The Signature of Positive Selection at Randomly Chosen Loci

Molly Przeworski[1]

Department of Statistics, University of Oxford, Oxford OX1 3TG, United Kingdom

Manuscript received June 4, 2001
Accepted for publication November 26, 2001

ABSTRACT

In Drosophila and humans, there are accumulating examples of loci with a significant excess of high-frequency-derived alleles or high levels of linkage disequilibrium, relative to a neutral model of a random-mating population of constant size. These are features expected after a recent selective sweep. Their prevalence suggests that positive directional selection may be widespread in both species. However, as I show here, these features do not persist long after the sweep ends: The high-frequency alleles drift to fixation and no longer contribute to polymorphism, while linkage disequilibrium is broken down by recombination. As a result, loci chosen without independent evidence of recent selection are not expected to exhibit either of these features, even if they have been affected by numerous sweeps in their genealogical history. How then can we explain the patterns in the data? One possibility is population structure, with unequal sampling from different subpopulations. Alternatively, positive selection may not operate as is commonly modeled. In particular, the rate of fixation of advantageous mutations may have increased in the recent past.

CONSIDERABLE debate has focused on what proportion of genetic changes is favored by natural selection, as well as what types of substitutions are most likely to have been selected (Andolfatto 2001; Fay and Wu 2001). Answers to these questions will help to elucidate the genetic basis of adaptation.

To infer that positive selection has acted on a particular genomic region, population geneticists usually sequence a number of individuals at a locus and test whether the pattern of polymorphism seen in the sample is unexpected under the standard neutral model of a random-mating population of constant size. Unfortunately, a departure from null model expectations can be due to one of many causes, so it is hard to establish that adaptation is responsible. In particular, an excess of rare variants may reflect a selected substitution at a closely linked site, but it may also be caused by population expansion or purifying selection, just to list a couple of alternatives. For this reason, an ideal "test of neutrality" would not only have high power to detect positive selection, but would also focus on an aspect of the data unlikely to be affected by demography or other factors. Such a test statistic (H) was recently proposed by Fay and Wu (2000), to detect a single, recent episode of positive selection (Otto 2000).

Since its introduction, significant H values have been reported for samples from *Acp26Aa* (Fay and Wu 2000), *achaete* (Fay and Wu 2000), Attacins A and B (Lazzaro and Clark 2001), and *desat2* (Takahashi *et al.* 2001) in *Drosophila melanogaster* and the *janA-ocn* region in *D. simulans* (Parsch *et al.* 2001). In humans, examples include FY (Hamblin *et al.* 2002), MAO-A (Gilad *et al.* 2001), and several noncoding loci: a subset of olfactory receptor pseudogenes (data from Gilad *et al.* 2000; M. Przeworski, unpublished results), psGBA (data from Martinez-Arias *et al.* 2001; M. Przeworski, unpublished results), the intron DMD7 (data from Nachman and Crowell 2000; M. Przeworski, unpublished results), and 3 out of 19 intergenic regions (Frisse *et al.* 2001; L. Frisse and A. Di Rienzo, personal communication). Considered together with multilocus evidence (*e.g.*, Aquadro *et al.* 1994; Andolfatto and Przeworski 2001; Nachman 2001) and an accumulating number of individual loci that show evidence of positive selection (reviewed in Andolfatto 2001), these frequency spectrum results suggest that a large fraction of genetic changes may be favored (Fay and Wu 2001).

In addition, patterns of linkage disequilibrium (LD) depart from the expectations of the standard neutral model in these species. There appears to be a genome-wide excess of intralocus linkage disequilibrium in *D. melanogaster* and non-African populations of *D. simulans* (Andolfatto and Przeworski 2000; J. D. Wall, P. Andolfatto and M. Przeworski, unpublished results) and there are numerous examples of pairwise linkage disequilibrium extending over unexpectedly large distances in humans (*e.g.*, Rieder *et al.* 1999; Taillon-Miller *et al.* 2000; Gilad *et al.* 2001; reviewed in Pritchard and Przeworski 2001). It is often argued that these patterns reflect the action of positive selection at or near the sampled region (*e.g.*, Taillon-Miller *et al.* 2000; Gilad *et al.* 2001; Parsch *et al.* 2001; other references in Andolfatto 2001), again suggesting that there are many targets for adaptation in the genome.

If so, patterns of polymorphism in many regions will

[1]*Address for correspondence:* Max Planck Institute for Evolutionary Anthropology, Inselstrasse 22, D-04103 Leipzig, Germany. E-mail: przewors@eva.mpg.de

have been shaped by repeated episodes of positive selection. However, as I show here, the *H* test has very low power to detect the effects of positive selection on a randomly chosen locus. Similarly, the effect of selection on LD is short-lived, so even neutral loci affected by multiple adaptive substitutions at linked sites are unlikely to show unusually high levels of allelic association.

METHODS

Frequency spectrum-based "tests of neutrality": The *H* statistic presented in Fay and Wu (2000) is the difference between two estimates of the population mutation rate $\theta = 4N\mu$, where N is the diploid effective population size of the species and μ the mutation rate per generation. The two estimates are the average number of pairwise differences in the sample, π (Tajima 1983) and $\theta_H = \Sigma_{i=1}^{n-1} p_i^2/\binom{n}{2}$, where n is the sample size and p_i the frequency of the derived (*i.e.*, nonancestral) allele at segregating site i (Fu 1995). *H* is negative when there is an excess of high-frequency-derived alleles relative to the standard neutral model.

This statistic is similar to one introduced by Tajima (1989a): Tajima's *D* is the (approximately) normalized difference between π and θ_w, an estimate of θ based on the number of segregating sites in the sample. In contrast to *H*, *D* does not use information about ancestral and derived states. Negative *D* values reflect a relative excess of rare alleles in a folded frequency spectrum. Here, both *H* and *D* are used as one-tailed tests of neutrality.

Simulations of positive selection: I estimate the power of *H* to detect a model of recurrent "selective sweeps" (*cf.* Kaplan *et al.* 1989; Stephan *et al.* 1992; Braverman *et al.* 1995). The model assumes a random-mating population of constant size. My implementation of this model follows the description in Braverman *et al.* (1995), except for two features. First, I use a fixed value of the population mutation rate, rather than a fixed number of segregating sites (Hudson 1993; Wall and Hudson 2001). Second, I allow for recombination within the neutral locus, both during neutral and selective phases (see below).

In the model, a neutral locus is affected by selective sweeps that occur at some random genetic distance c, where c is uniform on $(0, M)$ and M is the maximum distance at which a single sweep has an effect on diversity levels. (What is meant by genetic distance is the population recombination rate between the neutral and selected locus.) M is on the order of $4Ns$ (Kaplan *et al.* 1989); in this implementation, $M = 4Ns$ (s is the selective coefficient of the favored allele). In simulations of a single selective sweep, the value of c is specified, as is the time since the fixation of the beneficial allele. In the model of repeated sweeps, the rate of sweeps is constant and chosen so that there is a small probability that two or more would occur simultaneously [using 1 − Equation 6 in Braverman *et al.* (1995)—this is a slight overestimate as it ignores the effects of interference between selected loci]. When a sweep occurs, the location of the selected site is randomly assigned to one side of the neutral locus. Selection is additive, with fitnesses $1, 1 + s, 1 + 2s$ for the three genotypes. Crossing over occurs within the neutral locus at rate ρ, where $\rho = 4Nr$ (r is the crossover rate per generation). There is no gene conversion, and I assume a constant rate of crossing over per base pair. The neutral locus evolves according to the infinite-sites model.

This selective sweep model is implemented as a succession of neutral and selective phases (when there are two alleles at a selected site). The algorithm for the neutral phase is the standard coalescent with recombination (*cf.* Hudson 1993). The selective sweep phase is implemented as in Braverman *et al.* (1995), with the addition of intralocus recombination. During a sweep, there are effectively two subpopulations at the neutral locus: lineages carrying the favored allele at the selected site and lineages carrying the unfavored allele. Three types of events can occur: (1) Two lineages in the same subpopulation can coalesce, (2) a lineage can recombine onto the same selective background, and (3) a lineage can recombine onto a different selective background. Patterns of polymorphism at the neutral locus are affected by events of type (2) only if the recombination breakpoint is within the neutral locus.

During the sweep, time changes in small increments, Δt. Within Δt, the probabilities of the events of interest are given by

$$\Pr\{\text{event (1)}\} = \left[\frac{\binom{i}{2}}{x(t)} + \frac{\binom{j}{2}}{(1 - x(t))}\right]\Delta t,$$

where $x(t)$ is the frequency of the favored allele at time t, i is the number of lineages carrying the favored allele, and j is the number of lineages carrying the unfavored allele (Braverman *et al.* 1995),

$$\Pr\{\text{event (2)}\} = [i\rho x(t) + j\rho(1 - x(t))]\Delta t$$

and

$$\Pr\{\text{event (3)}\} = [i(\rho + c)(1 - x(t)) + j(\rho + c)x(t)]\Delta t.$$

The change in frequency of the favored allele is modeled deterministically, from frequency ε to $1 - \varepsilon$, using Equation 3a in Stephan *et al.* (1992). I set $\varepsilon = 1/2N$ (as do Fay and Wu 2000) so $x(t)$ is given explicitly for all t. A path x can also be found by simulating the rise of a selected allele forward in time, thereby allowing for a fully stochastic treatment of the selective sweep. Modeling the rise in frequency by binomial sampling or a diffusion approximation does not change the qualitative results (results not shown).

Call the sum of the probabilities of all possible events

within a time interval S_t; $(1 - S_t)$ is approximately the probability that no event occurs, when the probabilities of all events are small. To calculate the time to the next event, I solve $\Pi^{y}(1 - S_t) < U$ for y, where U is a uniform random variable on (0, 1) and the product is taken over successive time intervals. Which event occurs at time y is chosen randomly with probability Pr{event $|t = y\}/S_y$.

If the event is of type 3, then with probability $\rho/(\rho + c)$ the crossover event occurs within the neutral locus and with probability $c/(\rho + c)$ between the selected and neutral locus. When a crossing-over event occurs within the neutral locus, a breakpoint b is chosen uniformly on $[0, L]$ where L is the length of the neutral locus. Assume, as an illustration, that the selected locus is to the left to the neutral locus and that the lineage carries the favored allele. Segments in the neutral locus right of b would then "migrate" to the subpopulation of the unfavored background. The number of lineages in both subpopulations has to be updated accordingly for those segments. Other cases are treated analogously.

The computer code for these simulations is written in C and based on coalescent programs kindly provided by R. Hudson (available at http://home.uchicago.edu/~rhudson1/). The program was error checked by comparing the output to the results in Figure 3 of Fay and Wu (2000; for which $\rho = 0$).

Power tests: The H and D tests are implemented as in Fay and Wu (2000). (For ease of comparison, note that the results in Fay and Wu are actually for a selective sweep model with fitnesses 1, $1 + 0.5s$, $1 + s$ for the three genotypes.) First, the 5% significance levels for H (or D) are determined by simulations of the standard neutral model with no recombination. I make the latter assumption for ease of comparison with Fay and Wu (2000) and because researchers have used critical values of H established for no recombination. The neutral model is implemented for a fixed number of segregating sites; *i.e.*, I generate genealogies and then place a fixed number of segregating sites on the tree. Second, data sets are generated under the alternative model for a given θ value (with or without recombination). If the value of H for a data set is more extreme than the significance level established for that number of segregating sites under the null model, the null model is rejected.

This procedure is meant to mimic what researchers would do in practice, when they come across a region with low diversity. Since the population mutation rate is unknown, one might ask to what extent the locus is consistent with the neutral model and a low mutation rate by testing if H is more extreme than expected for the *observed number of segregating sites*. If no segregating sites were found, no test would be performed. When estimating power, I exclude all runs in which there are no segregating sites. [For sake of comparison, note that Fay and Wu (2000) do not.] This procedure turns out to have roughly the right nominal rejection probability for a wide range of θ values (results not shown). The same is true for D, as well as other tests of neutrality (Wall and Hudson 2001).

The H test relies on identification of the ancestral allele. In practice, this is done with one or more outgroups, and the inference may be incorrect if there are mutations at the same site on the outgroup lineage(s). How likely this is depends on the mutation rate and on the extent of mutation rate variability across sites. Fay and Wu (2000) introduce a correction for the probability of an incorrect inference by assuming a constant mutation rate and the use of one outgroup, while I assume a known ancestral state.

Linkage disequilibrium: There are many possible summaries of LD and none is an obvious choice. Here, I consider two measures of linkage disequilibrium. The first is r^2 (*cf.* Weir 1996), a commonly used summary of the extent of allelic association between a pair of sites. I plot the decay of r^2 with distance for all polymorphisms with a frequency of the minor allele ≥ 0.1. A relative excess of LD is sometimes characterized as a deficiency in the number of distinct haplotypes for the observed number of segregating sites (*e.g.*, Parsch *et al.* 2001; Wall 2001; other references in Andolfatto 2001). To examine this aspect of the data, I consider a second summary of LD: the number of haplotypes normalized by the number of segregating sites, $n_{\text{Haps}}/(S + 1)$ (n_{Haps} is the number of distinct haplotypes in the sample and S the number of segregating sites). With no recombination, the maximum value of $n_{\text{Haps}}/(S + 1)$ is 1. Under the standard neutral model, lower levels of recombination result in a smaller $E(n_{\text{Haps}}/(S + 1))$. A total of 10^4 simulations were run for each set of parameters. In simulations used to examine levels of LD, crossing over occurs within the neutral locus at rate $\rho > 0$.

RESULTS

Selective sweeps with recombination: Most of the theoretical attention paid to models of positive selection has focused on the "selective sweep" or "hitch hiking" model (Maynard Smith and Haigh 1974). This model describes the rapid increase in frequency (and ultimate fixation in the population) of an initially rare and strongly favored allele. The effects of a selective sweep on the frequency spectrum of linked neutral sites can be understood as follows: Imagine first that there is no recombination and that we draw a sample of chromosomes from the present. They all bear a particular favored mutation, A. This allele increased in frequency very rapidly, such that, not very long ago, there were only a few copies in the population. As the number of copies of the favored allele decreases (going backward in time), coalescences between lineages ancestral to our sample happen faster and faster. This means that members of a sample from this region are much more closely

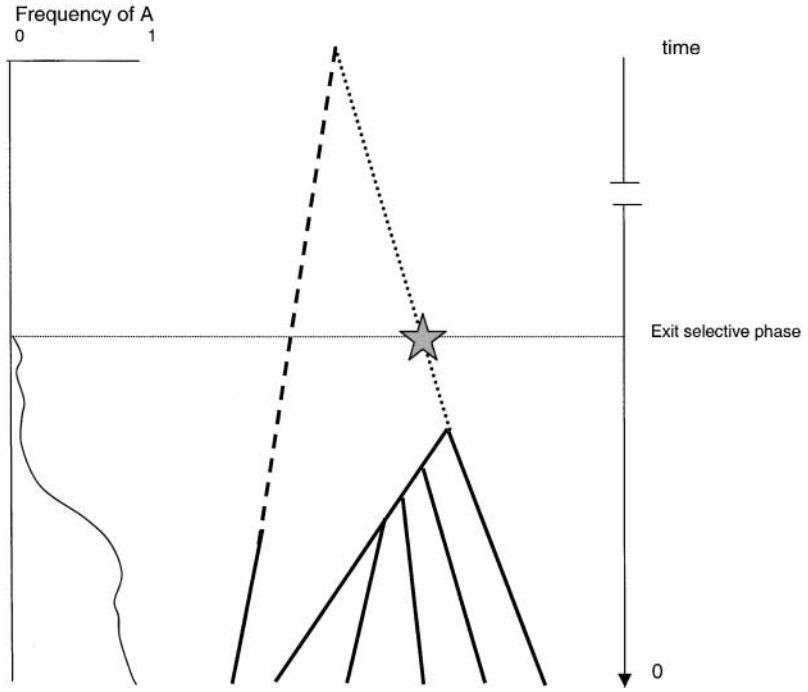

FIGURE 1.—One possible genealogical tree for a sample of six at a neutral site linked to a selected site. The frequency of the favored allele, A, is illustrated on the graph to the left, with time on the *x*-axis. As the frequency of the favored allele decreases, the rate of coalescence increases. However, if one of the neutral lineages (shown as long dashes) recombines onto a nonfavored background (going backward in time), it may have to wait (at least) until after the original mutation from A to a (represented by the gray star), to coalesce with other lineages. Any mutation on the dotted branch will be at high frequency in the sample.

related than they would be at an unlinked neutral site. The genealogy is close to star-shaped, so, as in the case of population growth (TAJIMA 1989b), we expect an excess of rare variants in our sample relative to the standard neutral model.

With recombination, selective sweeps can no longer be treated as population size reductions (BARTON 1998). As we go back in time, the frequency of the favored allele decreases, but the frequency of the unfavored allele increases. One way to think of this is as a subdivided population model, where the two populations are changing size over time (BARTON 1998). Consider the genealogy of a neutral site linked to the selected site. Suppose that a lineage is currently associated with the advantageous allele A, but (going backward in time) recombines onto a chromosome with the unfavored allele, a. For that lineage to coalesce with the other lineages still associated with A, one of two things must happen: Either it must recombine back onto an A background, or we have to wait until after the original mutation from A to a (represented by a star in Figure 1). If the latter, two lineages will be present at the beginning of the sweep, as in Figure 1; their mean time to coalescence is given by the neutral expectation, $2N$. At the neutral site, we will obtain an unbalanced tree that looks like Figure 1 (note that this drawing is not to scale). Any mutation on the dotted line will be at high frequency in our sample. Thus, in the presence of recombination, selective sweeps will produce not only rare variants, but also high-frequency ones (in practice, high- and low-frequency variants can be distinguished by using outgroups to infer which allele is ancestral). While population growth and purifying selection also predict an excess of rare alleles, they do not predict excess high-frequency-derived alleles.

***H* has low power to detect old sweeps:** On the basis of these insights, FAY and WU (2000) constructed a test, *H*, which focuses on the number of high-frequency-derived alleles (see METHODS). They demonstrated that the power of *H* to detect a sweep that ended at time $t = 0$ can be high. Thus, if we consider a "candidate locus" where there is independent evidence for the action of recent positive selection (*e.g.*, TAKAHASHI *et al.* 2001), we can be fairly confident that a significant *H* test is indicative of positive selection. However, this model is unlikely to describe the situation where researchers apply the *H* test to a randomly chosen locus.

Instead, sweeps might be thought of as occurring at random locations and times. In this case, the power of *H* is much reduced. First, the power of *H*, $P(H)$, decreases rapidly with the time since the fixation of the favored allele, as the high-frequency variants fix in the population and no longer contribute to polymorphism (KIM and STEPHAN 2000). For example, in Figure 2, if $N = 10^6$, the power is roughly equal to the nominal rejection probability after 5×10^5 generations or one-eighth of the mean time to coalescence under neutrality, $4N$ ($t = 0.125$ in Figure 2). For *D. melanogaster*, assuming 10 generations a year (and if $N = 10^6$), this corresponds to 5×10^4 years. For some time after the sweep, the power is actually <0.05 (see also KIM and STEPHAN 2001): Of the variation that preexisted the sweep event, all the high-frequency variants have fixed (at least in the sample) so that any remaining alleles are at lower frequency; those that arose after the sweep are young and therefore also at low frequency. As a result, there

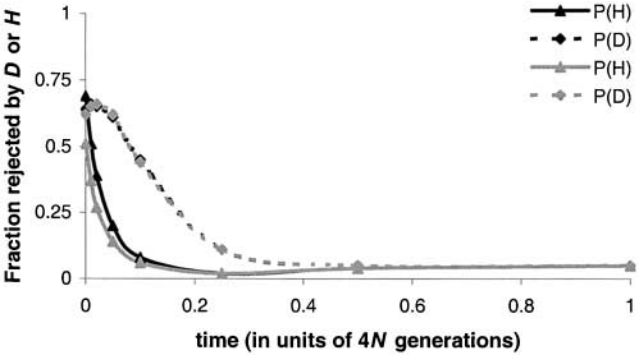

FIGURE 2.—The power of H and D as a function of the time since the fixation of the favored allele, as estimated from 10^4 simulations (see METHODS). Black lines are for an effective population size $N = 10^6$ and a selection coefficient $s = 0.005$ (as in Figure 3 of FAY and WU 2000) and gray lines are for $N = 10^4$ and $s = 0.05$. The sample size is 50, the population mutation rate $\theta = 5$, and the genetic distance to the selected locus, c, is chosen such that $c/s = 0.01$. There is no recombination within the neutral locus. The powers of H (triangles) and D (diamonds) are shown as solid and dashed lines, respectively. The two lines for $P(D)$ are essentially superimposed.

are fewer high-frequency-derived alleles than expected under the null model (for a given number of segregating sites).

The D test retains substantial power for a much longer period of time since the sweep than does H. These results suggest that D might be a better test for detecting selective sweeps. When selection is recent, however, the use of D and H is not redundant. For example, if the parameters are as in Figure 2 and $t = 0$, the proportion of runs where H is significant but D is not is 19% (for D but not H, it is 13%).

The effect of other parameters on $P(H)$: With a larger θ value, there is a higher probability of having a mutation on the dotted branch in Figure 1 and therefore more power to detect the effects of a sweep. For example, immediately after a sweep, $P(H|\theta = 10)$ is 79% (with $N = 10^6$, with other parameters as in Figure 2) while $P(H|\theta = 5)$ is 69%. The power of H also increases with larger sample size (results not shown).

Of fundamental importance in determining $P(H)$ is the number of lineages that recombine on to the unfavored background during the sweep. As can be seen in Figure 1, for the ancestral genealogy to have long internal branches requires at least one recombination event between selected classes. How likely this is depends on the strength of selection and on the recombination rate between the selected and neutral loci (c). If c is too small, there will be no recombination events, and all lineages will coalesce during the sweep. If c is very large, there will be many recombination events, and the neutral locus will not reflect the effects of selection. Thus, if the neutral locus is very close to the sweep, or too far away, $P(H)$ is substantially reduced (Figure 3 in FAY and WU 2000; results not shown).

The power of H depends on s and c, not just on their ratio. Keeping c/s constant does not produce the same number of recombinants for different sets of (c, s) values, because the total length of the tree (and hence the probability of a recombination event) does not depend linearly on s. In fact, for the same c/s value, stronger selection (and therefore larger c values) will result in higher $P(H)$. As an illustration, if $N = 10^4$, as might be the case for humans (LI and SADLER 1991), $c/s = 0.01$, and $s = 0.005$, then immediately after a sweep, $P(H)$ is only 10% while $P(D)$ is 58%. For the same c/s value, if $s = 0.05$, $P(H)$ is 51% and $P(D)$ is 62% (Figure 2).

The power of H in practice: Researchers have assessed the significance of the H test with critical values established under the assumptions of a constant population size and no recombination. In reality, however, there is recombination within the neutral locus. In the presence of recombination, the use of critical values for the case of no recombination is conservative; *i.e.*, the null model is rejected <5% of the time at the 5% level. This can be seen by comparing the $P(H|\text{no sweep})$ in Table 1 for different values of ρ, the population recombination rate for the neutral locus. Even though the H test is

TABLE 1

The power of H and D as a function of the time since the sweep ended

			$t = 0$	0.02	0.05	0.10	0.25	0.50	1.00	No sweep
$N = 10^6$, $s = 0.005$	$\rho = 0$	$P(H)$	0.69	0.39	0.19	0.08	0.02	0.04	0.05	0.05
		$P(D)$	0.65	0.65	0.62	0.45	0.12	0.05	0.05	0.05
	$\rho = 20$	$P(H)$	0.76	0.48	0.23	0.08	0.01	$<10^{-2}$	$<10^{-2}$	0.01
		$P(D)$	0.69	0.64	0.59	0.44	0.14	0.03	0.01	0.01
$N = 10^4$, $s = 0.05$	$\rho = 0$	$P(H)$	0.51	0.27	0.14	0.06	0.02	0.04	0.05	0.05
		$P(D)$	0.62	0.66	0.62	0.44	0.11	0.05	0.05	0.05
	$\rho = 5$	$P(H)$	0.57	0.32	0.16	0.07	0.01	$<10^{-2}$	0.02	0.03
		$P(D)$	0.63	0.65	0.59	0.45	0.14	0.04	0.03	0.03

The time t since the fixation of the beneficial mutation is scaled in units of $4N$ generations, where N is the effective population size, ρ is the population recombination rate for the neutral locus, and s is the selection coefficient of the favored allele. The sample size is 50 and the population mutation rate at the neutral locus, θ, is 5. A total of 10^4 simulations were run for each set of parameters.

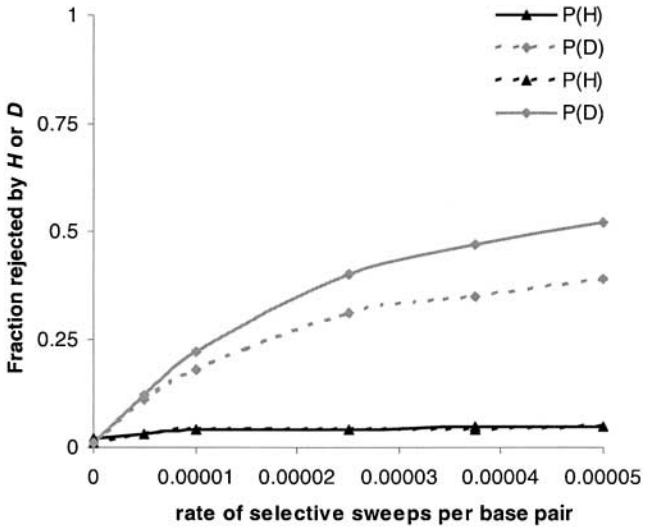

FIGURE 3.—The power of H and D (one-tailed) to detect repeated selective sweeps, as estimated from 10^4 simulations (see METHODS). The effective population size is $N = 10^6$ and the selection coefficient $s = 0.01$. The sample size is 50 chromosomes. The population recombination rate for the neutral locus, ρ, is 20. On the x-axis is the expected number of selective sweeps per base pair per $4N$ generations, assuming a recombination rate of 5×10^{-9}/bp/generation. Dashed lines are for a population mutation rate $\theta = 5$ and solid ones are for $\theta = 10$. The two lines for $P(H)$ are essentially superimposed.

conservative in the presence of intralocus recombination, some recombination increases the power to detect a sweep at a linked site. (Obviously this is true only up to a point: If there is a very high level of recombination, the neutral locus will no longer reflect selection at linked sites.) As can be seen in Table 1, the increase in power is slight, and $P(H)$ still decreases extremely quickly with t.

In humans, the violation of a second assumption will lead one to overestimate the power of H to detect a sweep. The human population size has increased dramatically in the recent past. The effect of population growth is to increase the rate of coalescences going backward in time. For the same average diversity levels, the tree in Figure 1 would therefore have shorter internal branches than it does under a constant-size model. This will reduce the number of high-frequency-derived alleles found at neutral sites linked to a selective sweep. Thus, the finding of numerous loci with extreme H values is even more surprising when this aspect of human demography is taken into account.

The power to detect sweeps at a randomly chosen locus: Results for the recurrent selective sweep model are shown in Figure 3. There is essentially no power to detect the effects of selection using H and the power does not increase with the strength of selection or the frequency of selective sweeps. This is to be expected: The power of H is high for very recent sweeps at a suitable distance from the neutral site. Simulations suggest that, if $N = 10^6$, $s = 0.005$, and the sample size is 50, the maximum distance at which sweeps have an effect on diversity levels is $c/s \approx 0.25$ (results not shown). For these parameters, $P(H) > 20\%$ for a distance between $0.00035 < c/s < 0.02$ (Figure 3 in FAY and WU 2000). If sweeps occur at a distance chosen uniformly such that c/s is between 0 and 0.25, 8% of sweeps will be within the relevant range. In addition, the beneficial allele will have fixed at some random time in the past, $t \gg 0$, and the power of H decreases with increasing t. In contrast to H, the power of D increases with both s and the rate of sweeps.

The effect of a single sweep on LD: As shown above, a significant H value is a short-lived signature of a selec-

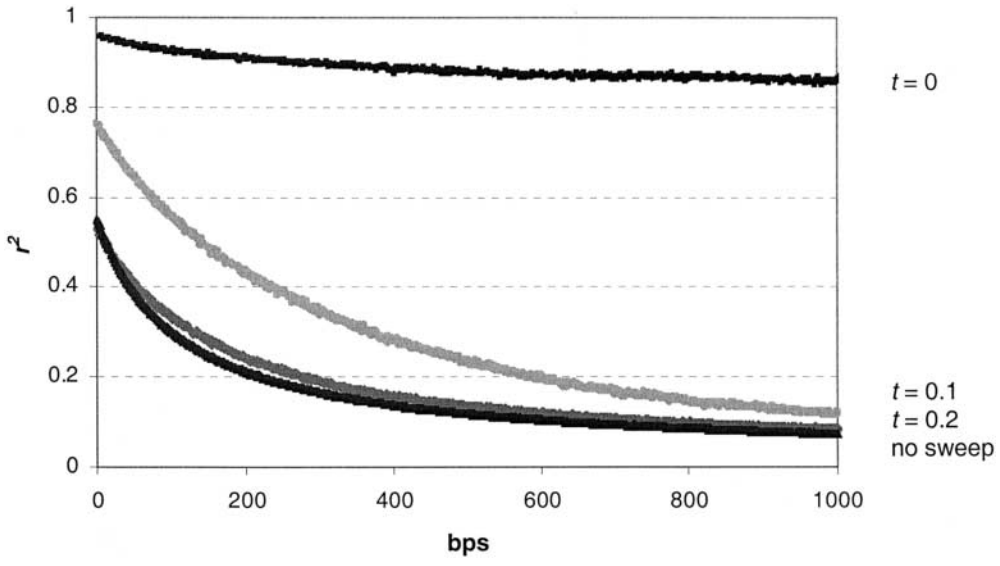

FIGURE 4.—The effect of selective sweeps on the expected decay of pairwise linkage disequilibrium. The effective population size is $N = 10^6$, the selection coefficient $s = 0.01$, the population mutation rate $\theta = 40$, and the sample size is 50. The population recombination rate for the neutral locus, ρ, is 20 (which corresponds to 1 kb for a recombination rate of 5×10^{-9}/bp/generation). The genetic distance to the sweep, c, is chosen so that $c/s = 0.005$. The time since the fixation of the favored allele, t, is scaled in units of $4N$ generations. A total of 10^4 simulations were run for each value of t. Only segregating sites with a minor allele frequency ≥0.1 are included.

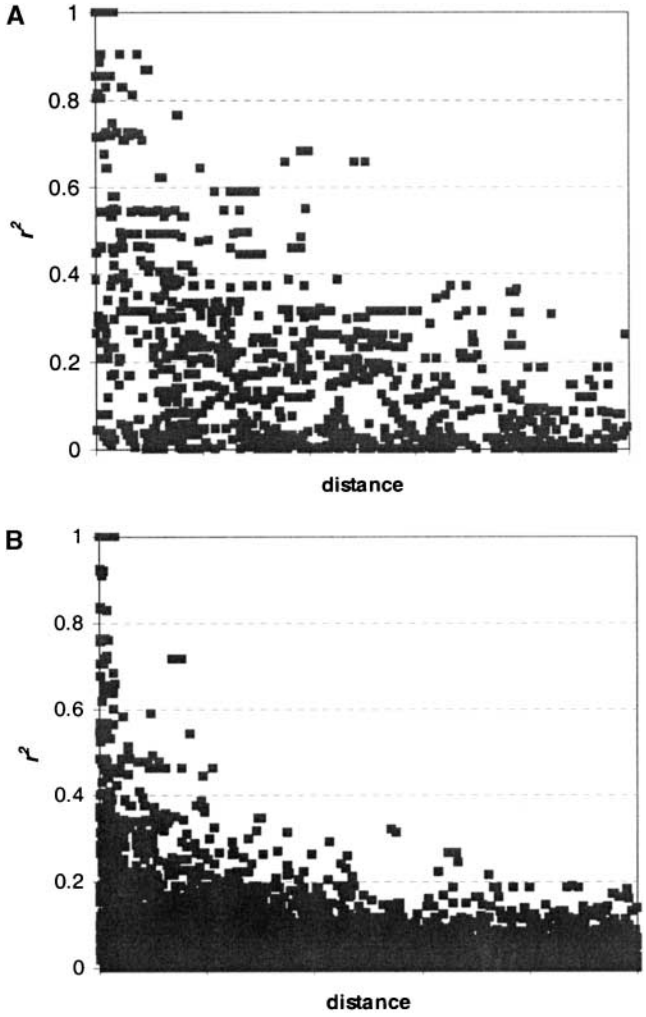

FIGURE 5.—An illustration of the effect of a selective sweep on a neutral locus: a scatterplot of r^2 for one simulated data set. Only segregating sites with a minor allele frequency ≥0.1 are included. The effective population size $N = 10^4$, the selection coefficient $s = 0.05$, and the population mutation rate $\theta = 40$. The population recombination rate for the neutral locus, ρ, is 200 (which corresponds to 1 Mb for a recombination rate of 0.5 cM/Mb/generation). The sweep occurs immediately adjacent to the neutral locus. The sample size is 50, so points >0.0768 are in significant linkage disequilibrium by a χ^2 test (*cf.* PRITCHARD and PRZEWORSKI 2001). (A) The beneficial allele fixed at time $t = 0$. (B) No sweep.

tive sweep. This is also true of another feature of the data, levels of linkage disequilibrium. In both Drosophila and humans, numerous loci appear to exhibit unexpectedly high levels of LD. In Drosophila, this is usually quantified as a paucity of haplotypes (*e.g.*, PARSCH *et al.* 2001; further references in ANDOLFATTO 2001) or a lower than expected estimate of the population recombination rate, ρ (ANDOLFATTO and PRZEWORSKI 2000; WALL 2001). In particular, in *D. melanogaster* and *D. simulans*, it appears that one estimate of ρ, C_{hud} (HUDSON 1987), is systematically lower than would be expected from independent estimates of the mutation and recombination rates. In humans, it is the distance over which LD extends in many regions that is unusual (*e.g.*, RIEDER *et al.* 1999; GILAD *et al.* 2001; reviewed in PRITCHARD and PRZEWORSKI 2001). For a couple of regions, ρ has also been shown to be lower than expected for European samples (PRITCHARD and PRZEWORSKI 2001). These patterns have not yet been explained.

As is illustrated in Figures 4 and 5, a recent sweep can substantially increase levels of LD. In Figure 4, I plot the expected decay of a summary of pairwise LD, r^2, for alleles with a minor allele frequency ≥0.1. Parameters are chosen to be plausible for *D. melanogaster*. If the beneficial allele fixed at time $t = 0$, there is a much slower rate of decay with distance than under the standard neutral model. Note, however, that fewer alleles satisfy the frequency cutoff after a sweep, so long sequences may be required for this pattern to be apparent in actual data. Figure 5 presents scatterplots of r^2 *vs.* distance for parameters germane to humans; as can be seen, a selected substitution at a linked site increases the number of distant pairs in significant LD.

The effect of a sweep on levels of LD dissipates quickly, depending on the summary of LD used and particularly on the sensitivity of the measure to changes in allele frequencies. Consider first the effect of a single sweep on the mean number of haplotypes normalized by the number of segregating sites, $E(n_{\text{Haps}}/(S+1))$. As can be seen in Table 2, a neutral locus affected by a very recent sweep can exhibit a paucity of haplotypes relative to a standard neutral model (depending on the values of s and c). This suggests an increase in LD. However, the summary $E(n_{\text{Haps}}/(S+1))$ becomes *greater* than expected under neutrality shortly after the sweep (see Table 2). This is easily understood: As the high-frequency variants fix and new mutations arise, most alleles are now rare and many form new haplotypes.

When only intermediate-frequency variants are considered, the effect of selective sweeps on allelic associations is clearer. In the last two rows of Table 2, I report $E(n_{\text{Haps}}/(S+1))$ *excluding singletons*. This statistic loosely corresponds to what is sometimes referred to as "haplotype structure" in the literature (*e.g.*, PARSCH *et al.* 2001). The ratio is sharply decreased by a sweep and monotonically increases to the neutral expectation with increasing time since the sweep. These results suggest that this statistic might be useful for detecting positive selection. Nonetheless, the effect of the selective sweep has all but vanished by $t = 0.1$, unless selection is very strong (*e.g.*, $Ns = 5 \times 10^3$). Pairwise linkage disequilibrium exhibits a similar behavior to the number of haplotypes: For example, in Figure 4, a sweep that ended at $t = 0.2$ has an undetectable effect on r^2. For these parameters, there is still a relative excess of LD by $t = 0.1$; however, this would be hard to discern in any one data set, because r^2 varies greatly from one locus to another under neutrality (PRITCHARD and PRZEWORSKI 2001).

One implication of these results is that selection would have to be strong and recent for selective sweeps

TABLE 2

The effect of a selective sweep on the mean $n_{Haps}/(S + 1)$

t	0	0.01	0.05	0.10	0.25	0.50	1.00	No sweep
$N = 10^6$, $s = 0.005$	(0.44)	(0.60)	(0.82)	0.89	0.98	0.99	0.93	0.89
$N = 10^4$, $s = 0.01$	0.76	0.79	0.83	0.85	0.83	0.76	0.70	0.67
$N = 10^4$, $s = 0.05$	(0.58)	0.69	0.81	0.84	0.84	0.77	0.70	0.67
$N = 10^6$, $s = 0.005^a$	(0.53)	(0.57)	(0.81)	(0.93)	(1.10)	(1.19)	(1.23)	1.25
$N = 10^4$, $s = 0.01^a$	(0.70)	(0.76)	(0.84)	(0.88)	(0.90)	(0.90)	0.93	0.93

The time since the fixation of the favored allele, t, is scaled in units of $4N$ generations, where N is the effective population size. n_{Haps} is the number of distinct haplotypes and S is the number of segregating sites. The sample size is 50, the population mutation rate for the neutral locus, θ, is 5, and the genetic distance to the selected locus, c, is chosen such that $c/s = 0.01$ (where s is the selection coefficient of the favored allele). In simulations where $N = 10^6$, the population recombination rate for the neutral locus, ρ, is 20 (corresponding to 1 kb if the recombination rate is 5×10^{-9}/bp/generation); where $N = 10^4$, ρ is 5 (corresponding to ~25 kb if the recombination rate is 0.5 cM/Mb/generation). In parentheses are those entries for which $E(n_{Haps}/(S + 1))$ is less than the neutral expectation.

[a] $n_{Haps}/(S + 1)$ is calculated excluding singleton mutations.

to account for the unexpectedly large distances over which LD sometimes extends in humans. This said, recent evidence suggests that most crossing-over events in humans may occur within narrow recombination hotspots, with most of the genome experiencing very low rates of crossing over (*e.g.*, Jeffreys *et al.* 2001). If so, "recombination coldspots" may preserve allelic associations longer than suggested by these simulations.

The effect of repeated sweeps on LD: Because the increase in LD is short-lived, anonymous loci subject to repeated selective sweeps do not show a marked excess of LD. In fact, summaries of LD that are highly sensitive to the frequency spectrum, such as C_{hud} or $E(n_{Haps}/(S + 1))$, suggest *less* LD under this model of recurrent sweeps than under neutrality. C_{hud}, in particular, is smaller when the sample variance in the number of pairwise differences is larger. Selective sweeps skew the frequency spectrum toward rare alleles, leading to a smaller variance in pairwise differences and larger values of C_{hud} (results not shown). Thus, repeated sweeps cannot account for the low values of C_{hud} found at most loci in both species of Drosophila (Andolfatto and Przeworski 2000), at least as modeled.

Repeated sweeps do produce a relative excess of LD when attention is restricted to intermediate frequency variants. For example, in 10^4 simulations, $E(n_{Haps}/(S + 1))$ excluding singletons is 1.24 in the absence of sweeps, 1.05 for $\lambda = 10^{-5}$, and 0.90 for $\lambda = 5 \times 10^{-5}$ (λ is the rate of sweep per base pair per $4N$ generations). Figure 6 plots the expected decay of r^2 with distance for these two rates of sweeps, with the other parameter values chosen to be plausible for *D. melanogaster*. The increase relative to a neutral model is slight. Note further that the rate $\lambda = 5 \times 10^{-5}$ is probably unrealistically high. For $s = 0.01$, and assuming a fixation probability of $2s$ (*cf.* Crow and Kimura 1970, p. 426), roughly one in every

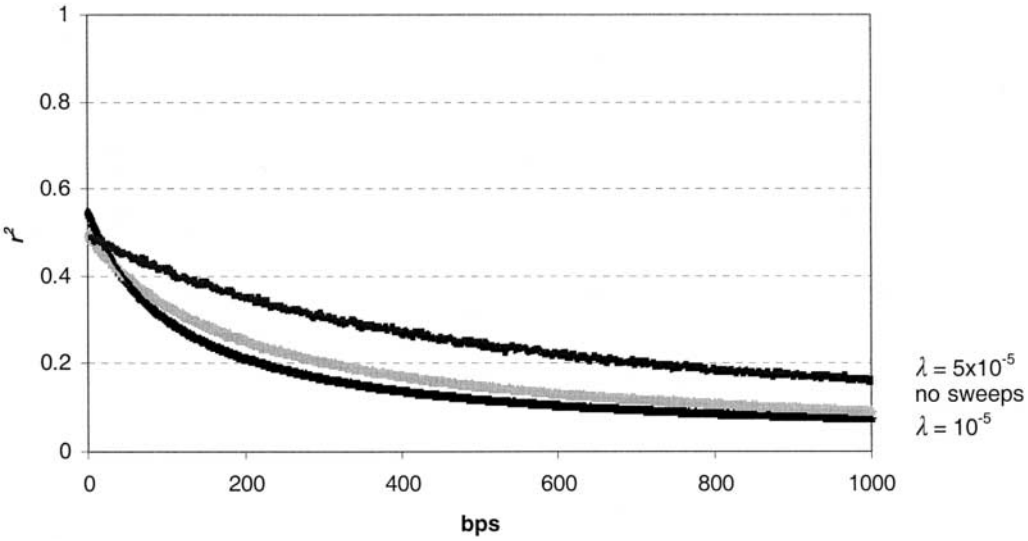

Figure 6.—The effect of repeated selective sweeps on the expected rate of decay of pairwise linkage disequilibrium. The effective population size $N = 10^6$, the selection coefficient $s = 0.01$, the population mutation rate $\theta = 40$, the population recombination rate $\rho = 20$, and the sample size is 50. The neutral locus is affected by repeated sweeps occurring at rate λ/bp/$4N$ generations (assuming a recombination rate of 5×10^{-9}/bp/generation).

TABLE 3

The power of *H* and *D* to detect a symmetric two-island model

	Panmixia	$4Nm = 1$ (sampled 48/2)	$4Nm = 1$ (sampled 50/0)	$4Nm = 0.5$ (sampled 50/0)	$4Nm = 2$ (sampled 50/0)
$P(H)$	0.05	0.14	0.14	0.19	0.09
$P(D)$	0.05	0.09	0.06	0.09	0.05

The power of *H* and *D* is estimated from 10^4 simulations, as described in METHODS. $4Nm$ is the number of migrants per deme per generation. The sample size is 50. The population mutation rate per deme is 2.5. There is no intralocus recombination.

three newly arising mutations would have to be advantageous to obtain this rate of selective sweeps (if the neutral mutation rate is taken to be 2×10^{-9}/generation/bp; McVEAN and VIEIRA 2001). Thus, for plausible parameters, the decay of LD is barely less steep than under a neutral model. Randomly chosen loci are therefore not expected to show strikingly high levels of LD, even if there have been multiple selective sweeps at linked sites.

DISCUSSION

The possible effect of population structure: If old or recurrent sweeps lead neither to high levels of LD nor to significant *H* tests, how do we interpret these features of the data? One possibility is that they were produced by a demographic departure from model assumptions. To examine this, I estimated the power of *H* (implemented as described for the sweep models) to detect a symmetric island model (WRIGHT 1951) when samples were drawn unequally from the different demes. In all cases reported here, θ for the whole population is 5, so for *k* demes, it is θ/k per deme. First, I consider a two-island model, each of size $N/2$, with 0.5–2 migrants per deme per generation; under this particular model, this migration rate corresponds to an F_{ST} value of ~0.11–0.33 (HUDSON *et al.* 1992). As can be seen in Table 3, if samples are drawn very unequally (*e.g.*, 48 and 2), we would reject the neutral model >5% of the time (at the 5% level) using *H*, even in the absence of selection. Even if samples are collected from only one locality, $P(H) > 5\%$, as the samples sometimes contain individuals whose ancestors were migrants from other demes. If levels of differentiation are higher (*e.g.*, $F_{ST} = 0.33$, corresponding to 0.5 migrant per deme per generation in a two-island model), $P(H)$ can be as high as 19%. If there are more than two islands, then, for approximately the same F_{ST} value, the power is similar (results not shown). In general, the power of *H* to detect population structure increases with higher θ or lower migration rates (results not shown). In summary, the null model can be rejected by the *H* test at substantially higher than the nominal rejection probability when samples are drawn unequally from different islands in an island model. In addition, population structure can produce high levels of LD (LI and NEI 1974; WALL 1999).

This particular model is likely to be unrealistic for both Drosophila and humans. However, the purpose of these simulations is simply to illustrate that a demographic model that produces trees such as Figure 1 more often than the standard neutral model will have the same effect on *H* as a selective sweep. In fact, recent bottlenecks (results not shown) and a metapopulation model (WAKELEY and ALICAR 2001) can also lead to high-frequency-derived alleles more often than expected under the standard neutral model. In other words, such alleles are not a unique signature of positive selection. In addition, in humans, most of the regions with a significant *H* test are noncoding, so there may be good reasons to search for demographic rather than selective explanations. It remains to be seen whether a more realistic model of demography can also produce extreme *H* values and levels of LD as high as are observed. One model worth investigating might be ancient structure, with unequal contributions of different subpopulations to the current gene pool.

Does selection operate as modeled? An alternative to demographic explanations is that positive selection does not operate as is commonly modeled. One assumption made by this model of recurrent positive selection is that a neutral locus is affected by at most one selected substitution at a time. The validity of this assumption depends crucially on the rate at which advantageous mutations arise and sweep to fixation. NACHMAN (2001) and ANDOLFATTO (2001) have estimated the rate of selective sweeps needed to account for the positive correlation between diversity levels and crossing-over rates observed in humans and in Drosophila, respectively. The probability of overlap can be estimated from Equation 6 in BRAVERMAN *et al.* (1995). On the basis of these rough calculations, it appears that in both species, selective sweeps will often occur concurrently (results not shown).

When two or more alleles are simultaneously favored, interference between them might alter the patterns of polymorphism relative to the predictions of a single-site model of positive selection (KIRBY and STEPHAN 1996). However, the selected sites would have to be very close

to one another on the chromosome for interference to have an effect. If the locations of the selected substitutions are chosen uniformly, as in this model, this condition is unlikely to be met. Under an alternative model, where several adaptive changes occur in a small region in short succession, interference between sweeps may be more likely. It is unknown whether such a scenario would lead to higher levels of LD or more high-frequency-derived alleles. Even so, the effects are likely to be short-lived, as recombination will rapidly break down allelic associations after the sweeps, and high-frequency alleles will drift to fixation. Thus, occasional overlaps are unlikely to explain the observed patterns.

More problematic is the assumption that the rate of selective sweeps is constant. If, instead, there has been an increase in the rate of genetic adaptations toward the present, many loci may reflect recent sweeps. In the case of cosmopolitan species of Drosophila, this time frame could reflect recent colonization of temperate habitats. Similarly, anatomically modern humans are thought to have left Africa and spread across the globe starting ~50 thousand years ago, and there have been major changes in population density over the past 10 kya (Jones *et al.* 1994). The emergence of modern humans and their spread through the world may have coincided with a burst of genetic adaptations.

Note further that the sojourn time of a selected allele in a random-mating population of constant size is $\sim 2\ln(2N)/s$ (assuming that the allele was selected when first introduced), where N is the diploid effective population size and s the selection coefficient of the favored allele (*cf.* Stephan *et al.* 1992). With the N values assumed throughout and a selection coefficient of 1%, this translates into $\approx 2 \times 10^3$ generations for humans and 2.9×10^3 generations for Drosophila (respectively, 4×10^4 years assuming 20 years per generation and 300 years assuming 10 generations a year). The demographic assumptions behind this calculation are likely to be invalid for the recent past of many cosmopolitan species. However, they suggest that if there has been an increase in the rate of sweeps in the recent past, a subset of loci may reflect incomplete sweeps—ones that are still ongoing or where the selected variant is no longer favored.

An additional assumption of this sweep model that is likely to be untrue in both *D. melanogaster* and humans is that of random mating. Indeed, there is evidence for population structure in both *D. melanogaster* (*e.g.*, Hale and Singh 1991; Begun and Aquadro 1993) and humans (*e.g.*, Cavalli-Sforza *et al.* 1994) as well as for geographic differences in selective pressures at particular loci (reviewed in Andolfatto 2001). Departures from random mating could distort the signature of selection relative to our expectations for a panmictic population, resulting in high levels of LD and, perhaps, in the maintenance of high-frequency-derived alleles.

In summary, the H test is a useful tool to confirm with polymorphism data that a candidate locus has undergone a recent sweep (*e.g.*, Parsch *et al.* 2001; Takahashi *et al.* 2001). However, it has low power to detect the effects of positive selection at a randomly chosen locus. In addition, it may not be conservative if there is hidden population structure. Similarly, while sweeps increase LD between intermediate frequency variants, the effect is short-lived. Thus, randomly chosen data sets with significant H values and high levels of LD may reflect demography rather than adaptation. Alternatively, positive selection may not operate as it is most commonly modeled.

I thank P. Andolfatto, A. Di Rienzo, P. Donnelly, J. Fay, I. Gordo, R. Griffiths, J. Pritchard, and J. Wall for helpful discussions and P. Andolfatto, Y. Gilad, R. Hudson, G. McVean, and J. Wall as well as D. Charlesworth and two anonymous reviewers for comments on the manuscript. M.P. is supported by a National Science Foundation Bioinformatics postdoctoral fellowship.

LITERATURE CITED

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

Andolfatto, P., and M. Przeworski, 2000 A genome-wide departure from the standard neutral model in natural populations of Drosophila. Genetics **155:** 257–268.

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

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*, edited by B. Golding. Chapman & Hall, New York.

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

Begun, D. J., and C. F. Aquadro, 1993 African and North American populations of *Drosophila melanogaster* are very different at the DNA level. Nature **365:** 548–550.

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.

Cavalli-Sforza, L. L., P. Manozzi and A. Piazza, 1994 *The History and Geography of Human Genes.* Princeton University Press, Princeton, NJ.

Crow, J. F., and M. Kimura, 1970 *An Introduction to Population Genetics Theory.* Alpha Editions, Edina, MN.

Fay, J. C., and C.-I Wu, 2000 Hitchhiking under positive Darwinian selection. Genetics **155:** 1405–1413.

Fay, J. C., and C.-I Wu, 2001 The neutral theory in the genomic era. Curr. Opin. Genet. Dev. **11:** 642–646.

Frisse, L., R. R. Hudson, A. Bartoszewicz, J. D. Wall, J. Donfack *et al.*, 2001 Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. Am. J. Hum. Genet. **69:** 831–843.

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

Gilad, Y., D. Segre, K. Skorecki, M. Nachman, D. Lancet *et al.*, 2000 Dichotomy of single-nucleotide polymorphism haplotypes in olfactory receptor genes and pseudogenes. Nat. Genet. **26:** 221–224.

Gilad, Y., S. Rosenberg, M. Przeworski, D. Lancet and K. Skorecki, 2001 Evidence for positive selection and population structure at the human MAO-A gene. Proc. Natl. Acad. Sci. USA **99:** 862–867.

Hale, L. R., and R. S. Singh, 1991 A comprehensive study of genic variation in natural populations of *Drosophila melanogaster.* IV. Mitochondrial DNA variation and the role of history *vs.* selection in the genetic structure of geographic populations. Genetics **129:** 103–117.

Hamblin, M. T., E. E. Thompson and A. Di Rienzo, 2002 Complex

signatures of natural selection at the duffy blood group locus. Am. J. Hum. Genet. **70:** 369–383.

Hudson, R. R., 1987 Estimating the recombination parameter of a finite population model without selection. Genet. Res. **50:** 245–250.

Hudson, R. R., 1993 The how and why of generating gene genealogies, pp. 23–36 in *Mechanisms of Molecular Evolution*, edited by N. Takahata and A. G. Clark. Japan Scientific Society, Tokyo.

Hudson, R. R., M. Slatkin and W. P. Maddison, 1992 Estimation of levels of gene flow from DNA sequence data. Genetics **132:** 583–589.

Jeffreys, A. J., L. Kauppi and R. Neumann, 2001 Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nat. Genet. **29:** 217–222.

Jones, S., R. Martin and D. Pilbeam (Editors), 1994 *The Cambridge Encyclopedia of Human Evolution*. Cambridge University Press, Cambridge.

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

Kim, Y., and W. Stephan, 2000 Joint effects of genetic hitchhiking and background selection on neutral variation. Genetics **155:** 1415–1427.

Kim, Y., and W. Stephan, 2001 Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics **160:** 765–777.

Kirby, D. A., and W. Stephan, 1996 Multi-locus selection and the structure of variation in a segment of the *white* gene of *Drosophila melanogaster*. Genetics **144:** 635–645.

Lazzaro, B. P., and A. G. Clark, 2001 Evidence for recurrent paralogous gene conversion and exceptional allelic divergence in the Attacin genes of *Drosophila melanogaster*. Genetics **159:** 659–671.

Li, W. H., and M. Nei, 1974 Stable linkage disequilibrium without epistasis in subdivided populations. Theor. Popul. Biol. **6:** 173–183.

Li, W. H., and L. Sadler, 1991 Low nucleotide diversity in man. Genetics **129:** 513–523.

Martinez-Arias, R., F. Calafell, E. Mateu, D. Comas, A. Andres *et al.*, 2001 Sequence variability of a human pseudogene. Genome Res. **11:** 1071–1085.

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

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

Nachman, M. W., and S. L. Crowell, 2000 Contrasting evolutionary histories of two introns of the duchenne muscular dystrophy gene, *Dmd*, in humans. Genetics **155:** 1855–1864.

Otto, S. P., 2000 Detecting the form of selection from DNA sequence data. Trends Genet. **16:** 526–529.

Parsch, J., C. D. Meiklejohn and D. Hartl, 2001 Patterns of DNA sequence variation suggest the recent action of positive selection in the janus-ocnus region of *Drosophila simulans*. Genetics **159:** 647–657.

Pritchard, J. K., and M. Przeworski, 2001 Linkage disequilibrium in humans: models and data. Am. J. Hum. Genet. **69:** 1–14.

Rieder, M. J., S. L. Taylor, A. G. Clark and D. A. Nickerson, 1999 Sequence variation in the human angiotensin converting enzyme. Nat. Genet. **22:** 59–62.

Stephan, W., T. H. E. Wiehe and M. Lenz, 1992 The effect of strongly selected substitutions on neutral polymorphism: analytic results based on diffusion theory. Theor. Popul. Biol. **41:** 237–254.

Taillon-Miller, P., I. Bauer-Sardina, N. L. Saccone, J. Putzel, T. Laitinen *et al.*, 2000 Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28. Nat. Genet. **25:** 324–328.

Tajima, F., 1983 Evolutionary relationships of DNA sequences in finite populations. Genetics **105:** 437–460.

Tajima, F., 1989a Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585–595.

Tajima, F., 1989b The effect of change in population size on DNA polymorphism. Genetics **123:** 597–601.

Takahashi, A., S. C. Tsaur, J. A. Coyne and C.-I Wu, 2001 The nucleotide changes governing cuticular hydrocarbon variation and their evolution in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **98:** 3920–3925.

Wakeley, J., and N. Alicar, 2001 Gene genealogies in a metapopulation. Genetics **159:** 893–905.

Wall, J. D., 1999 Recombination and the power of statistical tests of neutrality. Genet. Res. **73:** 65–79.

Wall, J. D., 2001 Insights from linked single nucleotide polymorphisms: what we can learn from linkage disequilibrium. Curr. Opin. Genet. Dev. **11:** 647–651.

Wall, J. D., and R. R. Hudson, 2001 Coalescent simulations and statistical tests of neutrality. Mol. Biol. Evol. **18:** 1134–1135.

Weir, B. S., 1996 *Genetic Data Analysis II*. Sinauer Associates, Sunderland, MA.

Wright, S., 1951 The genetical structure of populations. Ann. Eugen. **15:** 323–354.

Communicating editor: D. Charlesworth