Recent codon preference reversals in the *Drosophila melanogaster* lineage

Haruka Yamashita [a,b], Tomotaka Matsumoto [a,b,†], Kent Kawashima [a,§], Hassan Sibroe Abdulla Daanaa [a,b], Ziheng Yang [c], Hiroshi Akashi [a,b,*]

[a] Evolutionary Genetics Laboratory, National Institute of Genetics, Mishima, Shizuoka 411-8540, Japan
[b] Department of Genetics, The Graduate University for Advanced Studies, SOKENDAI, Mishima, Shizuoka 411-8540, Japan
[c] Department of Genetics, Evolution and Environment, University College London, London WC1E 6BT, United Kingdom

Present addresses:
[†] T.M. … SRL and Shizuoka Cancer Center Collaborative Laboratories Inc., Sunto-gun, Shizuoka, 411-8777, Japan
[§] K.K. … Division of Evolutionary Studies of Complex Systems, Research Center for Integrative Evolutionary Science, The Graduate University for Advanced Studies, SOKENDAI, Hayama, Kanagawa 240-0193, Japan

* Correspondence: hiakashi@nig.ac.jp

Abstract

Nearly neutral evolutionary dynamics may be prevalent in genome evolution but the functional effects of weakly selected mutations generally lie outside the range of direct measurement. Synonymous codon usage bias is well-suited for population genetic inference; under major codon preference, translationally superior "major" codons confer fitness benefits relative to their less efficiently and/or accurately decoded synonymous counterparts. The fitness benefits of major codons are likely vanishingly small, but common selective forces among genes justify pooling of mutations with predicted fitness effects to confer statistical power to detect footprints of directional forces. In naturally occurring variation from *Drosophila simulans,* minor to major codon mutations segregate at higher frequencies within populations than major to minor changes. In contrast, in *Drosophila melanogaster*, codon family-specific polymorphism patterns reveal a reduced efficacy of natural selection in most synonymous families, but surprisingly, support reversals of favored states in the four codon families encoded by NAY. These codon families show accelerated synonymous fixations in favor of NAT codons deeper in the gene tree, within the ancestral *D. melanogaster* lineage. Differences for both allele frequencies and fixation rates are greater among X-linked, relative to autosomal, loci. These four lines of evidence are best explained by fitness effect reversals and illustrate how fine-scale sequence pattern analyses can reveal ongoing genome-wide adaptation and motivate experimental investigations of novel function. The correspondence between reversed selection and wobble position queuosine modification in NAY cognate tRNAs is intriguing, but the phenotypic basis of reversals remains unclear.

Introduction

Codon usage bias is a well-developed system for studying evolutionary dynamics under nearly neutral genome evolution. A combination of sequence patterns and experimental findings support "major codon preference" (MCP), co-adaptation between synonymous codon usage and cognate tRNA abundances and their modifications to allow efficient use of the translational machinery (*i.e.*, greater elongation rates and reduced misincorporation) [reviewed in (Andersson and Kurland 1991)]. A wide range of microbes, as well as multi-cellular eukaryotes, show elevated usage of particular codons within synonymous families in highly expressed genes [reviewed in (Akashi 2001; Duret 2002)]. Such codons are generally recognized by abundant and/or non-wobble-pairing cognate tRNAs and are referred to as "major" codons (Ikemura 1985). Biochemical studies support both elevated elongation rates (Varenne *et al.* 1984; Curran and Yarus 1989; Dana and Tuller 2014) and reduced misincorporations (Precup and Parker 1987; Kramer and Farabaugh 2007) at major codons relative to their minor counterparts. Putative major codons can be identified through characterizations of tRNA pools [*e.g.*, (Ikemura 1981, 1982)], experimental measures of translation rate [*e.g.*, (Varenne *et al.* 1984)], or by compositional trends among genes [*e.g.*, (Grantham *et al.* 1981; Shields *et al.* 1988; Sharp and Devine 1989; Lloyd and Sharp 1992; Stenico *et al.* 1994; Akashi 1995; Kanaya *et al.* 1999)].

The main features of MCP can be captured in relatively simple evolutionary models. Observed levels of codon usage bias are consistent with a balance among weak evolutionary forces including mutation, genetic drift and weak natural selection (Li 1987; Bulmer 1991). Under MCP, mutations from minor to major codons confer small fitness benefits and mutations in the opposite direction are deleterious with similar magnitudes (Akashi 1995); prediction of fitness classes of synonymous changes is a key advantage for evolutionary analyses. In addition, because translational effects of

synonymous codon usage should be similar among genes, data can be pooled among loci to enhance the statistical power to detect minute fixation biases (*i.e.*, forces that alter expected allele frequencies from generation to generation in a consistent direction). These features facilitate the study of evolutionary dynamics under a balance of weak forces and previous population genetic analyses have demonstrated directional forces that favor major codons over their synonymous counterparts [*e.g.*, (Akashi 1995; Akashi and Schaeffer 1997; Kliman 1999; Sharp *et al.* 2010; Jackson *et al.* 2017)]. GC-biased gene conversion can also underlie base compositional biases (Marais 2003; Duret and Galtier 2009) and distinguishing between natural selection and non-selective fixation biases has proven challenging. In addition, a number of empirical studies support beneficial effects of translational pauses in regulating protein folding or membrane insertion [reviewed in (Komar 2021)]. Major codons may be disadvantageous in some contexts but such cases may be more strongly selected and relatively rare.

Here, we take advantage of genome-scale data from population samples of closely related Drosophila species to pursue evolutionary analyses at an increased resolution. We test fixation biases within individual synonymous families both among mutations segregating in populations and among fixations in ancestral lineages. Our analyses support MCP favoring predominantly G- and C-ending codons in *Drosophila simulans* but reveal appreciable heterogeneity in the magnitude of selective forces among synonymous families. In contrast, the *Drosophila melanogaster* lineage shows little evidence of fixation biases on synonymous mutations for most synonymous families. Population genetics theory predicts long-term stability of codon preferences under co-adaptation between codon usage and tRNA pools, but surprisingly, we found that a subset of synonymous families, those encoded by NAY codons, show strong evidence for recent codon preference reversals in the *D. melanogaster* lineage. These findings, together with greater contrasts in both the site frequency

spectrum and fixation patterns at X-linked, relative to autosomal, loci make a compelling case for genome-wide shifts in NAY codon family adaptation.

Results and Discussion

We analyzed available population genomic and outgroup DNA sequence data among closely related species from the *Drosophila melanogaster* subgroup. We constructed CDS and intron sequence alignments for 21 *D. simulans* lines from a Madagascar population (Rogers *et al.* 2014; Jackson *et al.* 2017) and 14 *D. melanogaster* lines from a Rwandan population (Pool *et al.* 2012). Reference sequences for *D. yakuba* and *D. erecta* (Drosophila 12 Genomes Consortium 2007), serve as outgroups (Fig. 1a). Classifying synonymous mutations by their predicted fitness effects, a key step for population genetic analyses of codon usage (Akashi 1995), requires inference of ancestral and derived states at polymorphic sites. The relatively short branch lengths in our gene tree allows reliable inference using a likelihood based method that incorporates both biased and fluctuating base composition [(Matsumoto *et al.* 2015; Matsumoto and Akashi 2018); see Materials and Methods].

Inferring fixation biases from D. simulans *and* D. melanogaster *polymorphism*

Directional forces impact allele frequencies of mutations segregating within populations. We can infer such forces by examining distributions of allele frequencies among variants segregating in a population, the site frequency spectrum (SFS). Direct comparisons of SFS between classes of mutations that are interspersed within DNA (Bulmer 1971; Sawyer *et al.* 1987) can be a robust statistical approach that is sensitive to weak forces (Akashi and Schaeffer 1997; Akashi 1999) including natural selection and GC-biased gene conversion (gBGC).

SFS comparisons support directional forces favoring G/C over A/T at both 2-fold and 4-fold redundant sites in *D. simulans* (Supplemental Table S4). The results expand on previous findings that combined data from 2-fold and 4-fold synonymous families for smaller numbers of genes (Akashi and Schaeffer 1997; Kliman 1999) and that employed only 4-fold synonymous families in

larger data sets (Jackson *et al.* 2017). These comparisons distinguish between mutational and fixation biases because the former affects the *numbers* of segregating mutations but not their frequencies within the population (assuming that mutation biases remain constant over the relevant time period). The results are also consistent with compositional trend analyses (Shields *et al.* 1988; Akashi 1995; Vicario *et al.* 2007) that supported G- and/or C-ending major codons in all synonymous families in *D. melanogaster* (compositional trends are similar in *D. simulans*).

Genome-scale data allow us to refine the SFS analyses to a resolution of individual synonymous families. The following analyses will focus on 2-fold redundant sites and the term "synonymous" will refer to this class unless noted otherwise. SFS analyses for 4-fold redundant sites in CDS are presented in Supplemental Material Section 2. In *D. simulans*, each of the 10 synonymous families shows SFS differences similar to the general pattern of elevated within-population sample frequencies of GC-increasing (W→S) compared with AT-increasing (S→W) mutations (Fig. 2a; Supplemental Table S5). The results are consistent with MCP predictions, but SFS for intron mutations also indicate a GC fixation bias (Supplemental Table S3) as noted in previous studies (Kliman 2014; Jackson *et al.* 2017; Jackson and Charlesworth 2021; Yıldırım and Vogl 2024). GC fixation biases that are common to synonymous and intron mutations can include biased gene conversion and/or fitness differences unrelated to MCP.

We employ a summary statistic to capture the magnitude of differences in SFS comparisons in order to distinguish among underlying directional forces. Glémin *at al.* (2015) estimated a fixation bias statistic (selection coefficient or conversion parameter scaled to population size) from SFS comparisons between forward and reverse mutations (see Materials and Methods). We employ GC pref γ as a measure of the magnitude of SFS differences between W→S and S→W mutations (Supplemental Fig. S1). Directional forces favor G and/or C-ending codons across synonymous

families in *D. simulans* (Fig. 3a) and the magnitude of GC fixation bias is heterogeneous among synonymous families. Intron sequences also show GC fixation biases consistent with gBGC and/or natural selection. Larger GC pref γ at synonymous sites than at intron sites supports additional directional forces favoring G and/or C-ending codons in coding regions (such as natural selection) as predicted under MCP. Kliman (2014) found similar patterns among synonymous families as well as introns in polymorphism data from an autosome in *Drosophila pseudoobscura*.

The numbers of polymorphisms from genes on the X chromosome are sufficient for comparisons to autosomal loci and stronger fixation biases at X-linked loci are a conspicuous feature in *D. simulans* that holds across synonymous families (Fig. 3a; Wilcoxon signed-rank test $p = 0.0051$). Potential causes are difficult to resolve, however, and include greater efficacy of natural selection acting on partially recessive fitness effects (Charlesworth *et al.* 1987), GC elevation related to dosage compensation (Alekseyenko *et al.* 2012) or a greater impact of gBGC on the X chromosome.

In contrast to trends that generally conform to expectations under MCP in *D. simulans*, synonymous family-specific analyses reveal remarkable patterns in *D. melanogaster* polymorphism. Several synonymous families show SFS supporting fixation bias in the opposite direction to that observed in *D. simulans*: S→W mutations are segregating at higher frequencies than W→S mutations. The pattern holds for the four NAY synonymous families (codons encoding Asp, Asn, His, and Tyr) but is not found outside of these families (Fig. 3b; Supplemental Table S6). The Lys synonymous family (AAR) shows SFS supporting GC fixation bias; W→S mutations segregate at higher frequencies than S→W mutations but the magnitude of fixation bias is small (Fig. 2b; Supplemental Table S6). All other 2-fold synonymous families show indistinguishable SFS at both autosomal and X-linked loci (Supplemental Table S6) consistent with previous analyses of pooled data (Akashi 1995).

D. melanogaster polymorphisms reveal differences in evolutionary processes for X-linked vs autosomal loci, an "X effect", for NAY families. Fig. 3b shows consistent support, across the four NAY synonymous families, for stronger fixation biases favoring NAT over NAC at X-linked than at autosomal genes. Thus, *D. simulans* and *D. melanogaster* both show elevated fixation biases at X-linked genes, but in opposing directions, *i.e.*, favoring GC and AT, respectively. AT preference at NAY codons in *D. melanogaster* may reflect partially recessive fitness effects and/or other forces that override fixation biases toward GC (*i.e.*, gBGC or dosage compensation constraints). The cause(s) of X chromosome vs autosome differences in GC fixation biases in *D. simulans* are ambiguous as discussed above. However, the X effect for NAT fixation bias in *D. melanogaster* is difficult to explain in the absence of lineage-specific fitness benefits for NAT over NAC codons. Mutation-driven scenarios require highly specified changes in mutational patterns (*i.e.*, specific to NAY codons, greater on the X chromosome, and timed near the MRCA of the species polymorphism) to explain NAY polymorphism patterns in *D. melanogaster*.

The analyses above support genome-wide differences in the direction of fixation biases at NAY codons in *D. melanogaster* and *D. simulans* but do not address the location of the preference shift within the evolutionary tree. Compositional trend analyses indicate NAC preference at all NAY families in species within the *melanogaster* group as well as in the more distantly related *obscura* group [(Akashi and Schaeffer 1997; Vicario *et al.* 2007); see also below]. SFS analyses further support NAC preference in *D. pseudoobscura* (Kliman 2014). The fixation bias patterns discussed above are most simply explained by a reversal of codon preference specific to NAY families in the *D. melanogaster* lineage after its split with *D. simulans*.

Fixation biases in the ancestral lineages of D. simulans *and* D. melanogaster

The SFS analyses above focused on polymorphisms (*i.e.*, mutations segregating within populations) in *D. simulans* and *D. melanogaster* and revealed strong evidence for AT fixation biases specific to NAY codons in *D. melanogaster*. We also examined patterns deeper in the evolutionary histories of these species; fixations on the *D. simulans* ancestral (*ms-s'*) and *D. melanogaster* ancestral (*ms-m'*) branches can reveal changes in mutation and/or fixation biases in the relatively short lineages prior to the *s'* and *m'* nodes (Fig. 1b). Previous analyses of pooled synonymous families showed weak AT content increases in the *D. simulans* ancestral lineage (Begun 2001; Kern and Begun 2005; Akashi *et al.* 2006; Singh *et al.* 2009; Jackson *et al.* 2017) and stronger AT content increases in the *D. melanogaster* ancestral lineage consistent with reduced GC fixation biases (Akashi 1995, 1996; Kern and Begun 2005; Akashi *et al.* 2006; Singh *et al.* 2009; Poh *et al.* 2012; Jackson *et al.* 2017).

We employ a summary statistic to capture the direction and extent of departures from GC content equilibrium: GC fixation skew, $d_{WS,SW} = (a - b) / (a + b)$, where $a = N_{W\rightarrow S}$, the number of W→S fixations and $b = N_{S\rightarrow W}$, the number of S→W fixations. Expected GC fixation skew is zero at steady-state GC content and differs in sign, but is scaled symmetrically, for GC-increasing and GC-decreasing departures from equilibrium.

GC content decline is prevalent in the ancestral *D. simulans* lineage among mutation classes. However, GC fixation skew is *positive* for the presumed best candidates for neutral evolution, introns with the lowest GC content (Fig. 4a; Supplemental Fig. S4). We attribute GC elevation at putatively neutrally evolving sites to an increase in mutational pressure toward GC. From coding regions, we analyzed synonymous changes within NAY and "non-NAY" families (synonymous families encoding Phe, Cys, Ser_2, Lys, Gln, and Glu) separately. For all three mutation classes (*i.e.*,

intron, non-NAY and NAY), loss of GC intensifies with ancestral GC content (Fig. 4a; Supplemental Fig. S6c). Such near-linear trends are consistent with simple MCP scenarios of non-stationary mutation ratio and/or fixation bias (Akashi *et al.* 2007). In particular, the *D. simulans* GC fixation skew patterns appear roughly consistent with a combination of increase in GC mutation pressure and reduction in GC fixation bias (Supplemental Fig. S6).

We tested whether GC content changes are consistent with shared parameter fluctuations among mutation classes. Data were partitioned so that comparisons are conducted within similar ancestral GC-content ranges (Supplemental Table S10). Intron and NAY classes show indistinguishable GC fixation skews in *D. simulans* (Supplemental Table S10). Non-NAY codons show a weak signal toward less AT-biased fixation skews than introns and NAY codons (Supplemental Table S10), but low fixation counts on the *ms-s'* branch (Fig. 1) prevent firm conclusions.

The ancestral lineage for *D. melanogaster* (Fig. 1) shows strong negative relationships between GC change and ancestral GC content for intron, non-NAY, and NAY classes. The slopes of these trends are steeper than in *D. simulans* (Fig. 4). Low-GC introns, our assumed neutrally evolving class, show an elevation of GC similar to the pattern in *D. simulans* (Fig. 4b) consistent with a shift in GC-elevating mutation bias that may predate the *ms* node. Stronger negative GC fixation skew slopes can be explained by larger fixation bias reductions in the ancestral *D. melanogaster* lineage (Akashi *et al.* 2007). Notably, NAY codons show accelerated AT content shifts compared to non-NAY codons within this lineage (Supplemental Table S10). Extents of GC content change were statistically indistinguishable between intron sites and non-NAY codons (Supplemental Table S10). Such patterns are predicted under a combination of fixation biases *favoring* S→W mutations at NAY codons *and* reduced GC fixation biases at intron sites and non-NAY codons.

X effects are a compelling feature in the SFS analyses (see above) and GC fixation skew comparisons between X-linked vs. autosomal genes reveal similarly notable patterns. We focus on

D. melanogaster because the low number of fixations on the *D. simulans* lineage limit the exploration of X effects (findings are discussed in Supplemental Material Section 3.2). In *D. melanogaster*, all four mutation classes show small, but statistically significant, elevations of AT gains at X-linked loci compared to autosomal loci (Supplemental Fig. S7e-h and Table S11). Such patterns could reflect X-specific changes in the mutation rate ratio (elevated AT bias) and/or fixation biases (greater relaxation of GC bias) that are shared among mutation classes. Notably, NAY codons show significantly greater AT gains at X-linked, than at autosomal loci, than other mutation classes (Supplemental Fig. S7h and Table S12). Enhanced efficacy of NAT codon preference at X-linked, relative to autosomal, loci for both polymorphism (post-*m'* node) and fixations (*ms-m'* nodes) in the *D. melanogaster* lineage provides a simple explanation for this pattern and would imply partially recessive fitness advantages for NAC to NAT mutations.

Functional basis of **D. melanogaster** *preference reversals at NAY codons*

Shared properties of NAY codons raise several possibilities for the biological cause(s) of the reversal of fitness benefits of NAT over NAC codons in the *D. melanogaster* lineage. These codons have common nucleotide composition at the 2nd and 3rd codon positions, show relatively weak fixation biases in *D. simulans* in SFS analyses (Fig. 3a), and are recognized by cognate tRNAs that undergo a particular chemical modification: queuosine (Q) modification at the anticodon 3rd (wobble) position G nucleoside. Q modification is observed in the majority of eukaryotes (Harada and Nishimura 1972; White *et al.* 1973a; Kasai *et al.* 1975; Zallot *et al.* 2014) and is not known to occur for cognate tRNAs for other synonymous families (Fergus *et al.* 2015). Many eubacteria can synthesize the Q precursor, queine, but eukaryotes lack the required biochemical pathways and sequester the micronutrient from their diet and/or gut microbiota. Because queine levels are strongly

dependent on diet in *Drosophila* (White *et al.* 1973b; Jacobson *et al.* 1981; Zaborske *et al.* 2014), lineage-specific modification levels (related to nutrient availability) have been proposed to explain codon preference differences (Chiari *et al.* 2010; Zaborske *et al.* 2014).

Fitness benefits of Q modification are clear from the retention of this pathway in a wide range of eukaryotes, but the functional basis of the benefit remains difficult to determine and may differ markedly among taxa [reviewed in (Fergus *et al.* 2015)]. Such benefits could lie outside of impacts on translational elongation/accuracy; for example, tRNA-modification affects rates of breakdown of isoacceptors into tRNA-derived small RNAs that function in gene regulation and stress response (Wang *et al.* 2018; Muthukumar *et al.* 2024). Experimental studies of protein synthesis support similar translation rates at NAC and NAT codons for unmodified cognate tRNAs and higher translational elongation rates at NAC codons for Q-modified tRNAs (Meier *et al.* 1985). Changes in Q modification can have measurable effects on translation rates at both cognate and non-cognate codons and tissue- and sex- specific effects in mice (Cirzi *et al.* 2023). Importantly, compositional trend analyses consistently support translational preference of NAC over NAT in organisms that employ Q modification [*Escherichia coli* (Kanaya *et al.* 1999), *Bacillus subtilis* (Kanaya *et al.* 1999), *Saccharomyces pombe* (Kanaya *et al.* 2001), and *Caenorhabditis elegans* (Stenico *et al.* 1994; Duret and Mouchiroud 1999)] as well as in those that lack the modification [*Saccharomyces cerevisiae* (Akashi 2003), *Candida albicans* (Lloyd and Sharp 1992) and *Arabidopsis thaliana* (Duret and Mouchiroud 1999; Wright *et al.* 2004)]. These patterns argue against a simple relationship between Q modification levels and major codon identity; in the limited number of established cases of loss of Q modification, translation selection appears to favor NAC codons regardless of Q modification.

We tested for signals for major codon reassignment (reversals in translational advantages) as a cause of NAT preference in *D. melanogaster*. Such a scenario seems unlikely given the general pattern of reduced selection intensity for major codons in this lineage; TTY (Phe) and AAR (Lys), the most strongly selected 2-fold families in *D. simulans* (Fig. 3a) and *D. pseudoobscura* (Kliman 2014), show weakly differentiated or non-distinguishable SFS in *D. melanogaster* (Fig. 3b; Supplemental Table S6). We also tested for evidence for increases in fitness benefits as a function of translation rate, a hallmark feature of MCP. We expect greater fixation biases for minor to major codon mutations among genes that show the strongest patterns of ancestral MCP. In *D. simulans*, both NAY and non-NAY codons show greater GC fixation biases in genes with higher major codon usage at the *ms* node (Supplemental Fig. S18). If NAT preference in *D. melanogaster* is driven by translational selection, then we expect similarly larger absolute magnitudes of fixation bias in genes under stronger translational selection. We detected no signal for such a trend; magnitudes of fixation bias are indistinguishable between high and low MCU genes (Supplemental Fig. S18). These results compel us to consider whether benefits of NAT codon usage may fall outside of standard MCP notions in *D. melanogaster*. Functional interactions with tRNA-derived small RNAs are an intriguing possibility but remain speculative and difficult to test in sequence patterns.

Codon compositional trends among distantly related Drosophila: no indications of major codon transitions among 2-fold synonymous families

Under MCP, the fitness benefits of major codons increase with the number of translation events at a given codon. Because this number is shared among codons within a gene, putative major codons within each synonymous family can be identified as those with elevated representation in genes under stronger translational selection (*i.e.*, those that have strong codon usage bias and/or that show

high estimates of translation rate or associated measures such as transcript abundance). However, potential contributions of mutational variation (*e.g.*, transcription associated mutation) and/or fixation biases other than translational selection (*e.g.*, gBGC) to such trends need to be considered carefully [*e.g.*, (Kanaya *et al.* 1999; Bénitière *et al.* 2024); Supplemental Fig. S3].

Compositional trend analyses have consistently supported G- and C-ending codon preference within non-NAY synonymous families in a wide range of Drosophila species (Akashi and Schaeffer 1997; Heger and Ponting 2007b; Vicario *et al.* 2007) but have yielded heterogeneous and, in some cases, ambiguous results for major codons in NAY families. We examined support for NAT preference in four distantly related Drosophila species (*D. willistoni*, *D. mojavensis*, *D. grimshawi*, and *D. virilis*), for which major codon shifts within NAY families have been reported (Heger and Ponting 2007b; Vicario *et al.* 2007; Zaborske *et al.* 2014). We also include *D. pseudoobscura* as a distant relative of *D. melanogaster* with strong support for MCP from population genetic studies (Akashi and Schaeffer 1997; Kliman 2014; Fuller *et al.* 2014). We employed G/C preference at non-NAY codons as a predictor of intensity for translational selection; this notion is supported by elevated usage of G- or C-ending codons at putatively highly expressed genes (*i.e.*, *D. melanogaster* high transcript abundance genes and their 1-to-1 OrthoFinder orthologs; Supplemental Fig. S16). NAC codon usage (*i.e.*, $CUB_{Chi/L}$ NAY; see Materials and Methods) increases as a function of this proxy for the strength of MCP (Fig. 5). The associations differ considerably among genomes with steep slopes in *D. melanogaster*, *D. pseudoobscura*, and *D. mojavensis* and more moderate slopes in *D. grimshawi*, *D. willistoni*, and *D. virilis*. However, all correlations are statistically significant and we did not find any support for NAT preference in these analyses of pooled synonymous families (Fig. 5; Supplemental Fig. S8). We found similar patterns for individual synonymous families (Supplemental Figs. S9-S15).

We considered factors that potentially confound evidence for NAT preference. Intron base composition could reflect mutational biases that may be related to transcription rates (or that vary among genes for other reasons). In addition, as we noted previously [see also (Jackson and Charlesworth 2021)], intron sites show a weak GC fixation bias in *D. simulans* with elevated bias among introns showing higher ancestral GC (Supplemental Figs. S2 and S4). Intron GC content is positively correlated with $CUB_{Chi/L}$ for non-NAY families in all species we examined (Supplemental Fig. S3). GC-elevating forces within introns, if they are shared with synonymous sites within coding regions of the genes in which they reside, could mask the signal of AT preference acting on synonymous mutations. To control for this possibility, we employed bin-specific GC content of introns in the CUB_{Chi/L_binexp} calculations. Signals of GC elevation in high $CUB_{Chi/L}$ genes remained strong in non-NAY families (Supplemental Figs. S8-S14), but patterns for several NAY families changed from strongly significant GC elevation in high $CUB_{Chi/L}$ genes to no association in CUB_{Chi/L_binexp} analyses (Supplemental Figs. S11, S12, and S14). We did not detect statistical support for NAT elevation in high $CUB_{Chi/L}$ genes in any of the examined genomes (Supplemental Fig S15). Note that if GC-elevating forces are not shared between introns and synonymous sites, CUB_{Chi/L_binexp} associations will be biased toward negative relationships.

The discussion above assumed dual translational advantages, faster elongation (Dana and Tuller 2014) and elevated accuracy (Sun and Zhang 2022), at major codons compared to minor codons within synonymous families. Zaborske and co-workers (2014) proposed that such parallel fitness benefits only hold at NAY codons under conditions of high Q modification levels in cognate tRNAs populations. They argued that fitness benefits conflict under reduced Q modification levels; misincorporation rates are lower at NAT codons than at NAC codons but elongation rates remain higher at NAC codons. Our findings suggest that translational accuracy-driven NAT preference is unlikely to underlie codon preference reversals in the *D. melanogaster* lineage. Fixation biases for

major codon usage have been ineffective in this lineage (both in the polymorphism and fixation analyses above). In addition, we did not detect the associations between constraint at the protein level and magnitudes of fixation bias predicted under accuracy-driven codon preference (Akashi 1994) in both NAY polymorphism (SFS) and fixation analyses (Supplemental Fig. S19 and Table S13).

The switch from NAC to NAT preference in *D. melanogaster* is surprising because MCP is a scenario of co-adaptation between tRNA pools and genome-wide codon usage that should be stable over evolutionary time (Bulmer 1987). A reversal in major/minor codon identity will have a fitness impact equivalent to thousands of slightly deleterious major to minor codon mutations. The total impact will depend on the numbers of major to minor codons in genome and the fitness difference between the states. Two factors likely mitigated the mainly conservative nature of major codon identity under MCP. First, in *D. simulans* (Fig. 3a) and *D. pseudoobscura* (Kliman 2014), GC fixation biases are weaker at NAY than at other synonymous families. In addition, major codon usage was lower at these codons in the common ancestor compared to non-NAY synonymous families (Fig. 4a; Supplemental Table S5). These factors likely reduced the overall deleterious effect of a genome-wide conversion of translationally advantageous NAC codons to less favored states.

Increased resolution of micro-evolutionary analyses revealed prominent signals of codon preference reversals at NAY codons in the *D. melanogaster* lineage. Genome-scale resequencing data allowed us to partition data, both to synonymous families and to short evolutionary lineages. This was essential because the main patterns of interest were restricted to a subset of synonymous families within short branches. Previous analyses of fixations in the ancestral *D. melanogaster* lineage yielded little support for AT preference for pooled synonymous sites (with a notable exception at the *Notch* locus) using a statistical approach that assumed shared fixation bias

parameters among synonymous families (Bauer DuMont *et al.* 2004; Nielsen *et al.* 2007; Singh *et al.* 2007).

Sequence-based evolutionary analyses strongly support recent genome-wide preference reversals at NAY codons in the *D. melanogaster* lineage and illustrate how population genetic inference can inform functional and evolutionary ecological analyses. NAY preference reversals appear to be a previously undetected, recent adaptation in a model organism and provides an opportunity to characterize its functional basis, potentially unrelated to MCP, using a well-developed array of molecular and genetic tools.

Materials and Methods

Abbreviations are summarized in Supplemental Table S1.

DNA sequences

We employed available genome data for isofemale lines established from natural populations of *D. melanogaster* and *D. simulans*. Data from a Rwanda population of *D. melanogaster* (Pool *et al.* 2012) was downloaded from http://www.dpgp.org/dpgp2/DPGP2.html (last downloaded on 22nd July 2014). 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 proportions of ambiguous nucleotides (RG4N, RG7 and RG33). We extracted predicted CDS and intron sequences from the genomic sequences for a total of 13,691 protein-coding genes using FlyBase r6.24 *D. melanogaster* genome annotations (last downloaded on 12th July 2019; Supplemental Material Sections 6.1 and 7).

We reconstructed genome sequences for *D. simulans* within-species samples using available data (last downloaded on 1st March 2017; Supplemental Material Sections 6.1 and 7). We analyzed the DNA variant information for 21 lines from a Madagascar population (Rogers *et al.* 2014; Jackson *et al.* 2017); 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 William Ballard, respectively, as described in Jackson *et al.* (2017). We extracted

predicted CDS and intron sequences from the genome sequences for a total of 13,831 protein-coding genes using FlyBase r2.02 *D. simulans* genome annotations (last downloaded on 9th March 2020).

Genome sequences from species within the *D. melanogaster* subgroup, *D. yakuba* and *D. erecta* were employed as outgroup data. We extracted predicted CDS and intron sequences from the genome sequences using genome annotations for FlyBase r1.05 *D. yakuba* (last downloaded on 12th July 2019) and FlyBase r1.05 *D. erecta* (last downloaded on 12th July 2019).

We constructed CDS and intron sequence alignments including *D. melanogaster* and *D. simulans* within-species samples and outgroup sequences (see Supplemental Material Section 8). Custom Python codes used in the alignment pipeline are available at (github link to be added). The data set includes 10,122 CDS and 18,719 intron alignments for autosomal loci and 1,746 CDS and 2,705 intron alignments for X-linked loci which are available at (zenodo link to be added).

Inference of polymorphic changes

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

We estimated probabilities of nucleotides at ancestral nodes in the gene tree shown in Fig. 1b. 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 composition is changing within 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 incorporates uncertainty in ancestral inference. We inferred ancestral nucleotides for both within- and between-species variation using a combination of BASEML (Yang 2007) and the Bifurcating Tree with Weighting (BTW) approach (Matsumoto and Akashi 2018). For BASEML ancestral inference, we employed the GTR-NH_b nucleotide substitution model (Tavaré 1986; Matsumoto *et al.* 2015) with 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) collapsed sequence pairs in *D. melanogaster* and *D. simulans* (terminal branches leading to m_{c1}, m_{c2} and s_{c1}, s_{c2}, respectively, in Fig. 1b). Ancestral inference was conducted separately for data from autosomal and X-linked loci. We refer to inferred changes on the m'-m_{c1} and m'-m_{c2} branches as polymorphisms in *D. melanogaster* and those on the s'-s_{c1} and s'-s_{c2} branches as polymorphisms in *D. simulans*. We employed probabilities of changes as counts for the numbers of polymorphic mutations for each of 12 mutation classes (Supplemental Material Section 9.2).

Polymorphism analysis

We analyzed SFS for forward and reverse mutations between pairs of nucleotides (*e.g.*, A→G vs G→A). We use the term "forward" mutations for W→S changes (where W indicates A or T, and S indicates G or C) among GC-altering mutations and for T→A and C→G among GC-conservative mutations. Mutations in the opposite direction are termed "reverse" mutations (designations of the terms "forward" and "reverse" are arbitrary). 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). Mann-Whitney U (MWU) tests were employed to test for SFS differences among synonymous and intron mutations. 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 the magnitude of fixation biases using a maximum likelihood method (Glémin *et al.* 2015) that fits observed SFS to theoretical expectations (Fisher 1930; Wright 1938). SFS of putatively neutral mutations (here, intron GC-conservative mutations; see also Supplemental Table S3) 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 "GC pref γ", the fixation bias statistic. Positive and negative values of GC pref γ indicate fixation biases that elevate and reduce GC content, respectively. This statistic 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, γ estimates are strongly correlated with MWU test statistics scaled to sample size (Supplemental Fig. S1); thus γ can be considered a summary statistic for the magnitude of difference between SFS that can be utilized for comparisons across mutation classes, X-linked and autosomal loci, and across species.

Inference of fixations in coding regions

We inferred synonymous and replacement changes within the *D. simulans* ancestral (*ms-s'*) and *D. melanogaster* ancestral (*ms-m'*) lineages by combining an approach introduced in Akashi *et al.* (2006) with the method described above (BTW using the GTR-NH_b model). This approach reconstructs internal node codon configurations (INCC) by combining separate estimations of internal node nucleotide configurations (INNC) for three codon positions, assuming independent processes among positions (Akashi *et al.* 2006).

We started with the base alignments for autosomal loci (see Supplemental Material Section 8). We inferred ancestral codons at internal nodes in the tree shown in Fig. 1b. A pair of ancestral and derived codons that differ at a single position will be referred to as a "codon change" (*e.g.*, AAA at the *ms* node and AAG at the *m'* node at a given codon position in the aligned CDS is an AAA→AAG codon change in the *ms-m'* branch). The probability of INCC is used as a "count" of the change. For the case of codon pairs that differ at two or three positions, we calculated the relative probability for each possible minimal step paths from ancestral and derived codons. These probabilities were weighted by the numbers of synonymous and nonsynonymous changes involved in a path (Nei and Gojobori 1986) using the nonsynonymous-to-synonymous substitution rate ratio calculated from codon pairs that differ at a single position (d_N / d_S = 0.1234 for autosomal loci; d_N / d_S = 0.1434 for X-linked loci). We used the Nei and Gojobori (1986) method to count the numbers of synonymous and nonsynonymous sites.

In our analyses, "fixations" refer to changes inferred on the *ms-s'* and *ms-m'* branches and are thus "fixed in the sample" and may include mutations segregating at high frequencies within the populations (these changes do not overlap with data for the SFS analysis). We also note that our approach of fixation counting does not consider multiple changes at a site within a branch. This can

cause underestimation of fixation counts, but the effect is expected to be minor given the low divergence levels between *D. simulans* and *D. melanogaster* (Matsumoto *et al.* 2015; Matsumoto and Akashi 2018).

GC fixation skew analysis

To infer changes in fixation bias and/or mutation rate bias, we compared $d_{\mathrm{WS,SW}}$ among genes with different GC content at the *ms* node (GC_{ms}). GC_{ms} is the sum of probabilities for INNC with G or C nucleotides at the *ms* node divided by the number of "sites" (*i.e.*, the total probabilities for INNC) for introns and serves as a proxy for ancestral fixation bias. We used a similar approach to calculate GC_{ms} at synonymous sites, the proportion of G- and/or C-ending codons at the *ms* node. GC_{ms} are used for binning introns and CDS into non-overlapping bins with similar numbers of sites (nucleotide positions for intron and codons for CDS). We filtered introns and CDS that include fewer than 10 aligned sites. Ancestral reconstructions were resampled in units of intron and CDS. Statistical approaches to detect differences in GC fixation skews between mutation classes are described in Supplemental Material Section 3.

Measure of codon usage bias

We employed a variant of the "scaled χ^2" statistic (Shields *et al.* 1988; Akashi 1995) as a measure of the deviation of synonymous codon usage from a putatively neutral expectation. Goodness-of-fit tests are designed to compare G+C counts and A+T counts at synonymous positions to expectations based on observed GC content at short introns. The χ^2 statistics from such tests are signed to indicate whether the proportion of G- or C-ending codons among synonymous codons is

greater (positive) or smaller (negative) than expected GC. χ^2 values are summed across synonymous families and the sum is divided by the number of codons to give a measure that is not dependent on sample size (*i.e.*, numbers of codons), codon usage bias ($CUB_{Chi/L}$).

Compositional trend analysis

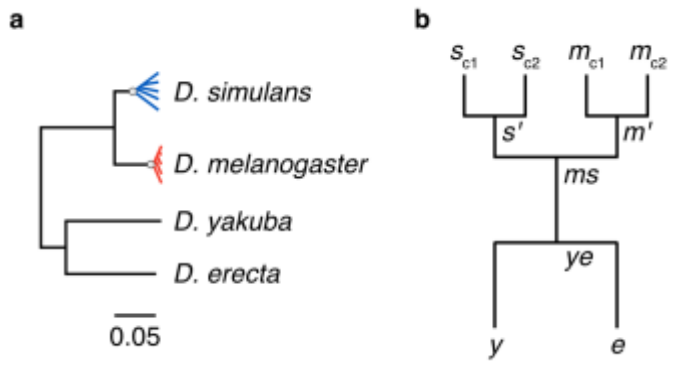
We employed six distantly related Drosophila species: *D. melanogaster*, *D. pseudoobscura*, *D. willistoni*, *D. grimshawi*, *D. mojavensis*, and *D. virilis*. For *D. melanogaster* and *D. pseudoobscura*, we used CDS and intron sequences to construct the *D. melanogaster* subgroup data set. For other species, we extracted predicted CDS and intron sequences from the genome-scale DNA sequences using annotations. We downloaded DNA sequences and annotations from NCBI (https://www.ncbi.nlm.nih.gov/assembly/) for *D. willistoni* (101, GCF_000005925.1), for *D. grimshawi* (103, GCF_018153295.1), for *D. mojavensis* (102, GCF_018153725.1), and for *D. virilis* (103, GCF_003285735.1). These data from NCBI were downloaded on 13th October 2021. For genes with multiple CDS isoforms, we used the longest CDS. See Supplemental Material Section 10 for details of filtering.

We obtained short intron sequences using transcript annotations. We excluded introns in untranslated regions and intron sites that are included in coding regions in one or more of transcript isoforms. Within introns, we filtered 10 bases at 5′ and 30 bases at 3′ splice junctions to reduce the contribution of functional constraint (Halligan *et al.* 2004; Halligan and Keightley 2006; Parsch *et al.* 2010). Only introns with 10 or more nucleotides remaining after these filtering steps were included in the analyses.

Major codons increase in usage within their synonymous family in genes experiencing higher MCP-related fixation biases. We tested associations between synonymous family-specific $CUB_{Chi/L}$ and non-NAY $CUB_{Chi/L}$ (excluding the synonymous family being tested). Our proxy for MCP-related fixation biases is supported by previous findings of G- or C-ending major codons for non-NAY synonymous families in the six Drosophila species under analysis (Heger and Ponting 2007a; Vicario *et al.* 2007). We ranked genes by $CUB_{Chi/L}$ non-NAY and assigned genes into 15 bins (minimum 70,000 non-NAY codons per bin). $CUB_{Chi/L}$ for each bin was calculated from combined codon frequencies among CDS within the bin. We employed Spearman's rank correlation to assess the associations.

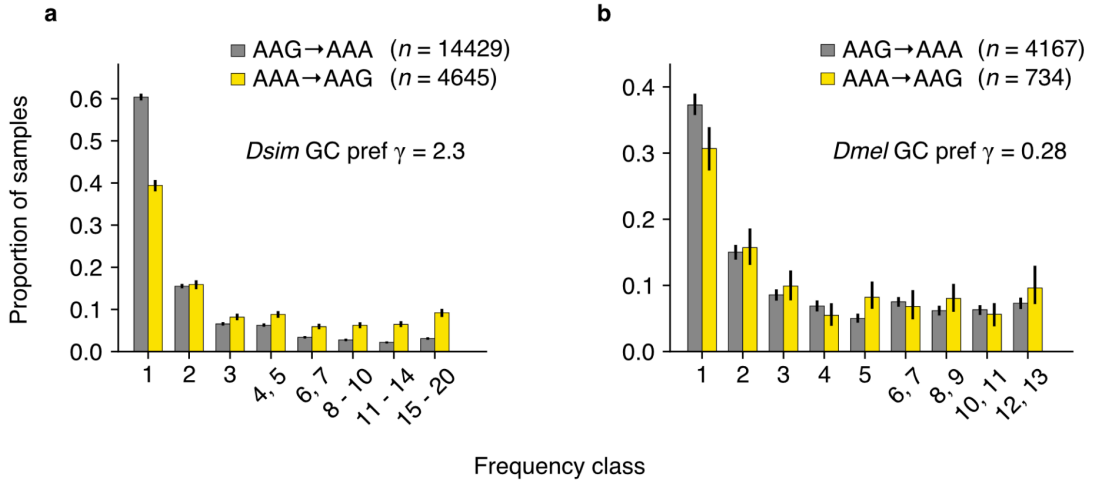
Figures

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



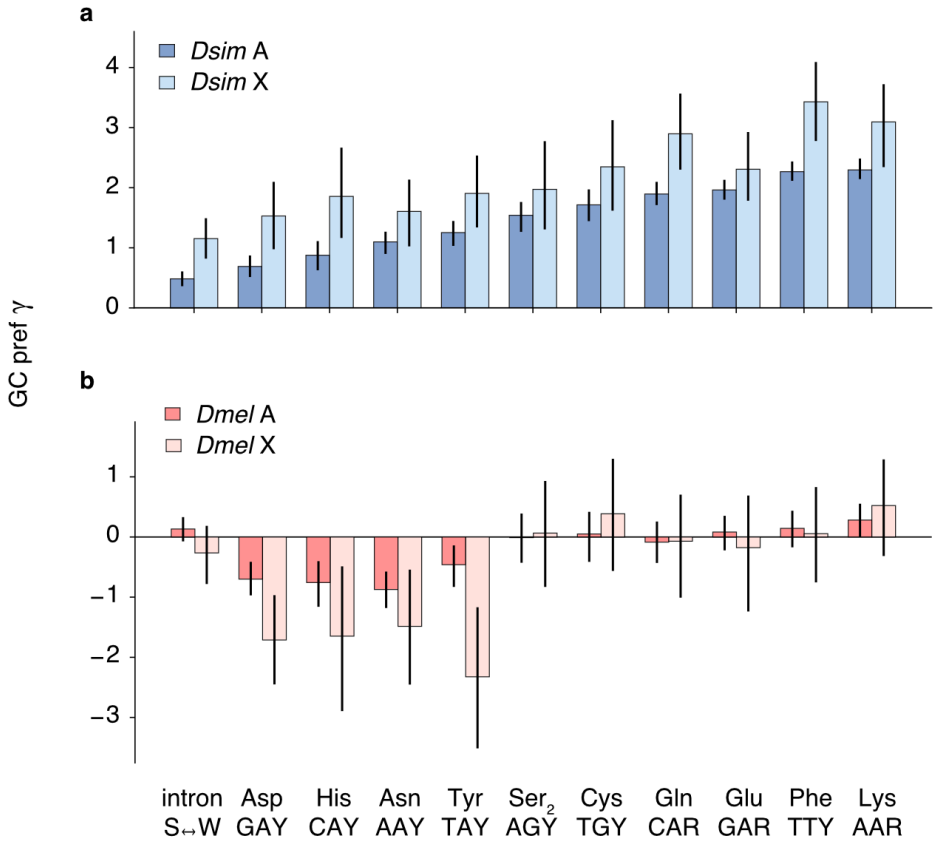
Population genetic analyses employed data from *D. simulans* (*Dsim*), *D. melanogaster* (*Dmel*), and two outgroups, *D. yakuba* (*Dyak*) and *D. erecta* (*Dere*); other species in the subgroup are not shown. (a) Genetic distances among DNA sequences from the four species. Branch lengths are numbers of nucleotide differences per site at autosomal short introns. The within-species gene trees are rough depictions based on the strong and weaker excesses of rare polymorphisms ("star-like" trees) compared to neutral equilibrium expectations in *Dmel* and *Dsim*, respectively. Gray circles indicate approximate positions of the MRCA within each population sample. If a single allele is sampled from a population, a proportion of the changes inferred on the terminal branch will occur at polymorphic sites; the post-MRCA distances reflect the expected numbers of such changes assuming independence among observed polymorphisms. (b) Tree topology employed for ancestral inference. Within-species variation was assigned to two sequences (m_{c1} and m_{c2} for *Dmel*, s_{c1} and s_{c2} for *Dsim,* see Materials and Methods). Ancestral nodes for within-species sequences are indicated as *s'* for *Dsim* and *m'* for *Dmel*. Ancestral nodes for between-species sequences are labeled *ms* and *ye* for the pairs *Dmel* / *Dsim* and *Dyak* / *Dere*, respectively. This topology is assumed for ancestral inference (note that branch lengths are estimated in the process).

Fig. 2. Site frequency spectra and fixation bias estimates: an example from Lysine codons in *D. simulans* and *D. melanogaster.*



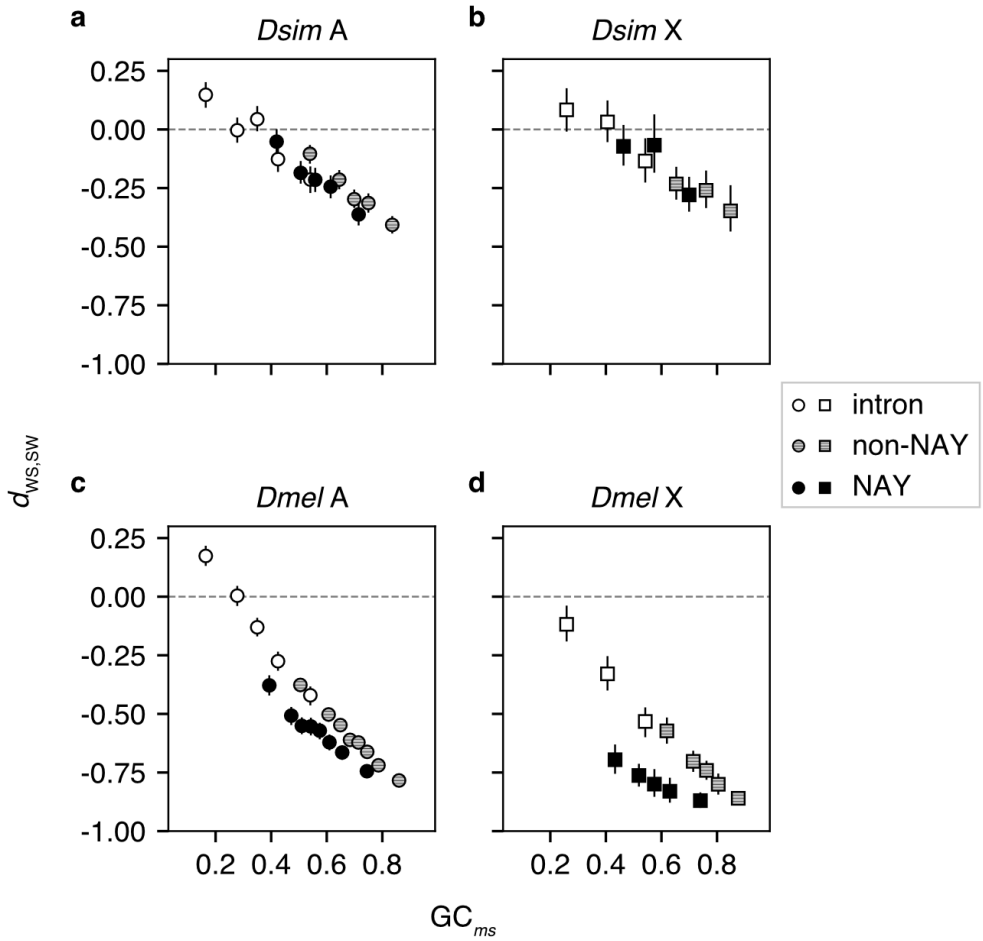
SFS for AAG→AAA (gray) is compared to that for AAA→AAG (yellow). (a) SFS for *D. simulans* (*Dsim*). (b) SFS for *D. melanogaster* (*Dmel*). See Supplemental Tables S5 and S6 for the results of statistical tests for *D. simulans* and *D. melanogaster*, respectively. Counts (*n*) for the two polymorphism classes are shown (values are rounded to integers). Data from some frequency classes are pooled as indicated on the x-axis labels. "GC pref γ" indicate maximum likelihood estimates of G- and C-favoring fixation biases using the approach of Glémin *et al.* (2015). SFS for intron GC-conservative changes are employed as a putatively neutral reference for the GC pref γ estimation. Independent ancestral inference was performed for bootstrap replicates. Error bars indicate 95% CIs among 300 bootstrap replicates.

Fig. 3. Synonymous family-specific fixation biases inferred from *D. simulans* and *D. melanogaster* site frequency spectra.



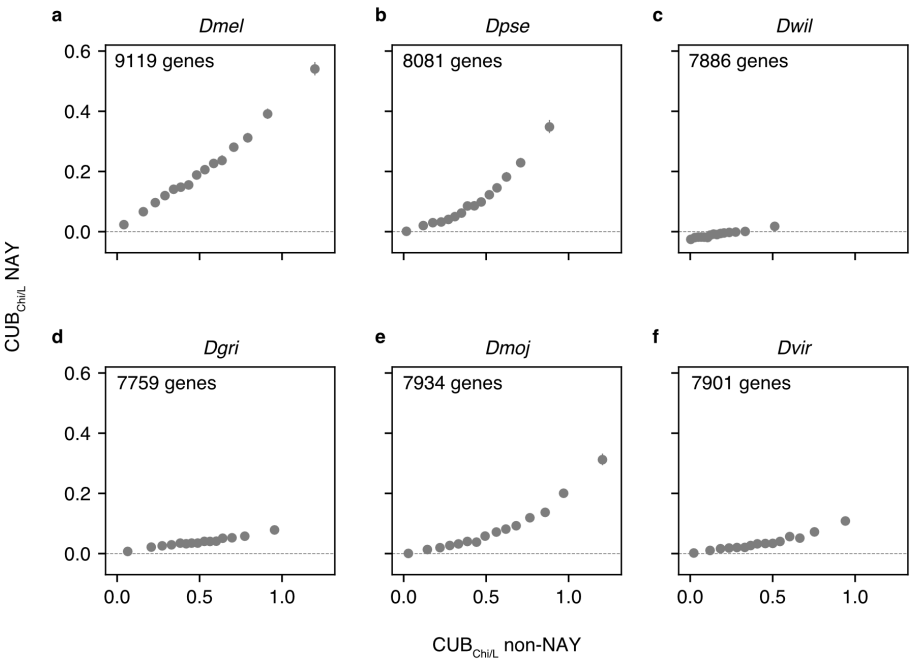
Fixation biases for GC-altering changes, GC pref γ, estimated using SFS data are shown for (a) *D. simulans* (*Dsim*) and (b) *D. melanogaster* (*Dmel*). Negative GC pref γ values indicate fixation bias favoring AT. "intron" refers to short introns. Autosomal (darker colors; "A") and X-linked (lighter colors; "X") loci are analyzed separately. SFS for GC-conservative changes within short introns from autosomal and X-linked loci are employed as required neutral references for GC pref γ estimation for the corresponding coding region data. Synonymous families are arranged in the order of γ values for *D. simulans* autosomal loci. GC pref γ values are presented in Supplemental Tables S5 and S6, for *D. simulans* and *D. melanogaster*, respectively. X-axis labels are shared between the top and bottom panels. Independent ancestral inference was performed for each bootstrap replicate. Error bars indicate 95% CIs among 300 bootstrap replicates.

Fig. 4. GC content changes in the *D. simulans* and *D. melanogaster* ancestral lineages.



GC fixation skew statistics, $d_{WS,SW} = (N_{W\rightarrow S} - N_{S\rightarrow W}) / (N_{W\rightarrow S} + N_{S\rightarrow W})$, where $N_{W\rightarrow S}$ is W→S fixation count and $N_{S\rightarrow W}$ is S→W fixation count. $d_{WS,SW}$ indicates the direction and magnitude of the departure from GC content equilibrium ($d_{WS,SW} = 0$). *Changes* in GC content are plotted against GC content at *ms* node, GC_{ms}. Fixations are inferred changes within internal branches (a and b) *ms-s for D. simulans* fixation data and (c and d) *ms-m'* for *D. melanogaster* fixation data (Fig. 1b). Autosomal (a and c) and X-linked (b and d) are analyzed separately. "intron" indicates short introns. Synonymous families are pooled into two classes: non-NAY (Gln, Glu, Lys, Phe, Cys, and Ser_2) and NAY (Asp, Asn, His, and Tyr). For binning, introns or CDS are ranked by GC_{ms} for the mutation class and assigned to bins with similar numbers of intron sites or codons, respectively. Statistical analyses are presented in Supplemental Table S10. GC fixation skews are compared between A vs X in Supplemental Fig. S7. Error bars indicate 95% CIs among 1000 bootstrap replicates.

Fig. 5. Compositional trends at NAY codons among distantly related Drosophila.



Compositional biases toward NAC at NAY codons ($CUB_{Chi/L}$ NAY) is plotted against $CUB_{Chi/L}$ for non-NAY codons, a proxy for magnitudes of translational selection. Data are from autosomal loci for (a) *D. melanogaster* (*Dmel*) and (b to f) their 1-to-1 OrthoFinder-orthologs for the other species: (b) *D. pseudoobscura* (*Dpse*), (c) *D. willistoni* (*Dwil*), (d) *D. grimshawi* (*Dgri*), (e) *D. mojavensis* (*Dmoj*), and (f) *D. virilis* (*Dvir*). Expected GC content was calculated using sites within short introns. Positive $CUB_{Chi/L}$ values indicate greater G-ending or C-ending codon usage than expected. Genes are ranked by $CUB_{Chi/L}$ non-NAY and divided into 15 bins with similar numbers of non-NAY codons. Codon frequencies are pooled among CDS within a bin to calculate $CUB_{Chi/L}$ NAY and $CUB_{Chi/L}$ non-NAY. Error bars indicate 95% CIs among 1000 bootstrap replicates.

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., 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., 2001 Gene expression and molecular evolution. Curr. Opin. Genet. Dev. 11: 660–666. https://doi.org/10.1016/s0959-437x(00)00250-1

Akashi H., 2003 Translational selection and yeast proteome evolution. Genetics 164: 1291–1303.

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

Akashi H., P. Goel, and A. John, 2007 Ancestral inference and the study of codon bias evolution: implications for molecular evolutionary analyses of the *Drosophila melanogaster* subgroup. PLoS One 2: e1065. https://doi.org/10.1371/journal.pone.0001065

Alekseyenko A. A., J. W. K. Ho, S. Peng, M. Gelbart, M. Y. Tolstorukov, *et al.*, 2012 Sequence-specific targeting of dosage compensation in Drosophila favors an active chromatin context. PLoS Genet. 8: e1002646. https://doi.org/10.1371/journal.pgen.1002646

Andersson G. E., and C. G. Kurland, 1991 An extreme codon preference strategy: codon reassignment. Mol. Biol. Evol. 8: 530–544. https://doi.org/10.1093/oxfordjournals.molbev.a040666

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

Bénitière F., T. Lefébure, and L. Duret, 2024 Variation in the fitness impact of translationally optimal codons among animals. bioRxiv 2024.07.22.604600. [accessed 2024 Aug 30]

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

Bulmer M., 1987 Coevolution of codon usage and transfer RNA abundance. Nature 325: 728–730. https://doi.org/10.1038/325728a0

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

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

Chiari Y., K. Dion, J. Colborn, A. Parmakelis, and J. R. Powell, 2010 On the possible role of tRNA base modifications in the evolution of codon usage: queuosine and Drosophila. J. Mol. Evol. 70: 339–345. https://doi.org/10.1007/s00239-010-9329-z

Cirzi C., J. Dyckow, C. Legrand, J. Schott, W. Guo, *et al.*, 2023 Queuosine-tRNA promotes sex-dependent learning and memory formation by maintaining codon-biased translation elongation speed. EMBO J. 42: e112507. https://doi.org/10.15252/embj.2022112507

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

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

Dana A., and T. Tuller, 2014 The effect of tRNA levels on decoding times of mRNA codons. Nucleic Acids Res. 42: 9171–9181. https://doi.org/10.1093/nar/gku646

Drosophila 12 Genomes Consortium, 2007 Evolution of genes and genomes on the Drosophila phylogeny. Nature 450: 203–218. https://doi.org/10.1038/nature06341

Duret L., and D. Mouchiroud, 1999 Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. Proc. Natl. Acad. Sci. U. S. A. 96: 4482–4487. https://doi.org/10.1073/pnas.96.8.4482

Duret L., 2002 Evolution of synonymous codon usage in metazoans. Curr. Opin. Genet. Dev. 12: 640–649. https://doi.org/10.1016/s0959-437x(02)00353-2

Duret L., and N. Galtier, 2009 Biased gene conversion and the evolution of mammalian genomic landscapes. Annu. Rev. Genomics Hum. Genet. 10: 285–311. https://doi.org/10.1146/annurev-genom-082908-150001

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

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

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

Fuller Z. L., G. D. Haynes, D. Zhu, M. Batterton, H. Chao, *et al.*, 2014 Evidence for stabilizing selection on codon usage in chromosomal rearrangements of *Drosophila pseudoobscura*. G3 4: 2433–2449. https://doi.org/10.1534/g3.114.014860

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

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

Halligan D. L., A. Eyre-Walker, P. Andolfatto, and P. D. Keightley, 2004 Patterns of Evolutionary Constraints in Intronic and Intergenic DNA of Drosophila. Genome Res. 14: 273–279. https://doi.org/10.1101/gr.1329204

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

Harada F., and S. Nishimura, 1972 Possible anticodon sequences of $tRNA^{His}$, $tRNA^{Asn}$, 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

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

Heger A., and C. P. Ponting, 2007b Variable strength of translational selection among 12 Drosophila species. Genetics 177: 1337–1348. https://doi.org/10.1534/genetics.107.070466

Ikemura T., 1981 Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: A proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. J. Mol. Biol. 151: 389–409. https://doi.org/10.1016/0022-2836(81)90003-6

Ikemura T., 1982 Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes: Differences in synonymous codon choice patterns of yeast and *Escherichia coli* with reference to the abundance of isoaccepting transfer RNAs. J. Mol. Biol. 158: 573–597. https://doi.org/10.1016/0022-2836(82)90250-9

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

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: jkab240. https://doi.org/10.1093/g3journal/jkab240

Jacobson K. B., W. R. Farkas, and J. R. Katze, 1981 Presence of queuine in Drosophila melanogaster: correlation of free pool with queuosine content of tRNA and effect of mutations in pteridine metabolism. Nucleic Acids Res. 9: 2351–2366. https://doi.org/10.1093/nar/9.10.2351

Kanaya S., Y. Yamada, Y. Kudo, and T. Ikemura, 1999 Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. Gene 238: 143–155. https://doi.org/10.1016/s0378-1119(99)00225-5

Kanaya S., Y. Yamada, M. Kinouchi, Y. Kudo, and T. Ikemura, 2001 Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. J. Mol. Evol. 53: 290–298. https://doi.org/10.1007/s002390010219

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

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

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., 2014 Evidence that natural selection on codon usage in *Drosophila pseudoobscura* varies across codons. G3 4: 681–692. https://doi.org/10.1534/g3.114.010488

Komar A. A., 2021 A code within a code: How codons fine-tune protein folding in the cell. Biochemistry (Mosc.) 86: 976–991. https://doi.org/10.1134/S0006297921080083

Kramer E. B., and P. J. Farabaugh, 2007 The frequency of translational misreading errors in *E. coli* is largely determined by tRNA competition. RNA 13: 87–96. https://doi.org/10.1261/rna.294907

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

Lloyd A. T., and P. M. Sharp, 1992 Evolution of codon usage patterns: the extent and nature of divergence between *Candida albicans* and *Saccharomyces cerevisiae*. Nucleic Acids Res. 20: 5289–5295. https://doi.org/10.1093/nar/20.20.5289

Marais G., 2003 Biased gene conversion: implications for genome and sex evolution. Trends Genet. 19: 330–338. https://doi.org/10.1016/S0168-9525(03)00116-1

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 8: 1755–1769. https://doi.org/10.1534/g3.117.300512

Meier F., B. Suter, H. Grosjean, G. Keith, and E. Kubli, 1985 Queuosine modification of the wobble base in $tRNA^{His}$ influences "*in vivo*" decoding properties. EMBO J. 4: 823–827. https://doi.org/10.1002/j.1460-2075.1985.tb03704.x

Muthukumar S., C.-T. Li, R.-J. Liu, and C. Bellodi, 2024 Roles and regulation of tRNA-derived small RNAs in animals. Nat. Rev. Mol. Cell Biol. 25: 359–378. https://doi.org/10.1038/s41580-023-00690-z

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). Mol. Biol. Evol. 28: 2695–2706. https://doi.org/10.1093/molbev/msr104

Nei M., and T. Gojobori, 1986 Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. 3: 418–426. 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

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

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

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. PLoS Genet. 8: e1003080. https://doi.org/10.1371/journal.pgen.1003080

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

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

Sawyer S. A., D. E. Dykhuizen, and D. L. Hartl, 1987 Confidence interval for the number of selectively neutral amino acid polymorphisms. Proc. Natl. Acad. Sci. U. S. A. 84: 6225–6228. https://doi.org/10.1073/pnas.84.17.6225

Sharp P. M., and K. M. Devine, 1989 Codon usage and gene expression level in *Dictyostelium discoideum*: highly expressed genes do “prefer” optimal codons. Nucleic Acids Res. 17: 5029–5039. https://doi.org/10.1093/nar/17.13.5029

Sharp P. M., L. R. Emery, and K. Zeng, 2010 Forces that influence the evolution of codon bias. Philos. Trans. R. Soc. Lond. B Biol. Sci. 365: 1203–1212. https://doi.org/10.1098/rstb.2009.0305

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

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

Stenico M., A. T. Lloyd, and P. M. Sharp, 1994 Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases. Nucleic Acids Res. 22: 2437–2446. https://doi.org/10.1093/nar/22.13.2437

Sun M., and J. Zhang, 2022 Preferred synonymous codons are translated more accurately: Proteomic evidence, among-species variation, and mechanistic basis. Sci Adv 8: eabl9812. https://doi.org/10.1126/sciadv.abl9812

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.

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

Wang X., Z. Matuszek, Y. Huang, M. Parisien, Q. Dai, *et al.*, 2018 Queuosine modification protects cognate tRNAs against ribonuclease cleavage. RNA 24: 1305–1313. https://doi.org/10.1261/rna.067033.118

White B. N., G. M. Tener, J. Holden, and D. T. Suzuki, 1973a 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 B. N., G. M. Tener, J. Holden, and D. T. Suzuki, 1973b Analysis of tRNAs during the development of Drosophila. Dev. Biol. 33: 185–195. https://doi.org/10.1016/0012-1606(73)90173-5

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

Wright S. I., C. B. K. Yau, M. Looseley, and B. C. Meyers, 2004 Effects of gene expression on molecular evolution in *Arabidopsis thaliana* and *Arabidopsis lyrata*. Mol. Biol. Evol. 21: 1719–1726. https://doi.org/10.1093/molbev/msh191

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

Yıldırım B. C., and C. Vogl, 2024 The influence of GC-biased gene conversion on nonadaptive sequence evolution in short introns of *Drosophila melanogaster*. J. Evol. Biol. 37: 383–400. https://doi.org/10.1093/jeb/voae015

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. 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