Genomic Effects of Nucleotide Substitutions in *Drosophila simulans*

Andrew D. Kern,[1,2] Corbin D. Jones[2] and David J. Begun

Center for Population Biology, University of California, Davis, California 95616

Manuscript received April 22, 2002
Accepted for publication September 26, 2002

ABSTRACT

Selective fixation of beneficial mutations reduces levels of linked, neutral variation. The magnitude of this "hitchhiking effect" is determined by the strength of selection and the recombination rate between selected and neutral sites. Thus, depending on the values of these parameters and the frequency with which directional selection occurs, the genomic scale over which directional selection reduces levels of linked variation may vary widely. Here we present a permutation-based analysis of nucleotide polymorphisms and fixations in *Drosophila simulans*. We show evidence of pervasive small-scale hitchhiking effects in this lineage. Furthermore, our results reveal that different types of fixations are associated with different levels of linked variation.

FIXATION of beneficial mutations results in reductions of linked, neutral variation. The scale of this hitchhiking effect (Maynard-Smith and Haigh 1974) reflects the relative strengths of selection and recombination during a selected mutant's sojourn through the population (Hill and Robertson 1966; Kaplan *et al.* 1989). In general, a reduction of linked variation is expected to occur over a physical region defined by the ratio of the recombination rate per base pair per generation and the selection coefficient. Therefore, if many beneficial nucleotide fixations result from weak selection, the associated hitchhiking effects may occur over only tens or hundreds of bases (Kaplan *et al.* 1989). Despite recent work demonstrating that regional levels of nucleotide polymorphism across the genome in Drosophila and other organisms are affected by selection at linked sites (Berry *et al.* 1991; Begun and Aquadro 1992; Langley *et al.* 1993; Aquadro *et al.* 1994; Nachman 1997, 2001; Nachman *et al.* 1998), the possibility of small-scale selective perturbations of polymorphism has not been investigated.

We can study the importance of small-scale hitchhiking effects by determining if levels of polymorphism are reduced near nucleotide sites that have fixed in the recent past. If some fraction of such sites fixed under directional selection, we may observe less heterozygosity in regions flanking fixations compared to randomly selected regions of DNA. Moreover, *a priori* categorization of fixations allows us to investigate possible heterogeneity of the substitution process across mutant classes. For example, if a greater fraction of amino acid fixations result from selection (compared to silent or noncoding fixations), the level of polymorphism in regions near amino acid fixations may be reduced relative to that in regions near silent fixations. More generally, sites experiencing stronger or more recent directional selection should be associated with regions of lower heterozygosity.

Here we develop a permutation-based test for detecting small-scale reductions of heterozygosity. Using standard methods from meta-analysis (Fisher 1935, 1954; Glass 1976; Sokal and Rohlf 1995; Good 2000), we applied our permutation test to a sample of *Drosophila simulans* genes to see if particular types of fixations are associated with small-scale regions of reduced DNA polymorphism.

MATERIALS AND METHODS

Sequence data: The names, physical locations, and summary statistics of variation for the loci used in our analysis are in Table 1. Most of the sequence data we used are from Begun and Whitley (2000) and references therein (accession nos. AF204277–AF204290, AF256057–AF256078, AF252637–AF252824, AF255311–AF255314, AF255316–AF255320, AF255322–AF255327, AF255329, and AF256079). We did not use genes from regions of low recombination because our power to detect local reductions of polymorphism depends on the presence of several polymorphic sites per locus. Sequences for *crq* were provided by T. A. Schlenke (accession nos. AF544231 and AF544232–AF544239). Sequence data from *Est-6* and *Adh* were obtained from GenBank [*Est-6*—(*D. simulans*) L34263, L34265, L10670, L34264, (*D. melanogaster*) AF147102, and (*D. yakuba*) AJ279007; *Adh*—(*D. simulans*) X57361–X57364, M36581, M19263, X00607, (*D. melanogaster*) X60792, and (*D. yakuba*) X57365].

All *D. simulans*-specific fixations were identified by parsimony using *D. melanogaster* and *D. yakuba* outgroup data. Parsimony should reliably identify ancestral states given the relatively low levels of sequence divergence between these species (Yang *et al.* 1995). Moreover, misidentified fixations should

Sequence data from this article have been deposited with the EMBL/GenBank Data Libraries under accession nos. AF544231 and AF544232–AF544239.

[1]*Corresponding author:* Center for Population Biology, University of California, 1 Shields Ave., Davis, CA 95616.
E-mail: adkern@ucdavis.edu

[2] These authors contributed equally to this work.

TABLE 1

Estimates of nucleotide heterozygosity for the 27 *D. simulans* loci studied

Locus (*n*)	Chrm.	Cyt.	θ	π
Adh (7)	2L	35B	0.0110	0.0108
AP-50 (8)	3R	94A	0.0160	0.0161
boss (5)	3R	96F	0.0116	0.0118
Cen190 (7)	3R	88E	0.0073	0.0091
crq (8)	2L	21C	0.0051	0.0049
dec-1 (7)	X	7C	0.0068	0.0071
eld (7)	3R	90B	0.0082	0.0107
Est-6 (4)	3R	68F	0.0178	0.0182
G6pd (8)	X	18E	0.0031	0.0041
g (7)	X	12B	0.0055	0.0038
Hsc70-1 (7)	3R	70C	0.0032	0.0031
hyd (8)	3R	92E	0.0032	0.0031
mei-218 (8)	X	15E	0.0056	0.0059
mira (6)	3R	86A	0.0044	0.0056
Osbp (8)	3R	96B	0.0112	0.0113
per (6)	X	3B	0.0102	0.0101
Pgd (7)	X	2D	0.0126	0.0114
Rel (7)	3Q	85C	0.0077	0.0069
Rh3 (5)	3R	85F	0.0127	0.0127
ry (18)	3R	90A	0.0186	0.0192
sn (8)	X	7D	0.0074	0.0087
sog (8)	X	13E	0.0025	0.0021
T-cpl (8)	3R	94A	0.0029	0.0028
Tpi (9)	3R	99D	0.0087	0.0068
v (8)	X	9F	0.0085	0.0100
CG3585 (7)	X	5E	0.0065	0.0082
z (6)	X	3A	0.0079	0.0078
		Average	0.0085	0.0087

n, sample size; Chrm., chromosome arm; Cyt., cytological location; θ and π, estimates of nucleotide heterozygosity based on number of segregating sites and number of pairwise differences. Note, CG3585 corresponds to the locus *X-gene* from Begun and Whitley (2000).

TABLE 2

Mean θ and π across classes of fixations

Fixation category	θ	π
Replacement	0.0081	0.0089
Preferred	0.0081	0.0083
Unpreferred	0.0082	0.0084
Unpreferred (w/o repl.)	0.0082	0.0080

Fixation category, type of nucleotide change that occurred (see text for details); unpreferred (w/o repl.), genes that had unpreferred fixations but no replacement fixations; θ and π, average estimates of heterozygosity for these loci.

simply add noise across all of our analyses. To be conservative, we restricted our analyses to fixations having a single, most parsimonious reconstruction of the ancestral state. Synonymous codons were assigned to preferred *vs.* unpreferred classes according to Sharp and Lloyd (1993). Both π and θ (Watterson 1975; Nei 1987) were used as estimators of heterozygosity. Table 2 presents mean levels of heterozygosity for genes included in our analysis that contain at least one of the indicated class of fixation. Note that there does not seem to be a relationship between fixation class and polymorphism at the whole gene level.

Permutation analysis: The goal of the permutation analysis was to use DNA polymorphism data to empirically generate null distributions of nucleotide heterozygosities for windows of defined size. The null hypothesis is that levels of DNA polymorphism in regions near sites that fixed in a gene along the *D. simulans* lineage ("test sites") are the same as levels of polymorphism observed in randomly selected regions within the same gene. The alternative hypothesis is that regions near test sites have reduced heterozygosity. Each test site was assigned to a category of fixation (*i.e.*, unpreferred, preferred, replacement, silent). For each gene, we estimated π and θ for a window of 200 bp centered on each test site of a given category and then calculated the mean heterozygosity across test sites. Test site windows could overlap if fixations were close to one another. We then permuted the locations of test sites within each gene by centering windows of the same size on "randomly" selected sites (permuting locations of test sites rather than polymorphic sites preserves any underlying physical heterogeneity of polymorphic sites in a gene). Given that third positions of codons tend to be more variable than first and second positions, randomly selected sites in a gene may not reflect the underlying distribution of test sites across codon positions in real data. Thus, not all sites are exchangeable (Good 2000). To address this statistical problem, the fraction of first, second, and third positions sampled during the permutation step was made to equal the fraction of first, second, and third positions found at test sites. The number of sites selected was equal to the number of test sites in the actual data. As was the case for test sites, windows of random sites could overlap. Average π and θ were then estimated for these randomly selected windows. This process was repeated several hundred times for each gene to generate a distribution of heterozygosities for windows of a given size. The total number of permutations depended on the number of exchangeable sites, which is approximately equal to the number of codons in the sample.

The observed mean heterozygosity across test sites for each gene was compared to the permutation-generated distribution. This yields the probability that the observed π and θ for windows around test sites were lower than expected under the null hypothesis (thus, this is a one-tailed test). A potential problem with our approach could be that the ends of surveyed DNA sequences were undersampled during the permutations as windows exceeding the ends of the sequences were excluded from the analysis. However, this is not a major concern as heterozygosities at the 5′ and 3′ ends of the surveyed regions were not significantly different from heterozygosities at other regions (analysis not shown).

The choice of window size is a complex and important issue that may affect the picture of variation within a gene and the power of our analysis (Silverman 1986). If window sizes are too small, then most randomly selected windows will have no segregating sites. This would reduce our power to detect a significant association of test sites with regions of reduced heterozygosity. Alternatively, large windows may "smooth" interesting local variation in polymorphism if the extent of the hitchhiking effect was substantially smaller than that of the window size (similar to the case of a large window width parameter in Silverman 1986). *A priori* knowledge of the effect of window size on the power of rejecting the null hypothesis *vs.* some alternative for each gene would be ideal. In the absence of such information, however, we chose the window size on the basis of the following empirically based rationale. Although *D.*

TABLE 3

Analysis of replacement fixations

Locus	No. of windows	Test θ	θ *P* value	Test π	π *P* value	No. of permutations
Adh	1	0.0020	0.040	0.0014	0.040	248
Cen190	1	0.0000	0.097	0.0000	0.097	429
crq	4	0.0058	0.653	0.0049	0.502	352
dec-1	4	0.0041	0.148	0.0033	0.088	497
eld	1	0.0102	0.702	0.0110	0.563	323
Est-6	4	0.0089	0.021	0.0088	0.021	460
G6pd	16	0.0028	0.329	0.0034	0.210	570
hyd	2	0.0058	0.581	0.0051	0.457	597
mei-218	22	0.0049	0.970	0.0061	0.965	410
mira	1	0.0120	0.767	0.0110	0.602	400
Osbp	1	0.0039	0.280	0.0050	0.308	389
Rel	24	0.0056	0.082	0.0050	0.101	918
ry	1	0.0247	0.990	0.0321	0.990	454
v	1	0.0058	0.331	0.0070	0.331	411
CG3585	1	0.0039	0.408	0.0043	0.368	498
z	2	0.0066	0.438	0.0083	0.585	475
	Test window average	0.0067	$P = 0.150$	0.0073	$P = 0.080$	
	Gene average	0.0087		0.0094		

No. of windows, test θ, and test π: estimates of polymorphism within test windows; no. of permutations and test window average: mean across loci of the estimated polymorphism of the test windows; gene average, mean across loci of the estimated polymorphism for the entire gene.

simulans is a relatively highly polymorphic species, most sites are monomorphic. The mean number of base pairs between polymorphic sites in our samples is ~75 bp, with a standard deviation of 25 bp. A 200-bp window is approximately twice the mean number of base pairs between polymorphic sites (75 bp) plus one standard deviation (25 bp). This means that most randomly chosen windows of 200 bp will include at least one polymorphic site. We used the mean plus a standard deviation because using only the mean would have reduced our power in genes with less polymorphism, whereas increasing the window size was likely not biased. Regardless, the results using a window size based on the mean were not significantly different from those using a window size based on the mean plus a standard deviation (data not shown). Software and source code implementing this method are available from the authors and at http://limulus.ucdavis.edu/~cojo/.

Statistical analysis: Failure to reject the null hypothesis at a locus may reflect a lack of polymorphism at that locus or other factors limiting the power of our analysis. We used Fisher's combined probability test to effectively increase our ability to detect a significant trend in the data. This test is suitable when separate statistical tests on different data sets test the same scientific hypothesis (Fisher 1954; Sokal and Rohlf 1995) and is especially applicable when the same significance tests are used on all data sets, yet a joint statistical analysis is impossible (as is the case here). Moreover, our use of Fisher's combined probability test is likely to be unbiased because all loci were subjected to an identical analysis that yields exact probability values (Sokal and Rohlf 1995).

One of the limitations of Fisher's test is that it cannot distinguish between several tests with consistent weak effects *vs.* a mixture of tests with strong effects and tests with no effect. To address this limitation, we used a standard test statistic from metaanalysis, Glass's g:

$$g = (\bar{x}^E - \bar{x}^C)/s^c$$

(Glass 1976). Here, $\bar{x}^E$ is the mean level of polymorphism in windows surrounding test sites, $\bar{x}^C$ is the mean level of polymorphism in windows surrounding all sites of the gene, and s^c is the standard deviation of polymorphism in windows surrounding all sites of the gene. Thus, g is a unitless measure of the reduction (or inflation) of polymorphism surrounding test sites in a gene.

Using g has two main advantages. First, it allows us to estimate the relative magnitude of the reduction in heterozygosity near test sites from different categories of fixation. Second, g scores can be used to compare the relative reduction of polymorphism surrounding test sites across loci. If, on average, there is no difference in the levels of heterozygosity surrounding test sites and the levels of heterozygosity at all sites then the mean g across loci should be zero. A t-test can be used to test the null hypothesis that g is zero (*i.e.*, there is no effect). This is an improvement over Fisher's test in that the null is less likely to be rejected if there are only a few genes of strong effect and many genes of no effect.

RESULTS

Tables 3 and 4 show the results of permutation analyses of polymorphism in 200-bp windows centered on replacement and silent fixations, respectively. Heterozygosity near replacement fixations ($\theta = 0.0067$, $\pi = 0.0073$) is slightly, though not significantly, reduced compared to overall levels of heterozygosity ($\theta = 0.0087$, $\pi = 0.0094$) in sequenced regions of individual genes (Fisher's combined probability, d.f. = 16, $P = 0.15$ and $P = 0.08$ for θ and π, respectively). To avoid the confounding effects of pooling data across loci, we calcu-

TABLE 4

Analysis of silent fixations

Locus	No. of windows	Test θ	θ *P* value	Test π	π *P* value	No. of permutations
AP-50	3	0.0148	0.279	0.0195	0.753	466
boss	3	0.0088	0.223	0.0088	0.263	551
Cen190	7	0.0052	0.093	0.0065	0.093	429
crq	7	0.0044	0.235	0.0038	0.176	352
Est-6	3	0.0118	0.108	0.0117	0.097	460
dec-1	13	0.0071	0.776	0.0077	0.814	497
G6pd	10	0.0023	0.177	0.0033	0.229	570
g	7	0.0052	0.197	0.0037	0.178	421
Hsc70-1	5	0.0033	0.472	0.0034	0.476	430
hyd	4	0.0067	0.765	0.0079	0.830	597
mei-218	11	0.0035	0.363	0.0044	0.297	410
mira	6	0.0144	0.970	0.0133	0.937	400
Osbp	2	0.0125	0.688	0.0131	0.773	389
per	5	0.0118	0.349	0.0103	0.262	626
Pgd	5	0.0078	0.401	0.0087	0.233	304
Rel	7	0.0041	0.046	0.0031	0.020	918
Rh3	2	0.0180	0.736	0.0173	0.664	376
ry	10	0.0185	0.220	0.0203	0.583	454
sn	6	0.0061	0.223	0.0068	0.122	483
sog	7	0.0028	0.579	0.0026	0.669	411
T-cpl	11	0.0026	0.550	0.0024	0.495	332
Tpi	3	0.0025	0.018	0.0023	0.022	268
v	7	0.0080	0.349	0.0101	0.403	498
CG3585	5	0.0058	0.275	0.0069	0.212	475
z	3	0.0051	0.069	0.0058	0.075	333
	Test window average	0.0077	$P = 0.060$	0.0081	$P = 0.054$	
	Gene average	0.0085		0.0087		

See Table 3 for explanation.

lated Glass's *g* statistic for each locus [*g* is a dimensionless measure of the difference between heterozygosity at our test sites and that of the gene as a whole (Glass 1976)]. An average *g* of 0 is expected under neutrality, whereas a negative *g* is expected if heterozygosity has been reduced. Consistent with the statistical results of average heterozygosity, mean *g*, while negative, is not statistically different from 0 ($g = -0.42$, $t = -1.502$, $P = 0.15$ for θ; $g = -0.379$, $t = -1.195$, $P = 0.25$ for π). Similarly, heterozygosity in 200-bp windows centered on silent fixations (θ = 0.0077; π = 0.0081) is not significantly reduced compared to overall levels of heterozygosity (θ = 0.0084, π = 0.0086) in sequenced regions of individual genes (Fisher's combined probability, d.f. = 50, $P = 0.06$ and $P = 0.054$ for θ and π, respectively). Interestingly, mean *g* across all loci for silent fixations is significant for θ ($g = -0.484$, $t = -2.726$, $P = 0.0118$) and is marginally significant for π ($g = -0.385$, $t = -2.049$, $P = 0.0516$). None of the individual genes shows significantly reduced heterozygosity near replacement or silent fixations when critical values are Bonferroni corrected for multiple tests.

Genomic patterns of codon usage in Drosophila suggest that silent mutations can be placed into at least two categories, preferred and unpreferred (Sharp and Lloyd 1993). Therefore, we can investigate whether patterns of linked variation differ between mutant classes at silent sites. Mean heterozygosity and mean *g* (Table 5) reveal no reduction of heterozygosity for windows centered on unpreferred fixations (θ = 0.0079, π = 0.0086). Furthermore, no individual genes showed a significant reduction of polymorphism in windows centered on unpreferred fixations. Windows centered on *preferred* fixations, however, show a highly significant reduction of polymorphism (Table 6; d.f. = 46, $P = 0.0096$ and $P = 0.0041$ for θ and π, respectively). Average polymorphism in regions near preferred codon fixations (θ = 0.0065, π = 0.0065) is ~25% lower compared to the genes from which they were sampled (θ = 0.0084, π = 0.0086). This effect is confirmed by the mean *g* across all loci with preferred fixations ($g = -0.612$, $t = -2.851$, $P = 0.0093$ for θ; $g = -0.631$, $t = -3.159$, $P = 0.0045$ for π), which also indicates a significant reduction in levels of heterozygosity flanking preferred fixations. Figure 1 provides a visual comparison of the distributions of *g* across loci for preferred and unpreferred fixations. Although five loci (*AP-50*, *Cen190*, *crq*, *mei-218*, and *Pgd*) show large reductions of

TABLE 5

Analysis of unpreferred fixations

Locus	No. of windows	Test θ	θ *P* value	Test π	π *P* value	No. of permutations
AP-50	2	0.0193	0.673	0.0267	0.978	466
boss	1	0.0048	0.194	0.0050	0.185	551
Cen190	2	0.0092	0.669	0.0119	0.676	429
crq	1	0.0077	0.900	0.0077	0.923	352
dec-1	8	0.0064	0.497	0.0071	0.619	497
Est-6	1	0.0164	0.393	0.0167	0.363	460
G6pd	4	0.0033	0.635	0.0044	0.561	570
g	2	0.0051	0.472	0.0036	0.472	421
Hsc70-1	2	0.0010	0.314	0.0007	0.239	430
hyd	2	0.0058	0.544	0.0066	0.591	597
mei-218	4	0.0048	0.846	0.0061	0.802	410
mira	3	0.0136	0.875	0.0130	0.852	400
Osbp	1	0.0116	0.555	0.0132	0.653	389
per	2	0.0088	0.155	0.0073	0.089	626
Pgd	4	0.0092	0.671	0.0105	0.592	304
Rel	1	0.0041	0.433	0.0029	0.298	918
Rh3	2	0.0180	0.736	0.0173	0.664	376
Ry	4	0.0171	0.121	0.0185	0.229	454
sn	2	0.0058	0.455	0.0064	0.323	483
sog	4	0.0029	0.593	0.0025	0.588	411
T-cpl	9	0.0031	0.732	0.0030	0.740	400
Tpi	1	0.0000	0.134	0.0000	0.134	268
v	5	0.0081	0.399	0.0105	0.485	498
CG3585	3	0.0071	0.528	0.0088	0.480	475
z	2	0.0055	0.225	0.0057	0.132	333
	Test window average	0.0079	$P = 0.810$	0.0086	$P = 0.740$	
	Gene average	0.0085		0.0087		

See Table 3 for explanation.

π for windows centered on preferred fixations, none are individually significant when critical values are conservatively adjusted for multiple tests.

Figure 2 illustrates the effect of window size on our analysis. Figure 2A shows the results of an expanding window analysis for a strongly significant result, in this case a significant reduction around preferred sites. Clearly, the reduction in heterozygosity is statistically detectable for a variety of window sizes. Figure 2, B and C, shows results typical of nonsignificant genes. Both genes lack the long stretch of significant window sizes seen in Figure 2A. In Figure 2C, preferred fixations drop below $P = 0.10$ for windows of ~140 bases, but only briefly, which suggests that this dip was due to chance.

One concern regarding our analysis is that we have assumed that four-codon families can be represented as having only two fitness classes, preferred and unpreferred. However, conserved patterns of rank order of codon usage within codon families across widely divergent Drosophila (Kreitman and Antezana 2000) suggest that some four-codon families may have as many as four potential fitness classes. Although it is unclear under what circumstances this fact is problematic for our inference, we can avoid this complication by investigating silent fixations of mutants belonging to twofold codon families, which alternate between preferred and unpreferred states by reversible mutation. Congruent with the results from all preferred and unpreferred fixations, preferred and unpreferred classes in twofold codons are dramatically different in our estimate of their levels of linked heterozygosity (preferred, $\theta = 0.0055$, $\pi = 0.0054$; unpreferred, $\theta = 0.0091$, $\pi = 0.0099$), although this difference is not statistically significant (perhaps as the consequence of reduced power in this restricted data set).

DISCUSSION

Our analysis of *D. simulans* polymorphism and divergence data revealed no evidence of hitchhiking effects associated with replacement fixations. One possible explanation is that the proteins in our sample evolve by genetic drift in *D. simulans*. Alternatively, replacement fixations may be composed of a large class of neutral mutants and a small class of strongly selected mutations. If this were the case we might not observe an overall

TABLE 6

Analysis of preferred fixations

Locus	No. of windows	Test θ	θ *P* value	Test π	π *P* value	No. of permutations
AP-50	2	0.0058	0.055	0.0052	0.021	466
boss	2	0.0108	0.437	0.0108	0.462	551
Cen190	4	0.0031	0.018	0.0037	0.016	429
crq	4	0.0029	0.090	0.0019	0.042	352
dec-1	3	0.0095	0.917	0.0095	0.849	497
Est-6	2	0.0095	0.100	0.0092	0.080	460
G6pd	4	0.0012	0.100	0.0020	0.107	570
g	4	0.0051	0.261	0.0036	0.228	421
Hsc70-1	2	0.0071	0.841	0.0079	0.867	430
mei-218	2	0.0010	0.039	0.0013	0.039	410
mira	3	0.0152	0.942	0.0137	0.882	400
Osbp	1	0.0135	0.709	0.0130	0.650	389
per	2	0.0131	0.581	0.0100	0.338	626
Pgd	1	0.0020	0.026	0.0014	0.019	304
Rel	3	0.0041	0.120	0.0033	0.099	918
ry	1	0.0145	0.118	0.0181	0.337	454
sn	3	0.0071	0.525	0.0079	0.443	483
sog	3	0.0026	0.564	0.0026	0.630	411
T-cpl	1	0.0019	0.567	0.0013	0.567	332
Tpi	2	0.0037	0.085	0.0035	0.108	268
v	2	0.0077	0.415	0.0090	0.285	498
CG3585	1	0.0039	0.437	0.0043	0.378	475
z	1	0.0044	0.270	0.0060	0.354	333
	Test window average	0.0065	$P = 0.0096$	0.0065	$P = 0.0041$	
	Gene average	0.0084		0.0086		

See Table 3 for explanation.

association of replacement fixations with reduced heterozygosity. Finally, the power of our analyses of replacement fixations could be compromised if the physical scale of reduced variation near replacement fixations were greater than the size of the windows or gene regions used in our analyses. This explanation, however, seems unlikely because loci that have fixed at least one amino acid are roughly as polymorphic as those that have fixed only unpreferred mutations (Table 2).

We observed a reduction of linked polymorphism near preferred fixations, but not near unpreferred fixations (Table 7). One might expect this result under the

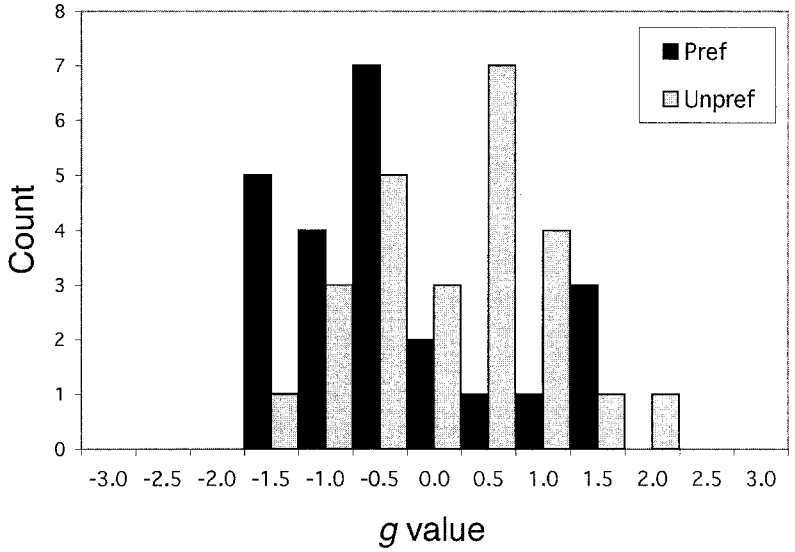

FIGURE 1.—Comparison of loss of heterozygosity at preferred *vs.* unpreferred fixations. This histogram shows values of Glass's *g* for preferred (solid bars) and unpreferred (shaded bars) fixations at the per gene level. The distribution for preferred fixations is shifted toward negative values relative to the unpreferred distribution, indicating a reduction in local polymorphism and implicating the action of selection.

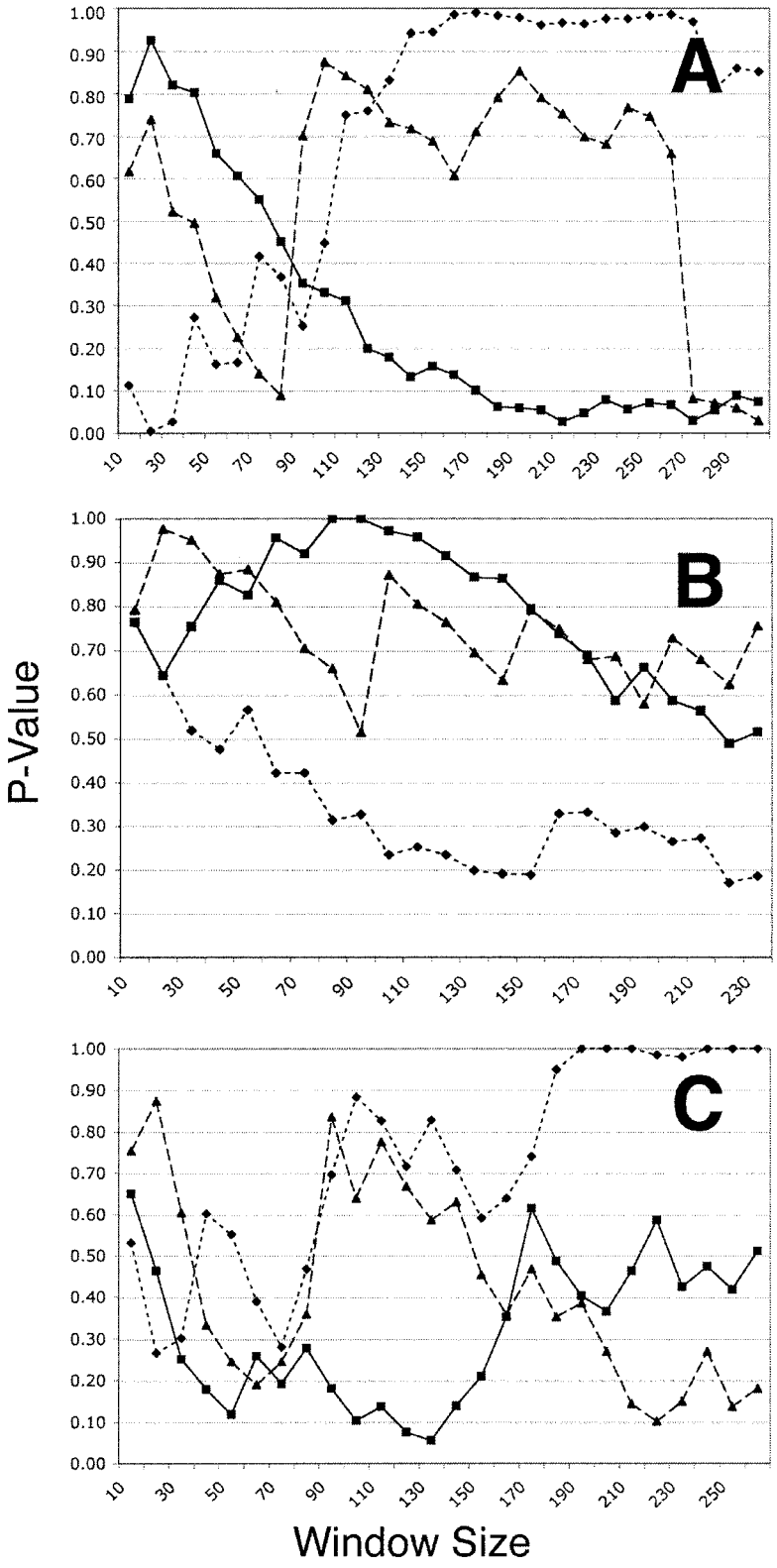

FIGURE 2.—*P* values produced by permutation analysis for a series of window sizes. Window size began at 10 bases and increased by 10 bases until a width of 300 bases was reached or until the window around a test site extended beyond the ends of the sequence. Genes with all three classes of fixations [replacement (◆), unpreferred (▲), and preferred (■)] were used in this analysis. (A) Results from *mei-218*, which showed a significant reduction in variation around preferred fixations in our earlier analysis and is intermediate in average heterozygosity. (B) Results from *Osbp*, which was not significant for any class of fixation and is intermediate in average heterozygosity. (C) Results from *ry*, which was not significant for any class of fixation and had the highest average heterozygosity of the 26 genes studied.

simple premise that preferred and unpreferred mutations are slightly beneficial and slightly deleterious alleles, respectively. However, this expectation is probably incorrect because we are examining a special set of mutations, namely those that have fixed. MARUYAMA (1974) showed that conditional on fixation, the mean and variance of the sojourn time are the same for fixed beneficial and deleterious mutations having identical selection coefficients. Although counterintuitive, this result can be understood in the following way: given

TABLE 7

Results summary

Fixation category	θ *P* value	π *P* value	Avg. Glass's $g\,\theta$	Avg. Glass's $g\,\pi$
Replacement	0.1500	0.0800	−0.4196	−0.3793
Silent	0.0600	0.0540	−0.4839	−0.3846
Preferred	0.0096	0.0041	−0.6116	−0.6312
Unpreferred	0.8100	0.7400	−0.1855	−0.0571

θ *P* value and π *P* value, the results of Fisher's combined probability test for each class of fixation; Avg. Glass's $g\,\theta$ and Avg. Glass's $g\,\pi$, dimensionless measures of the degree to which the average polymorphism of test windows deviates from that of the rest of the gene.

that a deleterious mutant will fix, the more strongly negative the mutant, the more quickly it must fix by drift to escape the selection that opposes its spread. Under this model, the hitchhiking effect associated with preferred and unpreferred fixations should be the same, *all else being equal*. Thus, one interpretation of our results is that some factor affecting the expected heterozygosity in a recurrent hitchhiking model differs among unpreferred and preferred fixations. For example, more recent fixations should be associated with greater reductions of linked polymorphism, all else being equal (Kaplan *et al.* 1989; Simonsen *et al.* 1995; Kim and Stephan 2002; Przeworski 2002), because there has been less time for neutral variation to accumulate following the fixation event. Therefore, one hypothesis for the greater reduction of heterozygosity near preferred fixations is that, on average, preferred fixations have occurred more recently than other types of fixations.

Previous analyses suggested that the *D. simulans* lineage has fixed significantly more unpreferred mutations (Begun 2001; McVean and Vieira 2001) than expected under the mutation-selection-drift model of silent-site evolution (Bulmer 1991). An ancient accumulation of unpreferred mutations followed by a more recent, compensatory accumulation of preferred mutations in *D. simulans* is consistent with our data. What might cause such dynamics remains a matter for speculation. It is possible that *D. simulans* genes accumulated unpreferred fixations by drift when population size was small and subsequently fixed preferred codons by directional selection when population size increased. However, the timescale of changes in population size must be slower than that of the substitution rate for such a model to be viable. Alternatively, the intensity of selection favoring preferred mutations may vary over time. The potential impact of interactions among weakly selected polymorphisms on neutral variation (*e.g.*, McVean and Charlesworth 2000; Comeron and Kreitman 2002) also merits further consideration.

As is the case for silent sites, the lack of hitchhiking effects associated with amino acid fixations could be explained by invoking episodic evolution if these fixation events occurred in the more distant past compared to preferred fixations. This hypothesis may be testable if data from *D. mauritiana* and *D. sechellia* allow us to identify which mutations in the *D. simulans* lineage fixed in the more recent *vs.* more ancient past. It will also be interesting to investigate whether the spatial distribution of polymorphisms across *D. melanogaster* genes is similar to what we have observed in *D. simulans*.

Although we are not in a position to strongly favor a particular substitution model for our data, the results reported here certainly provide motivation for additional analyses of linked selection. For example, we have little understanding of how different population genetic parameters affect our permutation test or the population genetic scenarios under which we may be able to detect the local footprint of selection. Finally, our results underscore Akashi's (1995, 1999) cautionary notes regarding the dangers of making population genetics inferences on the causes of protein evolution under the premise that silent mutations are neutral (*e.g.*, Smith and Eyre-Walker 2002; Fay *et al.* 2002).

We thank P. Awadalla, A. Betancourt, J. Gillespie, C. Langley, M. Lawniczak, M. Przeworski, S. Schaeffer, D. Weinreich, and two anonymous reviewers for comments on drafts of this manuscript. A.D.K. is a Howard Hughes Medical Institute predoctoral fellow. C.D.J. and D.J.B. were funded by the National Science Foundation.

LITERATURE CITED

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

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

Aquadro, C. F., D. J. Begun and E. C. Kindahl, 1994 Selection, recombination and DNA polymorphism in Drosophila, pp. 46–56 in *Non-neutral Evolution: Theories and Molecular Data*, edited by B. Golding. Chapman & Hall, New York.

Begun, D. J., 2001 The frequency distribution of nucleotide variation in *Drosophila simulans*. Mol. Biol. Evol. **18**: 1343–1352.

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

Begun, D. J., and P. Whitley, 2000 Reduced X-linked nucleotide polymorphism in *Drosophila simulans*. Proc. Natl. Acad. Sci. USA **97:** 5960–5965.

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

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

Comeron, J. M., and M. Kreitman, 2002 Population, evolutionary and genomic consequences of interference selection. Genetics **161:** 380–410.

Fay, J. C., G. J. Wyckoff and C.-I Wu, 2002 Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. Nature **415:** 1024–1026.

Fisher, R. A., 1935 *The Design of Experiments*. Hafner Press, New York.

Fisher, R. A., 1954 *Statistical Methods for Research Workers*, Ed. 12. Hafner Press, New York.

Glass, G. V., 1976 Primary, secondary, and meta-analysis of research. Educ. Res. **5:** 3–8.

Good, P., 2000 *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses.* Springer, New York.

Hill, W. G., and A. Robertson, 1966 The effect of linkage on limits to artificial selection. Genet. Res. **8:** 269–294.

Kaplan, N. L., R. R. Hudson and C. H. Langley, 1989 The "hitchhiking effect" revisited. Genetics **123:** 887–899.

Kim, Y., and W. Stephan, 2002 Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics **160:** 765–777.

Kreitman, M., and M. Antezana, 2000 The population and evolutionary genetics of codon bias, pp. 82–101 in *Evolutionary Genetics: From Molecules to Morphology*, edited by R. S. Singh and C. B. Krimbas. Cambridge University Press, Cambridge, UK.

Langley, C. H., J. MacDonald, N. Miyashita and M. Aguade, 1993 Lack of correlation between interspecific divergence and intraspecific polymorphism at the suppressor of forked region in *Drosophila melanogaster* and *Drosophila simulans*. Proc. Natl. Acad. Sci. USA **90:** 1800–1803.

Maruyama, T., 1974 The age of an allele in a finite population. Genet. Res. **23:** 137–143.

Maynard Smith, J., and J. Haigh, 1974 The hitch-hiking effect of a favourable gene. Genet. Res. **23:** 23–35.

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

McVean, G. A., and J. Vieira, 2001 Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in Drosophila. Genetics. **157:** 245–257.

Nachman, M. W., 1997 Patterns of DNA variability at *X*-linked loci in *Mus domesticus*. Genetics **147:** 1303–1316.

Nachman, M. W., 2001 Single nucleotide polymorphisms and recombination rate in humans. Trends Genet. **17:** 481–485.

Nachman, M. W., V. L. Bauer, S. L. Crowell and C. F. Aquadro, 1998 DNA variability and recombination rates at *X*-linked loci in humans. Genetics **150:** 1133–1141.

Nei, M., 1987 *Molecular Evolutionary Genetics.* Columbia University Press, New York.

Przeworski, M., 2002 The signature of positive selection at randomly chosen loci. Genetics **160:** 1179–1189.

Sharp, P. M., and A. T. Lloyd, 1993 Codon usage, pp. 378–397 in *An Atlas of Drosophila Genes: Sequences and Molecular Features*, edited by G. Maroni. Oxford University Press, Oxford.

Silverman, B. W., 1986 *Density Estimation for Statistics and Data Analysis.* Chapman & Hall, London.

Simonsen, K. L., G. A. Churchill and C. F. Aquadro, 1995 Properties of statistical tests of neutrality for DNA polymorphism data. Genetics **141:** 413–429.

Smith, N. G. C., and A. Eyre-Walker, 2002 Adaptive protein evolution in *Drosophila*. Nature **415:** 1021–1024.

Sokal, R. R., and F. J. Rohlf, 1995 *Biometry*, Ed. 3. W. H. Freeman, New York.

Watterson, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256–276.

Yang, Z., S. Kumar and M. Nei, 1995 A new method of inference of ancestral nucleotide and amino acid sequences. Genetics **141:** 1641–1650.

Communicating editor: S. W. Schaeffer