Letter

Hitchhiking effects of recurrent beneficial amino acid substitutions in the *Drosophila melanogaster* genome

Peter Andolfatto[1]

Division of Biological Sciences, University of California San Diego, La Jolla, California 92093, USA

Several recent studies have estimated that a large fraction of amino acid divergence between species of *Drosophila* was fixed by positive selection, using statistical approaches based on the McDonald-Kreitman test. However, little is known about associated selection coefficients of beneficial amino acid mutations. Recurrent selective sweeps associated with adaptive substitutions should leave a characteristic signature in genome variability data that contains information about the frequency and strength of selection. Here, I document a significant negative correlation between the level and the frequency of synonymous site polymorphism and the rate of protein evolution in highly recombining regions of the X chromosome of *D. melanogaster*. This pattern is predicted by recurrent adaptive protein evolution and suggests that adaptation is an important determinant of patterns of neutral variation genome-wide. Using a maximum likelihood approach, I estimate the product of the rate and strength of selection under a recurrent genetic hitchhiking model, $\bar{\lambda}2N_e s \sim 3 \times 10^{-8}$. Using an approach based on the McDonald-Kreitman test, I estimate that ~50% of divergent amino acids were driven to fixation by positive selection, implying that beneficial amino acid substitutions are of weak effect on average, on the order of 10^{-5} (i.e., $2N_e s \sim 40$). Two implications of these results are that most adaptive substitutions will be difficult to detect in genome scans of selection and that population size (and genetic drift) may be an important determinant of the evolutionary dynamics of protein adaptation.

[Supplemental material is available online at www.genome.org. The sequence data from this study have been submitted to GenBank under accession nos. EU216760–EU218523.]

Rates of evolution vary considerably among proteins (Li 1997), and the causes of this variation are the topic of continuing controversy. In the view of the neutral or nearly neutral theories, the rate of protein evolution largely reflects a balance between the level of selective constraint on a protein (i.e., negative or purifying selection) and the efficacy of selection (Kimura 1983; Ohta 2002). An opposing view posits that rates of protein evolution depend primarily on the rate at which new advantageous amino acid substitutions arise and how efficiently they can become incorporated (Gillespie 1991).

Several recent studies have used a population genetic approach based on the McDonald-Kreitman (hereafter, MK) test to distinguish between these two models (McDonald and Kreitman 1991). The MK approach is based on comparing levels of polymorphism and divergence of a putatively neutral class of sites (often synonymous sites) with a putatively selected class (e.g., amino acids). The general conclusion has been that a large fraction of the protein divergence between *Drosophila* species is driven by positive selection (Fay et al. 2002; Smith and Eyre-Walker 2002; Sawyer et al. 2003, 2007; Bierne and Eyre-Walker 2004; Andolfatto 2005; Welch 2006). However, little is known about the associated selection coefficients (s) of beneficial amino acid mutations. This is an important quantity in population genetics, as the magnitude of s determines the efficacy of natural selection relative to genetic drift in incorporating beneficial mutations and our ability to detect positive selection in genome scans using closely linked neutral sites (Andolfatto 2001; Eyre-Walker 2006).

Models of genetic hitchhiking predict that regions closely linked to sites experiencing positive selection should harbor reduced levels of neutral polymorphism (Maynard Smith and Haigh 1974; Kaplan et al. 1989; Wiehe and Stephan 1993). In *Drosophila melanogaster*, levels of diversity are reduced in regions of the genome with low rates of crossing-over, suggesting an interaction between selection and linkage (Aguadé et al. 1989; Berry et al. 1991; Begun and Aquadro 1992). The correlation between levels of diversity and recombination rate has been used to estimate the frequency and strength of positive selection under a recurrent hitchhiking model (Wiehe and Stephan 1993; Stephan 1995; Kim and Stephan 2000; Andolfatto 2001; Innan and Stephan 2003; Eyre-Walker 2006; Kim 2006). However, as pointed out by Charlesworth et al. (1993), a similar relationship between levels of diversity and recombination rates is predicted by the effects of recurrent deleterious mutations—the so called "background selection" effect.

The relative importance of recurrent hitchhiking and background selection has been vigorously debated (for review, see Andolfatto 2001), and attempts to distinguish between these models continue to prove difficult (Braverman et al. 1995, 2005; Zurovcova and Eanes 1999; Kim and Stephan 2000; Langley et al. 2000; Andolfatto and Przeworski 2001; Carr et al. 2001; Jensen et al. 2002; Innan and Stephan 2003; but see Bachtrog 2004). This difficulty largely arises from the fact that the predictions of two models can be quite similar (Charlesworth et al. 1995; Gordo et al. 2002), depending on the precise values of parameters (such as recombination rates and the distribution of selection coefficients) that are generally not known. A further difficulty arises

[1]**Corresponding author.**
E-mail pandolfatto@ucsd.edu; fax (858) 534-7108.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.6691007.

from the recent finding that a large fraction of noncoding sites in the *Drosophila* genome shows evidence for being under direct selection (Andolfatto 2005; Bachtrog and Andolfatto 2006), calling into question the suitability of noncoding DNA as neutral sites from which to estimate selection parameters at linked sites.

If adaptive turnover of amino acids is common, more rapidly evolving genes are expected to harbor lower levels of linked synonymous site diversity due to hitchhiking effects associated with linked beneficial amino acid substitutions. I document such a pattern in *D. melanogaster*, and I use a maximum likelihood approach to estimate parameters of adaptive protein evolution under a recurrent hitchhiking model. I then discuss several alternative linkage-selection models that may predict a similar correlation between levels of synonymous diversity and rates of amino acid evolution.

Results

Reduced synonymous site diversity at rapidly evolving proteins

To detect predicted hitchhiking effects associated with recurrent fixation of beneficial amino acid substitutions, I surveyed levels of nucleotide polymorphism in 137 coding regions scattered across the X chromosome of *D. melanogaster* and their level of divergence to *D. simulans*. Since a strong effect of recombination rate on both polymorphism levels (Begun and Aquadro 1992) and amino acid divergence (Betancourt and Presgraves 2002; Haddrill et al. 2007) has been documented in *D. melanogaster*, the regions surveyed here are drawn from the highly recombining portion of the X, for which there is little evidence of recombination rate variation (Charlesworth 1996). To detect hitchhiking effects, I used linked synonymous sites. Previous work has noted selection for codon usage in *Drosophila*, and thus their use as a neutral marker in the genome has been questioned (Akashi 1995; Akashi and Schaeffer 1997). However, there is little evidence for ongoing selection on codon usage or selective constraint on synonymous sites in the *D. melanogaster* genome (Akashi 1995; McVean and Vieira 2001; Andolfatto 2005; Nielsen et al. 2007), suggesting that synonymous polymorphisms may be a suitable "almost neutral" marker for tracking hitchhiking events in this species.

In Figure 1A, I show that levels of synonymous site diversity (π_s) are reduced in coding regions with higher levels of amino acid substitution (K_a; $R = -0.19$; $P \leq 0.025$, rank correlation test). K_a is significantly positively correlated with rates of substitution at synonymous sites, K_s (data not shown; $R = 0.44$; $P \leq 3.4 \times 10^{-7}$, rank correlation test), raising the possibility that variation in mutation rate among genes drives rates of protein evolution (see Discussion). However, variation in mutation rates (quantified as K_s) weakens the negative relationship between π_s and K_a, rather than strengthening it. Indeed, the strength of the correlation between π_s and K_a improves when variation in K_s is removed as a confounding factor (Fig. 1B; partial regression residuals of π_s and K_a; $R = -0.45$; $P < 1.3 \times 10^{-7}$; see also Supplemental materials 2). When loci are binned into three equally sized groups based on levels of amino acid divergence (partial regression residuals), a consistent reduction in average levels of π_s is observed with increasing average K_a ($P \leq 0.01$ for all comparisons, Wilcoxon two-sample test; see Fig. 4 below), suggesting

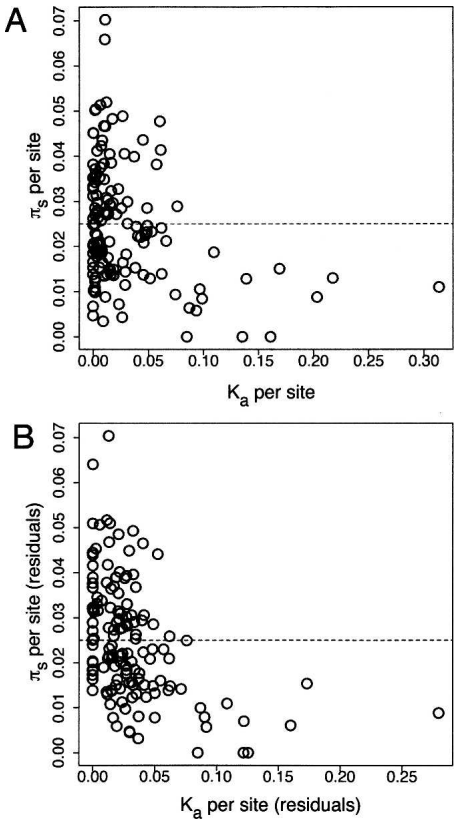

Figure 1. Reduced levels of synonymous site diversity associated with in high rates of protein evolution in *D. melanogaster*. (*A*) Raw estimates of synonymous site diversity (π_s) vs. nonsynonymous divergence (K_a). (*B*) Partial correlation residuals of π_s versus K_a, correcting for K_s. Negative residuals have been set to zero. (Dashed lines) Weighted average diversity across loci.

the correlation between π_s and K_a is not driven by a few rapidly evolving genes.

Reduced codon usage bias at rapidly evolving proteins

The reduction in local N_e associated with repeated selective sweeps is predicted to reduce the efficacy of selection on very weakly selected sites (Hill and Robertson 1966), such as synonymous sites experiencing selection to maintain codon usage bias (Akashi 1995; Akashi and Schaeffer 1997). In line with this prediction, the level of codon usage bias as quantified using the *Fop* index (frequency of preferred codons; Ikemura 1981) is significantly negatively correlated with K_a (Fig. 2; *Fop* vs. partial regression residuals of K_a; $R = -0.32$, $P \leq 1.6 \times 10^{-4}$, Rank correlation test). This confirms previous findings of Betancourt and Presgraves (2002), who used an independent data set and a different codon usage table (see Methods). While there is also a strong negative correlation between *Fop* and K_s, accounting for variation in K_a among genes suggests this correlation is driven by the correlations between K_a and K_s and between *Fop* and K_a (*Fop* vs. partial regression residuals of K_s; $R = -0.09$, $P = 0.31$, Rank correlation test). Though *Fop* and π_s are also significantly correlated, no significant residual correlation remains after accounting for correlations between *Fop* and K_a and between π_s and K_a (data not shown). The observation of reduced levels of codon bias at rapidly evolving genes is consistent with a long-term reduction in

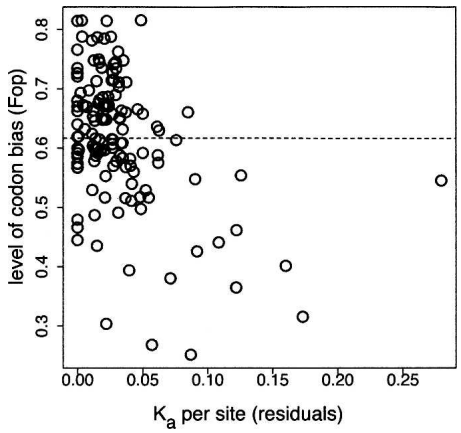

Figure 2. Reduced levels of codon usage bias associated with high rates of protein evolution. Plotted is the frequency of preferred codons (*Fop*) versus partial regression residuals of K_a (corrected for K_s). (Dashed line) Average *Fop* across loci.

N_e, and reduced efficacy of selection maintaining codon bias, associated with elevated K_a (Kim 2004). However, it should be noted that gene expression pattern is also a covariate of both K_a and codon usage bias and may also contribute to the observed correlation (Akashi 2001; Betancourt and Presgraves 2002; Marais et al. 2004).

Estimating parameters under a recurrent hitchhiking model

Reduced levels of linked synonymous site diversity and levels of codon bias in rapidly evolving proteins are predicted by a model that posits recurrent adaptive amino acid substitutions, with more such substitutions per unit time in rapidly evolving genes (Kaplan et al. 1989; Kim 2004). In what follows, I estimate parameters under such a model and then discuss alternative possibilities.

The expected reduction in levels of linked neutral variability under the recurrent hitchhiking model (see Methods) primarily depends on two properties of beneficial mutations: the rate at which they occur (λ) and their average selection coefficients (s). If we assume that all hitchhiking is caused by adaptive protein evolution, and that there is no heterogeneity in the fraction of adaptive divergence among genes (Sawyer et al. 2003, 2007; Bierne and Eyre-Walker 2004; Welch 2006), λ can be directly related to α, the fraction of protein divergence driven to fixation by positive selection (i.e., $\lambda = \alpha K_a/2T$, where T is the species divergence time; see Methods). I implement a maximum likelihood approach to estimate parameters under a recurrent hitchhiking model, using observed levels of synonymous site diversity (π_s) as a function of K_a (see Methods). In Figure 3, I show a joint likelihood surface for θ, the neutral population mutation rate, and the product αs. The maximum likelihood estimates for $\theta = 0.031$ (95% CI 0.028–0.037) per site, and $\alpha s = 6 \times 10^{-6}$ (95% CI $4–12 \times 10^{-6}$). Figure 4 shows the fit of the recurrent hitchhiking model to the observed data, using maximum likelihood estimates of θ and αs and locus-specific parameters.

In the recurrent hitchhiking model employed here, s and α (or λ) are estimated jointly. The relationship between α and s (Fig. 5) implies that the average genome-wide diversity reduction due to recurrent hitchhiking is compatible with either very strong but infrequent selection (i.e., large s and small α) or very common but weak selection (i.e., small s and large α). The average intensity of selection on fixed amino acid mutations, s, can be estimated given an independent estimate of α. Using the maximum likelihood method of Bierne and Eyre-Walker (2004), I estimate $\alpha = 50\%$ for my data set (95% CI 38%–59%; see Supplemental materials 4), a value that is in agreement with several previous estimates of α in *D. melanogaster* (for review, see Eyre-Walker 2006). This estimated value of α implies that $s = 1.2 \times 10^{-5}$, and the scaled intensity of selection $2N_e s \sim 40$ (estimating $\hat{N}_e = 1.78 \times 10^6$ from $\hat{\theta} = 3N_e\mu = 0.031$ and assuming $\mu = 5.8 \times 10^{-9}$ per site per generation; see Methods). While the 95% confidence interval on the estimate of α is relatively large, it is consistent with a narrow range of selection coefficients (i.e., $1.0–1.6 \times 10^{-5}$; Fig. 5). If α is closer to 1, as suggested by Sawyer et al. (2003, 2007), we would infer $s \sim 6 \times 10^{-6}$, and $2N_e s \sim 20$ (Fig. 5). Since $\bar{\lambda} = \alpha\bar{K}_a/2T$, we can also estimate $\bar{\lambda}2N_e s = 3 \times 10^{-8}$ ($\hat{\alpha} = 0.50$, $\bar{K}_a = 0.028$ per site, $\bar{K}_s = 0.147$ per site, $\hat{T} = 10^7$ generations). This estimate is remarkably similar to estimates of $\bar{\lambda}2N_e s$ inferred from the relationship between levels of polymorphism and recombination rate in *D. melanogaster* (Wiehe and Stephan 1993; Stephan 1995; Innan and Stephan 2003; Kim 2006).

Discussion

Distinguishing positive and negative selection models

Strongly reduced diversity at synonymous sites closely linked to rapidly evolving proteins is unexpected under a strictly neutral model and points to an interaction between linkage and selection. A difficulty arises in correctly interpreting this pattern, however, because this correlation is qualitatively consistent with several, very different, types of linkage–selection interaction models. On one hand, elevated K_a in rapidly evolving genes may reflect the recurrent fixation of adaptive amino acid substitutions (the recurrent hitchhiking model). Under this model, amino acid substitutions themselves contribute to local reductions in linked neutral variability (Maynard-Smith and Haigh 1974; Kaplan et al. 1989) and relaxed purifying selection at closely linked synonymous sites (Kim 2004). In support of this view, approaches based on the McDonald-Kreitman framework (McDonald and Kreitman 1991) have estimated that a large fraction (~50% or more) of amino acid divergence between *Drosophila* species was driven to fixation by positive selection (for review, see Eyre-Walker 2006).

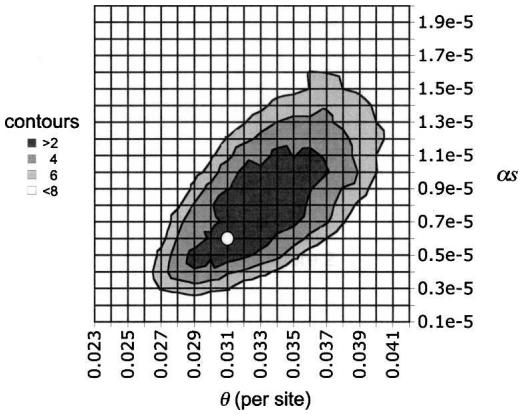

Figure 3. Joint likelihood surface for θ and αs under the recurrent hitchhiking model. (White circle) Maximum likelihood estimate. Each contour represents two units of log likelihood.

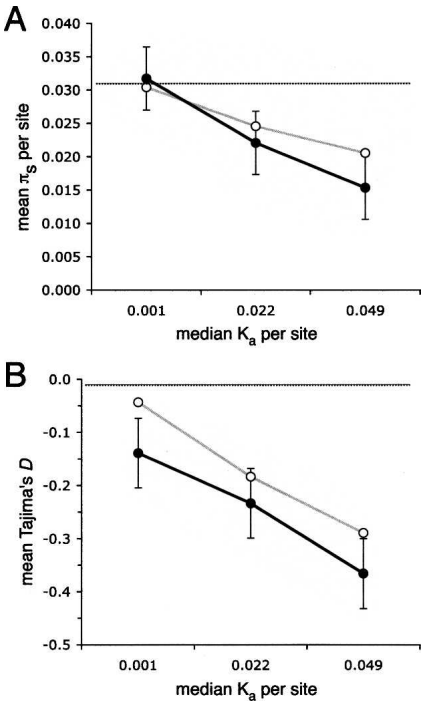

Figure 4. The fit of the recurrent hitchhiking model to the data using maximum likelihood estimates of θ and αs, and locus-specific parameters. (*A*) The fit to average levels of synonymous site diversity (π_s). (*B*) The fit to average Tajima's *D*. Observed and simulated data are binned into three categories based on levels of amino acid divergence (K_a). Plotted are the mean of 1000 simulated replicates (white circles, gray lines) and the mean for the observed data (black circles, solid lines) with standard errors. (Horizontal dotted lines) Expectation under the standard neutral model with recombination.

On the other hand, the removal of weakly deleterious mutations from a population can also result in reductions in the local N_e of a genomic region (the background selection model; Manning and Thompson 1984; Charlesworth et al. 1993). This model, in turn, predicts local reductions in levels of linked neutral diversity and the differential accumulation of very weakly deleterious mutations including unpreferred synonymous mutations (reducing codon bias) and some fraction of amino acid substitutions (potentially elevating K_a; Charlesworth 1994). Thus, it is theoretically possible that elevated levels of amino acid substitution are not the *cause* of reductions in local levels of synonymous variability and codon usage bias per se, but rather all three are a secondary by-product of the removal of linked deleterious mutations by negative selection.

A third model lies somewhere in between the pure recurrent hitchhiking and background selection models and posits an interaction between positively and negatively selected amino acid mutations. The fixation of a positively selected mutation can interfere with purifying selection at closely linked sites by dragging slightly deleterious mutations either to high frequency or to fixation. Thus, like background selection, recurrent positive selection can reduce the local N_e of a genomic region, and thus the efficacy of purifying selection on linked weakly deleterious mutations (Hill and Robertson 1966). I invoked a version of this model above to account for reductions in levels of codon bias at synonymous sites closely linked to rapidly evolving proteins under the pure recurrent hitchhiking interpretation of the data. If there is also a sizeable class of very weakly deleterious amino acid polymorphisms (i.e., $-10 < 2N_e s < -1$), then interference between positively selected and weakly deleterious amino acid mutations may reduce the efficacy of selection on the latter, increasing their probability of fixation. Though mechanistically distinct from background selection, this model makes the similar qualitative prediction that slightly deleterious mutations may accumulate faster in rapidly evolving genes.

Distinguishing among these models is not a simple task because their relative importance depends on little-known parameters such as the distribution of selection coefficients and the scale at which mutation and recombination rates vary in the genome. In what follows, I point to several empirical patterns suggesting that variation in the efficacy of purifying selection on amino acid mutations, whether due to background selection or recurrent positive selection, is probably of limited importance in explaining variation in K_a among genes.

Variation in recombination rate is predicted to result in heterogeneity in levels of background selection and, thus, variation in local N_e and the efficacy of selection across a chromosome (see equations 8 and 9 of Hudson and Kaplan 1995). While there is little evidence for recombination rate heterogeneity across the X chromosome interval surveyed (Charlesworth 1996), K_a is not significantly correlated with recombination rates estimated for these regions using several other approaches (see Supplemental materials 3). Even if there were some cryptic variation in recombination rates, proteins do not tend to evolve faster in regions of low but non-zero recombination, where the efficacy of selection is predicted to be reduced (Betancourt and Presgraves 2002; Haddrill et al. 2007). One interpretation of this finding is that there are too few amino acid mutations with selection coefficients small enough (i.e., $-10 < 2N_e s < -1$) to lead to significant accumulations in such regions. Another possibility is that gene conversion (Langley et al. 2000; Andolfatto and Wall 2003) or low levels of crossing-over are sufficient to prevent the accumulation of slightly deleterious amino acid mutations (Haddrill et al. 2007).

Even in the absence of variation in recombination rates,

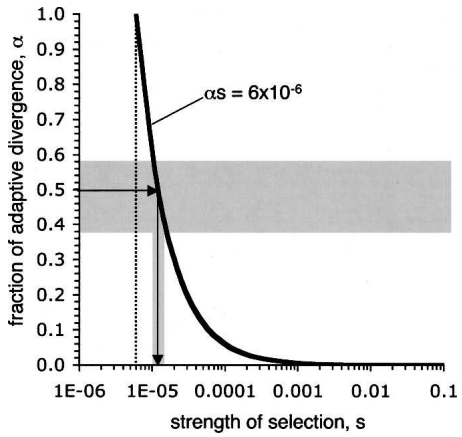

Figure 5. The fraction of amino acid mutations fixed by positive selection (α) versus their selective advantage s for $\alpha s = 6 \times 10^{-6}$ (solid line). (Arrow) Maximum likelihood estimate of α using the approach of Bierne and Eyre-Walker (2004) (see Supplemental materials 6), (shaded area) 95% CI for the estimate of α, (vertical dotted line) minimum possible selection coefficent ($\alpha = 1$; $s = 6 \times 10^{-6}$).

variation in N_e (and the efficacy of selection) due to background selection could be induced by variation in the deleterious mutation rate (Hudson and Kaplan 1995). At first glance, the significant positive correlation between K_s and K_a (Supplemental materials 2) would seem to support this view: Regions of the genome with elevated mutation rates (indicated by elevated K_s) may experience greater levels of background selection (as indicated by reduced π_s and elevated K_a). However, the correlation between π_s and elevated K_a actually becomes *stronger* when accounting for variation in K_s (see Results and Supplemental materials 2). Thus, rather than driving the correlation between π_s and elevated K_a (as predicted under the background selection model), variation in mutation rate appears to obscure the pattern by inflating π_s at rapidly evolving proteins. Spatial genomic variation in the deleterious mutation rate (and thus the level of background selection) could be induced by variation in gene density or the fraction of the local genome that is functionally important, even in the absence of mutation rate variation. However, I find no evidence for an effect of gene density on levels of π_s or K_a in this data set (see Supplemental materials 4), suggesting that variation in the local deleterious mutation rate has little effect on K_a.

Implications of recurrent weak positive selection

While the relative contributions of other forms of linked selection to the observed correlation between π_s and K_a remain unclear, recurrent hitchhiking due to a high rate of weakly positively selected amino acid substitution is sufficient to account for the observed pattern (Fig. 4). Gillespie (2000) posited that, if adaptation is common, neutral variation in the genome may be more strongly influenced by selection through the effects of genetic linkage than by genetic drift—the "genetic draft" model. If valid, levels of neutral genome variability may be an unreliable predictor of species population size (Maynard Smith and Haigh 1974; Gillespie 2000). My estimates suggest a modest average reduction in diversity due to recurrent hitchhiking (i.e., $\hat{\theta} = 3.1\%$ and the observed $\overline{\pi}_s = 2.5\%$). While not quite in the realm of "genetic draft," adaptive amino acid evolution in *Drosophila* does appear to be common enough to have a sizeable impact on average levels of genome variability, even in regions of high recombination.

In addition, a recurrent hitchhiking model predicts a genome-wide skew of the neutral polymorphism frequency spectrum toward low-frequency polymorphisms (Braverman et al. 1995; Przeworski 2002; Kim 2006). In fact, the polymorphism frequency spectrum is generally negative for synonymous sites in *D. melanogaster* (Tajima's $\overline{D} = -0.28$ for the loci studied here) and several other *Drosophila* species (Kliman et al. 2000; Machado et al. 2002; Kopp and Barmina 2005; Bachtrog and Andolfatto 2006; Bachtrog et al. 2006), a skew that has largely been interpreted as a signature of population growth. In *D. melanogaster*, however, this negative skew in the frequency spectrum is more severe at rapidly evolving genes, as predicted under a recurrent hitchhiking model (Fig. 4). Moreover, in simulations of a recurrent hitchhiking model using estimated parameters for θ and αs, the expected Tajima's $\overline{D}$ is -0.19 with 95% CI $[-0.31, -0.05]$, suggesting that recurrent adaptive amino acid substitutions are sufficient to account for the negative skew in the frequency spectrum at linked synonymous sites.

Using estimates of adaptive divergence based on the MK approach and the fit to the recurrent hitchhiking model, the average strength of selection on adaptive amino acid substitutions is inferred to be quite weak (i.e., $s \sim 10^{-5}$). Thus, it may be difficult to detect most adaptive substitutions occurring in regions of high recombination using linked neutral variability. For example, if $s \sim 10^{-5}$ on average, the expected window of reduced variability surrounding most adaptive amino acid substitutions is expected to be <100 base pairs in highly recombining regions of the *Drosophila* genome (assuming $\rho = 2$ cM/Mb; Kaplan et al. 1989). This implies that most adaptive substitutions will be missed in standard genome-wide scans for selection, leading to an underestimate of the frequency of adaptive evolution. The few adaptive substitutions that are detected will tend to be drawn from the tail of a distribution of selection coefficients, and may not be representative of most adaptation (Teshima et al. 2006).

It is important to note that several simplifying assumptions of the recurrent hitchhiking model that I have employed may have led to an underestimate of s. The first is the assumption that positively selected mutations are drawn primarily from newly arising mutations rather than drawn from standing variation (Dykhuizen and Hartl 1980; Orr and Betancourt 2001; Innan and Kim 2004; Przeworski et al. 2005). Similarly, selection is assumed to be constant over the sojourn time of beneficial substitutions rather than fluctuating over time (Takahata et al. 1975; Gillespie 1991; Orr and Betancourt 2001; Mustonen and Lassig 2007). Since the expected sojourn time of a positively selected allele ($\sim 2\ln(2N)/s$ for a newly arising mutation) is longer when s is small, the likelihood that intensity or the direction of selection will change over the course of its fixation increases. Generally, this should cause the strength of selection to be underestimated using the method employed in this study. Future work could be aimed at estimating the timescale and intensity of these fluctuations (e.g., see Mustonen and Lassig 2007).

A second assumption made here is that there is no interference among selected mutations. This has already been discussed above in the context of interference between positively and negatively selected mutations (above). Interference between positively selected amino acid substitutions and weak purifying selection to maintain codon usage bias is hinted at by the relationship between K_a and the codon usage index *Fop* (see Results). However, given the high frequency of adaptive divergence and small selection coefficients inferred for adaptive amino acid substitutions, segregating advantageous mutations may be common enough to interfere with one another (Fisher 1930; Muller 1932), particularly in rapidly evolving proteins. Given rough estimates of the rate and strength of selection made here, a useful next step may be to revisit forward-simulation approaches to ask how interference may influence estimates of the frequency and strength of positive selection (e.g., Kim and Stephan 2003).

These caveats aside, positing much stronger selection coefficients for adaptive amino acid substitutions would imply that only a very small fraction of amino acid divergence between *D. melanogaster* and *D. simulans* is adaptive. For example, positing that $\bar{s} > 0.001$ would imply that <1% of amino acid divergence between these species was driven to fixation by positive selection, which is clearly at odds with inferences of the adaptive fraction of divergence (α) based on the framework of the McDonald-Kretiman test. Estimates of α by these methods would have to be off by more than an order of magnitude to change the conclusion that weakly beneficial amino acid mutations are common. The existence of a large class of weakly beneficial mutations implies that changes in population size may considerably limit the adaptive potential of a species (Ohta 2002). For example, it

appears that humans and chimpanzees have undergone less adaptive protein evolution than bacteria, viruses, and *Drosophila* (for review, see Eyre-Walker 2006). Since hominids are thought to have population sizes more than two orders of magnitude smaller than *Drosophila*, many of the adaptive amino acid substitutions occurring in *Drosophila* would be effectively neutral.

Methods

Survey of coding regions

The 137 loci surveyed in this study are X-linked coding regions with a sample size of 12 *D. melanogaster* alleles from a Zimbabwe, Africa population. A single *D. simulans* sequence was also surveyed, where possible, to provide estimates of divergence. A subset (31) of these genes was previously surveyed in Andolfatto (2005). For 13 regions that I could not amplify from *D. simulans*, I used available *D. simulans* genome sequence (http://rana.lbl.gov/drosophila). For four loci, *D. simulans* genome sequence was unavailable and so I used *D. sechellia* genome sequence as a surrogate. All genes were selected randomly with respect to gene function and fall between cytological positions 3C3 and 18F4 (see Supplemental materials, Fig. S1.1). This interval on the X chromosome comprises the most highly recombining portion of the X chromosome, and there is little evidence for heterogeneity in recombination rate (Supplemental materials, Fig. S1.1). Information about the specific loci surveyed and primers used can be found in the Supplemental materials, Table S1.1.

Each 700- to 800-bp region was PCR-amplified from genomic DNA extracted from single male flies, and primers and nucleotides were removed using exonuclease I and shrimp alkaline phosphatase. Cleaned products were sequenced on both strands using Big-Dye (version 3, Applied Biosystems) and run on an ABI 3730 capillary sequencer. Sequence traces were edited using Sequencher (Gene Codes) software, and multiple sequence alignments were generated using MUSCLE (http://www.drive5.com/muscle/) with protein-alignment-assisted adjustments to preserve reading frames. Sequences have been deposited in GenBank under accession nos. EU216760–EU218523, and FASTA alignments are available from the website http://www.biology.ucsd.edu/labs/andolfatto.

Classifying preferred codons in *D. melanogaster*

Following Bachtrog (2007), I classified synonymous codons into preferred (P) and unpreferred (U), based on the *D. melanogaster* genome annotation (Release 4.7). The ''scaled'' X^2 measure (Shields et al. 1988) was used to quantify the level of bias in a given gene. The X^2 deviation from expectations based on the average A+T content for short introns (introns <87-bp long, A+T = 64.7%) was calculated for each synonymous family (excluding the codon family being analyzed). The sum of X^2 values is divided by the total number of codons to yield a measure of codon bias that is roughly independent of gene length. Preferred codons are identified as those whose frequencies (within a synonymous family) show significant positive correlations with the degree of bias (Akashi 1995). The resulting codon preference table (see Supplemental materials 7) is identical to that for *D. pseudoobscura* (Bachtrog 2007), with the exception of glycine for which the "GGG" codon is not preferred in *D. melanogaster*.

Polymorphism and divergence analysis

The estimated number of synonymous sites, nonsynonymous sites, average pairwise diversity (π), average pairwise divergence to *D. simulans* (D_{xy}), as well as counts of the number of polymorphisms (S) and the summary of the frequency distribution of polymorphism frequencies, Tajima's D (Tajima 1989), were calculated using a library of Perl scripts ("Polymorphorama") written by me and D. Bachtrog. The number of nonsynonymous and synonymous sites was estimated using the method of Nei and Gojobori (1986). Average pairwise diversity (π) and divergence (D_{xy}) estimates were either corrected for multiple hits using a Jukes–Cantor correction (Jukes and Cantor 1969) or, in the case of synonymous divergence, the Kimura (1980) two-parameter model. Multiply hit sites were included in all analyses, but insertion–deletion polymorphisms and polymorphic sites overlapping alignment gaps were excluded.

Estimating recurrent hitchhiking parameters

I employ the relationship between synonymous site diversity (π_s) and amino acid divergence (K_a) to quantify recurrent hitchhiking parameters, such as the strength of selection (s) and the rate of adaptive substitution per site per generation (λ). To illustrate how recurrent hitchhiking is expected to impact levels of linked neutral variability, it is useful to consider the analytical approximation of Wiehe and Stephan (1993). The expected nucleotide diversity at neutral sites is

$$E(\pi) = \frac{\theta\rho}{\rho + k\gamma\lambda}, \quad (1)$$

where θ is the population mutation rate, ρ is the recombination rate per site per generation, k is a constant ≈ 0.075, and γ ($=2N_e s$) is the intensity of positive selection (where N_e is the effective population size of the species and s is the strength of selection). Here, I assume that the rate of selective sweeps in the vicinity of a focal neutral site is determined by the local rate of amino acid substitution at gene i (i.e., $K_{a,i}$). I assume that some fraction of divergent amino acids at each locus, α, was driven to fixation by positive selection, and that this fraction is constant across genes (as inferred in *Drosophila*; see Bierne and Eyre-Walker 2004; Welch 2006). Thus, if the local rate of selective sweeps due to recurrent adaptive amino acid substitutions is $\lambda_i = \alpha K_{a,i}/2T$, where T is the divergence time in generations between *D. melanogaster* and *D. simulans*, equation (1) can be rewritten as

$$E(\pi_i) = \frac{\theta\rho_i}{\rho_i + k2N_e s(\alpha K_{a,i}/2T)}, \quad (2)$$

where ρ_i is the local rate of recombination at locus i. Equation 2 allows levels of neutral diversity to be directly related to the extent of adaptive amino acid divergence at each locus (i.e., $\alpha K_{a,i}$). Estimating $T = (\overline{K}_s - \theta)/2\mu$ and $N_e = \theta/3\mu$ requires an estimate of the neutral mutation rate, μ. When needed, $\hat{\mu}$ is assumed to be 5.8×10^{-9} per generation, the estimated average rate for single nucleotide mutations from *D. melanogaster* mutation-accumulation lines (Haag-Liautard et al. 2007). However, Equation 2 can further be rearranged to show that it depends on the parameters θ and the product αs as

$$E(\pi_i) = \frac{\theta\rho_i}{\rho_i + \alpha s K_{a,i}(kN_e/T)}. \quad (3)$$

Since T is scaled in units of the effective population size, the expected reduction in variability under recurrent hitchhiking, $E(\pi_i)/\theta$, is independent of μ.

While the analytical theory serves as a useful guide, I use the recurrent selective sweep coalescent simulation machinery described in Jensen et al. (2007) to co-estimate θ and αs in a maximum likelihood framework. The simulation machinery was modified by K. Thornton to account for the stochastic trajecto-

ries of positively selected alleles in finite populations (following Coop and Griffiths 2004; Przeworski et al. 2005; K. Thornton, unpubl.) and is appropriately applied to cases where positive selection is potentially weak. For each locus, I simulated 10,000 replicates of the recurrent hitchhiking coalescent with locus-specific parameters: the sample size n_i, the locus length L_i, the population mutation rate θL_i, the population recombination rate $(\rho_i/\hat{\mu})\theta L_i$, and the local rate of selective sweeps $\lambda_i = \alpha K_{a,i}/2\hat{T}$ (where $\hat{T} = (\overline{K}_s - \theta)/2\hat{\mu}$). Global parameters included N_e and s. The dynamics of positively selected alleles prior to becoming established in a population depend on the strength of genetic drift (i.e., N_e) relative to selection (s). Since s is not known a priori, I repeated simulations over several values of N_e and s, to check the sensitivity of the inference procedure (data not shown).

Given little evidence for recombination rate variation across this interval of the X (Charlesworth 1996) and the lack of a correlation between π_s and several estimates of variation in the local recombination rate (see Discussion), I set $\rho_i = \hat{\rho} = 2.27 \times 10^{-8}$ per site per generation (Andolfatto and Przeworski 2001; Charlesworth 1996). Incorporating locus-specific estimates of ρ had little effect on the results, which follows from the fact that estimates of the local recombination rate are not strong predictors of π_s (Supplemental materials 3). Though variation in mutation rate might be expected across loci, I minimized this by performing a partial regression of π_s versus K_a, controlling for variation in K_s (implemented with the JMP5.1 software package, http://www.jmp.com; see Results). The residuals from this partial regression were used as the observed values of π_s and K_a for each locus. Negative residuals were set to zero.

Simulations covered a 20×20 grid of θ and αs values. For each of these grid points, the likelihood of locus i was approximated as $Lik(\theta,\alpha s|\pi_{s,i},P_i) \propto \Pr(\pi_{s,i} \pm \delta|\theta,\alpha s,P_i)$, where P_i represents locus-specific simulation parameters (including λ_i), and δ is a tolerance that was arbitrarily set to 0.1. θ and αs are estimated as the values that maximize the product over loci:

$$\prod Lik(\theta,\alpha s|\pi_{s,i},P_i). \quad (4)$$

The implicit assumption of independence among loci in this approach is supported by the lack of spatial autocorrelation in π_s estimates (see Supplemental materials 5). 95% credibility intervals are estimated as being within two log-likelihood units of the maximum likelihood estimate (i.e., using the standard X^2 approximation). I demonstrate the fit of the estimated parameters to the data by simulating 1000 replicates of the recurrent hitchhiking coalescent for each of the 137 loci using maximum likelihood estimates of θ and αs and locus-specific parameters and compare simulated and observed mean π and mean Tajima's D as a function of the rate of amino acid evolution.

When inferring selection parameters from levels of diversity under a recurrent hitchhiking model, α and s are confounded and only their product is estimated. To estimate α independently of s, I estimate α using the approach of Bierne and Eyre-Walker (2004) (see Supplemental materials 6). The number of divergent nonsynonymous and synonymous mutations (D) was estimated as $D_{xy} - \pi$, following Andolfatto (2005), and negative values of D were set to zero. I exclude singleton and doubleton mutations (i.e., all mutations with minor allele frequency $\leq 2/12$) to minimize the downward bias of the estimator due to segregating deleterious amino acid mutations (Andolfatto 2005).

Acknowledgments

I thank K. Thornton for generous help implementing the recurrent hitchhiking simulations and for donating computer cluster time. D. Bachtrog is coauthor of the Polymorphorama library and kindly provided code to classify preferred codons. D. Bachtrog, G. Coop, J. Jensen, M. Przeworski, K. Thornton, and four anonymous reviewers provided numerous insightful comments on the manuscript and analyses. I thank B. Fischman and K. Wong for technical assistance.

References

Aguadé, M., Miyashita, N., and Langley, C. 1989. Restriction-map variation at the zeste-tko region in natural populations of *Drosophila melanogaster*. *Mol. Biol. Evol.* **6:** 123–130.

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

Akashi, H. 2001. Gene expression and molecular evolution. *Curr. Opin. Genet. Dev.* **11:** 660–666.

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

Andolfatto, P. 2001. Adaptive hitchhiking effects on genome variability. *Curr. Opin. Genet. Dev.* **11:** 635–641.

Andolfatto, P. 2005. Adaptive evolution of non-coding DNA in *Drosophila*. *Nature* **437:** 1149–1153.

Andolfatto, P. and Przeworski, M. 2001. Regions of lower crossing over harbor more rare variants in African populations of *Drosophila melanogaster*. *Genetics* **158:** 657–665.

Andolfatto, P. and Wall, J.D. 2003. Linkage disequilibrium patterns across a recombination gradient in African *Drosophila melanogaster*. *Genetics* **165:** 1289–1305.

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

Bachtrog, D. 2007. Reduced selection for codon usage bias in *Drosophila miranda*. *J. Mol. Evol.* **64:** 586–590.

Bachtrog, D. and Andolfatto, P. 2006. Selection, recombination and demographic history in *Drosophila miranda*. *Genetics* **174:** 2045–2059.

Bachtrog, D., Thornton, K., Clark, A., and Andolfatto, P. 2006. Extensive introgression of mitochondrial DNA relative to nuclear gene flow in the *Drosophila yakuba* species group. *Evolution Int. J. Org. Evolution* **60:** 292–302.

Begun, D.J. and Aquadro, C.F. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *Drosophila melanogaster*. *Nature* **356:** 519–520.

Berry, A.J., Ajioka, J.W., and Kreitman, M. 1991. Lack of polymorphism on the *Drosophila* fourth chromosome resulting from selection. *Genetics* **129:** 1111–1117.

Betancourt, A.J. and Presgraves, D.C. 2002. Linkage limits the power of natural selection in *Drosophila*. *Proc. Natl. Acad. Sci.* **99:** 13616–13620.

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

Braverman, J., Hudson, R., Kaplan, N., Langley, C., and Stephan, W. 1995. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics* **140:** 783–796.

Braverman, J.M., Lazzaro, B.P., Aguade, M., and Langley, C.H. 2005. DNA sequence polymorphism and divergence at the *erect wing* and *suppressor of sable* loci of *Drosophila melanogaster* and *D. simulans*. *Genetics* **170:** 1153–1165.

Carr, M., Soloway, J.R., Robinson, T.E., and Brookfield, J.F. 2001. An investigation of the cause of low variability on the fourth chromosome of *Drosophila melanogaster*. *Mol. Biol. Evol.* **18:** 2260–2269.

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

Charlesworth, B. 1996. Background selection and patterns of genetic diversity in *Drosophila melanogaster*. *Genet. Res.* **68:** 131–149.

Charlesworth, B., Morgan, M.T., and Charlesworth, D. 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics* **134:** 1289–1303.

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

Coop, G. and Griffiths, R.C. 2004. Ancestral inference on gene trees under selection. *Theor. Popul. Biol.* **66:** 219–232.

Dykhuizen, D. and Hartl, D.L. 1980. Selective neutrality of 6PGD

allozymes in *E. coli* and the effects of genetic background. *Genetics* **96:** 801–817.
Eyre-Walker, A. 2006. The genomic rate of adaptive evolution. *Trends Ecol. Evol.* **21:** 569–575.
Fay, J.C., Wyckoff, G.J., and Wu, C.I. 2002. Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. *Nature* **415:** 1024–1026.
Fisher, R.A. 1930. *The genetical theory of natural selection*. Oxford University Press, Oxford.
Gillespie, J.H. 1991. *The causes of molecular evolution*. Oxford University Press, Oxford.
Gillespie, J.H. 2000. Genetic drift in an infinite population. The pseudohitchhiking model. *Genetics* **155:** 909–919.
Gordo, I., Navarro, A., and Charlesworth, B. 2002. Muller's ratchet and the pattern of variation at a neutral locus. *Genetics* **161:** 835–848.
Haag-Liautard, C., Dorris, M., Maside, X., Macaskill, S., Halligan, D.L., Charlesworth, B., and Keightley, P.D. 2007. Direct estimation of per nucleotide and genomic deleterious mutation rates in *Drosophila*. *Nature* **445:** 82–85.
Haddrill, P., Halligan, D., Tomaras, D., and Charlesworth, B. 2007. Reduced efficacy of selection in regions of the *Drosophila* genome that lack crossing over. *Genome Biol.* **8:** R18. doi: 10.1186/gb-2007-8-2-r18.
Hill, W.G. and Robertson, A. 1966. The effect of linkage on limits to artificial selection. *Genet. Res.* **8:** 269–294.
Hudson, R. and Kaplan, N. 1995. The coalescent process and background selection. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **349:** 19–23.
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* system. *J. Mol. Biol.* **151:** 389–409.
Innan, H. and Kim, Y. 2004. Pattern of polymorphism after strong artificial selection in a domestication event. *Proc. Natl. Acad. Sci.* **101:** 10667–10672.
Innan, H. and Stephan, W. 2003. Distinguishing the hitchhiking and background selection models. *Genetics* **165:** 2307–2312.
Jensen, M.A., Charlesworth, B., and Kreitman, M. 2002. Patterns of genetic variation at a chromosome 4 locus of *Drosophila melanogaster* and *D. simulans*. *Genetics* **160:** 493–507.
Jensen, J.D., Thornton, K.R., Bustamante, C.D., and Aquadro, C.F. 2007. On the utility of linkage disequilibrium as a statistic for identifying targets of positive selection in non-equilibrium populations. *Genetics* **176:** 2371–2379.
Jukes, T.H. and Cantor, C. 1969. Evolution of protein molecules. In *Mammalian protein metabolism* (ed. M.N. Munro), pp. 21–132. Academic Press, New York.
Kaplan, N., Hudson, R., and Langley, C. 1989. The "hitchhiking effect" revisited. *Genetics* **123:** 887–899.
Kim, Y., 2004. Effect of strong directional selection on weakly selected mutations at linked sites: Implication for synonymous codon usage. *Mol. Biol. Evol.* **21:** 286–294.
Kim, Y., 2006. Allele frequency distribution under recurrent selective sweeps. *Genetics* **172:** 1967–1978.
Kim, Y. and Stephan, W. 2000. Joint effects of genetic hitchhiking and background selection on neutral variation. *Genetics* **155:** 1415–1427.
Kim, Y. and Stephan, W. 2003. Selective sweeps in the presence of interference among partially linked loci. *Genetics* **164:** 389–398.
Kimura, M. 1980. A simple model for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16:** 111–120.
Kimura, M. 1983. *The neutral theory of molecular evolution*. Cambridge University Press, Cambridge.
Kliman, R., Andolfatto, P., Coyne, J., Depaulis, F., Kreitman, M., Berry, A., McCarter, J., Wakeley, J., and Hey, J. 2000. The population genetics of the origin and divergence of the *Drosophila simulans* complex species. *Genetics* **156:** 1913–1931.
Kopp, A. and Barmina, O. 2005. Evolutionary history of the *Drosophila bipectinata* species complex. *Genet. Res.* **85:** 23–46.
Langley, C.H., Lazzaro, B.P., Phillips, W., Heikkinen, E., and Braverman, J.M. 2000. Linkage disequilibria and the site frequency spectra in the *su(s)* and *su(w^a)* regions of the *Drosophila melanogaster* X chromosome. *Genetics* **156:** 1837–1852.
Li, W. 1997. *Molecular evolution*. Sinauer Associates, Inc., Sunderland, MA.
Machado, C., Kliman, R., Markert, J., and Hey, J. 2002. Inferring the history of speciation from multilocus DNA sequence data: The case of *Drosophila pseudoobscura* and close relatives. *Mol. Biol. Evol.* **19:** 472–488.
Manning, J.T. and Thompson, D.J. 1984. Mullers ratchet and the accumulation of favorable mutations. *Acta Biotheor.* **33:** 219–225.
Marais, G., Domazet-Loso, T., Tautz, D., and Charlesworth, B. 2004. Correlated evolution of synonymous and nonsynonymous sites in *Drosophila*. *J. Mol. Evol.* **59:** 771–779.
Maynard Smith, J. and Haigh, J. 1974. The hitch-hiking effect of a favourable gene. *Genet. Res.* **23:** 23–35.
McDonald, J.H. and Kreitman, M. 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351:** 652–654.
McVean, G.A. and Vieira, J. 2001. Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in *Drosophila*. *Genetics* **157:** 245–257.
Muller, H.J. 1932. Some genetic aspects of sex. *Am. Nat.* **66:** 118–138.
Mustonen, V. and Lassig, M. 2007. Adaptations to fluctuating selection in *Drosophila*. *Proc. Natl. Acad. Sci.* **104:** 2277–2282.
Nei, M. and Gojobori, T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3:** 418–426.
Nielsen, R., Bauer-DuMont, V. L., Hubisz, M.J., and C. F. Aquadro, 2007. Maximum likelihood estimation of ancestral codon usage bias parameters in *Drosophila*. *Mol. Biol. Evol.* **24:** 228–235.
Ohta, T. 2002. Near-neutrality in evolution of genes and gene regulation. *Proc. Natl. Acad. Sci.* **99:** 16134–16137.
Orr, H.A. and Betancourt, A.J. 2001. Haldane's sieve and adaptation from the standing genetic variation. *Genetics* **157:** 875–884.
Przeworski, M. 2002. The signature of positive selection at randomly chosen loci. *Genetics* **160:** 1179–1189.
Przeworski, M., Coop, G., and Wall, J.D. 2005. The signature of positive selection on standing genetic variation. *Evolution Int. J. Org. Evolution* **59:** 2312–2323.
Sawyer, S., Kulathinal, R., Bustamante, C., and Hartl, D. 2003. Bayesian analysis suggests that most amino acid replacements in *Drosophila* are driven by positive selection. *J. Mol. Evol.* **57S1:** 154–164.
Sawyer, S., Parsch, J., Zhang, Z., and Hartl, D. 2007. Prevalence of positive selection among nearly neutral amino acid replacements in *Drosophila*. *Proc. Natl. Acad. Sci.* **104:** 6504–6510.
Shields, D.C., Sharp, P.M., Higgins, D.G., and Wright, F. 1988. "Silent" sites in *Drosophila* genes are not neutral: Evidence of selection among synonymous codons. *Mol. Biol. Evol.* **5:** 704–716.
Smith, N.G. and Eyre-Walker, A. 2002. Adaptive protein evolution in *Drosophila*. *Nature* **415:** 1022–1024.
Stephan, W. 1995. An improved method for estimating the rate of fixation of favorable mutations based on DNA polymorphism data. *Mol. Biol. Evol.* **12:** 959–962.
Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123:** 585–595.
Takahata, N., Iishi, K., and Matsuda, H. 1975. Effect of temporal fluctuations of selection coefficient on gene frequency in a population. *Proc. Natl. Acad. Sci.* **72:** 4541–4545.
Teshima, K., Coop, G., and Przeworski, M. 2006. How reliable are empirical genomic scans for selective sweeps? *Genome Res.* **16:** 702–712.
Welch, J. 2006. Estimating the genomewide rate of adaptive protein evolution in *Drosophila*. *Genetics* **173:** 821–837.
Wiehe, T.H. and Stephan, W. 1993. Analysis of a genetic hitchhiking model, and its application to DNA polymorphism data from *Drosophila melanogaster*. *Mol. Biol. Evol.* **10:** 842–854.
Zurovcova, M. and Eanes, W. 1999. Lack of nucleotide polymorphism in the Y-linked sperm flagellar dynein gene *Dhc-Yh3* of *Drosophila melanogaster* and *D. simulans*. *Genetics* **153:** 1709–1715.

Received May 8, 2007; accepted in revised form October 2, 2007.