Hitchhiking Under Positive Darwinian Selection

Justin C. Fay* and Chung-I Wu*,†

**Committee on Genetics and †Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637*

Manuscript received August 7, 1999
Accepted for publication March 20, 2000

ABSTRACT

Positive selection can be inferred from its effect on linked neutral variation. In the restrictive case when there is no recombination, all linked variation is removed. If recombination is present but rare, both deterministic and stochastic models of positive selection show that linked variation hitchhikes to either low or high frequencies. While the frequency distribution of variation can be influenced by a number of evolutionary processes, an excess of derived variants at high frequency is a unique pattern produced by hitchhiking (derived refers to the nonancestral state as determined from an outgroup). We adopt a statistic, *H*, to measure an excess of high compared to intermediate frequency variants. Only a few high-frequency variants are needed to detect hitchhiking since not many are expected under neutrality. This is of particular utility in regions of low recombination where there is not much variation and in regions of normal or high recombination, where the hitchhiking effect can be limited to a small ($<$1 kb) region. Application of the *H* test to published surveys of Drosophila variation reveals an excess of high frequency variants that are likely to have been influenced by positive selection.

THE extent to which positive Darwinian selection shapes molecular evolution has been a dominant issue in the last three decades (Kimura 1983; Nei 1987). Under neutrality, the amount of DNA variation within a species is proportional to the amount of divergence between species (Kimura 1983). In contrast, the spread of a positively selected mutation through a population removes linked neutral variation without affecting neutral divergence (Maynard Smith and Haigh 1974). Surveys of DNA variation have revealed a correlation between levels of variation and rates of recombination in plants (Stephan and Langley 1998), flies (Stephan and Langley 1989; Begun and Aquadro 1992; Martín-Campos *et al.* 1992; Stephan and Mitchell 1992), mice (Nachman 1997), and humans (Nachman *et al.* 1998). Reduced levels of variation are also found within small regions of some genes (McDonald 1998). However, there are a number of alternative explanations for reduced levels of variation. In particular, "background selection," or the elimination of deleterious mutations segregating in a population, predicts low levels of variation in regions of low recombination (Charlesworth *et al.* 1993). Distinguishing positive selection from these alternative explanations has become a significant empirical and theoretical pursuit.

The frequency spectrum of variation provides a means of detecting positive selection independent of levels of variation. In the absence of recombination, hitchhiking eliminates all linked variation and a population in recovery is characterized by an excess of new mutations at low frequency. In the presence of recombination, hitchhiking is incomplete since not all variation is removed and its direct effect on the frequency spectrum is not known. Unless the hitchhiking effect is very strong, even intragenic variation will experience an incomplete rather than complete hitchhiking effect. For example, a selection coefficient of 10^{-3} and a recombination rate of 10^{-8}/bp (about the average in *Drosophila melanogaster*; Ashburner 1989) would still leave 10% of the heterozygosity 500 bp away from a site under selection (Stephan *et al.* 1992). In a region where recombination is reduced 10-fold, such as the tip of the X chromosome (Aguadé *et al.* 1989), the corresponding distance would be 5 kb.

The incomplete removal of variation may create a pattern of variation very different from the pattern produced after all variation is removed. For example, in the case of a brief reduction in population size, there is a striking difference between residual patterns of variation found immediately after a bottleneck and the excess of new mutations that subsequently accumulate (Tajima 1989a; Fay and Wu 1999). To understand the effect of positive selection on linked variation, we describe the hitchhiking effect on the frequency spectrum as a function of the rate of recombination. We then adopt a statistical test that can detect hitchhiking even when there are few segregating sites. Application of the test to published surveys of variation in regions of low and high recombination in Drosophila indicates a departure from neutrality in the direction predicted by hitchhiking.

Corresponding author: Chung-I Wu, Department of Ecology and Evolution, University of Chicago, 1101 E. 57th St., Chicago, IL 60637. E-mail: ciwu@uchicago.edu

MATERIALS AND METHODS

Polymorphism data: We restricted our analysis to published polymorphism surveys of Drosophila species. Recombination rates across the *D. melanogaster* genome have been estimated from the frequency of recombination per cytological band (Comerón *et al.* 1999). Regions of low recombination were designated as cytological divisions with rates of recombination $< 5 \times 10^{-9}$/bp: 1A–2D, 20C–20F, 21A, 38A–44C, 60C–60F, 61A–61B, and 75F–84F (excluding the fourth chromosome where recombination is absent). This cutoff value is ~10 times lower than the division with the highest estimated rate of recombination. There are six published polymorphism surveys in the designated regions: *achaete* (*ac*, 1B1; Martín-Campos *et al.* 1992), *asense* (*ase*, 1B4; Hilton *et al.* 1994), *Phosphogluconate dehydrogenase* (*Pgd*, 2D6; Begun and Aquadro 1994), *suppressor of forked* [*su(f)*, 20E; Langley *et al.* 1993], *fushi tarazu* (*ftz*, 84B1; Jenkins *et al.* 1995), and *Glucose dehydrogenase* (*Gld*, 84D1; Hamblin and Aquadro 1997). The *Accessory gland protein 26Aa* (*Acp26Aa*) is in a region of high recombination in *D. melanogaster* (26A, *c* estimated to be 5×10^{-8}/bp) and its intra- and interspecific variation has been extensively studied (Aguadé *et al.* 1992; Tsaur and Wu 1997; Aguadé 1998; Tsaur *et al.* 1998). *vermilion* (*v*; Stephan and Langley 1989; Stephan *et al.* 1998) is in a region of reduced recombination in *D. ananassae*.

The frequency spectrum of derived variants, including synonymous, nonsynonymous, and noncoding variants, was constructed for each survey using an outgroup. When multiple locations were sampled, the data were pooled for the determination of the frequency spectrum. For *achaete*, the European sample was scaled from 192 to 50, which is similar in size to the African and North American samples. All data are DNA polymorphism within *D. melanogaster* and divergence from *D. simulans*, except for *achaete* [which is a four-cutter restriction fragment length polymorphism (RFLP) study] and *vermilion*, which is a survey of polymorphism in *D. ananassae* with divergence from *D. pallidosa*.

Simulation methods: A standard coalescence algorithm was used to generate neutral genealogies (Hudson 1990). A modified coalescence algorithm, which tracks the genealogy of a neutral locus linked to an advantageous mutation, was used to generate hitchhiking genealogies. The algorithm is a simplification of that detailed in Braverman *et al.* (1995), such that the recombination rate (c), selection coefficient (s), and time at which an advantageous mutation arises are set parameters instead of random variables. The selected allele confers a fitness of $1 + s$ in the heterozygous state and $1 + 2s$ in the homozygous state and decreases deterministically from a frequency of $1 - 1/2N$ to $1/2N$ in a population size (N) of 10^6 individuals. Recombination events occur between the selected and neutral locus, and coalescent events occur between advantageous alleles or between nonadvantageous alleles, depending on the frequency of the advantageous mutation and the number of advantageous and nonadvantageous alleles. Once all alleles have coalesced, mutations are randomly distributed on the genealogy of the neutral locus. All results were generated from at least 1000 iterations of the algorithm.

Statistics: Tajima's D statistic (Tajima 1989b) is based on the difference between two commonly used estimators of $\theta = 4N\mu$, where N is the effective population size and μ is the mutation rate: $\hat{\theta}_\pi$ is calculated from average heterozygosity (Tajima 1983) and $\hat{\theta}_W$ is calculated from the number of segregating sites (Watterson 1975),

$$\hat{\theta}_\pi = \sum_{i=1}^{n-1} \frac{2S_i i(n-i)}{n(n-1)} \quad (1)$$

$$\hat{\theta}_W = \left(\sum_{i=1}^{n-1} \frac{1}{i}\right)^{-1} \sum_{i=1}^{n-1} S_i, \quad (2)$$

where S_i is the number of derived variants found i times in a sample of n chromosomes. Let $\hat{\theta}_H$ be an estimator of θ weighted by the homozygosity of the derived variants, as opposed to the ancestral variants,

$$\hat{\theta}_H = \sum_{i=1}^{n-1} \frac{2S_i i^2}{n(n-1)}, \quad (3)$$

where $\hat{\theta}_H$ is an unbiased estimator of θ and is a special case, notated as $L(-1)$, of a general class of recently derived estimators that vary in their weighting schemes (Fu 1995). Let H be the difference between $\hat{\theta}_\pi$ and $\hat{\theta}_H$. The null distributions of the D and H statistics were generated using 1000 iterations of a neutral coalescence algorithm conditioning on the observed number of segregating sites. P values were calculated as the probability of a neutral D or H value being less than the observed D or H value. The D and H statistics are probably conservative since critical values were generated in the absence of recombination (Wall 1999), unless otherwise noted. The power of the D and H statistics was calculated as the probability of rejecting a neutral genealogy given a hitchhiking genealogy. The critical D and H values ($\alpha = 0.05$, one-sided) were generated for any given number of segregating sites. For calculating the power of the H statistic, the probability of misinference was 0.00375/site (see below).

An outgroup was used to infer the derived and ancestral states for all polymorphism data analyzed. However, a backmutation would result in the incorrect inference of the derived polymorphic state. As compensation, the probability of misinference was incorporated into the null distribution of the H statistic by exchanging the frequency of the derived and ancestral state, with probability equal to that of misinference, for each segregating site. For nucleotide sequences, the probability of a backmutation is $d/3$, where d is the net divergence or the average number of fixed differences between the two species. The observation that transitions occur at twice the rate of transversions (Moriyama and Powell 1996) was incorporated and resulted in a probability of backmutation equal to $3d/8$ (see Appendix at http://home.uchicago.edu/~jfay/appendix.html). The probability of misinference was calculated separately for synonymous, nonsynonymous, and noncoding sites and then weighted by the number of synonymous, nonsynonymous, and noncoding polymorphic sites. For restriction site data, where the outgroup sequence and one of the ingroup sequences are known, the probability of misinference differs, depending on the state of the outgroup (presence or absence of the restriction site). If the site is present in the outgroup, then the derived state is inferred to be a loss and the probability of misinference is $d/(3 - 2d)$, whereas if the site is absent in the outgroup the probability of misinference is $4d/(1 + 3d)$ (see Appendix at http://home.uchicago.edu/~jfay/appendix.html). The probability of misinference was calculated assuming that only one backmutation could account for the observed data.

THEORY

Deterministic: The hitchhiking effect can be described by the change in allele frequency at a neutral polymorphic locus (B, b) due to the spread of a linked advantageous mutation, A, through a population. Clearly, B will increase or decrease in frequency depending on whether the B or b allele is originally linked to the advantageous mutation. In the absence of recom-

bination, the variant originally linked to the advantageous mutation is fixed, but when recombination is present but rare, all variants that remain segregating are most likely present at either very low or high frequencies.

The frequency spectrum after hitchhiking can be found, knowing the frequency of variations before selection and their change in frequency due to selection. To distinguish between low- and high-frequency variants, ancestral (old) and derived (new) states must be distinguished (this can be done empirically using an outgroup). Let $\phi(x)$ be the frequency spectrum of derived variants in a population. Under neutrality, the expected number of sites where the derived variant is between a frequency of x and $x + dx$ is given by

$$\phi(x)dx = \frac{\theta}{x}\,dx. \tag{4}$$

In Equation 4, $\theta = 4N\mu$, where N is the effective population size and μ is the mutation rate of the region (Watterson 1975). Deterministically, if a neutral linked variant, *B*, hitchhikes with *A*, its frequency is transformed from x to $1 - c^* + xc^*$; otherwise it is reduced to xc^*, where c^* is a scaled measure of recombination. c^* is given in its exact form in Maynard Smith and Haigh (1974) but can be approximated by $(c/s)\ln(1/p_0)$, where s is the haploid selection coefficient, c is the rate of recombination between the two sites, and p_0 is the initial frequency of the advantageous mutation. The probability that *B* hitchhikes with *A* is equal to its initial frequency, x, so a uniform number of derived variants over the entire frequency spectrum, $x\phi(x)dx = \theta dx$, are transformed to high frequencies, while the rest, $(\theta/x - \theta)dx$, are transformed to low frequencies (inset of Figure 1). The frequency spectrum after hitchhiking can be approximated after a linear transformation of expectations (subscripted L and H are for low and high frequencies, respectively),

$$\phi_L(x) = \theta\left(\frac{1}{x} - \frac{1}{c^*}\right) \quad \text{for } \frac{1}{2N} \le x < c^* \tag{5}$$

$$\phi_H(x) = \theta\left(\frac{1}{c^*}\right) \quad \text{for } 1 - c^* < x \le 1 - \frac{1}{2N}, \tag{6}$$

and zero otherwise (inset of Figure 1).

Stochastic: Coalescent simulations of single hitchhiking events, where the advantageous mutation has just reached fixation, produce a skew in the frequency spectrum similar to that predicted by the deterministic theory (Figure 1). The hitchhiking algorithm incorporates stochastic mutation and recombination events but assumes that the advantageous allele follows a deterministic increase in frequency (materials and methods). An exact treatment of the spread of an advantageous mutation through the population incorporates stochastic fluctuations in the frequency of the selected allele when it is at low or high frequencies but would produce

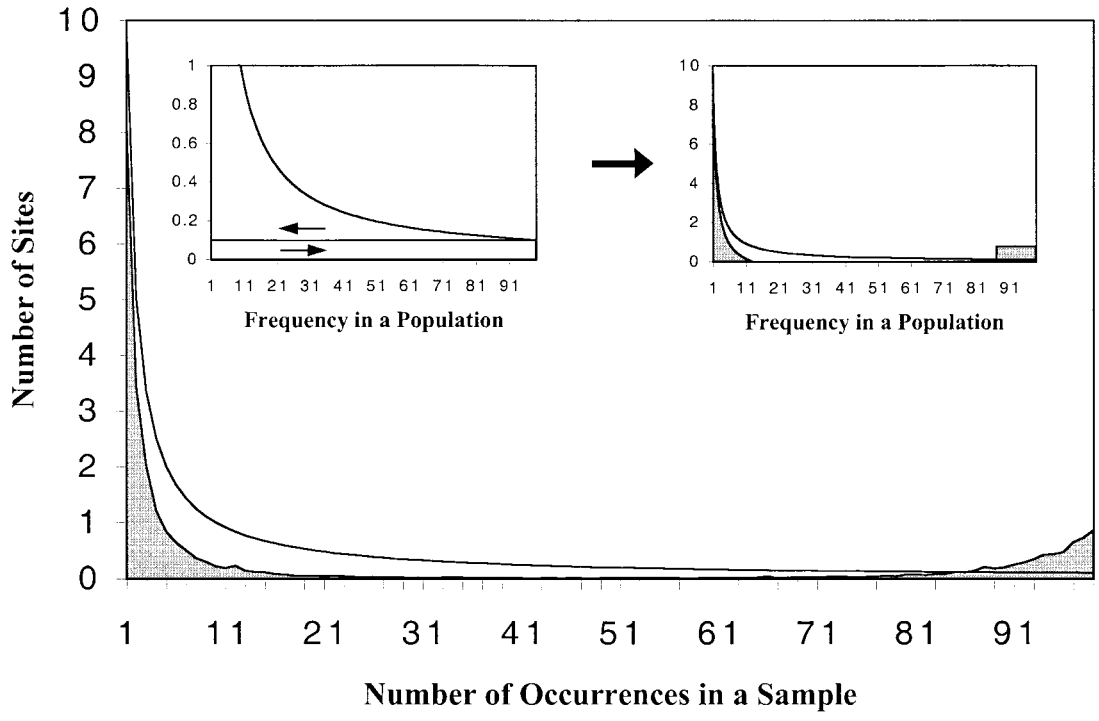

Figure 1.—The frequency spectrum of derived variants expected under neutrality (open areas) and subsequent to a hitchhiking event (shaded areas) with parameters $\theta = 10$, $n = 100$, $c/s = 0.01$, and $s = 10^{-3}$ for 1000 iterations of a coalescence algorithm. The inset shows the proportion of variants in a population that are expected to increase or decrease in frequency during a hitchhiking event and the resulting frequency spectrum found under the deterministic model.

TABLE 1

Expected numbers of low-, intermediate-, and high-frequency variants for different strengths of the hitchhiking effect

c/s	Low $0 < x \leq 0.1$	Intermediate $0.1 < x < 0.9$	High $0.9 \leq x < 1.0$	Relative abundance Low:intermediate:high
Neutral	29.29	21.42	1.06	27.6:20.2:1.0
0.100	25.55	18.65	1.62	15.7:11.5:1.0
0.010	17.93	3.31	4.50	4.0:0.7:1.0
0.001	11.52	0.31	1.27	9.1:0.2:1.0

The expected number of sites in a sample of 100 was calculated from 1000 iterations of a coalescence algorithm (materials and methods), where $\theta = 10$ and $s = 10^{-3}$.

similar results so long as selection is stronger than recombination (Barton 1998).

The strength of hitchhiking is determined by the ratio of the recombination rate to the selection coefficient. As the strength of the hitchhiking effect increases, the ratio of low- to intermediate- to high-frequency variants becomes skewed toward low and high frequencies (Table 1). The frequency spectrum approaches that predicted under neutrality as the strength of hitchhiking decreases. Most importantly, there is a striking difference between a genealogy produced by complete (no recombination events) and incomplete hitchhiking. In the extreme case when all samples except one have coalesced during hitchhiking, nearly all mutations are present at a frequency of either $(n - 1)/n$ or $1/n$, with equal probability. In contrast, when there is no recombination a star genealogy is produced and any new mutation would be found once in a sample (Figure 2). Therefore, when a neutral variant is tightly, but not completely, linked to a site that has just been selectively driven to fixation, the salient feature of hitchhiking is a bipartite spectrum of allele frequencies. Subsequent to a hitchhiking event, variation will be regained, first as low-frequency variants, and will eventually return to equilibrium.

Statistics: A standard statistical test for hitchhiking is Tajima's *D* statistic, which compares the number of rare to intermediate-frequency variants (Tajima 1989b). Tajima's *D* is the standardized difference between two commonly used measures of variability: $\hat{\theta}_W$, which is based on the total number of segregating sites and is influenced most by low-frequency variation (Watterson 1975), and $\hat{\theta}_\pi$, which is based on average heterozygosity and is influenced most by intermediate-frequency variants (Tajima 1989b). While the *D* statistic should be useful in the detection of a partial hitchhiking event, many other processes, such as background selection or a change in population size (Charlesworth *et al.* 1993), may generate *D* values that depart from neutrality due to an excess of low-frequency variants. On the other hand, strong deterministic forces are needed to drive variants to high frequencies. A statistic based on an excess of high-frequency variants would thus be uniquely sensitive to hitchhiking.

We define the test statistic, *H*, as the difference between $\hat{\theta}_\pi$, which is influenced most by variants at intermediate frequencies, and $\hat{\theta}_H$ (materials and methods), which is influenced most by high-frequency variants. This test is analogous to and was motivated by the $F(0, -1)$ test described in Fu (1995), the only difference being $F(0, -1)$ is the difference between $\hat{\theta}_\pi$ and $\hat{\theta}_H$ divided by the variance of the difference. Under neutrality the expected difference between two estimators of θ is zero, but following a hitchhiking event $\hat{\theta}_H$ and $\hat{\theta}_W$ should be $> \hat{\theta}_\pi$. This prediction is met over a range of c/s values (Figure 3).

The power of both the *D* and *H* tests in rejecting neutrality is similar and greatest for intermediate values of c/s. The power of these statistics compared with the percentage of simulated hitchhiking events where both the *D* and *H* tests reject neutrality shows that the two tests are complementary in their ability to detect a single hitchhiking event. For example, when $c/s = 0.01$, 51.7 and 61.1% of the hitchhiking simulations were rejected by the *D* and *H* tests, respectively, while 38.6% were rejected by both. Both *D* and *H* reject neutrality when there is an excess of low- and high-frequency variants, but only *D* can reject neutrality when there is an excess

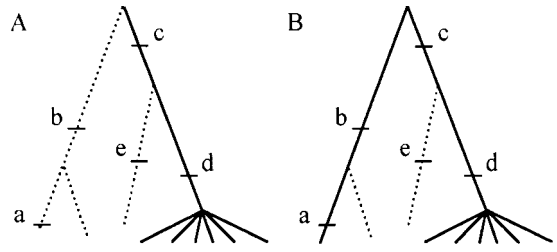

Figure 2.—Genealogies produced by positive selection (A) in the absence of recombination and (B) when there is a single recombinant in a sample. All variation is lost in the genealogy of A but in B mutations a and b produce low-frequency variants and mutations c and d produce high-frequency variants. Dotted lines indicate lineages lost due to hitchhiking.

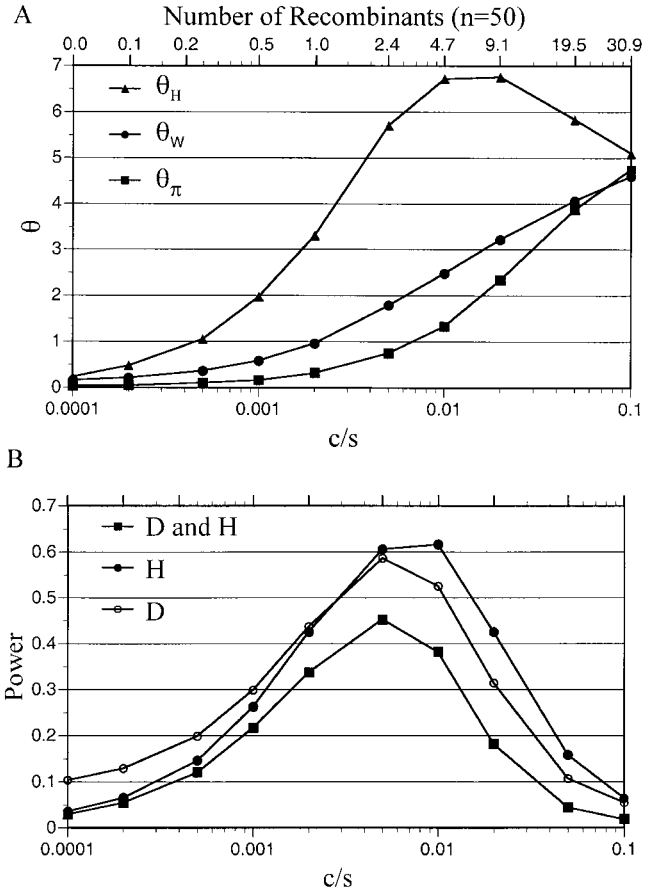

Figure 3.—(A) The mean for each of the three different estimators of θ as a function of recombination rate where each point is calculated from 5000 coalescence simulations for $s = 0.005$ and $\theta = 5$ (materials and methods). In the absence of hitchhiking, the means of θ_H, θ_W, and θ_π from the simulations are 4.94, 4.99, and 4.97, respectively. The number of recombinants is the number of alleles in a sample of 50 that escaped coalescence during hitchhiking by recombining onto an advantageous chromosome. (B) The corresponding power of the *D* and *H* statistics. The solid squares indicate the proportion of cases where both statistics reject neutrality and the two statistics overlap.

of low- but not high-frequency variants and vice versa for *H*. While Figure 3 gives an indication of slight differences between the power of the two tests, this difference should be interpreted with caution since it depends on a number of parameters not explored here, such as the number of recombination events, level of variation before hitchhiking, and time since the hitchhiking event. For instance, when one recombinant is in a sample and there are few segregating sites, only very low- or high-frequency variants would be expected since there is nearly an equal probability that a derived variant is found at low (a and b) or high (c and d) frequency (Figure 2). Under this scenario the *H* test but not Tajima's *D* may be significant (Table 2). Regardless of the relative powers of the two tests, it is the excess of high- but not low-frequency variants that is the unique signature of positive selection.

APPLICATION TO DATA

In regions of low recombination, the *H* test may distinguish hitchhiking from a neutral or background selection model, which is expected to have the same relative abundance of intermediate- to high-frequency variants (Charlesworth *et al.* 1995; Fu 1997). We identified seven published polymorphism surveys in four regions with greatly reduced rates of recombination (materials and methods). *Achaete*, *asense*, and *Pgd* are located on the tip of the X chromosome and a significant excess of high-frequency variants is found at *achaete* (Table 2). An excess of high-frequency variants is not found at *su(f)* at the base of the X chromosome or at *ftz* or *Gld*, near the centromere of the third chromosome. In *D. ananassae*, there is a nonsignificant excess of high-frequency variants at *vermilion* near the centromere of the X chromosome.

The levels of variation and excess of high-frequency variants found at *achaete*, *asense*, and *Pgd* suggest a weakening hitchhiking effect with their distance from the tip of the X chromosome. *Achaete* has a significant excess of high-frequency variants and the largest reduction in levels of variation. A weaker hitchhiking pattern is found at *asense* and it is weaker still at *Pgd*. Out of eight comparisons one significant association was found between variation at *achaete* and *Pgd* (Begun and Aquadro 1991). This is consistent with the hitchhiking effect, which produces linkage disequilibrium (Thomson 1977). In the absence of hitchhiking, no genealogical association would be expected between loci separated by a recombinational distance, $4Nc$, of >100 (Hudson 1983). The three loci should be independent, given that the rate of recombination between *achaete* and *asense* is estimated to be 1.5×10^{-4} and between *achaete* and *Pgd*, 6×10^{-3}, approximately (FlyBase 1999). Treating the three surveys as independent, the three *P* values can be combined (Sokal and Rohlf 1981) to show a significant departure from neutrality for the entire *ac-ase-Pgd* region ($P = 0.022$).

In regions of normal recombination, the hitchhiking effect may be limited to a small region so the location of the sequence under selection must be known. The *accessory gland protein* gene (*Acp26Aa*) has been shown to be under positive selection by interspecific comparisons (Aguadé *et al.* 1992; Tsaur and Wu 1997; Aguadé 1998; Tsaur *et al.* 1998). On the basis of the expected ratio of synonymous and nonsynonymous differences within and between species (McDonald and Kreitman 1991), only 29 nonsynonymous substitutions are expected between *D. melanogaster* and *D. simulans*, whereas 75 are observed (Tsaur *et al.* 1998). Thus, it is likely that a hitchhiking event occurred recently enough to influence patterns of intraspecific variation in *D. melanogaster*.

As previously noted (Tsaur *et al.* 1998), there is an excess of high-frequency variants in the *Acp26Aa* gene (Figure 4). Because the locus is in a region of high

TABLE 2

The probability of hypothetical and published DNA polymorphism surveys under the neutral model

Locus	Sample size	Variable sites	θ_W	θ_π	θ_H	d	P(BM)	No. of occurrences of each derived variant	$P(D)$	$P(H)$
Example	10	1	0.35	0.20	1.80	0.010	0.004	[9]	0.430	0.028
Example	10	3	1.06	0.60	5.40	0.010	0.004	[9 9 9]	0.128	0.002
Example	10	3	1.06	0.60	3.62	0.010	0.004	[1 9 9]	0.128	0.018
Example	10	3	1.06	0.60	1.84	0.010	0.004	[1 1 9]	0.128	0.090
Example	10	3	1.06	0.60	0.07	0.010	0.004	[1 1 1]	0.128	0.665
achaete	147	3	0.54	0.87	3.85	0.062	0.084	[84, 120, 141]	0.863	0.027
asense	6	6	2.63	2.20	4.60	0.038	0.014	[1, 1, 1, 4, 5, 5]	0.215	0.077
Pgd	13	14	4.51	5.44	6.06	0.069	0.026	[1(2), 3(3), 4(3), 5(3), 10(2), 11]	0.825	0.298
vermilion	45	6	1.37	0.81	2.55	0.022	0.008	[1(3), 3, 24, 44]	0.160	0.070
Acp26Aa	101	54	10.41	8.07	17.97	0.117	0.044	See Figure 4	0.083	0.043

d is divergence to the outgroup. P(BM) is the probability of a backmutation (materials and methods). $P(D)$ and $P(H)$ are the probability of observing a more negative D and H value under neutrality ($\alpha = 0.05$, one sided), respectively. Numbers in parentheses indicate the number of sites where the derived variant is observed; *i.e.*, *Pgd* has two singletons. For *Acp26Aa*, $4Nc/\text{bp} = 0.0725$ was used to generate the expected distribution of D and H values. At two of the three polymorphic sites at the *achaete* locus, the derived variant was inferred to be a loss of a restriction site.

recombination, c is estimated to be 4.4×10^{-8} (Comerón *et al.* 1999), and there is ample evidence of recombination between sites as close as 20 bp apart, a hitchhiking effect could be limited to a small region within the gene and each segregating site would experience a different strength of the hitchhiking effect. A sliding window plot of polymorphism *vs.* divergence shows an excess of high-frequency variants on both sides of a region where there is an absence of variation when compared to divergence (Figure 5). This is the expected pattern produced by positive selection over a continuous range of linked variation (Figure 3). Of the 13 variants found at a frequency of >50%, all lie within the region affected by hitchhiking and six lie on one side and seven lie on the other side of the region presumably affected by complete hitchhiking, which is indicated by the arrows. No such pattern is found in the *Acp26Ab* gene.

DISCUSSION

In addition to reducing levels of variation (Maynard Smith and Haigh 1974), positive selection skews the frequency spectrum of linked variation to low and high frequencies. This hitchhiking effect is a function of recombinational distance from the site under selection and time since the hitchhiking event(s) (Kaplan *et al.* 1989).

Recombination: Hitchhiking is complete or incomplete in the removal of variation depending on whether or not recombination has occurred. The size of the region affected by partial hitchhiking is at least two orders of magnitude larger than that affected by complete hitchhiking. For a c/s value of 2×10^{-4}, hitchhiking is nearly complete, heterozygosity is reduced by 99%, and there is at least 1 recombinant out of a sample of 50 in 10% of hitchhiking events (Figure 3). In contrast, for a c/s value of 0.02, hitchhiking is partial and heterozygosity is reduced by 57%. For humans and flies, where the rate of recombination between adjacent nucleotides is on the order of 10^{-8} (Ashburner 1989; Broman *et al.* 1998), complete hitchhiking ($c/s = 2 \times 10^{-4}$) is limited to a region of 20 ($s = 0.001$) or 200 bp ($s = 0.01$).

For incomplete hitchhiking, the degree to which variation is reduced corresponds to the degree to which the frequency spectrum is skewed. This is the case for *achaete*, *asense*, and *Pgd*, which span a 2-Mb region on the tip of the X chromosome (FlyBase 1999), and for the second exon of the *Acp26Aa* gene, where nearly all variation is eliminated from an ~350-bp region and the entire hitchhiking pattern is limited to an ~700-bp region (Figure 5). Taking c to be 5×10^{-8}/bp (Comerón *et al.* 1999) at the *Acp26Aa* gene, the selection coefficient can be estimated from the size of the region affected by complete hitchhiking. Assuming hitchhiking is complete ($c/s < 2 \times 10^{-3}$), ~175 bp away from the site under selection the selection coefficient is at least 0.004. The actual selection coefficient may be even larger since $4Nc$ estimated from polymorphism is 0.0725/bp (Hudson *et al.* 1987; Rozas and Rozas 1999) and hitchhiking produces linkage disequilibrium. In addition, population subdivision may weaken the hitchhiking effect (see below).

Time of hitchhiking: Previous studies have shown that in the absence of recombination hitchhiking can be detected for a brief period, ~0.5N generations, some time after a hitchhiking event (Simonsen *et al.* 1995; Fu 1997). We have shown that in the presence of recombination, hitchhiking can be detected immediately after the fixation of an advantageous mutation. Regardless of whether recombination is present or absent, there

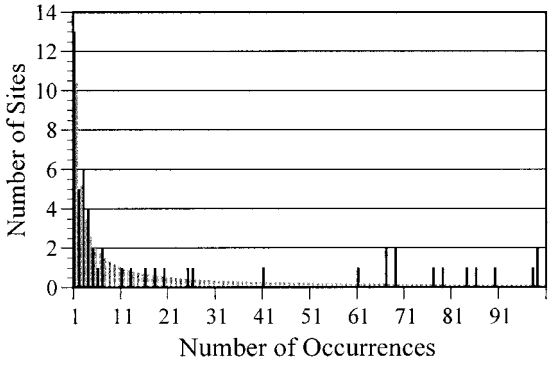

Figure 4.—The observed (■) and expected (□) numbers of derived variants that occur in a worldwide sample of 101 *Acp26Aa D. melanogaster* genes (Aguadé *et al.* 1992; Aguadé 1998; Tsaur *et al.* 1998).

is only a transient period after hitchhiking when the deviation from the neutral frequency spectrum is detectable. As a population recovers from hitchhiking, *D* can be expected to be more powerful than *H*. This is because *D* but not *H* is sensitive to the influx of new mutations during the recovery period. As the excess of high-frequency variants is less likely to be influenced by other forces such as population history or background selection, monitoring this portion of the frequency spectrum represents a trade-off between power and distinguishing hitchhiking from several different hypotheses.

The significant excess of high-frequency variants in regions of low recombination as reported in Table 2 therefore suggests hitchhiking may be quite frequent in these regions, an inference supported by uniformly low levels of variation in regions of low recombination. As a result, recovery to equilibrium levels of variation may never occur and the effects of multiple or overlapping hitchhiking events must be considered. Models of strong recurrent hitchhiking events (Braverman *et al.* 1995) and numerous sites concurrently under weak selection (Santiago and Caballero 1998) predict an excess of rare variants and negative *D* values. Although recurrent hitchhiking is not modeled in this study, a single strong hitchhiking event will push all variation to either very low or high frequencies, regardless of the frequency spectrum before selection. However, it is possible that a second round of hitchhiking may result in high-frequency variants being the only detectable polymorphism, as is observed at *achaete*. If the hitchhiking effect in the first round is strong and only low-frequency variants remain, during the second round of hitchhiking, low-frequency variants will either increase to high frequencies or decrease to nondetectable frequencies.

Population subdivision: While an excess of low-frequency variants can be explained by positive selection, background selection, a change in population size, and population subdivision (Charlesworth *et al.* 1993), an excess of high-frequency variants can only be explained by positive selection and specific demographic models. For example, if there are many fixed differences between populations and if only a few migrants are found in a sample, most variants would be at either low or high frequency. A more extreme scenario is that of introgression of alleles from one species into another, in which case all substitutions between the two species would appear to be either low- or high-frequency variants. Neither scenario is applicable to the data used in this article since ancient population subdivision is not supported in either *D. melanogaster* or *D. ananassae* (Stephan 1989; Stephan and Langley 1989; Moriyama and Powell 1996). However, population subdivision does affect the variance of all tests that assume an equilibrium nondifferentiated population (but see Charlesworth 1998; Stephan *et al.* 1998).

The pooling of subpopulation data makes the *H* test conservative and may obscure patterns of hitchhiking in a single subpopulation. For instance, at *su(f)* there is one low (2/47) and one high (43/47) frequency-derived variant found in a North American sample. Neither of the two variants are found in a sample of 49 African flies (Begun and Aquadro 1995). Thus, the pooled samples (2/96, 43/96) have no derived variants at a frequency of >50%. Population subdivision may

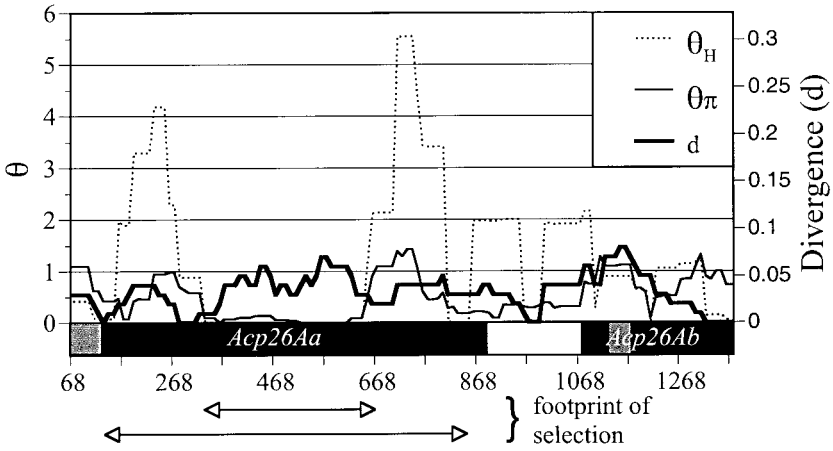

Figure 5.—A sliding window plot of polymorphism and divergence (d, thick line) between *D. melanogaster* line NC1 and *D. simulans* (Aguadé *et al.* 1992). Window size is 100 bp and step size is 10 bp. On the abscissa, solid boxes are exons and shaded boxes are introns. The two double arrows indicate the size of the region affected by hitchhiking: the smaller one (~350 bp) has lost nearly all variation and the larger one (~700 bp) encompasses the region where there is an excess of high-frequency variants.

also have obscured a clear hitchhiking pattern at *vermilion*. There is a nonsignificant excess of high-frequency variants when four subpopulations are pooled. Using a method independent of the frequency spectrum, which compares levels of variation within and between populations to divergence, hitchhiking was inferred to have homogenized two of the four populations, an observation incompatible with background selection (Stephan *et al.* 1998).

CONCLUSIONS

An excess of very-high-frequency variants is a distinct consequence of hitchhiking. Background selection cannot easily account for this pattern, which is observed in some regions of low recombination, as revealed by the *H* statistic. Although it has been previously suggested that the lack of significant *D* values in these regions is inconsistent with a hitchhiking explanation (Braverman *et al.* 1995; Charlesworth *et al.* 1995), cases of hitchhiking that yield significant *H* but not *D* and vice versa are not uncommon. The absence of significant *D* values can also be explained by the unexplored range of parameter space of the recurrent hitchhiking model, population subdivision, or by alternative models of selection, such as frequency-dependent selection (Gillespie 1991; Braverman *et al.* 1995). Further improvement of our ability to detect positive selection and distinguish its effects from other evolutionary processes will be made by incorporating population structure (Charlesworth 1998; Slatkin and Wiehe 1998; Stephan *et al.* 1998) and linkage disequilibrium (Thomson 1977; Grote *et al.* 1998) into models of selection.

We thank B. Charlesworth, R. Hudson, M. Jensen, C. Langley, T. Nagylaki, J. Spofford, C. Ting, J. Wall, and two reviewers for comments and suggestions. We thank S. Sun for help in deriving the probability of a back mutation. This work was supported by National Institutes of Health and National Science Foundation grants to C.-I Wu and a Genetics Training Grant to J. C. Fay.

LITERATURE CITED

Aguadé, M., 1998 Different forces drive the evolution of the *Acp26Aa* and *Acp26Ab* accessory gland genes in the *Drosophila melanogaster* species complex. Genetics **150:** 1079–1089.

Aguadé, M., N. Miyashita and C. H. Langley, 1989 Reduced variation in the *yellow-achaete-scute* region in natural populations of *Drosophila melanogaster*. Genetics **122:** 607–615.

Aguadé, M., N. Miyashita and C. H. Langley, 1992 Polymorphism and divergence in the *Mst26a* male accessory gland gene region. Genetics **132:** 755–770.

Ashburner, M., 1989 *Drosophila: A Laboratory Handbook*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Barton, N. H., 1998 The effect of hitch-hiking on neutral genealogies. Genet. Res. **72:** 123–133.

Begun, D. J., and C. F. Aquadro, 1991 Molecular population genetics of the distal portion of the X chromosome in Drosophila: evidence for genetic hitchhiking of the *yellow-achaete* region. Genetics **129:** 1147–1158.

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 C. F. Aquadro, 1994 Evolutionary inferences from DNA variation at the 6-phosphogluconate dehydrogenase locus in natural populations of *Drosophila*: selection and geographic differentiation. Genetics **136:** 155–171.

Begun, D. J., and C. F. Aquadro, 1995 Evolution at the tip and base of the X chromosome in an African population of *Drosophila melanogaster*. Mol. Biol. Evol. **12:** 382–390.

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

Broman, K. W., J. C. Murray, V. C. Sheffield, R. L. White and J. L. Weber, 1998 Comprehensive human genetic maps: individual and sex-specific variation in recombination. Am. J. Hum. Genet. **63:** 861–869.

Charlesworth, B., 1998 Measures of divergence between populations and the effect of forces that reduce variability. Mol. Biol. Evol. **15:** 538–543.

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

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

Comerón, J. M., M. Kreitman and M. Aguadé, 1999 Natural selection on synonymous sites is correlated with gene length and recombination in *Drosophila*. Genetics **151:** 239–249.

Fay, J. C., and C.-I Wu, 1999 A human population bottleneck is compatible with the discordance between patterns of mitochondrial vs. nuclear DNA variation. Mol. Biol. Evol. **16:** 1003–1005.

FlyBase, 1999 The FlyBase Database of the Drosophila Genome Projects and community literature. Nucleic Acids Res. **27:** 85–88.

Fu, Y. X., 1995 Statistical properties of segregating sites. Theor. Popul. Biol. **48:** 172–197.

Fu, Y. X., 1997 Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics **147:** 915–925.

Gillespie, J. H., 1991 *The Causes of Molecular Evolution*. Oxford University Press, Oxford.

Grote, M. N., W. Klitz and G. Thomson, 1998 Constrained disequilibrium values and hitchhiking in a three-locus system. Genetics **150:** 1295–1307.

Hamblin, M. T., and C. F. Aquadro, 1997 Contrasting patterns of nucleotide sequence variation at the Glucose Dehydrogenase (Gld) locus in different populations of Drosophila melanogaster. Genetics **145:** 1053–1062.

Hilton, H., R. M. Kliman and J. Hey, 1994 Using hitchhiking genes to study adaptation and divergence during speciation within the *Drosophila melanogaster* species complex. Evolution **48:** 1900–1913.

Hudson, R. R., 1983 Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. **23:** 183–201.

Hudson, R. R., 1990 Gene genealogies and the coalescent process, pp. 1–44 in *Oxford Surveys in Evolutionary Biology*, edited by D. Futuyama and J. Antonovics. Oxford University Press, New York.

Hudson, R. R., M. Kreitman and M. Aguadé, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116:** 153–159.

Jenkins, D. L., C. A. Ortori and J. F. Y. Brookfield, 1995 A test for adaptive change in DNA sequences controlling transcription. Proc. R. Soc. Lond. Ser. B **261:** 203–207.

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

Kimura, M., 1983 *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, UK.

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

Martín-Campos, J., J. M. Comerón, N. Miyashita and M. Aguadé, 1992 Intraspecific and interspecific variation at the *y-ac-sc* region of *Drosophila simulans* and *Drosophila melanogaster*. Genetics **130:** 805–816.

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

McDonald, J. H., 1998 Improved tests for heterogeneity across a

region of DNA sequence in the ratio of polymorphism to divergence. Mol. Biol. Evol. **15:** 377–384.
McDonald, J. H., and M. Kreitman, 1991 Adaptive protein evolution at the Adh locus in *Drosophila*. Nature **351:** 652–654.
Moriyama, E. N., and J. R. Powell, 1996 Intraspecific nuclear DNA variation in Drosophila. Mol. Biol. Evol. **13:** 261–277.
Nachman, M. W., 1997 Patterns of DNA variability at X-linked loci in *Mus domesticus*. Genetics **147:** 1303–1316.
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.
Rozas, J., and R. Rozas, 1999 DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics **15:** 174–175.
Santiago, E., and A. Caballero, 1998 Effective size and polymorphism of linked neutral loci in populations under directional selection. Genetics **149:** 2105–2117.
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.
Slatkin, M., and T. Wiehe, 1998 Genetic hitch-hiking in a subdivided population. Genet. Res. **71:** 155–160.
Sokal, R. R., and F. J. Rohlf, 1981 *Biometry*. W. H. Freeman and Co., New York.
Stephan, W., 1989 Molecular genetic variation in the centromeric region of the X chromosome in three Drosophila ananassae populations. II. The Om(1D) locus. Mol. Biol. Evol. **6:** 624–635.
Stephan, W., and C. H. Langley, 1989 Molecular genetic variation in the centromeric region of the x chromosome in three *Drosophila ananassae* populations. I. Contrasts between *vermilion* and *forked* loci. Genetics **121:** 89–99.
Stephan, W., and C. H. Langley, 1998 DNA polymorphism in *Lycopersicon* and crossing-over per physical length. Genetics **150:** 1585–1593.
Stephan, W., and S. J. Mitchell, 1992 Reduced levels of DNA polymorphism and fixed between-population differences in the centromeric region of *Drosophila ananassae*. Genetics **132:** 1039–1045.
Stephan, W., T. H. E. Wiehe and M. W. Lenz, 1992 The effect of strongly selected substitutions on neutral polymorphism-analytical results based on diffusion theory. Theor. Popul. Biol. **41:** 237–254.
Stephan, W., L. Xing, D. A. Kirby and J. M. Braverman, 1998 A test of background selection hypothesis based on nucleotide data from *Drosophila ananassae*. Proc. Natl. Acad. Sci. USA **95:** 5649–5654.
Tajima, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics **123:** 437–460.
Tajima, F., 1989a The effect of change in population size on DNA polymorphism. Genetics **123:** 597–601.
Tajima, F., 1989b Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585–595.
Thomson, G., 1977 The effect of a selected locus on linked neutral loci. Genetics **85:** 753–788.
Tsaur, S. C., and C.-I Wu, 1997 Positive selection and the molecular evolution of a gene of male reproduction, *Acp26Aa* of *Drosophila*. Mol. Biol. Evol. **14:** 544–549.
Tsaur, S. C., C. T. Ting and C.-I Wu, 1998 Positive selection driving the evolution of a gene of male reproduction, *Acp26Aa*, of *Drosophila*. II. Divergence versus polymorphism. Mol. Biol. Evol. **15:** 1040–1046.
Wall, J. D., 1999 Recombination and the power of statistical tests of neutrality. Genet. Res. **74:** 65–79.
Watterson, G. A., 1975 On the number of segregating sites. Theor. Popul. Biol. **7:** 256–276.

Communicating editor: W. Stephan