, respectively. The sequential Bonferroni method (Holm 1979) was employed in multiple test corrections within each table.

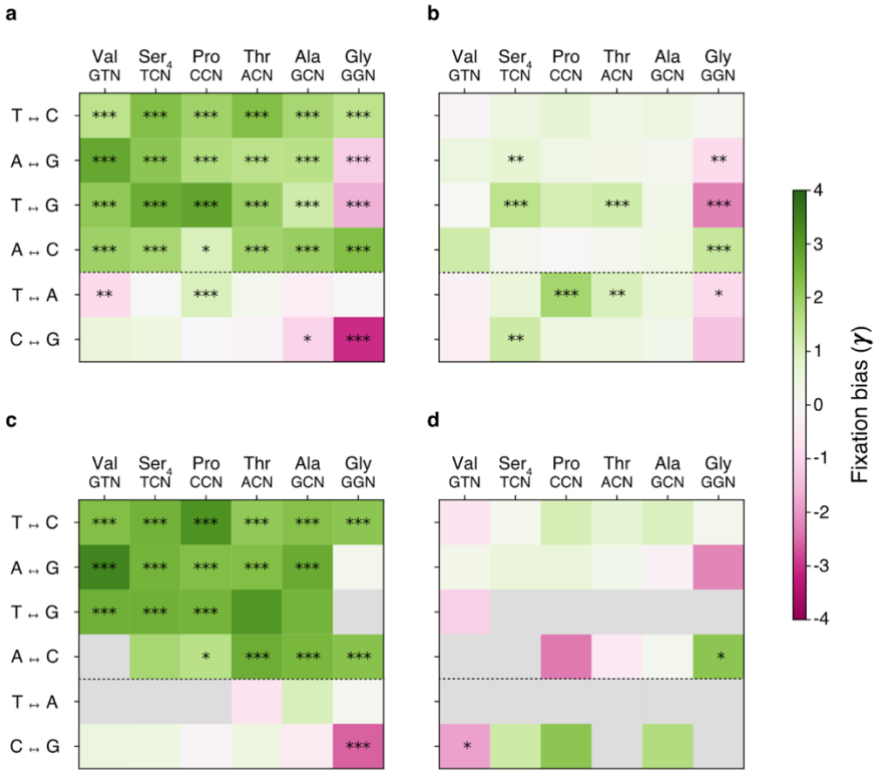


**Figure 8. Magnitude of fixation bias estimates for intronic and synonymous mutations at NAY codons.**

Bootstrap distributions of $\gamma_{W\leftrightarrow S}$ estimates for intronic and synonymous mutations in *D. simulans* autosomal loci. SI were employed for the intron analysis. Synonymous mutations were pooled among Asp, His, Asn and Tyr (labeled as "Qmod", see Table 1). Resampling of introns and CDS was conducted for 10,000 replicates. A vertical dotted line indicates $\gamma_{W\leftrightarrow S}$ = 0.

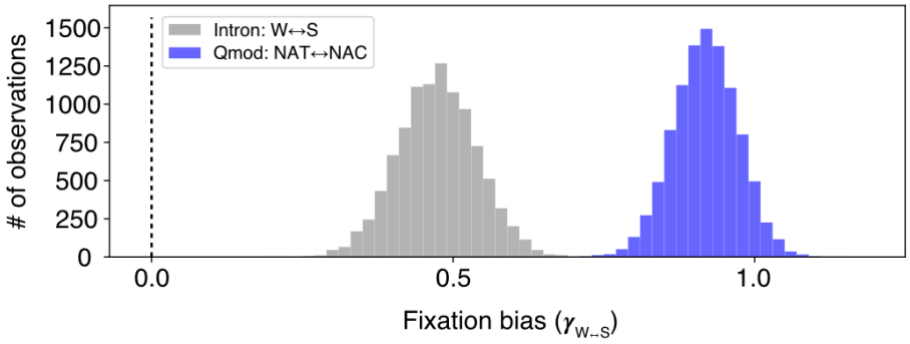
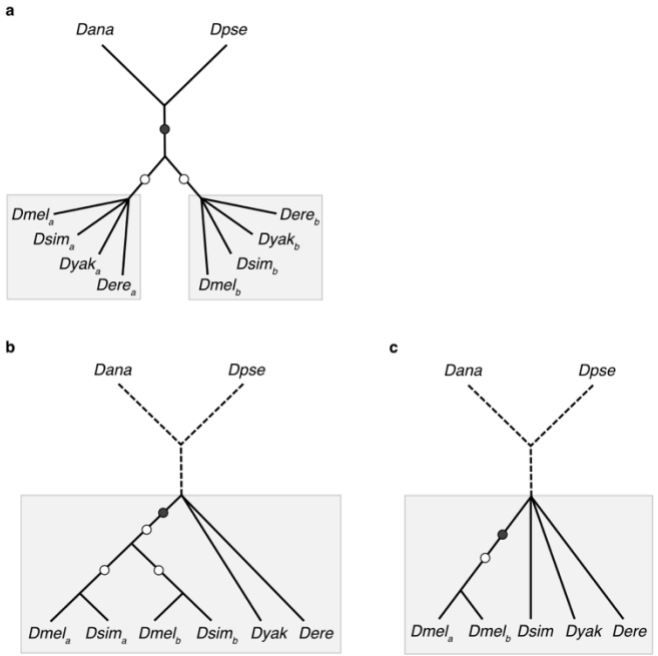


**Figure 9. Phylogenetic analysis for ortholog identification.**

Clades that contain candidate *msye* ortholog sets (gray boxes; candidate set-containing clades). Three gene duplication scenarios were considered: (a) gene duplication prior to the ancestor of the *D. melanogaster* subgroup, (b) gene duplication on an internal branch within the *D. melanogaster* subgroup [prior to the *D. melanogaster* (*Dmel*), *D. simulans* (*Dsim*) split or prior to the *D. yakuba* (*Dyak*), *D. erecta* (*Dere*) split] and (c) gene duplication on a terminal branch within the *D. melanogaster* subgroup. Locations of inferred gene duplications are shown as filled circles. Clades requiring bootstrap support (≥50%, *n*=1000) are indicated with open circles. Relationships among the *D. melanogaster* subgroup members are shown as star trees when the topology is not considered for extraction. *D. ananassae* (*Dana*) /*D. pseudoobscura* (*Dpse*) outgroup lineages are shown as dotted lines when outgroup support is not required for the identification of duplication lineages. We extracted cases with mixtures of (a), (b), and (c) duplication types if the extant gene configuration can be explained by a single most parsimonious scenario that combines such gene duplications.



# Tables

**Table 1. Fixation bias tests: short introns.**

| Species | Chromosome | Mutation type | # forward changes[a] | # reverse changes[b] | MWU $z$[c] |
|---|---|---|---|---|---|
| *Dsim* | A | T ↔ C | 5154.0 | 5675.0 | 7.42 *** |
| | | A ↔ G | 4941.7 | 4865.3 | 5.45 *** |
| | | T ↔ G | 2186.2 | 2723.8 | 3.79 *** |
| | | A ↔ C | 2203.7 | 3313.3 | 2.92 * |
| | | T ↔ A | 5253.7 | 5120.3 | −0.836 |
| | | C ↔ G | 1521.0 | 1323.0 | −0.106 |
| | X | T ↔ C | 675.8 | 762.2 | 4.09 *** |
| | | A ↔ G | 627.3 | 549.7 | 4.69 *** |
| | | T ↔ G | 281.0 | 310.0 | 3.23 ** |
| | | A ↔ C | 262.0 | 357.0 | 4.27 *** |
| | | T ↔ A | 476.3 | 460.7 | 1.67 |
| | | C ↔ G | 236.4 | 172.6 | 0.915 |
| *Dmel* | A | T ↔ C | 882.7 | 1382.3 | 1.40 |
| | | A ↔ G | 863.6 | 1214.4 | 0.118 |
| | | T ↔ G | 400.4 | 551.6 | 0.661 |
| | | A ↔ C | 386.6 | 697.4 | −1.77 |
| | | T ↔ A | 810.2 | 863.8 | 0.469 |
| | | C ↔ G | 357.7 | 317.3 | −1.48 |
| | X | T ↔ C | 169.4 | 359.6 | −2.13 |
| | | A ↔ G | 157.9 | 254.1 | 0.558 |
| | | T ↔ G | 69.2 | 117.8 | −0.209 |
| | | A ↔ C | 55.5 | 134.5 | −0.380 |
| | | T ↔ A | 157.4 | 148.6 | 1.15 |
| | | C ↔ G | 78.7 | 53.3 | −0.798 |

[a] Numbers of forward mutations found polymorphic within species. Forward refers to "left" to "right" changes (e.g., T→C for T↔C) and reverse refers to the opposite direction.

[b] Numbers of reverse mutations found polymorphic within species.

[c] MWU test statistics for SFS comparisons between forward and reverse changes. Positive $z$ values indicate SFS of forward changes skewed toward higher frequencies compared with reverse changes. *, ** and *** indicate $p < 0.05$, $< 0.01$ and $< 0.001$, respectively. The sequential Bonferroni method (Holm 1979) was employed in multiple test corrections within chromosome class within each species.

**Table 2. Fixation bias tests: synonymous mutations pooled within 2-fold and within 4-fold families.**

| Species | Site class[a] | # W→S changes[b] | # S→W changes[c] | MWU $z$[d] |
|---|---|---|---|---|
| *Dsim* | 2-fold | 39447.5 | 103446.5 | 56.2 *** |
| | 4-fold | 33847.6 | 105169.4 | 54.4 *** |
| *Dmel* | 2-fold | 6346.2 | 30863.8 | −2.36 * |
| | 4-fold | 6572.7 | 30582.3 | 7.01 *** |

[a] Class of mutable sites. "2-fold" and "4-fold" indicate 2-fold and at 4-fold redundant sites in autosomal loci, respectively.

[b] Numbers of W→S changes found polymorphic within species.

[c] Numbers of S→W changes found polymorphic within species.

[d] MWU test statistics from SFS comparisons of W→S vs S→W changes. Positive $z$ values indicate SFS of W→S changes skewed toward higher frequencies compared with S→W changes. *, ** and *** indicate $p < 0.05$, $< 0.01$ and $< 0.001$, respectively.

**Table 3. Fixation bias tests: synonymous mutations for individual 2-fold families.**

| Species | Chromosome | Site class[a] | Mutation type | # W→S changes[b] | # S→W changes[c] | MWU $z^d$ |
|---|---|---|---|---|---|---|
| *Dsim* | A | Phe | TTT ↔ TTC | 4063.8 | 16559.2 | 27.0 *** |
| | | Tyr | TAT ↔ TAC | 3367.2 | 7966.8 | 13.7 *** |
| | | His | CAT ↔ CAC | 2971.4 | 5914.6 | 8.44 *** |
| | | Gln | CAA ↔ CAG | 2898.9 | 12357.1 | 18.9 *** |
| | | Asn | AAT ↔ AAC | 4995.3 | 8933.7 | 12.6 *** |
| | | Lys | AAA ↔ AAG | 4645.1 | 14428.9 | 29.5 *** |
| | | Asp | GAT ↔ GAC | 6609.2 | 9070.8 | 8.36 *** |
| | | Glu | GAA ↔ GAG | 4537.4 | 16454.6 | 22.8 *** |
| | | Cys | TGT ↔ TGC | 2315.3 | 5746.7 | 15.9 *** |
| | | $Ser_2$ | AGT ↔ AGC | 2983.7 | 6074.3 | 14.3 *** |
| | | Intron | W ↔ S | 14485.6 | 16577.4 | 10.4 *** |
| | X | Phe | TTT ↔ TTC | 515.1 | 1772.9 | 13.9 *** |
| | | Tyr | TAT ↔ TAC | 383.8 | 803.2 | 7.75 *** |
| | | His | CAT ↔ CAC | 416.5 | 641.5 | 6.21 *** |
| | | Gln | CAA ↔ CAG | 386.3 | 1365.7 | 12.4 *** |
| | | Asn | AAT ↔ AAC | 692.0 | 839.0 | 7.42 *** |
| | | Lys | AAA ↔ AAG | 463.0 | 1411.0 | 14.2 *** |
| | | Asp | GAT ↔ GAC | 864.3 | 926.7 | 7.96 *** |
| | | Glu | GAA ↔ GAG | 492.0 | 1809.0 | 11.3 *** |
| | | Cys | TGT ↔ TGC | 245.6 | 555.4 | 8.05 *** |
| | | $Ser_2$ | AGT ↔ AGC | 373.1 | 756.9 | 6.09 *** |
| | | Intron | W ↔ S | 1846.0 | 1979.0 | 8.12 *** |
| *Dmel* | A | Phe | TTT ↔ TTC | 696.1 | 4655.9 | 2.32 |
| | | Tyr | TAT ↔ TAC | 524.4 | 2601.6 | −1.31 |
| | | His | CAT ↔ CAC | 453.2 | 1973.8 | −2.94 * |
| | | Gln | CAA ↔ CAG | 480.5 | 3641.5 | −0.763 |
| | | Asn | AAT ↔ AAC | 789.8 | 2854.2 | −4.75 *** |
| | | Lys | AAA ↔ AAG | 733.5 | 4166.5 | 3.37 ** |
| | | Asp | GAT ↔ GAC | 1004.3 | 2884.7 | −3.97 *** |
| | | Glu | GAA ↔ GAG | 746.8 | 4746.2 | 0.277 |
| | | Cys | TGT ↔ TGC | 422.0 | 1667.0 | 0.231 |
| | | $Ser_2$ | AGT ↔ AGC | 529.8 | 1638.2 | 0.259 |
| | | Intron | W ↔ S | 2533.4 | 3845.6 | 0.497 |
| | X | Phe | TTT ↔ TTC | 91.9 | 1037.1 | 0.426 |
| | | Tyr | TAT ↔ TAC | 69.1 | 566.9 | −3.50 ** |
| | | His | CAT ↔ CAC | 63.2 | 480.8 | −2.69 |
| | | Gln | CAA ↔ CAG | 73.5 | 881.5 | 0.449 |
| | | Asn | AAT ↔ AAC | 105.1 | 647.9 | −2.13 |
| | | Lys | AAA ↔ AAG | 96.2 | 886.8 | 1.37 |
| | | Asp | GAT ↔ GAC | 176.3 | 661.7 | −3.91 ** |
| | | Glu | GAA ↔ GAG | 75.6 | 1062.4 | −1.92 |
| | | Cys | TGT ↔ TGC | 52.1 | 352.9 | 1.74 |
| | | $Ser_2$ | AGT ↔ AGC | 77.8 | 369.2 | 0.818 |
| | | Intron | W ↔ S | 452.1 | 865.9 | −1.30 |

(continue from the previous page)

[a] Class of mutable sites. Amino acid abbreviations indicate 2-fold synonymous families. “Intron” indicates sites within SI.

[b] Numbers of W→S changes found polymorphic within species.

[c] Numbers of S→W changes found polymorphic within species.

[d] MWU test statistics from SFS comparisons of W→S vs S→W changes. Positive *z* values indicate SFS of W→S changes skewed toward higher frequencies compared with S→W changes. *, ** and *** indicate $p < 0.05$, $< 0.01$ and $< 0.001$, respectively. The sequential Bonferroni method (Holm 1979) was employed in multiple test corrections within chromosome class within each species.

**Table 4. Fixation bias tests: synonymous mutations for individual 2-fold families based on ancestral inference for pooled families.**

| Species | Site class[a] | Mutation type | # W→S changes[b] | # S→W changes[c] | MWU $z^d$ |
|---|---|---|---|---|---|
| *Dsim* | Phe | TTT ↔ TTC | 4248.1 | 16374.9 | 29.9 *** |
| | Tyr | TAT ↔ TAC | 3349.4 | 7984.6 | 13.2 *** |
| | His | CAT ↔ CAC | 2972.2 | 5913.8 | 8.47 *** |
| | Gln | CAA ↔ CAG | 2932.1 | 12323.9 | 19.5 *** |
| | Asn | AAT ↔ AAC | 4986.8 | 8942.2 | 12.6 *** |
| | Lys | AAA ↔ AAG | 4589.3 | 14484.7 | 28.3 *** |
| | Asp | GAT ↔ GAC | 6609.1 | 9070.9 | 8.54 *** |
| | Glu | GAA ↔ GAG | 4573.9 | 16418.1 | 23.5 *** |
| | Cys | TGT ↔ TGC | 2246.8 | 5815.2 | 14.0 *** |
| | $Ser_2$ | AGT ↔ AGC | 2939.8 | 6118.2 | 13.2 *** |
| *Dmel* | Phe | TTT ↔ TTC | 703.1 | 4648.9 | 2.10 |
| | Tyr | TAT ↔ TAC | 522.2 | 2603.8 | −1.47 |
| | His | CAT ↔ CAC | 452.0 | 1975.0 | −2.75 * |
| | Gln | CAA ↔ CAG | 489.5 | 3632.5 | −0.485 |
| | Asn | AAT ↔ AAC | 788.7 | 2855.3 | −4.82 *** |
| | Lys | AAA ↔ AAG | 715.1 | 4184.9 | 2.77 * |
| | Asp | GAT ↔ GAC | 1014.1 | 2874.9 | −3.84 ** |
| | Glu | GAA ↔ GAG | 751.0 | 4742.0 | 0.556 |
| | Cys | TGT ↔ TGC | 396.7 | 1692.3 | −0.423 |
| | $Ser_2$ | AGT ↔ AGC | 513.7 | 1654.3 | −0.228 |

*Note*: Polymorphisms at autosomal loci are analyzed. Ancestral states were inferred using redundant sites pooled among 10 families and polymorphic mutations were inferred using terminal node nucleotide configurations for each family. Column titles are labeled similarly as Table 3. *, ** and *** indicate $p < 0.05$, $< 0.01$ and $< 0.001$ from MWU tests, respectively. The sequential Bonferroni method (Holm 1979) was employed in multiple test corrections within chromosome class within each species.

**Table 5. Fixation bias tests: synonymous mutations for individual 4-fold families in autosomal loci.**

| | | *Dsim* | | | *Dmel* | | |
|---|---|---|---|---|---|---|---|
| Site class[a] | Mutation type | # forward changes[b] | # reverse changes[c] | MWU $z$[d] | # forward changes[b] | # reverse changes[c] | MWU $z$[d] |
| Val | GTT ↔ GTC | 1452.8 | 4286.2 | 9.76 *** | 279.9 | 1376.1 | 0.526 |
| | GTA ↔ GTG | 1403.9 | 8413.1 | 18.7 *** | 291.3 | 2614.7 | 3.00 |
| | GTT ↔ GTG | 997.2 | 4929.8 | 12.4 *** | 161.2 | 1558.8 | 1.08 |
| | GTA ↔ GTC | 294.2 | 1524.8 | 6.63 *** | 68.8 | 425.2 | 2.30 |
| | GTT ↔ GTA | 1149.8 | 1062.2 | −3.75 ** | 238.8 | 288.2 | −0.643 |
| | GTC ↔ GTG | 1479.8 | 2370.2 | 2.56 | 333.1 | 704.9 | −1.15 |
| Ser4 | TCT ↔ TCC | 1527.5 | 6309.5 | 16.8 *** | 234.1 | 1936.9 | 1.49 |
| | TCA ↔ TCG | 1304.4 | 4918.6 | 13.5 *** | 344.2 | 1513.8 | 3.60 ** |
| | TCT ↔ TCG | 522.2 | 2292.8 | 9.97 *** | 95.1 | 481.9 | 4.17 *** |
| | TCA ↔ TCC | 684.0 | 2752.0 | 8.08 *** | 153.7 | 863.3 | 0.266 |
| | TCT ↔ TCA | 760.6 | 762.4 | 0.767 | 172.2 | 176.8 | 1.49 |
| | TCC ↔ TCG | 1945.8 | 1816.2 | 2.39 | 633.5 | 376.5 | 3.99 ** |
| Pro | CCT ↔ CCC | 1677.5 | 5566.5 | 13.7 *** | 289.4 | 1564.6 | 3.07 |
| | CCA ↔ CCG | 2473.0 | 5946.0 | 13.2 *** | 520.8 | 1609.2 | 1.39 |
| | CCT ↔ CCG | 646.9 | 1878.1 | 11.4 *** | 169.9 | 442.1 | 2.89 |
| | CCA ↔ CCC | 1207.2 | 3339.8 | 2.85 * | 227.9 | 1003.1 | 0.161 |
| | CCT ↔ CCA | 1190.0 | 1355.0 | 4.96 *** | 334.7 | 261.3 | 5.12 *** |
| | CCC ↔ CCG | 1941.2 | 1343.8 | 0.877 | 573.0 | 325.0 | 1.40 |
| Thr | ACT ↔ ACC | 1633.7 | 6000.3 | 17.7 *** | 259.1 | 1792.9 | 0.914 |
| | ACA ↔ ACG | 1581.8 | 4813.2 | 11.5 *** | 317.7 | 1531.3 | 1.56 |
| | ACT ↔ ACG | 698.4 | 2148.6 | 8.09 *** | 149.8 | 521.2 | 4.36 *** |
| | ACA ↔ ACC | 863.8 | 3645.2 | 9.01 *** | 169.0 | 1039.0 | 1.26 |
| | ACT ↔ ACA | 1576.4 | 1219.6 | 0.750 | 440.7 | 261.3 | 3.89 ** |
| | ACC ↔ ACG | 1742.5 | 1098.5 | 1.08 | 514.1 | 265.9 | 1.64 |
| Ala | GCT ↔ GCC | 2398.2 | 9659.8 | 16.3 *** | 417.9 | 2764.1 | 2.09 |
| | GCA ↔ GCG | 1757.1 | 4519.9 | 10.0 *** | 357.2 | 1274.8 | 2.03 |
| | GCT ↔ GCG | 900.8 | 1959.2 | 5.46 *** | 139.6 | 525.4 | 0.440 |
| | GCA ↔ GCC | 1077.6 | 4860.4 | 12.1 *** | 219.5 | 1413.5 | 1.64 |
| | GCT ↔ GCA | 1910.5 | 1665.5 | −1.27 | 471.1 | 368.9 | 1.86 |
| | GCC ↔ GCG | 2418.8 | 1000.2 | −2.94 * | 638.9 | 268.1 | 0.474 |
| Gly | GGT ↔ GGC | 3552.4 | 8284.6 | 17.9 *** | 656.3 | 2467.7 | 1.26 |
| | GGA ↔ GGG | 2797.6 | 1955.4 | −8.94 *** | 597.8 | 586.2 | −4.13 ** |
| | GGT ↔ GGG | 774.3 | 818.7 | −5.38 *** | 161.4 | 249.6 | −5.56 *** |
| | GGA ↔ GGC | 1634.4 | 4333.6 | 16.9 *** | 290.8 | 1027.2 | 5.81 *** |
| | GGT ↔ GGA | 1595.7 | 2843.3 | 1.39 | 369.3 | 676.7 | −3.37 * |
| | GGC ↔ GGG | 1766.6 | 563.4 | −9.77 *** | 457.0 | 145.0 | −1.86 |

[a] Class of mutable sites. Amino acid abbreviations indicate 4-fold synonymous families.

[b] Numbers forward mutations found polymorphic within species.

[c] Numbers of reverse mutations found polymorphic within species.

[d] MWU test statistics for SFS comparisons between forward and reverse changes. Positive $z$ values indicate SFS of forward changes skewed toward higher frequencies compared with reverse changes. *, ** and *** indicate $p < 0.05$, $< 0.01$ and $< 0.001$, respectively. The sequential Bonferroni method (Holm 1979) was employed in multiple test corrections within chromosome class within each species.

**Table 6. Fixation bias tests: synonymous mutations for individual 4-fold families in X-linked loci.**

| | | *Dsim* | | | *Dmel* | | |
|---|---|---|---|---|---|---|---|
| Site class[a] | Mutation type | # forward changes[b] | # reverse changes[c] | MWU $z$[d] | # forward changes[b] | # reverse changes[c] | MWU $z$[d] |
| Val | GTT ↔ GTC | 164.9 | 505.1 | 5.40 *** | 45.7 | 296.3 | −1.46 |
| | GTA ↔ GTG | 151.6 | 904.4 | 9.91 *** | 39.0 | 560.0 | 0.605 |
| | GTT ↔ GTG | 102.4 | 568.6 | 6.06 *** | 18.8 | 322.2 | −1.02 |
| | GTA ↔ GTC | 35.5 | 153.5 | 2.68 | 10.1 | 78.9 | 0.937 |
| | GTT ↔ GTA | 113.1 | 82.9 | −1.86 | 40.3 | 37.7 | −0.944 |
| | GTC ↔ GTG | 198.3 | 283.7 | 0.649 | 62.9 | 158.1 | −3.53 * |
| Ser$_4$ | TCT ↔ TCC | 125.3 | 614.7 | 5.12 *** | 26.8 | 395.2 | 1.02 |
| | TCA ↔ TCG | 158.3 | 643.7 | 7.29 *** | 45.4 | 389.6 | 0.896 |
| | TCT ↔ TCG | 49.6 | 242.4 | 4.31 *** | 19.0 | 94.0 | 1.81 |
| | TCA ↔ TCC | 65.9 | 264.1 | 2.20 | 9.7 | 169.3 | −0.187 |
| | TCT ↔ TCA | 53.3 | 52.7 | 2.29 | 42.2 | 16.8 | 0.947 |
| | TCC ↔ TCG | 257.2 | 259.8 | 1.75 | 125.0 | 79.0 | 1.65 |
| Pro | CCT ↔ CCC | 141.9 | 580.1 | 5.50 *** | 35.7 | 302.3 | 0.270 |
| | CCA ↔ CCG | 344.2 | 845.8 | 7.62 *** | 98.9 | 427.1 | 1.09 |
| | CCT ↔ CCG | 53.0 | 253.0 | 4.49 *** | 19.7 | 91.3 | 1.27 |
| | CCA ↔ CCC | 151.9 | 324.1 | 3.38 * | 27.1 | 215.9 | −2.67 |
| | CCT ↔ CCA | 77.2 | 116.8 | 2.72 | 42.4 | 37.6 | 3.06 |
| | CCC ↔ CCG | 284.3 | 211.7 | −0.406 | 137.6 | 63.4 | 3.11 |
| Thr | ACT ↔ ACC | 156.8 | 652.2 | 6.82 *** | 37.3 | 375.7 | 1.28 |
| | ACA ↔ ACG | 184.0 | 620.0 | 7.39 *** | 55.8 | 365.2 | −0.217 |
| | ACT ↔ ACG | 53.1 | 233.9 | 2.84 | 20.8 | 83.2 | 1.50 |
| | ACA ↔ ACC | 75.4 | 354.6 | 6.68 *** | 25.2 | 202.8 | −1.63 |
| | ACT ↔ ACA | 107.3 | 126.7 | −1.95 | 59.0 | 44.0 | 2.38 |
| | ACC ↔ ACG | 216.7 | 148.3 | 0.654 | 101.0 | 58.0 | 2.20 |
| Ala | GCT ↔ GCC | 239.0 | 1163.0 | 8.90 *** | 68.8 | 608.2 | 2.08 |
| | GCA ↔ GCG | 214.9 | 581.1 | 5.95 *** | 60.8 | 340.2 | −1.11 |
| | GCT ↔ GCG | 77.7 | 250.3 | 2.09 | 19.2 | 109.8 | −2.28 |
| | GCA ↔ GCC | 135.8 | 546.2 | 6.18 *** | 33.5 | 269.5 | 0.580 |
| | GCT ↔ GCA | 147.9 | 173.1 | 1.14 | 66.6 | 47.4 | 0.631 |
| | GCC ↔ GCG | 359.4 | 163.6 | −0.915 | 146.6 | 73.4 | 2.47 |
| Gly | GGT ↔ GGC | 543.9 | 1132.1 | 10.6 *** | 145.9 | 693.1 | −0.567 |
| | GGA ↔ GGG | 339.9 | 246.1 | −0.711 | 77.1 | 148.9 | −3.17 |
| | GGT ↔ GGG | 109.2 | 82.8 | 0.833 | 22.2 | 49.8 | −1.19 |
| | GGA ↔ GGC | 200.9 | 504.1 | 5.71 *** | 48.4 | 207.6 | 3.25 * |
| | GGT ↔ GGA | 183.1 | 258.9 | 0.332 | 52.2 | 147.8 | −1.23 |
| | GGC ↔ GGG | 260.7 | 87.3 | −4.52 *** | 93.3 | 22.7 | −2.60 |

*Note*: Column titles are labeled similarly as Table 5. *, ** and *** indicate $p < 0.05$, $< 0.01$ and $< 0.001$ from MWU tests, respectively. The sequential Bonferroni method (Holm 1979) was employed in multiple test corrections within chromosome class within each species.

**Table 7. Data filtering statistics.**

| Site class[a] | Chromosome[b] | # alignments[c] | KH[d] | TE overlap[e] | Alternatively spliced | Transcript overlap[f] | # alignments[g] |
|---|---|---|---|---|---|---|---|
| CDS | A | 10,122 | 761 | 491 | 334,393 | 446,299 | 8,166 |
| | X | 1,746 | 133 | 304 | 67,528 | 100,456 | 1,382 |
| Intron | A | 18,719 | 1,475 | - | 8,765 | 77,852 | 15,927 |
| | X | 2,705 | 141 | - | 5,370 | 13,584 | 2,350 |

[a] "Intron" indicates data for SI.

[b] Chromosome class: autosomal (A) and X-linked chromosomal loci.

[c] Numbers of CDS or introns prior to filtering.

[d] Numbers of CDS or introns located in heterochromatic/lowly recombining regions defined by (Kliman and Hey 1993).

[e] Numbers of codons overlapping with transposable elements.

[f] Numbers of codons overlapping with transcripts for other genes.

[g] Numbers of CDS or introns after filtering.

# 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 divergence at "silent" sites in Drosophila DNA. Genetics 139: 1067–1076.

Akashi H., 1996 Molecular evolution between Drosophila melanogaster and D. simulans: reduced codon bias, faster rates of amino acid substitution, and larger proteins in D. melanogaster. Genetics 144: 1297–1307.

Akashi H., and S. W. Schaeffer, 1997 Natural selection and the frequency distributions of "silent" DNA polymorphism in Drosophila. Genetics 146: 295–307.

Akashi H., 1997 Distinguishing the effects of mutational biases and natural selection on DNA sequence variation. Genetics 147: 1989–1991.

Akashi H., 1999 Inferring the fitness effects of DNA mutations from polymorphism and divergence data: statistical power to detect directional selection under stationarity and free recombination. Genetics 151: 221–238.

Akashi H., W.-Y. Ko, S. Piao, A. John, P. Goel, *et al.*, 2006 Molecular evolution in the Drosophila melanogaster species subgroup: frequent parameter fluctuations on the timescale of molecular divergence. Genetics 172: 1711–1726. https://doi.org/10.1534/genetics.105.049676

Andersson S. G., and C. G. Kurland, 1990 Codon preferences in free-living microorganisms. Microbiol. Rev. 54: 198–210.

Andolfatto P., 2001 Contrasting patterns of X-linked and autosomal nucleotide variation in Drosophila melanogaster and Drosophila simulans. Mol. Biol. Evol. 18: 279–290. https://doi.org/10.1093/oxfordjournals.molbev.a003804

Aota S., and T. Ikemura, 1986 Diversity in G + C content at the third position of codons in vertebrate genes and its cause. Nucleic Acids Res. 14: 6345–6355. https://doi.org/10.1093/nar/14.16.6345

Assaf Z. J., S. Tilk, J. Park, M. L. Siegal, and D. A. Petrov, 2017 Deep sequencing of natural and experimental populations of Drosophila melanogaster reveals biases in the spectrum of new mutations. Genome Res. 27: 1988–2000. https://doi.org/10.1101/gr.219956.116

Avery P. J., 1984 The population genetics of haplo-diploids and X-linked genes. Genet. Res. 44: 321–341. https://doi.org/10.1017/S0016672300026550

Baker B. S., M. Gorman, and I. Marín, 2003 DOSAGE COMPENSATION IN DROSOPHILA. https://doi.org/10.1146/annurev.ge.28.120194.002423

Bauer DuMont V., J. C. Fay, P. P. Calabrese, and C. F. Aquadro, 2004 DNA Variability and Divergence at the Notch Locus in Drosophila melanogaster and D. simulans: A Case of Accelerated Synonymous Site Divergence. Genetics 167: 171–185. https://doi.org/10.1534/genetics.167.1.171

Begun D. J., and C. F. Aquadro, 1993 African and North American populations of Drosophila melanogaster are very different at the DNA level. Nature 365: 548–550. https://doi.org/10.1038/365548a0

Begun D. J., and P. Whitley, 2000 Reduced X-linked nucleotide polymorphism in Drosophila simulans. Proc. Natl. Acad. Sci. U. S. A. 97: 5960–5965. https://doi.org/10.1073/pnas.97.11.5960

Begun D. J., 2001 The frequency distribution of nucleotide variation in Drosophila simulans. Mol. Biol. Evol. 18: 1343–1352. https://doi.org/10.1093/oxfordjournals.molbev.a003918

Bengtsson B. O., 1986 Biased conversion as the primary function of recombination. Genet. Res. 47: 77–80. https://doi.org/10.1017/s001667230002454x

Bennetzen J. L., and B. D. Hall, 1982 Codon selection in yeast. J. Biol. Chem. 257: 3026–3031.

Bernardi G., and G. Bernardi, 1985 Codon usage and genome composition. J. Mol. Evol. 22: 363–365. https://doi.org/10.1007/BF02115693

Bienz M., and E. Kubli, 1981 Wild-type tRNATyrG reads the TMV RNA stop codon, but Q base-modified tRNATyrQ does not. Nature 294: 188–190. https://doi.org/10.1038/294188a0

Birdsell J. A., 2002 Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. Mol. Biol. Evol. 19: 1181–1197. https://doi.org/10.1093/oxfordjournals.molbev.a004176

Brown T. C., and J. Jiricny, 1988 Different base/base mispairs are corrected with different efficiencies and specificities in monkey kidney cells. Cell 54: 705–711. https://doi.org/10.1016/s0092-8674(88)80015-1

Brown T. C., and J. Jiricny, 1989 Repair of base-base mismatches in simian and human cells. Genome 31: 578–583. https://doi.org/10.1139/g89-107

Bulmer M. G., 1971 Protein polymorphism. Nature 234: 410–411. https://doi.org/10.1038/234410b0

Bulmer M., 1991 The selection-mutation-drift theory of synonymous codon usage. Genetics 129: 897–907.

Campos J. L., B. Charlesworth, and P. R. Haddrill, 2012 Molecular evolution in nonrecombining regions of the Drosophila melanogaster genome. Genome Biol. Evol. 4: 278–288. https://doi.org/10.1093/gbe/evs010

Campos J. L., K. Zeng, D. J. Parker, B. Charlesworth, and P. R. Haddrill, 2013 Codon Usage Bias and Effective Population Sizes on the X Chromosome versus the Autosomes in Drosophila melanogaster. Mol. Biol. Evol. 30: 811–823. https://doi.org/10.1093/molbev/mss222

Charlesworth B., J. A. Coyne, and N. H. Barton, 1987 The Relative Rates of Evolution of Sex Chromosomes and Autosomes. Am. Nat. 130: 113–146. https://doi.org/10.1086/284701

Chovnick A., G. H. Ballantyne, D. L. Baillie, and D. G. Holm, 1970 Gene conversion in higher organisms: half-tetrad analysis of recombination within the rosy cistron of Drosophila melanogaster. Genetics 66: 315–329. https://doi.org/10.1093/genetics/66.2.315

Clément Y., and P. F. Arndt, 2011 Substitution patterns are under different influences in primates and rodents. Genome Biol. Evol. 3: 236–245. https://doi.org/10.1093/gbe/evr011

Clément Y., and P. F. Arndt, 2013 Meiotic recombination strongly influences GC-content evolution in short regions in the mouse genome. Mol. Biol. Evol. 30: 2612–2618. https://doi.org/10.1093/molbev/mst154

Clément Y., G. Sarah, Y. Holtz, F. Homa, S. Pointet, *et al.*, 2017 Evolutionary forces affecting synonymous variations in plant genomes., (G. P. Copenhaver, and G. P. Copenhaver, Eds.). PLoS Genet. 13: e1006799. https://doi.org/10.1371/journal.pgen.1006799

Clemente F., and C. Vogl, 2012 Unconstrained evolution in short introns? - an analysis of genome-wide polymorphism and divergence data from Drosophila. J. Evol. Biol. 25: 1975–1990. https://doi.org/10.1111/j.1420-9101.2012.02580.x

Collins T. M., P. H. Wimberger, and G. J. P. Naylor, 1994 Compositional Bias, Character-State Bias, and Character-State Reconstruction Using Parsimony. Syst. Biol. 43: 482–496. https://doi.org/10.1093/sysbio/43.4.482

Comeron J. M., R. Ratnappan, and S. Bailin, 2012 The many landscapes of recombination in Drosophila melanogaster. PLoS Genet. 8: e1002905. https://doi.org/10.1371/journal.pgen.1002905

Crow J. F., and M. Kimura, 1965 Evolution in Sexual and Asexual Populations. Am. Nat. 99: 439–450. https://doi.org/10.1086/282389

Curran J. F., and M. Yarus, 1989 Rates of aminoacyl-tRNA selection at 29 sense codons in vivo. J. Mol. Biol. 209: 65–77. https://doi.org/10.1016/0022-2836(89)90170-8

Darwin C., 1859 *The Origin of Species*. Jorn Murray, London.

Doctor B. P., J. E. Loebel, M. A. Sodd, and D. B. Winter, 1969 Nucleotide sequence of Escherichia coli tyrosine transfer ribonucleic acid. Science 163: 693–695. https://doi.org/10.1126/science.163.3868.693

Drummond D. A., and C. O. Wilke, 2008 Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134: 341–352. https://doi.org/10.1016/j.cell.2008.05.042

DuMont V. L. B., N. D. Singh, M. H. Wright, and C. F. Aquadro, 2009 Locus-specific decoupling of base composition evolution at synonymous sites and introns along the Drosophila melanogaster and Drosophila sechellia lineages. Genome Biol. Evol. 1: 67–74. https://doi.org/10.1093/gbe/evp008

Duret L., and P. F. Arndt, 2008 The impact of recombination on nucleotide substitutions in the human genome. PLoS Genet. 4: e1000071. https://doi.org/10.1371/journal.pgen.1000071

Emms D. M., and S. Kelly, 2015 OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 16: 1–14. https://doi.org/10.1186/s13059-015-0721-2

Eyre-Walker A., 1993 Recombination and mammalian genome evolution. Proc. Biol. Sci. 252: 237–243. https://doi.org/10.1098/rspb.1993.0071

Eyre-Walker A., 1997 Differentiating between selection and mutation bias. Genetics 147: 1983–1987.

Eyre-Walker A., 1999 Evidence of selection on silent site base composition in mammals: potential implications for the evolution of isochores and junk DNA. Genetics 152: 675–683. https://doi.org/10.1093/genetics/152.2.675

Eyre-Walker A., M. Woolfit, and T. Phelps, 2006 The distribution of fitness effects of new deleterious amino acid mutations in humans. Genetics 173: 891–900. https://doi.org/10.1534/genetics.106.057570

Fan Q., F. Xu, and T. D. Petes, 1995 Meiosis-specific double-strand DNA breaks at the HIS4 recombination hot spot in the yeast Saccharomyces cerevisiae: control in cis and trans. Mol. Cell. Biol. 15: 1679–1688. https://doi.org/10.1128/MCB.15.3.1679

Felsenstein J., 1974 The evolutionary advantage of recombination. Genetics 78: 737–756.

Felsenstein J., 2004 PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. www. evolution. gs. washington. edu.

Fergus C., D. Barnes, M. A. Alqasem, and V. P. Kelly, 2015 The queuine micronutrient: charting a course from microbe to man. Nutrients 7: 2897–2929. https://doi.org/10.3390/nu7042897

Figuet E., M. Ballenghien, J. Romiguier, and N. Galtier, 2014 Biased gene conversion and GC-content evolution in the coding sequences of reptiles and vertebrates. Genome Biol. Evol. 7: 240–250. https://doi.org/10.1093/gbe/evu277

Filipski J., 1987 Correlation between molecular clock ticking, codon usage fidelity of DNA repair, chromosome banding and chromatin compactness in germline cells. FEBS Lett. 217: 184–186. https://doi.org/10.1016/0014-5793(87)80660-9

Fisher R. A., 1930 *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford.

Freese E., 1962 On the evolution of the base composition of DNA. J. Theor. Biol. 3: 82–101. https://doi.org/10.1016/S0022-5193(62)80005-8

Galtier N., E. Bazin, N. B. Genetics, and 2006, 2006 GC-biased segregation of noncoding polymorphisms in Drosophila. Genetics Soc America 172: 221–228. https://doi.org/10.1534/genetics.105.046524

Galtier N., C. Roux, M. Rousselle, J. Romiguier, E. Figuet, *et al.*, 2018 Codon Usage Bias in Animals: Disentangling the Effects of Natural Selection, Effective Population Size, and GC-Biased Gene Conversion, (N. Singh, Ed.). Mol. Biol. Evol. 35: 1092–1103. https://doi.org/10.1093/molbev/msy015

Gardin J., R. Yeasmin, A. Yurovsky, Y. Cai, S. Skiena, *et al.*, 2014 Measurement of average decoding rates of the 61 sense codons in vivo. Elife 3. https://doi.org/10.7554/eLife.03735

Gaur R., G. R. Björk, S. Tuck, and U. Varshney, 2007 Diet-dependent depletion of queuosine in tRNAs in Caenorhabditis elegans does not lead to a developmental block. J. Biosci. 32: 747–754. https://doi.org/10.1007/s12038-007-0074-4

Gerton J. L., J. DeRisi, R. Shroff, M. Lichten, P. O. Brown, *et al.*, 2000 Global mapping of meiotic recombination hotspots and coldspots in the yeast Saccharomyces cerevisiae.

Proc. Natl. Acad. Sci. U. S. A. 97: 11383–11390. https://doi.org/10.1073/pnas.97.21.11383

Glémin S., P. F. Arndt, P. W. Messer, D. Petrov, N. Galtier, *et al.*, 2015 Quantification of GC-biased gene conversion in the human genome. Genome Res. 25: 1215–1228. https://doi.org/10.1101/gr.185488.114

Goodman H. M., J. Abelson, A. Landy, S. Brenner, and J. D. Smith, 1968 Amber suppression: a nucleotide change in the anticodon of a tyrosine transfer RNA. Nature 217: 1019–1024. https://doi.org/10.1038/2171019a0

Gouy M., and C. Gautier, 1982 Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res. 10: 7055–7074. https://doi.org/10.1093/nar/10.22.7055

Grantham R., C. Gautier, M. Gouy, R. Mercier, and A. Pavé, 1980a Codon catalog usage and the genome hypothesis. Nucleic Acids Res. 8: r49–r62. https://doi.org/10.1093/nar/8.1.197-c

Grantham R., C. Gautier, and M. Gouy, 1980b Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type. Nucleic Acids Res. 8: 1893–1912. https://doi.org/10.1093/nar/8.9.1893

Grantham R., C. Gautier, M. Gouy, M. Jacobzone, and R. Mercier, 1981 Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res. 9: r43-74. https://doi.org/10.1093/nar/9.1.213-b

Grosjean H. J., S. de Henau, and D. M. Crothers, 1978 On the physical basis for ambiguity in genetic coding interactions. Proc. Natl. Acad. Sci. U. S. A. 75: 610–614. https://doi.org/10.1073/pnas.75.2.610

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. https://doi.org/10.1186/gb-2007-8-2-r18

Haddrill P. R., and B. Charlesworth, 2008 Non-neutral processes drive the nucleotide composition of non-coding sequences in Drosophila. Biol. Lett. 4: 438–441. https://doi.org/10.1098/rsbl.2008.0174

Halligan D. L., and P. D. Keightley, 2006 Ubiquitous selective constraints in the Drosophila genome revealed by a genome-wide interspecies comparison. Genome Res. 16: 875–884. https://doi.org/10.1101/gr.5022906

Hambuch T. M., and J. Parsch, 2005 Patterns of synonymous codon usage in Drosophila melanogaster genes with sex-biased expression. Genetics 170: 1691–1700. https://doi.org/10.1534/genetics.104.038109

Harada F., and S. Nishimura, 1972 Possible anticodon sequences of tRNA His , tRNA Asm , and tRNA Asp from Escherichia coli B. Universal presence of nucleoside Q in the first position of the anticodons of these transfer ribonucleic acids. Biochemistry 11: 301–308. https://doi.org/10.1021/bi00752a024

Harris H., 1966 Enzyme polymorphisms in man. Proc. R. Soc. Lond. B Biol. Sci. 164: 298–310. https://doi.org/10.1098/rspb.1966.0032

Hayes P., C. Fergus, M. Ghanim, C. Cirzi, L. Burtnyak, *et al.*, 2020 Queuine Micronutrient Deficiency Promotes Warburg Metabolism and Reversal of the Mitochondrial ATP Synthase in Hela Cells. Nutrients 12. https://doi.org/10.3390/nu12030871

Heger A., and C. P. Ponting, 2007 Evolutionary rate analyses of orthologs and paralogs from 12 Drosophila genomes. Genome Res. 17: 1837–1849. https://doi.org/10.1101/gr.6249707

Holloway A. K., D. J. Begun, A. Siepel, and K. S. Pollard, 2008 Accelerated sequence divergence of conserved genomic elements in Drosophila melanogaster. Genome Res. 18: 1592–1601. https://doi.org/10.1101/gr.077131.108

Holm S., 1979 A Simple Sequentially Rejective Multiple Test Procedure. Scand. Stat. Theory Appl. 6: 65–70.

Holmquist G. P., 1992 Chromosome bands, their chromatin flavors, and their functional features. Am. J. Hum. Genet. 51: 17–37.

Hu T. T., M. B. Eisen, K. R. Thornton, and P. Andolfatto, 2013 A second-generation assembly of the Drosophila simulans genome provides new insights into patterns of lineage-specific divergence. Genome Res. 23: 89–98. https://doi.org/10.1101/gr.141689.112

Hubby J. L., and R. C. Lewontin, 1966 A molecular approach to the study of genic heterozygosity in natural populations. I. The number of alleles at different loci in Drosophila pseudoobscura. Genetics 54: 577–594.

Ikemura T., 1981 Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes. J. Mol. Biol. 146: 1–21. https://doi.org/10.1016/0022-2836(81)90363-6

Ikemura T., 1985 Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. 2: 13–34. https://doi.org/10.1093/oxfordjournals.molbev.a040335

Ikemura T., and K. Wada, 1991 Evident diversity of codon usage patterns of human genes with respect to chromosome banding patterns and chromosome numbers; relation between nucleotide sequence data and cytogenetic data. Nucleic Acids Res. 19: 4333–4339. https://doi.org/10.1093/nar/19.16.4333

Jackson B. C., J. L. Campos, P. R. Haddrill, B. Charlesworth, and K. Zeng, 2017 Variation in the Intensity of Selection on Codon Bias over Time Causes Contrasting Patterns of Base Composition Evolution in Drosophila. Genome Biol. Evol. 9: 102–123. https://doi.org/10.1093/gbe/evw291

Jackson B., and B. Charlesworth, 2021 Evidence for a force favoring GC over AT at short intronic sites in Drosophila simulans and Drosophila melanogaster. G3 11. https://doi.org/10.1093/g3journal/jkab240

Jeffreys A. J., and R. Neumann, 2002 Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot. Nat. Genet. 31: 267–271. https://doi.org/10.1038/ng910

Kacser H., and J. A. Burns, 1981 The molecular basis of dominance. Genetics 97: 639–666. https://doi.org/10.1093/genetics/97.3-4.639

Kasai H., Y. Kuchino, K. Nihei, and S. Nishimura, 1975 Distribution of the modified nucleoside Q and its derivatives in animal and plant transfer RNA's. Nucleic Acids Res. 2: 1931–1939. https://doi.org/10.1093/nar/2.10.1931

Katoh K., K. Misawa, K. Kuma, and T. Miyata, 2002 MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30: 3059–3066. https://doi.org/10.1093/nar/gkf436

Katze J. R., B. Basile, and J. A. McCloskey, 1982 Queuine, a modified base incorporated posttranscriptionally into eukaryotic transfer RNA: wide distribution in nature. Science 216: 55–56. https://doi.org/10.1126/science.7063869

Katze J. R., U. Gündüz, D. L. Smith, C. S. Cheng, and J. A. McCloskey, 1984 Evidence that the nucleic acid base queuine is incorporated intact into tRNA by animal cells. Biochemistry 23: 1171–1176. https://doi.org/10.1021/bi00301a022

Kern A. D., and D. J. Begun, 2005 Patterns of polymorphism and divergence from noncoding sequences of Drosophila melanogaster and D. simulans: evidence for nonequilibrium processes. Mol. Biol. Evol. 22: 51–62. https://doi.org/10.1093/molbev/msh269

Kersten H., and W. Kersten, 1990 Chapter 2 Biosynthesis and Function of Queuine and Queuosine tRNAs, pp. B69–B108 in *Journal of Chromatography Library*, edited by Gehrke C. W., Kuo K. C. T. Elsevier.

Kimura M., 1968 Evolutionary rate at the molecular level. Nature 217: 624–626. https://doi.org/10.1038/217624a0

Kimura M., and T. Ohta, 1971 Protein polymorphism as a phase of molecular evolution. Nature 229: 467–469. https://doi.org/10.1038/229467a0

King J. L., and T. H. Jukes, 1969 Non-Darwinian evolution. Science 164: 788–798. https://doi.org/10.1126/science.164.3881.788

Kirkpatrick D. T., Y. H. Wang, M. Dominska, J. D. Griffith, and T. D. Petes, 1999 Control of meiotic recombination and gene expression in yeast by a simple repetitive DNA sequence that excludes nucleosomes. Mol. Cell. Biol. 19: 7661–7671. https://doi.org/10.1128/MCB.19.11.7661

Kirtland G. M., T. D. Morris, P. H. Moore, J. J. O'Brian, C. G. Edmonds, *et al.*, 1988 Novel salvage of queuine from queuosine and absence of queuine synthesis in Chlorella pyrenoidosa and Chlamydomonas reinhardtii. J. Bacteriol. 170: 5633–5641. https://doi.org/10.1128/jb.170.12.5633-5641.1988

Kliman R. M., and J. Hey, 1993 Reduced natural selection associated with low recombination in Drosophila melanogaster. Mol. Biol. Evol. https://doi.org/10.1093/oxfordjournals.molbev.a040074

Kliman R. M., and J. Hey, 1994 The effects of mutation and natural selection on codon bias in the genes of Drosophila. Genetics 137: 1049–1056. https://doi.org/10.1093/genetics/137.4.1049

Kliman R. M., 1999 Recent Selection on Synonymous Codon Usage in Drosophila. J. Mol. Evol. 49: 343–351. https://doi.org/10.1007/pl00006557

Kliman R. M., and J. Hey, 2003 Hill-Robertson interference in Drosophila melanogaster: reply to Marais, Mouchiroud and Duret. Genet. Res. 81: 89–90.

Ko W.-Y., R. M. David, and H. Akashi, 2003 Molecular Phylogeny of the Drosophila melanogaster Species Subgroup. J. Mol. Evol. 57: 562–573. https://doi.org/10.1007/s00239-003-2510-x

Kopp A., and J. R. True, 2002 Phylogeny of the Oriental Drosophila melanogaster Species Group: A Multilocus Reconstruction, (J. Whitfield, Ed.). Syst. Biol. 51: 786–805. https://doi.org/10.1080/10635150290102410

Kurtz S., A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway, *et al.*, 2004 Versatile and open software for comparing large genomes. Genome Biol. 5: 1–9. https://doi.org/10.1186/gb-2004-5-2-r12

Langley C. H., K. Stevens, C. Cardeno, Y. C. G. Lee, D. R. Schrider, *et al.*, 2012 Genomic Variation in Natural Populations of Drosophila melanogaster. Genetics 192: genetics.112.142018-598. https://doi.org/10.1534/genetics.112.142018

Lassalle F., S. Périan, T. Bataillon, X. Nesme, L. Duret, *et al.*, 2015 GC-Content evolution in bacterial genomes: the biased gene conversion hypothesis expands. PLoS Genet. 11: e1004941. https://doi.org/10.1371/journal.pgen.1004941

Lewontin R. C., and J. L. Hubby, 1966 A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of Drosophila pseudoobscura. Genetics 54: 595–609.

Lewontin R., 1974 The Genetic Basis of Evolutionary Change. Columbia Univ. Press. 346 p

Li W.-H., 1987 Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. J. Mol. Evol. 24: 337–345. https://doi.org/10.1007/bf02134132

Long H., W. Sung, S. Kucukyildirim, E. Williams, S. F. Miller, *et al.*, 2018 Evolutionary determinants of genome-wide nucleotide composition. Nat Ecol Evol 2: 237–240. https://doi.org/10.1038/s41559-017-0425-y

Mancera E., R. Bourgon, A. Brozzi, W. Huber, and L. M. Steinmetz, 2008 High-resolution mapping of meiotic crossovers and non-crossovers in yeast. Nature 454: 479–485. https://doi.org/10.1038/nature07135

Mank J. E., B. Vicoso, S. Berlin, and B. Charlesworth, 2010 Effective population size and the Faster-X effect: empirical results and their interpretation. Evolution 64: 663–674. https://doi.org/10.1111/j.1558-5646.2009.00853.x

Mansourian S., A. Enjin, E. V. Jirle, V. Ramesh, G. Rehermann, *et al.*, 2018 Wild African Drosophila melanogaster Are Seasonal Specialists on Marula Fruit. Curr. Biol. 28: 3960-3968.e3. https://doi.org/10.1016/j.cub.2018.10.033

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. https://doi.org/10.1073/pnas.091427698

Marais G., D. Mouchiroud, and L. Duret, 2003 Neutral effect of recombination on base composition in Drosophila. Genet. Res. 81: 79–87. https://doi.org/10.1017/s0016672302006079

Marín I., M. L. Siegal, and B. S. Baker, 2000 The evolution of dosage-compensation mechanisms. Bioessays 22: 1106–1114. https://doi.org/10.1002/1521-1878(200012)22:12<1106::AID-BIES8>3.0.CO;2-W

Maside X., A. W. Lee, and B. Charlesworth, 2004 Selection on Codon Usage in Drosophila americana. Curr. Biol. 14: 150–154. https://doi.org/10.1016/j.cub.2003.12.055

Matsumoto T., H. Akashi, and Z. Yang, 2015 Evaluation of Ancestral Sequence Reconstruction Methods to Infer Nonstationary Patterns of Nucleotide Substitution. Genetics 200: 873–890. https://doi.org/10.1534/genetics.115.177386

Matsumoto T., and H. Akashi, 2018 Distinguishing Among Evolutionary Forces Acting on Genome-Wide Base Composition: Computer Simulation Analysis of Approximate Methods for Inferring Site Frequency Spectra of Derived Mutations in Recombining Regions. G3 g3.300512.2017. https://doi.org/10.1534/g3.117.300512

McNamara A. L., and D. W. Smith, 1978 The function of the histidine tRNA isoaccepting species in hemoglobin synthesis. J. Biol. Chem. 253: 5964–5970. https://doi.org/10.1016/S0021-9258(17)34563-5

McVean G. A., and J. Vieira, 2001 Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in Drosophila. Genetics 157: 245–257. https://doi.org/10.1093/genetics/157.1.245

Meier F., B. Suter, H. Grosjean, G. Keith, and E. Kubli, 1985 Queuosine modification of the wobble base in tRNAHis influences ‘in vivo’ decoding properties. EMBO J. 4: 823–827. https://doi.org/10.1002/j.1460-2075.1985.tb03704.x

Meunier J., and L. Duret, 2004 Recombination drives the evolution of GC-content in the human genome. Mol. Biol. Evol. 21: 984–990. https://doi.org/10.1093/molbev/msh070

Mikhaylova L. M., and D. I. Nurminsky, 2011 Lack of global meiotic sex chromosome inactivation, and paucity of tissue-specific gene expression on the Drosophila X chromosome. BMC Biol. 9: 29. https://doi.org/10.1186/1741-7007-9-29

Morgan T. H., 1914 NO CROSSING OVER IN THE MALE OF DROSOPHILA OF GENES IN THE SECOND AND THIRD PAIRS OF CHROMOSOMES. Biol. Bull. 26: 195–204. https://doi.org/10.2307/1536193

Moriyama E. N., and D. L. Hartl, 1993 Codon usage bias and base composition of nuclear genes in Drosophila. Genetics 134: 847–858.

Moriyama E. N., and J. R. Powell, 1996 Intraspecific nuclear DNA variation in Drosophila. Mol. Biol. Evol. 13: 261–277. https://doi.org/10.1093/oxfordjournals.molbev.a025563

Moriyama E. N., and J. R. Powell, 1997 Codon Usage Bias and tRNA Abundance in Drosophila. J. Mol. Evol. 45: 514–523. https://doi.org/10.1007/pl00006256

Morris R. C., K. G. Brown, and M. S. Elliott, 1999 The Effect of Queuosine on tRNA Structure and Function. J. Biomol. Struct. Dyn. 16: 757–774. https://doi.org/10.1080/07391102.1999.10508291

Mount S. M., C. Burks, G. Hertz, G. D. Stormo, O. White, *et al.*, 1992 Splicing signals in Drosophila: intron size, information content, and consensus sequences. Nucleic Acids Res. 20: 4255–4262. https://doi.org/10.1093/nar/20.16.4255

Mugal C. F., P. F. Arndt, and H. Ellegren, 2013 Twisted signatures of GC-biased gene conversion embedded in an evolutionary stable karyotype. Mol. Biol. Evol. 30: 1700–1712. https://doi.org/10.1093/molbev/mst067

Muller H. J., 1932 Some Genetic Aspects of Sex. Am. Nat. 66: 118–138. https://doi.org/10.1086/280418

Müller M., C. Legrand, F. Tuorto, V. P. Kelly, Y. Atlasi, *et al.*, 2019 Queuine links translational control in eukaryotes to a micronutrient from bacteria. Nucleic Acids Res. 47: 3711–3727. https://doi.org/10.1093/nar/gkz063

Muyle A., L. Serres-Giardi, A. Ressayre, J. Escobar, and S. Glémin, 2011 GC-biased gene conversion and selection affect GC content in the Oryza genus (rice). academic. oup. com 28: 2695–2706. https://doi.org/10.1093/molbev/msr104

Nabholz B., A. Künstner, R. Wang, E. D. Jarvis, and H. Ellegren, 2011 Dynamic evolution of base composition: causes and consequences in avian phylogenomics. Mol. Biol. Evol. 28: 2197–2210. https://doi.org/10.1093/molbev/msr047

Nagylaki T., 1983 Evolution of a finite population under gene conversion. Proceedings of the National Academy of Sciences 80: 6278–6281.

Nei M., and T. Gojobori, 1986 Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. https://doi.org/10.1093/oxfordjournals.molbev.a040410

Nielsen R., V. L. Bauer DuMont, M. J. Hubisz, and C. F. Aquadro, 2007 Maximum likelihood estimation of ancestral codon usage bias parameters in Drosophila. Mol. Biol. Evol. 24: 228–235. https://doi.org/10.1093/molbev/msl146

Nishimura S., 1983 Structure, biosynthesis, and function of queuosine in transfer RNA. Prog. Nucleic Acid Res. Mol. Biol. 28: 49–73. https://doi.org/10.1016/s0079-6603(08)60082-3

Noguchi S., Y. Nishimura, Y. Hirota, and S. Nishimura, 1982 Isolation and characterization of an Escherichia coli mutant lacking tRNA-guanine transglycosylase. Function and biosynthesis of queuosine in tRNA. J. Biol. Chem. 257: 6544–6550.

Odenthal-Hesse L., I. L. Berg, A. Veselis, A. J. Jeffreys, and C. A. May, 2014 Transmission distortion affecting human noncrossover but not crossover recombination: a hidden source of meiotic drive. PLoS Genet. 10: e1004106. https://doi.org/10.1371/journal.pgen.1004106

Ogasawara N., 1985 Markedly unbiased codon usage in Bacillus subtilis. Gene 40: 145–150. https://doi.org/10.1016/0378-1119(85)90035-6

Ohgi T., T. Kondo, and T. Goto, 1979 SYNTHESIS OF 7-AMINOMETHYL-7-DEAZAGUANINE, ONE OF THE NUCLEOSIDE Q (QUEUOSINE) PRECURSORS FOR THE POST-TRANSCRIPTIONAL MODIFICATION OF tRNA. Chem. Lett. 8: 1283–1286. https://doi.org/10.1246/cl.1979.1283

Ohta T., 1972 Evolutionary rate of cistrons and DNA divergence. J. Mol. Evol. 1: 150–157. https://doi.org/10.1007/BF01659161

Ohta T., 1973 Slightly deleterious mutant substitutions in evolution. Nature 246: 96–98.

Okada N., and S. Nishimura, 1979 Isolation and characterization of a guanine insertion enzyme, a specific tRNA transglycosylase, from Escherichia coli. J. Biol. Chem. 254: 3061–3066. https://doi.org/10.1016/S0021-9258(17)30182-5

Ott G., H. Kersten, and S. Nishimura, 1982 Dictyostelium discoideum: a useful model system to evaluate the function of queuine and of the Q-family of tRNAs. FEBS Lett. 146: 311–314. https://doi.org/10.1016/0014-5793(82)80941-1

Owenby R. K., M. P. Stulberg, and K. B. Jacobson, 1979 Alteration of the Q family of transfer RNAs in adult Drosophila melanogaster as a function of age, nutrition, and genotype. Mech. Ageing Dev. 11: 91–103. https://doi.org/10.1016/0047-6374(79)90027-7

Parsch J., S. Novozhilov, S. S. Saminadin-Peter, K. M. Wong, and P. Andolfatto, 2010 On the Utility of Short Intron Sequences as a Reference for the Detection of Positive and Negative Selection in Drosophila. Mol. Biol. Evol. 27: 1226–1234. https://doi.org/10.1093/molbev/msq046

Pessia E., A. Popa, S. Mousset, C. Rezvoy, L. Duret, *et al.*, 2012 Evidence for widespread GC-biased gene conversion in eukaryotes. Genome Biol. Evol. 4: 675–682. https://doi.org/10.1093/gbe/evs052

Petes T. D., R. E. Malone, and L. S. Symington, 1991 Recombination in yeast. The Molecular and Cellular Biology of the Yeast Saccharomyces: Genome Dynamics, Protein Synthesis and Energetics 407–521.

Poh Y.-P., C.-T. Ting, H.-W. Fu, C. H. Langley, and D. J. Begun, 2012 Population Genomic Analysis of Base Composition Evolution in Drosophila melanogaster. Genome Biol. Evol. 4: 1245–1255. https://doi.org/10.1093/gbe/evs097

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, (B. F. McAllister, Ed.). PLoS Genet. 2: e173. https://doi.org/10.1371/journal.pgen.0020173

Pool J. E., R. B. Corbett-Detig, R. P. Sugino, K. A. Stevens, C. M. Cardeno, *et al.*, 2012 Population Genomics of Sub-Saharan Drosophila melanogaster: African Diversity and Non-African Admixture, (H. S. Malik, Ed.). PLoS Genet. 8: e1003080. https://doi.org/10.1371/journal.pgen.1003080

Pracana R., A. D. Hargreaves, J. F. Mulley, and P. W. H. Holland, 2020 Runaway GC Evolution in Gerbil Genomes. Mol. Biol. Evol. 37: 2197–2210. https://doi.org/10.1093/molbev/msaa072

Precup J., and J. Parker, 1987 Missense misreading of asparagine codons as a function of codon identity and context. J. Biol. Chem. 262: 11351–11355. https://doi.org/10.1016/S0021-9258(18)60966-4

Procé S. M. de, K. Zeng, A. J. Betancourt, and B. Charlesworth, 2012 Selection on codon usage and base composition in Drosophila americana. Biol. Lett. 8: 82–85. https://doi.org/10.1098/rsbl.2011.0601

RajBhandary U. L., S. H. Chang, H. J. Gross, F. Harada, F. Kimura, *et al.*, 1969 E. COLI TYROSINE TRANSFER RNA - PRIMARY SEQUENCE AND DlRECT EVIDENCE FOR BASE PAIRING BETWEEN THE TERMINAL SEQUENCES. Fed. Proc., Fed. Am. Soc. Exp. Biol. 28: 409.

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. https://doi.org/10.1126/science.1085881

Reyniers J. P., J. R. Pleasants, B. S. Wostmann, J. R. Katze, and W. R. Farkas, 1981 Administration of exogenous queuine is essential for the biosynthesis of the queuosine-containing transfer RNAs in the mouse. J. Biol. Chem. 256: 11591–11594. https://doi.org/10.1016/S0021-9258(19)68443-7

Robinson M., R. Lilley, S. Little, J. S. Emtage, G. Yarranton, *et al.*, 1984 Codon usage can affect efficiency of translation of genes in Escherichia coli. Nucleic Acids Res. 12: 6663–6671. https://doi.org/10.1093/nar/12.17.6663

Robinson M. C., E. A. Stone, and N. D. Singh, 2014 Population genomic analysis reveals no evidence for GC-biased gene conversion in Drosophila melanogaster. Mol. Biol. Evol. 31: 425–433. https://doi.org/10.1093/molbev/mst220

Rogers R. L., J. M. Cridland, L. Shao, T. T. Hu, P. Andolfatto, *et al.*, 2014 Landscape of standing variation for tandem duplications in Drosophila yakuba and Drosophila simulans. Mol. Biol. Evol. 31: 1750–1766. https://doi.org/10.1093/molbev/msu124

Rousselle M., A. Laverré, E. Figuet, B. Nabholz, and N. Galtier, 2019 Influence of Recombination and GC-biased Gene Conversion on the Adaptive and Nonadaptive Substitution Rate in Mammals versus Birds. Mol. Biol. Evol. 36: 458–471. https://doi.org/10.1093/molbev/msy243

Sawyer S. A., D. E. Dykhuizen, and D. L. Hartl, 1987 Confidence interval for the number of selectively neutral amino acid polymorphisms. Proceedings of the National Academy of Sciences 84: 6225–6228. https://doi.org/10.1073/pnas.84.17.6225

Sharp P. M., T. M. Tuohy, and K. R. Mosurski, 1986 Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res. 14: 5125–5143. https://doi.org/10.1093/nar/14.13.5125

Sharp P. M., and W. H. Li, 1987 The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol. Biol. Evol. 4: 222–230. https://doi.org/10.1093/oxfordjournals.molbev.a040443

Shields D. C., P. M. Sharp, D. G. Higgins, and F. Wright, 1988 Silent sites in Drosophila genes are not neutral: evidence of selection among synonymous codons. Mol. Biol. Evol. 5: 704–716. https://doi.org/10.1093/oxfordjournals.molbev.a040525

Siard T. J., K. B. Jacobson, and W. R. Farkas, 1991 Queuine metabolism and cadmium toxicity in Drosophila melanogaster. Biofactors 3: 41–47.

Singh N. D., P. F. Arndt, and D. A. Petrov, 2005 Genomic heterogeneity of background substitutional patterns in Drosophila melanogaster. Genetics 169: 709–722. https://doi.org/10.1534/genetics.104.032250

Singh N. D., V. L. Bauer DuMont, M. J. Hubisz, R. Nielsen, and C. F. Aquadro, 2007 Patterns of mutation and selection at synonymous sites in Drosophila. Mol. Biol. Evol. 24: 2687–2697. https://doi.org/10.1093/molbev/msm196

Singh N. D., P. F. Arndt, A. G. Clark, and C. F. Aquadro, 2009 Strong Evidence for Lineage and Sequence Specificity of Substitution Rates and Patterns in Drosophila. Mol. Biol. Evol. 26: 1591–1605. https://doi.org/10.1093/molbev/msp071

Singhal R. P., and V. N. Vakharia, 1983 The role of queuine in the aminoacylation of mammalian aspartate transfer RNAs. Nucleic Acids Res. 11: 4257–4272. https://doi.org/10.1093/nar/11.12.4257

Smith D. W., and A. L. McNamara, 1982 The effect of the Q base modification on the usage of tRNAHis in globin synthesis. Biochem. Biophys. Res. Commun. 104: 1459–1463. https://doi.org/10.1016/0006-291x(82)91414-0

Smith N. G., and A. Eyre-Walker, 2001 Synonymous codon bias is not caused by mutation bias in G+C-rich genes in humans. Mol. Biol. Evol. 18: 982–986. https://doi.org/10.1093/oxfordjournals.molbev.a003899

Sørensen M. A., C. G. Kurland, and S. Pedersen, 1989 Codon usage determines translation rate in Escherichia coli. J. Mol. Biol. 207: 365–377. https://doi.org/10.1016/0022-2836(89)90260-x

Staubach F., J. F. Baines, S. Künzel, E. M. Bik, and D. A. Petrov, 2013 Host species and environmental effects on bacterial communities associated with Drosophila in the laboratory and in the natural environment. PLoS One 8: e70749. https://doi.org/10.1371/journal.pone.0070749

Sueoka N., 1962 On the genetic basis of variation and heterogeneity of DNA base composition. Proc. Natl. Acad. Sci. U. S. A. 48: 582–592. https://doi.org/10.1073/pnas.48.4.582

Suzuki T., Y. Yashiro, I. Kikuchi, Y. Ishigami, H. Saito, *et al.*, 2020 Complete chemical structures of human mitochondrial tRNAs. Nat. Commun. 11: 4269. https://doi.org/10.1038/s41467-020-18068-6

Szostak J. W., T. L. Orr-Weaver, R. J. Rothstein, and F. W. Stahl, 1983 The double-strand-break repair model for recombination. Cell 33: 25–35. https://doi.org/10.1016/0092-8674(83)90331-8

Takano-Shimizu T., 2001 Local changes in GC/AT substitution biases and in crossover frequencies on Drosophila chromosomes. Mol. Biol. Evol. 18: 606–619. https://doi.org/10.1093/oxfordjournals.molbev.a003841

Tavaré S., 1986 Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences, pp. 57–86 in *Some Mathematical Questions in Biology: DNA Sequence Analysis*, edited by Miura R. M. Lectures on Mathematics in the Life Sciences.

Thurmond J., J. L. Goodman, V. B. Strelets, H. Attrill, L. S. Gramates, *et al.*, 2018 FlyBase 2.0: the next generation. Nucleic Acids Res. 47: D759–D765. https://doi.org/10.1093/nar/gky1003

Tuorto F., R. Liebers, T. Musch, M. Schaefer, S. Hofmann, *et al.*, 2012 RNA cytosine methylation by Dnmt2 and NSun2 promotes tRNA stability and protein synthesis. Nat. Struct. Mol. Biol. 19: 900–905. https://doi.org/10.1038/nsmb.2357

Tuorto F., F. Herbst, N. Alerasool, S. Bender, O. Popp, *et al.*, 2015 The tRNA methyltransferase Dnmt2 is required for accurate polypeptide synthesis during haematopoiesis. EMBO J. 34: 2350–2362. https://doi.org/10.15252/embj.201591382

Tuorto F., C. Legrand, C. Cirzi, G. Federico, R. Liebers, *et al.*, 2018 Queuosine-modified tRNAs confer nutritional control of protein translation. EMBO J. 37: e99777. https://doi.org/10.15252/embj.201899777

Varenne S., J. Buc, R. Lloubes, and C. Lazdunski, 1984 Translation is a non-uniform process. Effect of tRNA availability on the rate of elongation of nascent polypeptide chains. J. Mol. Biol. 180: 549–576. https://doi.org/10.1016/0022-2836(84)90027-5

Vicario S., E. N. Moriyama, and J. R. Powell, 2007 Codon usage in twelve species of Drosophila. BMC Evol. Biol. 7: 1–17. https://doi.org/10.1186/1471-2148-7-226

Walsh J. B., 1983 Role of biased gene conversion in one-locus neutral theory and genome evolution. Genetics 105: 461–468. https://doi.org/10.1093/genetics/105.2.461

White B. N., G. M. Tener, J. Holden, and D. T. Suzuki, 1973 Activity of a transfer RNA modifying enzyme during the development of Drosophila and its relationship to the su(s) locus. J. Mol. Biol. 74: 635–651. https://doi.org/10.1016/0022-2836(73)90054-5

White M. A., M. Wierdl, P. Detloff, and T. D. Petes, 1991 DNA-binding protein RAP1 stimulates meiotic recombination at the HIS4 locus in yeast. Proc. Natl. Acad. Sci. U. S. A. 88: 9755–9759. https://doi.org/10.1073/pnas.88.21.9755

White M. A., M. Dominska, and T. D. Petes, 1993 Transcription factors are required for the meiotic recombination hotspot at the HIS4 locus in Saccharomyces cerevisiae. Proc. Natl. Acad. Sci. U. S. A. 90: 6621–6625. https://doi.org/10.1073/pnas.90.14.6621

Williams A. L., G. Genovese, T. Dyer, N. Altemose, K. Truax, *et al.*, 2015 Non-crossover gene conversions show strong GC bias and unexpected clustering in humans. Elife 4. https://doi.org/10.7554/eLife.04637

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. https://doi.org/10.1016/j.ympev.2006.09.002

Wong A. C.-N., Q.-P. Wang, J. Morimoto, A. M. Senior, M. Lihoreau, *et al.*, 2017 Gut Microbiota Modifies Olfactory-Guided Microbial Preferences and Foraging Decisions in Drosophila. Curr. Biol. 27: 2397-2404.e4. https://doi.org/10.1016/j.cub.2017.07.022

Wright S., 1938 The Distribution of Gene Frequencies Under Irreversible Mutation. Proc. Natl. Acad. Sci. U. S. A. 24: 253–259. https://doi.org/10.1073/pnas.24.7.253

Wu T. C., and M. Lichten, 1994 Meiosis-induced double-strand break sites determined by yeast chromatin structure. Science 263: 515–518. https://doi.org/10.1126/science.8290959

Yang Z., 2007 PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24: 1586–1591. https://doi.org/10.1093/molbev/msm088

Yokoyama S., T. Miyazawa, Y. Iitaka, Z. Yamaizumi, H. Kasai, *et al.*, 1979 Three-dimensional structure of hyper-modified nucleoside Q located in the wobbling position of tRNA. Nature 282: 107–109. https://doi.org/10.1038/282107a0

Zaborske J. M., V. L. B. DuMont, E. W. J. Wallace, T. Pan, C. F. Aquadro, *et al.*, 2014 A nutrient-driven tRNA modification alters translational fidelity and genome-wide protein coding across an animal genus., (H. S. Malik, Ed.). PLoS Biol. 12: e1002015. https://doi.org/10.1371/journal.pbio.1002015

Zallot R., C. Brochier-Armanet, K. W. Gaston, F. Forouhar, P. A. Limbach, *et al.*, 2014 Plant, Animal, and Fungal Micronutrient Queuosine Is Salvaged by Members of the DUF2419 Protein Family. ACS Publications 9: 1812–1825. https://doi.org/10.1021/cb500278k

Zuckerkandl E., and L. Pauling, 1965 Evolutionary Divergence and Convergence in Proteins, pp. 97–166 in *Evolving Genes and Proteins*, edited by Bryson V., Vogel H. J. Academic Press.