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 (Stefan and Langley 1998), flies (Stefan and Langley 1989; Begun and Aquadro 1992; Martin-Campos et al. 1992; Stefan 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 (Stefan et al. 1992). In a region where recombination is reduced 10-fold, such as the tip of the X chromosome (Aguide 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.
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 (Comeron et al. 1999). Regions of low recombination were designated as cytological divisions with rates of recombination < 5 \times 10^{-4} bp: 1A–2D, 2C–2F, 2A, 3B–4A, 4C–6F, 6A–6B, and 7F–8F (excluding the fourth chromosome where recombination is absent). This cutoff value is about 10 times lower than the division with the highest estimated rate of recombination. There are six published polymorphism surveys in the designated regions: achate (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 (suf, 2O1; Langley et al. 1993), fushi tarazu (fz, 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 and intraspecific variation has been extensively studied (Aguadé et al. 1992; Tsaur and Wu 1997; Aquadé 1998; Tsaur et al. 1998), vermilion (v; Steph an and Langley 1989; Steph an 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 achate, 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 achate (which is a four-cutter restriction fragment length polymorphism [RFLP] study) and vermilion, which is a survey of polymorphism in D. ananassae and 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 Braver 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 \( \mu \) is the mutation rate: \( \hat{\theta}_u \) is calculated from average heterozygosity (Tajima 1983) and \( \hat{\theta}_w \) is calculated from the number of segregating sites (Waterson 1975),

\[
\hat{\theta}_w = \frac{\sum_{i=1}^{n-1} 2Si(n-i)}{n(n-1)}
\]

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

\[
\hat{\theta}_u = \frac{\sum_{i=1}^{n-1} 2Si(n-i)}{n(n-1)}
\]

where \( \hat{\theta}_n \) is an unbiased estimator of \( \theta \) and is a special case, noted 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 \( \hat{\theta}_u \) and \( \hat{\theta}_w \). 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, back mutation 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 back mutation 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 back mutation 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 back mutation 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-
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 \( x + \text{dx} \) is given by

\[
\phi(x) \text{dx} = \frac{\theta}{x} \text{dx}.
\]  
(4)

In Equation 4, \( \theta = 4N\mu \), where \( N \) is the effective population size and \( \mu \) 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, \( x\phi(x) \text{dx} = \text{dx} \), are transformed to high frequencies, while the rest, \( \text{(x/}x - \theta)\text{dx} \), 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).

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

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

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

![Image of frequency spectrum](image-url)
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: \(\theta_w\), which is based on the total number of segregating sites and is influenced most by low-frequency variation (Watterson 1975), and \(\theta_n\), 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 \(\theta_n\), which is influenced most by variants at intermediate frequencies, and \(\theta_w\) (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 \(\theta_n\) and \(\theta_w\) divided by the variance of the difference. Under neutrality the expected difference between two estimators of \(\theta\) is zero, but following a hitchhiking event \(\theta_w\) and \(\theta_n\) should be \(>\theta_n\). 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 \(c/s = 0.01, 0.1\), and 0.5, 1.7 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 of high-frequency variants.

**Table 1**

<table>
<thead>
<tr>
<th>(c/s)</th>
<th>Low (0 &lt; x \leq 0.1)</th>
<th>Intermediate (0.1 &lt; x &lt; 0.9)</th>
<th>High (0.9 &lt; x &lt; 1.0)</th>
<th>Relative abundance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neutral</td>
<td>29.29</td>
<td>21.42</td>
<td>1.06</td>
<td>27.6:20.2:1.0</td>
</tr>
<tr>
<td>0.100</td>
<td>25.55</td>
<td>18.65</td>
<td>1.62</td>
<td>15.7:11.5:1.0</td>
</tr>
<tr>
<td>0.010</td>
<td>17.93</td>
<td>3.31</td>
<td>4.50</td>
<td>4.0:0.7:1.0</td>
</tr>
<tr>
<td>0.001</td>
<td>11.52</td>
<td>0.31</td>
<td>1.27</td>
<td>9.1:0.2:1.0</td>
</tr>
</tbody>
</table>

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 \(s = 10^{-3}\).
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 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 $1.5 \times 10^{-5}$ and between achaete and Pgd, $6 \times 10^{-5}$, 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 a-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

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

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 (α = 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/bp = 0.0725 was used to generate the expected distribution of D and H values. At two of the three polymorphic sites at the achate locus, the derived variant was inferred to be a loss of a restriction site.

recombination, c is estimated to be 4.4 × 10^{-8} (Comeron 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 × 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 achate, 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^{-9}/bp (Comeron 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 × 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
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 hitchhik-
ing 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 fre-
quencies.

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 be-
tween 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 vari-
ants. 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 and Langley 1989; Moriyama and Powell 1996).
However, population subdivision does affect the variance of all tests that assume an equi-
librium nondifferentiated population (but see Charle-
sworth 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. Nei-
ther 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
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 (Stefan 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 consistent 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; Stefan 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


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


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.


Communicating editor: W. Stephan