Codon Usage Selection Can Bias Estimation of the Fraction of Adaptive Amino Acid Fixations

Tomotaka Matsumoto,[1] Anoop John,[‡,1] Pablo Baeza-Centurion,[§,¶,1] Boyang Li,[‖,1] and Hiroshi Akashi*[,1,2]

[1]Division of Evolutionary Genetics, National Institute of Genetics, Yata, Mishima, Shizuoka, Japan

[2]Department of Genetics, The Graduate University for Advanced Studies (SOKENDAI), Yata, Mishima, Shizuoka, Japan

[‡]Present address: CTO, Zyxware Technologies, Trivandrum—Kerala, India
[§]Present address: Centre for Genomic Regulation (CRG), Barcelona, Spain
[¶]Present address: Universitat Pompeu Fabra (UPF), Barcelona, Spain
[‖]Present address: School of Management Science and Engineering, Dongbei University of Finance and Economics, Dalian, P.R.China

***Corresponding author:** E-mail: hiakashi@nig.ac.jp.

Associate editor: John November

Abstract

A growing number of molecular evolutionary studies are estimating the proportion of adaptive amino acid substitutions (α) from comparisons of ratios of polymorphic and fixed DNA mutations. Here, we examine how violations of two of the model assumptions, neutral evolution of synonymous mutations and stationary base composition, affect α estimation. We simulated the evolution of coding sequences assuming weak selection on synonymous codon usage bias and neutral protein evolution, $\alpha = 0$. We show that weak selection on synonymous mutations can give polymorphism/divergence ratios that yield α-hat (estimated α) considerably larger than its true value. Nonstationary evolution (changes in population size, selection, or mutation) can exacerbate such biases or, in some scenarios, give biases in the opposite direction, α-hat $< \alpha$. These results demonstrate that two factors that appear to be prevalent among taxa, weak selection on synonymous mutations and non-steady-state nucleotide composition, should be considered when estimating α. Estimates of the proportion of adaptive amino acid fixations from large-scale analyses of *Drosophila melanogaster* polymorphism and divergence data are positively correlated with codon usage bias. Such patterns are consistent with α-hat inflation from weak selection on synonymous mutations and/or mutational changes within the examined gene trees.

Key words: **McDonald–Kreitman test, adaptive protein evolution, codon bias, base composition.**

Introduction

Mounting evidence supports that positive selection (i.e., fitness benefits of new amino acid-altering mutations) has played a predominant role in protein divergence in a wide range of taxa including microbes, insects, plants, and vertebrates (reviewed in Eyre-Walker 2006; Sella et al. 2009; Fay 2011). One of the main approaches for estimating the proportion of adaptive substitutions compares sampled numbers of polymorphic and fixed differences between functional classes of mutations interspersed in DNA (Sawyer et al. 1987; McDonald and Kreitman 1991; Akashi 1999). In general, a test mutation class (amino acid altering, or replacement changes) is compared to a presumably neutrally evolving control class of variation (usually synonymous changes). Reduced ratios of polymorphic to fixed amino acid replacement changes (i.e., excess amino acid fixations) are generally interpreted as evidence of adaptive protein evolution.

The proportion of beneficial amino acid substitutions, α, can be estimated from replacement and synonymous polymorphism/divergence ratios (Charlesworth 1994; Smith and Eyre-Walker 2002). Here, we will use $\hat{\alpha}_{MK}$ for α estimated (Smith and Eyre-Walker 2002) from 2 × 2 contingency tables of counts of synonymous and nonsynonymous polymorphic sites (P_s, P_n) and fixed differences (D_s, D_n) (McDonald and Kreitman 1991). Some key assumptions of the standard approach include 1) neutral nonsynonymous polymorphism (i.e., adaptive substitutions have a short transit time within populations and are unlikely to be sampled), 2) constant mutation rates within each mutation class over the time interval examined, and 3) neutral evolution of synonymous mutations. Several modifications of the simple $\hat{\alpha}_{MK}$ approach attempt to relax these assumptions. Filtering rare variants attempts to reduce the contribution of weakly deleterious polymorphism (Fay et al. 2001) and several methods have been developed to control for the combined effect of demographic history and weakly deleterious mutations (Andolfatto 2008; Ellegren 2008; Eyre-Walker and Keightley 2009; Parsch et al. 2009; Eilertson et al. 2012; Messer and Petrov 2013).

Here, we examine how two factors, selection on codon usage and nonstationary base composition (arising from changes in mutation and/or selection intensity) impact estimates of adaptive protein evolution. Such violations of the standard assumptions are supported in microbes (Ikemura 1985; Sharp et al. 2010; Agashe and Shankar 2014), *Drosophila* (Akashi 1995; Powell and Moriyama 1997; Moriyama and Powell 1997; Rodríguez-Trelles et al. 2000; Akashi et al.

© The Author 2016. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.
All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

doi:10.1093/molbev/msw027 Advance Access publication February 12, 2016

2006; Vicario et al. 2007), plants (Wright et al. 2006), and mammals (Duret et al. 2002). We show, through computer simulation, that estimates of adaptive protein evolution are sensitive to assumptions of neutrality of synonymous mutations as well as compositional equilibrium and that estimation biases can be quite strong ($\hat{\alpha}_{MK} - \alpha > 0.3$). $\hat{\alpha}_{MK}$ bias predicts a positive correlation between ancestral codon bias and estimates of adaptive evolution. We observe such associations in *Drosophila melanogaster* polymorphism and divergence data and argue that estimation biases related to selection and/or fluctuating mutation may have inflated $\hat{\alpha}_{MK}$. Quantifying the extent of bias will require a better understanding of lineage-specific evolutionary processes among the species examined.

Results

$\hat{\alpha}_{MK}$ under Selection and Nonstationary Evolution at Synonymous Sites: Simulation Results

Major codon preference inflates r_{pd} (ratio of the numbers of polymorphic sites and fixations) for synonymous changes compared with the neutral expectation and can cause false positives for adaptive protein evolution (Akashi 1995). Thus, we expected selection at synonymous sites to bias $\hat{\alpha}_{MK}$ as a function of selection intensity. Figure 1 shows $\hat{\alpha}_{MK}$ from seven different simulation scenarios (described in table 1). Details of the simulation are explained in Materials and Methods. In scenario *st*, major codon usage (MCU, the proportion of major codons) remains at steady-state (all parameters remain constant throughout the simulation). Selection at synonymous site causes overestimation of α and the extent of bias increases considerably with selection intensity at synonymous sites (fig. 1). At moderate codon bias, MCU $\approx$ 0.7, $\hat{\alpha}_{MK}$ was roughly 0.2 (20% adaptive amino acid fixations inferred under an assumption of neutral synonymous evolution). At high levels of codon bias, MCU $\approx$ 0.95, $\hat{\alpha}_{MK}$ reached 0.6 (>60% inferred adaptive amino acid fixations). Note that our simulations considered only 2-fold redundant sites and all synonymous mutations affect fitness. Higher redundancy codon families can include neutral synonymous mutations which will reduce bias in $\hat{\alpha}_{MK}$. Supplementary table S1, Supplementary Material online, shows estimated numbers of polymorphic and fixed differences and r_{pd} values for each simulation scenario. For scenario *st*, r_{pd} remains stable across MCU values for neutral replacement changes (recombination rates are high and linked selection from codon bias is negligible). However, r_{pd} increases for 1⟶0 mutations in the range of selection coefficients examined (Akashi 1995). In addition, the proportion of ancestrally preferred codons (and thus the per locus 1⟶0 mutation rate) increases with MCU. These two factors have a considerable effect on r_{pd} for synonymous changes (pooled preferred and unpreferred mutations) and, consequently, on $\hat{\alpha}_{MK}$.

We next consider nonstationary codon bias. Burn-in and initial parameters were identical to scenario *st* but mutation and/or selection parameter changes were introduced at time t_{15k} leading to increases or decreases in MCU. Figure 1A shows $\hat{\alpha}_{MK}$ in decreasing MCU scenarios. Scenario u_{i15k} implemented a 50% elevation of *u*, the 1⟶0 mutation rate, at time t_{15k}. The increased mutation rate increased numbers of polymorphic sites and also slightly elevated numbers of fixations; r_{pd} increased for both synonymous and replacement changes relative to scenario *st* (supplementary table S1, Supplementary Material online). Without MCU selection (MCU $\approx$ 0.4), r_{pd}'s for synonymous and replacement changes are similarly affected by the increased mutation rate because synonymous and replacement sites have similar base composition (fig. 1A). However, under MCU selection, synonymous sites have ancestrally higher proportions of 1's. Thus, the increased 1⟶0 mutation rate elevates synonymous r_{pd} (relative to replacement or neutral r_{pd}) and intensifies overestimation of α compared with scenario *st*.

Scenario N_{d15k} considers a population size decrease to one-third the initial value at time t_{15k}. In this scenario, the time to most recent common ancestor (TMRCA) for segregating sites becomes closer to the present and the numbers of fixations are elevated. In addition, expected numbers of polymorphic sites are reduced in smaller N_e, and r_{pd} decreases for both synonymous and replacement changes. Decreased population size also decreases selection intensity (N_es) at synonymous sites, which elevates synonymous, relative to replacement, polymorphism. Overall, the ratios of synonymous to replacement polymorphism and fixation are elevated to similar degrees and $\hat{\alpha}_{MK}$ is similar to scenario *st* (fig. 1A; supplementary table S1, Supplementary Material online). Note that this "robustness" may be particular to the degree and timing of population size change considered in scenario N_{d15k}. Scenario $N_{d15k}\_u_{i15k}$ considers a combination of mutation rate increase and population size decrease at time t_{15k}. In this scenario, r_{pd}'s were intermediate between those of u_{i15k} and N_{d15k} (supplementary table S1, Supplementary Material online).

Figure 1B shows $\hat{\alpha}_{MK}$ for scenarios of increasing codon bias. Decreased 1⟶0 mutation rate (scenario u_{d15k}) reduces the numbers of polymorphic sites and fixations relative to scenario *st* (supplementary table S1, Supplementary Material online). In high MCU, the reduction was larger for the numbers of synonymous than replacement polymorphism because a larger fraction of synonymous sites have ancestral favored states; the decreased mutation rate per site has a strong effect on per locus synonymous mutation rates. In contrast to polymorphism, the numbers of pooled synonymous fixations are similar to scenario *st*. This reflects the combination of reduced 1⟶0 fixations and elevated 0⟶1 fixations under increasing MCU. Overall, r_{pd}'s for synonymous changes are reduced more than for neutral changes (relative to scenario *st*) and the bias in $\hat{\alpha}_{MK}$ is reduced (fig. 1B).

The effect of increased population size is considered in scenario N_{i15k}. Here, the within-species TMRCA is increased; numbers of fixations decrease whereas the number of polymorphic sites increases. As shown in supplementary table S1, Supplementary Material online, relative to scenario *st*, the number of synonymous and replacement polymorphic sites in N_{i15k} are larger but the ratio of synonymous/replacement polymorphism is lower. At synonymous sites, elevated population size increases the efficacy of natural selection

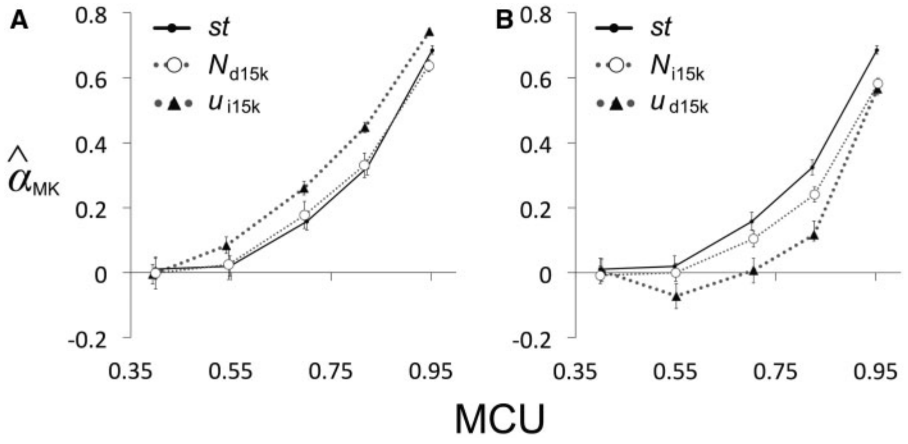

Fig. 1. α estimation under selection on synonymous mutations and nonstationary base composition evolution. MCU is the proportion of major codons, those that confer higher fitness. Scenario definitions are given in table 1. Bootstrap analyses were conducted by resampling codons from the simulated sequences. For each bootstrap sample, $\hat{\alpha}_{MK}$ was calculated using the numbers of polymorphic and divergent synonymous and replacement changes in the resampled data. Error bars show 95% confidence intervals from 1,000 replicates.

Table 1. Eleven Parameter Sets Used for the Simulation.

	Burn-in[a]			After t_0[a]			After t_{15k}[a]		
Scenario ID[b]	u [c]	v [c]	N [c]	u	v	N	U	v	N
st	3.0×10^{-6}	2.0×10^{-6}	1,000	3.0×10^{-6}	2.0×10^{-6}	1,000	3.0×10^{-6}	2.0×10^{-6}	1,000
N_{d15k}	—[d]	—	—	—	—	—	—	—	*332* [e]
N_{i15k}	—	—	—	—	—	—	—	—	*2,000*
u_{i15k}	—	—	—	—	—	—	6.0×10^{-6}	—	—
u_{d15k}	—	—	—	—	—	—	1.5×10^{-6}	—	—
$N_{d15k}\_u_{i15k}$	—	—	—	—	—	—	6.0×10^{-6}	—	*332*
$N_{i15k}\_u_{d15k}$	—	—	—	—	—	—	1.5×10^{-6}	—	*2,000*
N_{d0}	—	—	*2,000*	—	—	—	—	—	—
N_{i0}	—	—	*332*	—	—	—	—	—	—
$N_{d0}\_N_{d15k}\_u_{i15k}$	—	—	*2,000*	—	—	—	6.0×10^{-6}	—	*332*
$N_{i0}\_N_{d15k}\_u_{i15k}$	—	—	*332*	—	—	—	6.0×10^{-6}	—	*332*

[a] Simulation is separated into three sections, "burn-in" for 25,000 generations, "after t_0" for 15,000 generations from the end of the burn-in and "after t_{15k}" for 5,000 generations.
[b] Scenario IDs express the parameter switches considered in each scenario. *st*; stationary, *N*; population size switches, and *u*; mutation rate *u* switches. Subscripts show parameter value increases (*i*) or decreases (*d*) at generation t_0 or t_{15k}.
[c] *u* and *v* are 0→1 and 1→0 mutation rate per site per generation, respectively and *N* is population size.
[d] "—" means parameter value is the same as in scenario *st*.
[e] Parameter values that differ from scenario *st* are written in *Italic*.

(i.e., selection reduces the frequencies of 1→0 synonymous mutations within the population). Note that increased *N* also enhances 0→1 adaptive synonymous fixations but the short time between t_{15k} and the end of the simulation (t_{20k}) reduces the contribution of this effect. The lower ratio of synonymous/replacement polymorphism relative to that of scenario *st* reduces the ratio of r_{pd}'s for synonymous/replacement changes, and decreases $\hat{\alpha}_{MK}$ (it becomes less biased) (fig. 1*B*; supplementary table S1, Supplementary Material online).

The combined scenario of mutation rate decrease and population size increase ($N_{i15k}\_u_{d15k}$) showed r_{pd}'s intermediate between those of scenarios u_{d15k} and N_{i15k} (supplementary table S1, Supplementary Material online). Both u_{d15k} and N_{i15k} scenarios showed $\hat{\alpha}_{MK}$ reduction (fig. 1*B*) and, scenario $N_{i15k}\_u_{d15k}$ showed greater reduction in $\hat{\alpha}_{MK}$ (see the ratio of r_{pd}'s for synonymous and replacement changes in supplementary table S1, Supplementary Material online). In low MCU cases (MCU ≈ 0.55), this scenario is biased toward negative $\hat{\alpha}_{MK}$.

Filtering Rare Variants

The results shown in figure 1 employ all polymorphic sites observed in the population samples of 50 chromosomes from each simulated data set. However, most $\hat{\alpha}_{MK}$ approaches exclude rare variants (Fay et al. 2001, 2002; Bierne and Eyre-Walker 2004; Zhang and Li 2005; Charlesworth and Eyre-Walker 2006; Pröschel et al. 2006). We examined how rare variant filtering (RVF) affect $\hat{\alpha}_{MK}$. Figure 2 shows $\hat{\alpha}_{MK}$ after filtering polymorphic sites with minor allele frequencies ≤0.1 (i.e., 1:49 and 2:48 configurations in our samples of 50 alleles). Figure 2*A* shows the results for scenarios *st*, N_{d15k}, and u_{i15k}, and figure 2*B* shows the results for scenarios *st*, N_{i15k}, and u_{d15k}. RVF generally reduced overestimation of α (compared with results in fig. 1), but overestimation remained considerable in most scenarios. Interestingly, the magnitude of the effect of RVF on $\hat{\alpha}_{MK}$ differed among scenarios. In particular, scenario N_{d15k} was little affected by RVF, but scenario N_{i15k} showed substantially reduced $\hat{\alpha}_{MK}$. In the decreased

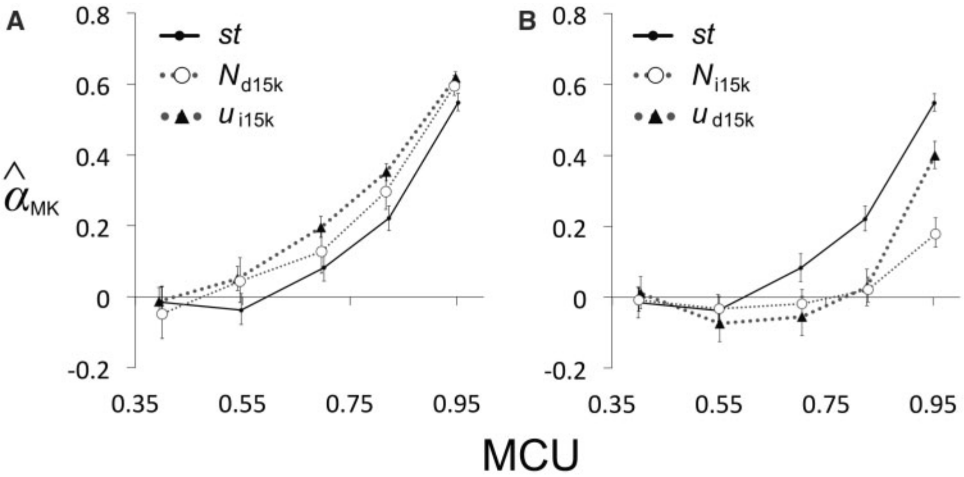

Fig. 2. Effect of rare variant filtering on α estimation under selection on synonymous mutations and nonstationary base composition evolution. Scenario definitions are given in table 1. Codons containing polymorphic sites with rare variants (frequency $\leq$0.1) were filtered from the simulated sequences prior to $\hat{\alpha}_{MK}$ calculation. Bootstrap analyses were conducted by resampling codons from the filtered simulated sequences. For each bootstrap sample, $\hat{\alpha}_{MK}$ was calculated using the numbers of polymorphic and divergent synonymous and replacement changes in the resampled data. Error bars show 95% confidence intervals from 1,000 replicates.

population size scenario, N_{d15k}, genetic drift decreased the proportion of rare polymorphic sites (Crow and Kimura 1970, p. 452) and the effect of RVF was small. On the other hand, in the increased population size scenario, N_{i15k}, we expect a large fraction of synonymous polymorphic sites with rare disadvantageous state 0. Because RVF removes such polymorphic sites, this scenario showed reduced $\hat{\alpha}_{MK}$ bias.

In the scenarios considered above, MCU evolved at steady-state prior to t_{15k}. We also considered cases of extended nonstationary evolution by introducing parameter changes at two time points, t_0 (i.e., immediately following the burn-in) and t_{15k}. The results are shown in the Supplementary Material online.

Adaptive Nonsynonymous Mutations

We examined α estimation in the presence of both estimation biases and adaptive replacement substitutions. We adjusted the *st* and $N_{d15k}\_u_{i15k}$ scenarios to include 20–30% adaptive amino acid substitutions by allowing replacement mutations to confer a fitness benefit, $s_{nonsyn} = 0.02$ at 1 in 100 codons ($s_{nonsyn} = 0.0$ at the remaining codons). These parameters gave $\alpha = 0.29$ and 0.23 in *st* and $N_{d15k}\_u_{i15k}$, respectively (reduced N after $t = 15{,}000$ lowers fixation probabilities for adaptive mutations in $N_{d15k}\_u_{i15k}$). For these simulated data, $\hat{\alpha}_{MK}$ reflects the effects of base composition selection on synonymous r_{pd}, positive selection on replacement r_{pd}, and, in the $N_{d15k}\_u_{i15k}$ case, N fluctuation on both classes. When codon bias selection is relatively strong, these factors combine to give considerably lower than additive effects on $\hat{\alpha}_{MK}$; that is, estimated α is less than the sum of $\hat{\alpha}_{MK}$ under neutral replacement evolution and true α (fig. 3).

$\hat{\alpha}_{MK}$ *and Base Composition in Drosophila*

Our simulation results predict associations between $\hat{\alpha}_{MK}$ and base composition for most scenarios of $\hat{\alpha}_{MK}$ inflation by codon bias selection and/or nonsteady-state base composition. We will refer to these effects collectively as synonymous site compositional bias effects (SCBE). Because translational selection favors G- and C-ending codons in *Drosophila melanogaster* (Akashi 1994; Akashi et al. 2006; Vicario et al. 2007), SCBE predicts positive correlations between GC content and $\hat{\alpha}_{MK}$. Large-scale analyses of *D. melanogaster* polymorphism and divergence data confirm the predicted association between $\hat{\alpha}_{MK}$ for GC_{4f} (GC content at third codon position of 4-fold redundant codons; fig. 4). The strong relationship (autosomal loci, $P = 0.957$, $P < 0.001$) between $\hat{\alpha}_{MK}$ and GC_{4f} is consistent with an inflated signal of adaptive evolution arising from selection and nonstationarity evolution of synonymous changes. Because low GC_{4f} genes show significantly elevated $\hat{\alpha}_{MK}$, the overall *D. melanogaster* pattern is similar to expectations for $\hat{\alpha}_{MK}$ under a combination of SCBE and adaptive protein evolution (fig. 3). This analysis employed genes from regions that experience moderate to high rates of recombination; the pattern is unlikely to reflect an indirect effect of low GC content in genes located in regions of restricted reduced crossing (Kliman and Hey 1993; Singh, Arndt, et al. 2005).

Several studies have reported enhanced adaptive evolution for genes located on the X chromosome relative to those on autosomes (reviewed in Llopart 2015). However, X-linked genes also show elevated GC content at synonymous sites (Singh, Davis, et al. 2005) and could thus experience greater $\hat{\alpha}_{MK}$ inflation by SCBE. Figure 4 shows that X-linked loci show greater $\hat{\alpha}_{MK}$ than autosomal genes of similar GC_{4f}. This difference is more apparent for the lower range of GC_{4f}. If we can assume similar SCBE for genes of similar base composition, these patterns are consistent with a combination of SCBE and "faster X" adaptive protein evolution. We attempted to control for SCBE to test evidence for other global factors that have been associated with adaptive protein evolution in *Drosophila* including recombination rate and expression intensity/patterns (Supplemental Material online).

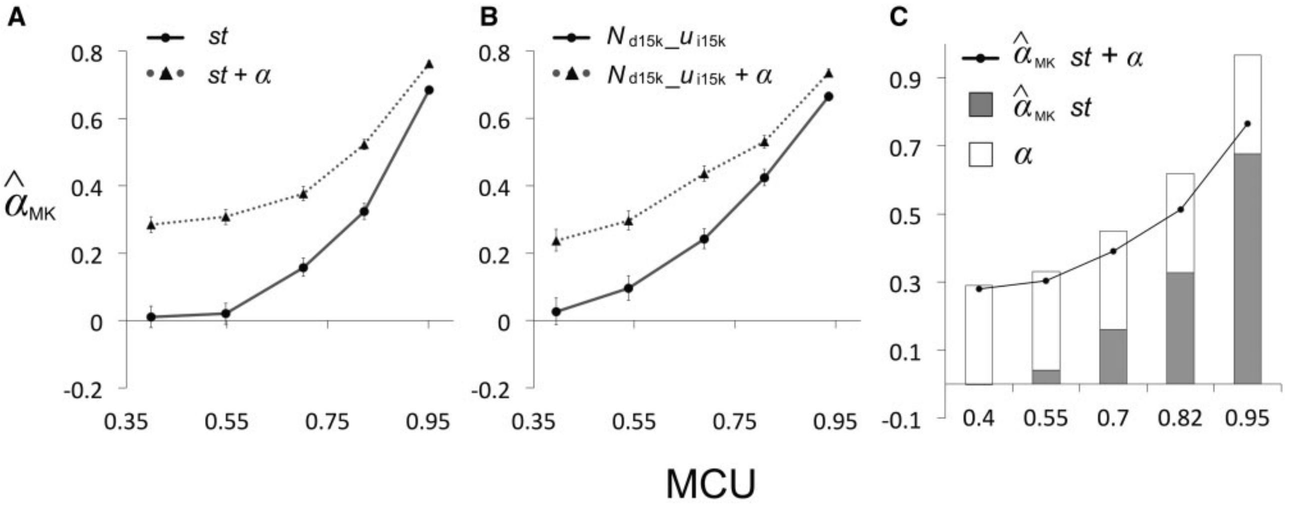

Fig. 3. α estimation under adaptive protein evolution. Panel *A* and *B* show estimated α for simulation scenarios *st* and $N_{d15k}\_u_{i15k}$, respectively for neutral replacement changes (filled circles) and 20–30% adaptive amino acid substitutions (filled triangles). In adaptive protein evolution simulations, we allow replacement mutations to confer a fitness benefit, selection coefficient (s_{nonsyn}) = 0.02 at 1 in 100 codons. All other simulations parameters were identical to the neutral protein evolution scenarios (table 1). Bootstrap analyses were conducted by resampling codons from the simulated sequences. For each bootstrap sample, $\hat{\alpha}_{MK}$ was calculated using the numbers of polymorphic and divergent synonymous and replacement changes in the resampled data. Error bars show 95% confidence intervals from 1,000 replicates. Note that high recombination rates in these simulations limit the effects of linkage and selective interference. Panel *C* shows theoretically predicted α estimation. Expected numbers of polymorphic and fixed mutations were calculated numerically using expressions from Sawyer and Hartl (1992) assuming steady-state codon bias evolution and independent evolution (i.e., free recombination) among sites. Dark bars show expected α under major codon preference and the white bars show the expected fraction of adaptive replacement fixations ($N_e s_{nonsyn} = 20$, $u/v = 0.6$). Dots show expected α estimated under the combined effect of biases related to selection at synonymous sites and true adaptive amino acid fixations. $\hat{\alpha}_{MK}$ was calculated according to Smith and Eyre-Walker (2002).

Discussion

MK 2 × 2 tables of numbers of polymorphic and fixed differences for DNA mutation classes are relatively simple to interpret if 1) the control class of variation evolves neutrally, 2) polymorphism/divergence mutation ratios are identical for the control and test classes of variation, and 3) polymorphism is neutral for the test class (no weakly selected mutations). If mutations are interspersed randomly within genetic regions, the test is robust to departures from equilibrium site frequency spectra for within population variation, even with some degree of genetic linkage among sites (Sawyer et al. 1987; McDonald and Kreitman 1991). Given widespread estimates of α based on MK tables, understanding the sensitively of the approach to violations of these assumptions is critical. Issue (3) from above, especially the combination of slightly deleterious test class mutations and changes in effective population size within or among the lineages, has been addressed in a number of studies (Charlesworth and Eyre-Walker 2008; Eyre-Walker and Keightley 2009; Wilson et al. 2011; Messer and Petrov 2013). This study considers violations of (1) and (2) from above and our simulation results show that selection on synonymous mutations and short-term nonstationary evolution of codon usage bias can substantially bias estimates of α from polymorphism/divergence counts. In particular, stationary and decreasing codon bias scenarios are prone to considerable elevations of $\hat{\alpha}_{MK}$ even when all replacement changes are neutral. "Weak" selection ($N_e s \approx 1$) on synonymous mutations can strongly bias $\hat{\alpha}_{MK}$.

$\hat{\alpha}_{MK}$ biases depend on evolutionary scenarios of synonymous site evolution. Recent relaxed selection and/or increases in mutation bias away from preferred states result in strong α overestimation which is little affected by filtering low frequency polymorphism. In our simulations, overestimation of α by synonymous site selection was reduced by decreased 1→0 mutation rates or increased population size, but such scenarios do not always decrease the risk of α estimation error. The effect of nonstationary codon usage bias is sensitive to the timing of parameter changes (fig. 1; supplementary fig. S2, Supplementary Material online) and can bias $\hat{\alpha}_{MK}$ in opposite directions (negative $\hat{\alpha}_{MK}$ in $N_{i15k}\ \_u_{d15k}$ and N_{i0}, fig. 1 and supplementary fig. S2, Supplementary Material online).

Our results are likely to be relevant for estimating adaptive protein evolution in the *D. melanogaster* lineage where both reduced selection and fluctuating mutation appear to have contributed to a dramatic change in codon bias (Akashi 1995, 1996; Akashi et al. 2006; Nielsen et al. 2007; Zeng and Charlesworth 2010). In comparisons to moderately distant species (e.g., *D. yakuba*), our simulation results predict a positive α estimation bias in the *D. melanogaster* lineage and a correlation between ancestral GC content and $\hat{\alpha}_{MK}$. The predicted association is supported (fig. 4), but elevated $\hat{\alpha}_{MK}$ among low codon bias genes is consistent with a substantial contribution of adaptive protein evolution.

Accurately quantifying the proportion of adaptively fixed replacement changes will likely require revised methods (discussed below). However, tests of biological factors associated

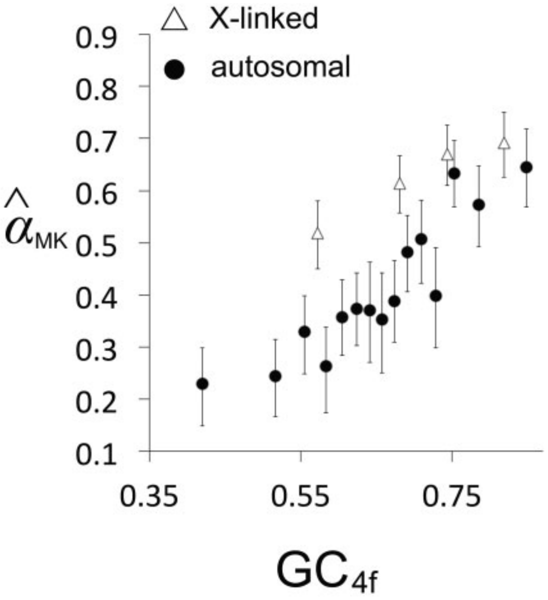

Fig. 4. $\hat{\alpha}_{MK}$ versus GC_{4f} in the *D. melanogaster* lineage. Welch's MKTest software with model 10ii (Welch 2006) was used for $\hat{\alpha}_{MK}$ calculation. Autosomal, non-male biased genes (3,582) and X-linked, non-male biased genes (791) were binned into GC_{4f} classes as described in the text. Only genes located in regions of moderate to high estimates of recombination (≥2.0 cM per Mb) in *D. melanogaster* (Fiston-Lavier *et al.* 2010) were included. Bootstrap analysis was conducted with 1,000 replicates resampling genes in the same bin and the average and 95% confidence interval are shown in the graph. Rare variants were filtered in these data; the correlation remains similar if singletons are included (data not shown).

with adaptive protein evolution may control for SCBE by accounting for base composition. Comparisons among genes with similar GC_{4f} support higher α for X-linked versus autosomal genes, for male-biased versus non-male-biased genes, and for genes that experience high rates of recombination (fig. 4; supplementary figs. S5 and S6, Supplementary Material online). These interpretations require the magnitude of SCBE to be shared genome-wide for genes of similar base composition. Region- or expression pattern-specific changes in evolutionary parameters should be tested to validate the approach.

Model fitting approaches can incorporate SCBE to improve α estimation. This will require complex models that include parameters for lineage-specific mutation and fixation bias (selection and biased gene conversion) as well as the distribution of deleterious effects among replacement mutations and both long and short-term N_e. Previous models have incorporated subsets of these factors, for example, selection on replacement changes and demography (Williamson et al. 2005; Eyre-Walker and Keightley 2009; Messer and Petrov 2013) and selection on base composition and unequal mutation rates (Nielsen et al. 2007; Singh et al. 2009; Zeng and Charlesworth 2009; Matsumoto et al. 2015). Parameter estimation under more complex models may be challenging, however, unless a large proportion of parameter values can be assumed to be shared across genes.

Filtering synonymous mutations may reduce SCBE in α estimation but can entail important tradeoffs. Employing only preferred synonymous changes provides a conservative approach to identifying adaptive protein evolution (Akashi 1995). Llopart and Comeron (2008) suggested filtering both positively and negatively selected synonymous changes at greater than 2-fold redundant codons to minimize effects of selection on the test class. Such approaches can be applied in cases where codon selection is well understood, but large reductions in the polymorphism and fixation counts for the control class will limit the ability to perform gene-specific parameter estimation (in our data analysis, models that include gene-specific evolutionary parameters were strongly preferred to uniform parameter models). In addition, controlling for codon selection may result in greater sensitivity to the mutational assumption of the MK approach. The expected numbers of polymorphic sites and fixed differences are directly proportional to per locus mutation rate (Sawyer and Hartl 1992). Different r_{pd}'s for test and control class reflect fitness differences if ratios of mutation rates are identical between the classes. However, changes in the mutation spectrum, even if the rate matrix is shared by test and control classes, can result in differential effects on the per locus mutation rate ratios if the classes differ in ancestral base composition or if they consist of different types of nucleotide transitions (e.g., transversions may be prevalent after filtering nonneutral synonymous changes). In our simulations, high MCU genes experienced elevated synonymous mutation rates (and greater $\hat{\alpha}_{MK}$ bias) under fluctuating 1→0 mutation rates (fig. 1, u_{i15k} scenario).

Alternative control class mutations for MK tests could also reduce SCBE. Noncoding DNA (e.g., introns or flanking regions) may include neutrally evolving sites but the sites should be chosen carefully both to avoid mutations that affect fitness (Halligan et al. 2004; Andolfatto 2005; Zeng and Charlesworth 2010) and for sufficient interspersion with nonsynonymous sites so that recent evolutionary histories (gene trees) do not differ for the test and control classes. In addition, base composition differences between coding regions and intron/flanking regions (e.g., GC content is generally lower in small introns than within coding regions) may invalidate the mutational assumption of the MK approach (see above).

SCBE biases can also be reduced by limiting comparisons to low codon bias genes (e.g., Andolfatto 2005). Analysis of genes with similar base composition at synonymous and nonsynonymous sites would also control for nonstationary mutation patterns (i.e., changes in mutation would affect per locus synonymous and replacement mutation rates similarly). In many lineages, this approach will considerably restrict the data both in the numbers and categories (functional classes, expression levels, etc.) of qualifying genes because even very weak selection can bias $\hat{\alpha}_{MK}$ (figs. 1–3). Genomes that show little effect of fixation biases on base composition at synonymous sites may be good candidates for current $\hat{\alpha}_{MK}$ approaches but such cases may not be common.

Our simulations based on codon bias evolution scenarios and genetic distances that are thought to be relevant for the

D. melanogaster subgroup revealed considerable $\hat{\alpha}_{MK}$ estimation biases for many of the scenarios examined (especially for high MCU genes). Because both selection on codon bias (and other fixation biases including biased gene conversion) and fluctuations in base composition may be widespread among taxa, these results may be broadly relevant when estimating the proportion of adaptive fixations in protein evolution from polymorphism and divergence counts.

Materials and Methods

Simulation Approach

To isolate the effects of biased mutation rate and codon usage selection on $\hat{\alpha}_{MK}$, we conducted "forward-running" computer simulations of sequence evolution within and between populations. We modeled haploid, hermaphrodite individuals in a population with discrete generations and employed a simplified genetic code: codons are composed of two nucleotide sites with two possible states at each site, "1" and "0" (i.e., two amino acids each encoded by 2-fold redundant codons). Mutations at the first codon position are nonsynonymous or replacement changes and second position changes are synonymous. Mutations occur at rates u for 1→0 and v for 0→1. "01" and "11" represent "major" codons with fitness 1 and "00" and "10" represent "minor" codons with fitness 1–s. MCU is the frequency of major codons among L codons in a sequence. All replacement changes (first codon position) are selectively neutral. The fitness of a genome (probability of survival till reproduction) is calculated assuming multiplicative effects; fitness for individual i with n minor codons is $f_i = (1-s)^n$. Each generation, the N genomes are paired randomly and recombination occurs with crossover probability r between neighboring nucleotide sites in each pair.

Parameter Values and Simulation Method

We set parameter values at $N = 1{,}000$, $u = 3.0 \times 10^{-6}$, $v = 2.0 \times 10^{-6}$, $r = 0.01$, and $L = 500{,}000$ codons as the base condition. Mutation bias was set to give a neutral equilibrium MCU of 40%, a value roughly consistent with G + C content in *D. melanogaster* noncoding regions (Moriyama and Hartl 1993; Singh, Arndt, et al. 2005; Vicario et al. 2007). High rates of recombination ($Nr = 10$) reduces selective interference and emulates free recombination. Supplementary figure S1, Supplementary Material online, shows simulated versus expected equilibrium MCU under free recombination and suggests close to independent evolution among sites for this simulation scenario. We employed fitness benefits, $s = 0$, 0.0003, 0.000634, 0.00098, and 0.0017 to give a range of equilibrium MCU values: 0.4, 0.55, 0.7, 0.83, and 0.95, respectively (see Li 1987 for basic model).

The starting condition for most simulations at t_0 was near-equilibrium for both the site frequency spectrum and codon bias under the basal parameter set. This was achieved through a 25,000-generation "burn-in" period prior to t_0. At the start of the burn-in, $t_{\text{-}25k}$, we set the MCU for all genomes in the population at the expected equilibrium MCU value for a given parameter set. A sequence was constructed by randomly sampling codons given their expected frequencies and populations were started as monomorphic for this sequence. The populations then evolved for 25,000 generations so that both the codon usage and the site frequency spectrum within the populations should be at, or near, stochastic equilibrium by t_0.

Following the burn-in (i.e., from t_0 forward), each population was evolved for 20,000 generations and all fixations of new mutations were recorded. At the end of a simulation, we sampled 50 sequences from the population to calculate $\hat{\alpha}_{MK} = 1 - \frac{\overline{D}_s}{\overline{D}_n}\overline{\left(\frac{P_n}{P_s+1}\right)}$. We used the total numbers of synonymous and replacement polymorphic sites and fixations (nonancestral states shared by all alleles in the sample) occurred on L codons as the counts $\overline{P}_s$, $\overline{P}_n$, $\overline{D}_s$, and $\overline{D}_n$ in the above equation. We calculated average and 95% confidence intervals of $\hat{\alpha}_{MK}$ using bootstrap analysis, that is, by resampling L codons with replacement 1,000 times.

In some scenarios, parameter changes were introduced during the simulation (between t_0 and t_{20k}). Such changes were implemented at the 15,000th generation (t_{15k}) to emulate the relative location of the split between *D. melanogaster* and *D. simulans* relative to divergence from *D. yakuba:* an extensive change in genome base composition appears to have occurred around this time (Akashi 1995, 1996; Akashi et al. 2006; Nielsen et al. 2007; Singh et al. 2009). In table 1, we show parameter values for six scenarios examined in this study. In each of these six scenarios, we consider five different s values (explained above). Simulation programs are available from the authors upon request.

DNA Polymorphism and Divergence Analysis in the *D. melanogaster* Subgroup

DNA Sequences

We calculated $\hat{\alpha}_{MK}$ using polymorphism data from *D. melanogaster* and fixations estimated within the subgroup. The *D. melanogaster* genome sequence (Adams et al. 2000) and annotation data (release 5.28, June 4, 2010) and genome sequences for *D. yakuba* (release 1.3) and *D. erecta* (release 1.3) (*Drosophila* 12 Genomes Consortium 2007) were obtained from Flybase. We obtained gene orthology information from Flybase (Tweedie et al. 2009, ftp://ftp.flybase.net/genomes/; version: gene_orthologs_fb_2010_05, last accessed May 14, 2010). For 1-1-1 orthologs among the three genomes, protein-coding sequences for *D. melanogaster*, *D. yakuba*, and *D. erecta* were aligned using MUSCLE (Edgar 2004) and back-translated to create DNA alignments. Only genes predicted to produce a single protein isoform in *D. melanogaster* were included (8,933 genes). Codons aligning to gaps or N's in any of the sequences (in any codon positions) were filtered from the alignment (i.e., all codons aligned to N- or gap-containing codons were removed).

For *D. melanogaster* polymorphism data, we followed the methods described in Campos et al. (2014). We obtained Q31 (equivalent to a Phred Q of 48) FASTA files from the DPGP project website (dpgp2_v2.ID5.primarycore.nohets.masked.fasta.bz2, available at http://www.dpgp.org/dpgp2/DPGP2.html, last accessed July 22, 2014; Pool et al. 2012). We analyzed data from the Rwandan population sample, the DPGP population for which the largest number of alleles had been sequenced. We analyzed sequences from 15 of the 23 Rwanda strains: RG2, RG3, RG4N, RG5, RG9, RG18N, RG19, RG22, RG24, RG25, RG28, RG32N, RG34, RG36, RG38N. The following strains were excluded because they show evidence for a high proportion of admixture with European populations (Pool et al. 2012): RG10, RG11N, RG15, RG21N, RG35. In addition, the RG33 and RG7 genome sequences contain large fractions of ambiguous nucleotides and were filtered.

DPGP sequences were added to the three species alignments using the *D. melanogaster* genome coordinates provided by Pool et al. (2012). The initial *D. melanogaster* 5.28 sequences were removed to give alignments of 15 DPGP *D. melanogaster* alleles and 1 allele each for *D. yakuba* and *D. erecta*. Codons aligning to N's in any of the sequences were filtered as described above. This process resulted in 8,014 protein-coding gene alignments of 17 sequences (15 *D. melanogaster*, 1 *D. yakuba*, and 1 *D. erecta*). Three further filtering steps were employed prior to analyses. For all analyses, we filtered codons with more than two segregating nucleotides in *D. melanogaster*. Such codons cannot be accommodated in our ancestral reconstruction method (see below), but were rare (about 2.3% of polymorphic codons and 0.1% of all codons). We also filtered aligned codons that included a stop codon in any of the sequences (about 0.005% of the remaining codons). Finally, we filtered genes with small numbers ($<$20) of 4-fold codons to reduce error in assigning genes to GC-classes for ancestral inference. These processes resulted in a data set of 7,917 protein-coding gene alignments of 17 sequences with 3,537,702 codons (1,236,915 4-fold codons) in total.

For some analyses, we "filtered" rare variants in *D. melanogaster* samples to reduce the effects of sequencing errors and slightly deleterious mutations (Eyre-Walker 2002; Fay et al. 2002; Keinan and Clark 2012). We converted 14:1 *D. melanogaster* codon configurations to 15:0 (i.e., singletons were converted to the *D. melanogaster* consensus state) and the resulting sequences were used for the ancestral reconstruction and $\hat{\alpha}_{MK}$ calculation as explained below. We estimated numbers of polymorphic and divergent sites in *D. melanogaster* and calculated $\hat{\alpha}_{MK}$.

Ancestral Reconstruction

We employed a maximum likelihood approach implemented in BASEML in PAML (Yang 2007) to estimate counts of polymorphic and fixed nucleotide changes. We employed probabilities of ancestral states among the 15 *D. melanogaster* alleles to determine counts of synonymous and replacement polymorphism. Because BASEML requires a phylogeny as input, the approach requires an adjustment for sequences that have undergone recombination because different sites or regions within the sequence may have different evolutionary histories (i.e., trees). Here, we employed a method described in (Akashi et al. 2006) to collapse or convert the sequence data for 15 *D. melanogaster* alleles and two outgroup sequences (*D. yakuba* and *D. erecta*) to four extant nodes, m_a, m_b, y, e. For codons that are variable among the *D. melanogaster* alleles, the two segregating codons are assigned randomly to m_a and m_b (codons with greater than two polymorphic codons were filtered as explained above). We will refer to the ancestral node for m_a and m_b as m_am_b and the ancestral node for y and e as ye.

The BASEML analysis employed a general time reversible model with lineage-specific base composition parameters, GTR-NH$_b$ (Matsumoto et al. 2015). Genes were grouped by GC$_{4f}$ into 10 bins of roughly 354,000 codons each. 4-fold redundant codons were used for binning, but all codons were included in the ancestral inference and $\hat{\alpha}_{MK}$ calculations. BASEML estimates posterior probabilities of ancestral states of each site at each node of the tree. Substitution parameters and posterior probabilities of ancestral states are estimated independently for first, second, and third codon position nucleotides within each bin. We calculated the ancestral probabilities of codons by multiplying the probabilities of first, second, and third position nucleotides of the codon (Akashi et al. 2007, Matsumoto et al. 2015). Methods that assume that states with the highest posterior probabilities are "correct" can be error prone (Matsumoto et al. 2015). To account for uncertainty in ancestral inference, we constructed 100 ancestral m_am_b sequences for each gene given the probabilities of ancestral codons. At each codon position of each ancestral sequence, we sampled codons based on their posterior probabilities. For example, at a given position, if BASEML estimated posterior probabilities of ancestral codons AAA and AAG as 0.9 and 0.1, respectively, codons in our ancestral sequence at this position were assigned AAA with probability 0.9 and AAG with probability 0.1. This sampling process was replicated to give 100 reconstructed m_am_b sequences for each gene.

Calculating Numbers of Polymorphic and Fixed Mutations

We used BASEML posterior probabilities to estimate the counts of synonymous and replacement polymorphism. Probabilities of particular codon changes are treated as their counts. Consider a codon position where m_a encodes "AAG" and m_b encodes "AAA" and at m_am_b, BASEML gives ancestral codon "AAA" with probability 0.9 and "AAG" with probability 0.1. We calculate 0.9 synonymous mutations (AAA→AAG) and 0.1 synonymous mutation (AAG→AAA) segregating among the *D. melanogaster* alleles. For our analysis, we pool all synonymous mutations (major→major, major→minor, minor→major, minor→minor), so the total at this codon would be one synonymous polymorphism.

We calculated the numbers of synonymous and replacement fixations using the reconstructed m_am_b sequences. For each gene, we employed CODEML in PAML (Yang 2007) to estimate the numbers of synonymous and replacement fixations that occurred between m_am_b and ye. We used the average numbers of substitutions determined for 100 m_am_b sequences for each gene (i.e., distances were calculated for each m_am_b sequence to orthologous *D. yakuba* and *D. erecta* sequences). The site model M1a (Yang 2007) was used to count the numbers of synonymous (S) and replacement (N) sites, as well as distances, d_S and d_N. The variance in these estimates was very small (data not shown) and we used the average $S \times d_S$ and $N \times d_N$ among 100 replicates as estimates of the numbers of synonymous and replacement substitutions, respectively.

$\hat{\alpha}_{MK}$ Estimation

Using the estimated numbers of polymorphic mutations and substitutions in each gene, we calculated $\hat{\alpha}_{MK}$ using Welch's software, which analyzes multilocus McDonald–Kreitman tables (McDonald and Kreitman 1991; Welch 2006). As recommended in Welch (2006), we calculated Akaike Information Criteria (Akaike 1974) and Bayesian Information Criteria (Schwarz 1978) for each bin in each analysis (autosomal, non-male-biased genes, autosomal, male-biased genes, X-linked non-male-biased genes and X-linked male biased genes). Both methods supported model 10ii for each bin. This model allows gene-specific parameters for theta, divergence time, and fraction of constrained sites but assumes a single α across genes within a bin.

GC Content versus $\hat{\alpha}_{MK}$

For $\hat{\alpha}_{MK}$ versus GC_{4f} analysis, we analyzed autosomal and X-linked, non-male biased genes (other categories and factors are examined in the Supplementary Material online). In this analysis, we filtered genes with low recombination rate ($<$ 2.0 cM per Mb, Fiston-Lavier et al. 2010). The remaining genes show no correlation between GC_{4f} and recombination rate. We classified autosomal and X-linked genes after recombination rate filtering into 15 and 4 bins, respectively, based on GC_{4f}. Because Welch's software conducts bootstrap analysis by resampling genes, we set each bin to have similar numbers of genes. $\hat{\alpha}_{MK}$ calculations were performed done on each bin with 1,000 bootstrap replicates. Spearman's rank correlation coefficient was calculated using R (v3.1.3) from the average GC_{4f} value and $\hat{\alpha}_{MK}$ in each bin. Computer programs and data employed in this study are available from the authors upon request.

Supplementary Material

Supplementary figures S1–S5 and tables S1–S3 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors are grateful to Naoki Osada, Ziheng Yang, Tomoko Ohta, Neha Mishra, and Kent Kawashima for discussions and advice. The authors also thank three anonymous reviewers for their insights and suggestions that helped to improve this work. This work was funded by the Systems Resilience project of the Research Organization of Information and Systems.

References

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, et al. 2000. The genome sequence of *Drosophila melanogaster*. *Science* 287(5461):2185–2195.

Agashe D, Shankar N. 2014. The evolution of bacterial DNA base composition. *J Exp Zool B Mol Dev Evol*. 322(7):517–528.

Akaike H. 1974. A new look at the statistical model identification. *IEEE Trans Automat Control*. 19(6):716–723.

Akashi H. 1994. Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. *Genetics* 136(3):927–935.

Akashi H. 1995. Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in *Drosophila* DNA. *Genetics* 139(2):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(3):1297–1307.

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(1):221–238.

Akashi H, Goel P, John A. 2007. Ancestral inference and the study of codon bias evolution: implications for molecular evolutionary analyses of the *Drosophila melanogaster* subgroup. *PLoS One* 2(10):e1065–e1014.

Akashi H, Ko WY, Piao S, John A, Goel P, Lin CF, Vitins AP. 2006. Molecular evolution in the Drosophila melanogaster species subgroup: frequent parameter fluctuations on the timescale of molecular divergence. *Genetics* 172(3):1711–1726.

Andolfatto P. 2005. Adaptive evolution of non-coding DNA in *Drosophila*. *Nature* 437(7062):1149–1152.

Andolfatto P. 2008. Controlling type-I error of the McDonald-Kreitman test in genomewide scans for selection on noncoding DNA. *Genetics* 180(3):1767–1771.

Bierne N, Eyre-Walker A. 2004. The genomic rate of adaptive amino acid substitution in *Drosophila*. *Mol Biol Evol*. 21(7):1350–1360.

Campos JL, Halligan DL, Haddrill PR, Charlesworth B. 2014. The relation between recombination rate and patterns of molecular evolution and variation in *Drosophila melanogaster*. *Mol Biol Evol*. 31(4):1010–1028.

Charlesworth B. 1994. The effect of background selection against deleterious mutations on weakly selected, linked variants. *Genet Res*. 63(3):213–227.

Charlesworth J, Eyre-Walker A. 2006. The rate of adaptive evolution in enteric bacteria. *Mol Biol Evol*. 23(7):1348–1356.

Charlesworth J, Eyre-Walker A. 2008. The McDonald-Kreitman test and slightly deleterious mutations. *Mol Biol Evol*. 25(6):1007–1015.

Crow JF, Kimura M. 1970. *An introduction to population genetics theory*. Caldwell: The Blackburn Press. p. 452.

Drosophila 12 Genomes Consortium. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* 450(7167):203–218.

Duret L, Semon M, Piganeau G, Mouchiroud D, Galtier N. 2002. Vanishing GC-rich isochores in mammalian genomes. *Genetics* 162(4):1837–1847.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 32(5):1792–1797.

Eilertson KE, Booth JG, Bustamante CD. 2012. SnIPRE: Selection inference using a Poisson random effects model. *PLoS Comp Biol*. 8(12):e1002806–e1002814.

Ellegren H. 2008. A selection model of molecular evolution incorporating the effective population size. *Evolution* 63(2):301–305.

Eyre-Walker A. 2002. Changing effective population size and the McDonald-Kreitman test. *Genetics* 162(4):2017–2024.

Eyre-Walker A. 2006. The genomic rate of adaptive evolution. *Trends Ecol Evol*. 21(10):569–575.

Eyre-Walker A, Keightley PD. 2009. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Mol Biol Evol*. 26(9):2097–2108.

Fay JC. 2011. Weighing the evidence for adaptation at the molecular level. *Trends Genet*. 27(9):343–349.

Fay JC, Wyckoff GJ, Wu CI. 2001. Positive and negative selection on the human genome. *Genetics* 158(3):1227–1234.

Fay JC, Wyckoff GJ, Wu CI. 2002. Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. *Nature* 415(6875):1024–1026.

Fiston-Lavier AS, Singh ND, Lipatov M, Petrov DA. 2010. *Drosophila melanogaster* recombination rate calculator. *Gene* 463(1-2):18–20.

Halligan D, Eyre-Walker A, Andolfatto P, Keightley P. 2004. Patterns of evolutionary constraints in intronic and intergenic DNA of *Drosophila*. *Genome Res*. 14(2):273–279.

Ikemura T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol*. 2(1):13–34.

Keinan A, Clark AG. 2012. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* 336(6082):740–743.

Kliman RM, Hey J. 1993. DNA sequence variation at the period locus within and among species of the *Drosophila Melanogaster* complex. *Genetics* 133(2):375–387.

Li WH. 1987. Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. *J Mol Evol*. 24(4):337–345.

Llopart A. 2015. Parallel faster-X evolution of gene expression and protein sequences in Drosophila: beyond differences in expression properties and protein interactions. *PLoS One* 10(3):e0116829.

Llopart A, Comeron JM. 2008. Recurrent events of positive selection in independent Drosophila lineages at the Spermatogenesis Gene roughex. *Genetics* 179(2):1009–1020.

Matsumoto T, Akashi H, Yang Z. 2015. Evaluation of ancestral sequence reconstruction methods to infer nonstationary patterns of nucleotide substitution. *Genetics* 200(3):873–890.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* 351(6328):652–654.

Messer PW, Petrov DA. 2013. Frequent adaptation and the McDonald-Kreitman test. *Proc Natl Acad Sci U S A*. 110(21):8615–8620.

Moriyama EN, Hartl DL. 1993. Codon usage bias and base composition of nuclear genes in *Drosophila*. *Genetics* 134(3):847–858.

Moriyama EN, Powell JR. 1997. Codon usage bias and tRNA abundance in *Drosophila*. *J Mol Evol*. 45(5):514–523.

Nielsen R, Bauer DuMont VL, Hubisz MJ, Aquadro CF. 2007. Maximum likelihood estimation of ancestral codon usage bias parameters in *Drosophila*. *Mol Biol Evol*. 24(1):228–235.

Parsch J, Zhang Z, Baines JF. 2009. The influence of demography and weak selection on the McDonald-Kreitman test: an empirical study in *Drosophila*. *Mol Biol Evol*. 26(3):691–698.

Pool JE, Corbett-Detig RB, Sugino RP, Stevens KA, Cardeno CM, Crepeau MW, Duchen P, Emerson JJ, Selao P, Begun DJ, et al. 2012. Population genomics of sub-Saharan *Drosophila Melanogaster*: African diversity and non-African admixture. *PLoS Genet*. 8(12):e1003080–e1003024.

Powell JR, Moriyama EN. 1997. Evolution of codon usage bias in *Drosophila*. *Proc Natl Acad Sci U S A*. 94(15):7784–7790.

Pröschel M, Zhang Z, Parsch J. 2006. Widespread adaptive evolution of *Drosophila* genes with sex-biased expression. *Genetics* 174(2):893–900.

Rodríguez-Trelles F, Tarrío R, Ayala FJ. 2000. Evidence for a high ancestral GC content in *Drosophila*. *Mol Biol Evol*. 17(11):1710–1717.

Sawyer SA, Dykhuizen DE, DuBose RF, Green L, Mutangadura-Mhlanga T, Wolczyk DF, Hartl DL. 1987. Distribution and abundance of insertion sequences among natural isolates of *Escherichia coli*. *Genetics* 115(1):51–63.

Sawyer SA, Hartl DL. 1992. Population genetics of polymorphism and divergence. *Genetics* 132(4):1161–1176.

Schwarz G. 1978. Estimating the dimension of a model. *Ann Statist*. 6(2):461–464.

Sella G, Petrov DA, Przeworski M, Andolfatto P. 2009. Pervasive natural selection in the Drosophila genome? *PLoS Genet*. 5(6):e1000495.

Sharp PM, Emery LR, Zeng K. 2010. Forces that influence the evolution of codon bias. *Philos Trans R Soc Lond B Biol Sci*. 365(1544):1203–1212.

Singh ND, Arndt PF, Clark AG, Aquadro CF. 2009. Strong evidence for lineage and sequence specificity of substitution rates and patterns in *Drosophila*. *Mol Biol Evol*. 26(7):1591–1605.

Singh ND, Arndt PF, Petrov DA. 2005. Genomic heterogeneity of background substitutional patterns in *Drosophila Melanogaster*. *Genetics* 169(2):709–722.

Singh ND, Davis JC, Petrov DA. 2005. Codon bias and noncoding gc content correlate negatively with recombination rate on the *Drosophila* X chromosome. *J Mol Evol*. 61(3):315–324.

Smith NGC, Eyre-Walker A. 2002. Adaptive protein evolution in *Drosophila*. *Nature* 415(6875):1022–1024.

Tweedie S, Ashburner M, Falls K, Leyland P, McQuilton P, Marygold S, Millburn G, Asumi-Sutherland D, Schroeder A, Seal R, et al. 2009. FlyBase: enhancing *Drosophila* gene ontology annotations. *Nucleic Acids Res*. 37(Database issue):D555–D559.

Vicario S, Moriyama EN, Powell JR. 2007. Codon usage in twelve species of *Drosophila*. *BMC Evol Biol*. 7(1):226–217.

Welch JJ. 2006. Estimating the genomewide rate of adaptive protein evolution in *Drosophila*. *Genetics* 173(2):821–837.

Williamson SH, Hernandez R, Fledel-Alon A, Zhu L, Nielsen R, Bustamante CD. 2005. Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proc Natl Acad Sci U S A*. 102(22):7882–7887.

Wilson DJ, Hernandez RD, Andolfatto P, Przeworski M. 2011. A population genetics-phylogenetics approach to inferring natural selection in coding sequences. *PLoS Genet*. 7(12):e1002395–e1002313.

Wright SI, Iorgovan G, Misra S, Mokhtari M. 2006. Neutral evolution of synonymous base composition in the Brassicaceae. *J Mol Evol*. 64(1):136–141.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 24(8):1586–1591.

Zeng K, Charlesworth B. 2009. Estimating selection intensity on synonymous codon usage in a nonequilibrium population. *Genetics* 183(2):651–662.

Zeng K, Charlesworth B. 2010. Studying patterns of recent evolution at synonymous sites and intronic sites in *Drosophila melanogaster*. *J Mol Evol*. 70(1):116–128.

Zhang L, Li WH. 2005. Human SNPs reveal no evidence of frequent positive selection. *Mol Biol Evol*. 22(12):2504–2507.