DOI: 10.1534/genetics.103.025387

Linkage Disequilibrium as a Signature of Selective Sweeps

Yuseob Kim[1] and Rasmus Nielsen

Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York 14853

Manuscript received December 5, 2003
Accepted for publication April 29, 2004

ABSTRACT

The hitchhiking effect of a beneficial mutation, or a selective sweep, generates a unique distribution of allele frequencies and spatial distribution of polymorphic sites. A composite-likelihood test was previously designed to detect these signatures of a selective sweep, solely on the basis of the spatial distribution and marginal allele frequencies of polymorphisms. As an excess of linkage disequilibrium (LD) is also known to be a strong signature of a selective sweep, we investigate how much statistical power is increased by the inclusion of information regarding LD. The expected pattern of LD is predicted by a genealogical approach. Both theory and simulation suggest that strong LD is generated in narrow regions at both sides of the location of beneficial mutation. However, a lack of LD is expected across the two sides. We explore various ways to detect this signature of selective sweeps by statistical tests. A new composite-likelihood method is proposed to incorporate information regarding LD. This method enables us to detect selective sweeps and estimate the parameters of the selection model better than the previous composite-likelihood method that does not take LD into account. However, the improvement made by including LD is rather small, suggesting that most of the relevant information regarding selective sweeps is captured by the spatial distribution and marginal allele frequencies of polymorphisms.

THE amount and pattern of genetic variation in a population is influenced by the evolutionary history of the population. Population genetics theory provides various tools by which important events in the past can be inferred from polymorphism data. One evolutionary process of great interest is the occurrence of a beneficial mutation and its fixation in the population, which is the basis of adaptation and phenotypic evolution of organisms. However, the rate, strength, and genomic distribution of beneficial mutations in natural populations are not well known, although there recently has been some progress in this area (Stephan 1995; Fay *et al.* 2001; Bustamante *et al.* 2002; Smith and Eyre-Walker 2002; Nielsen and Yang 2003; Piganeau and Eyre-Walker 2003). Identifying particular loci where a recent fixation of beneficial mutation occurred has been subject to much recent scientific interest (Harr *et al.* 2002; Kim and Stephan 2002; Vigouroux *et al.* 2002).

Because beneficial mutations occur very infrequently such that it is practically impossible to observe an ongoing substitution, investigation on them heavily depends on finding "footprints" of positive selection that occurred in the past. Recent theoretical investigations showed that a position on a chromosome where a recent fixation of a beneficial mutation occurred can be identified from a sample of DNA sequences (Maynard Smith and Haigh 1974; Fay and Wu 2000; Kim and Stephan 2002; Przeworski 2002, 2003). A local reduction of genetic variation, commonly called a "selective sweep," is caused by the rapid fixation of a beneficial mutation. Selective sweeps cause drastic changes in (i) the spatial distribution of polymorphic sites, (ii) the frequency spectrum, and (iii) linkage disequilibrium (LD) around the site where the substitution of a beneficial allele took place. A statistical method for detecting selective sweeps based on the information in i and ii was devised by Kim and Stephan (2002). They constructed a composite-likelihood function by multiplying together the likelihood functions from individual sites and showed that reasonably good estimates of the strength and the location of directional selection are obtained by this method. However, this method ignores information from allelic association among polymorphic sites. A number of theoretical and empirical studies have shown that LD is an important signature of selective sweeps (Parsch *et al.* 2001; Kim and Stephan 2002; Przeworski 2002; Sabeti *et al.* 2002; Wootton *et al.* 2002). Therefore, it was hypothesized that a large amount of information has been lost by not taking LD into account in the calculation of the composite likelihood.

In this study, we propose a new composite-likelihood method for detecting signatures of selection that incorporates information from measures of linkage disequilibrium. We also discuss how a unique pattern of LD in DNA sequence data is generated by selective sweep and examine the relative importance of LD compared to other signatures of a selective sweep, *i.e.*, the spatial

[1]*Corresponding author:* Department of Biology, Hutchison Hall, University of Rochester, Rochester, NY 14627.
E-mail: ykim@mail.rochester.edu

distribution of polymorphic sites and the frequency spectrum. Przeworski (2002) investigated the effect of selective sweeps, both recurrent and nonrecurrent, on LD at a neutral locus partially linked to the site of selection. This article focuses on a slightly different situation: A very recent fixation of a beneficial allele occurs in the middle of the region investigated and no additional selection in the past is assumed. This study is also distinguished from the investigation of LD caused by an incomplete sweep of a beneficial mutation in a population (Parsch *et al.* 2001; Sabeti *et al.* 2002; Quesada *et al.* 2003), which leaves a very distinct pattern of genetic variation that is relatively easy to detect.

PATTERNS OF LD REVEALED BY SIMPLE SUMMARY STATISTICS

In this section we briefly examine the patterns of LD caused by selective sweep using simulations and simple summary statistics. Simulations under the model of selective sweep and that of neutral equilibrium were performed following the method of Kim and Stephan (2002). Briefly, an ancestral history of n chromosomes, each of which is 10 kb long, was constructed by a coalescent with recombination algorithm backward in time ($T_{\text{limit}} = 7.0$). The scaled recombination rate over the chromosome is given by $R = 4NL\rho$, where L ($= 10^4$) is the length of the chromosome and ρ is the recombination rate per base per generation. Mutations are mapped on marginal trees (coalescent trees corresponding to individual nucleotide sites) with a fixed parameter $\theta = 4N\mu$, where N is the diploid population size and μ is the mutation rate per generation per nucleotide site. If multiple mutations occur on a marginal tree, they are not distinguished from each other in the sample but grouped as derived alleles. For the selective sweep simulations, the strength of directional selection is given by $\alpha = 2Ns$, where s is the selection coefficient. The length of time between the sampling (present) and the fixation of the beneficial allele is given by τ, which is measured in units of $2N$ generations. LD in the simulated data was measured by either K, the number of distinct haplotypes formed by a given number of consecutive polymorphic sites (Depaulis and Veuille 1998), or r^2 (Hill and Robertson 1968), a common measure of LD between two polymorphic sites.

First, we examine the spatial pattern of LD in relation to the site of selection. K was counted for a sliding window of l consecutive polymorphic sites. A low value of K indicates a strong LD among these sites. Figure 1 shows two examples of selective sweep simulations in which the selected mutation arises in the center of the chromosome. The spatial pattern of K ($l = 14$) is compared with those of Tajima's ($\hat{\theta}_\pi$) and Fay and Wu's ($\hat{\theta}_H$) estimates of θ (Tajima 1983; Fay and Wu 2000). Low values of K were observed at some distance away from the site of selection, in the same regions where

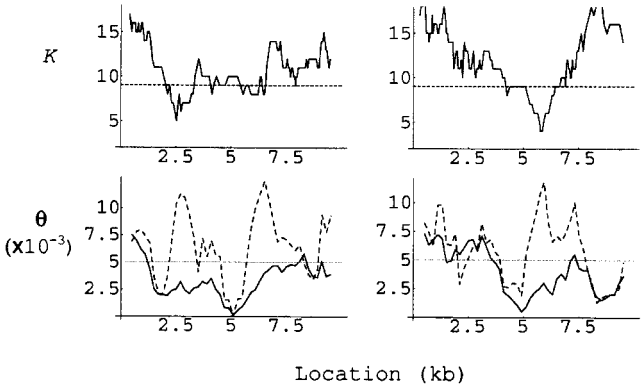

Figure 1.—Patterns of variation produced by recent selective sweep ($n = 25$, $\tau = 0.001$, $R = 1000$, $\alpha = 1000$, $\theta = 0.005$). Results of two replicates of simulation are shown. Top graphs are the plots of K *vs.* the location of the sliding window ($l = 14$) over which K was counted. *x*-axis represents the position along the sequence measured by kilobases. The location of a window is defined as the arithmetic mean of the nucleotide positions of l polymorphic sites. Dashed horizontal lines indicate the threshold of K ($= 9$) determined by neutral simulations (2.5th percentile). Below each graph of K is the corresponding plot of $\hat{\theta}_\pi$ (joined by solid lines) and $\hat{\theta}_H$ (joined by broken lines) over sliding windows (window size, 1 kb; step size, 0.2 kb). Gray horizontal lines represent the standing level of variation at neutrality ($\theta = 0.005$). It should be noted that, as two different kinds of windows (fixed number of polymorphic sites *vs.* fixed number of base pairs) are used, only an approximate comparison of spatial patterns in K and θ can be made.

an excess of high-frequency-derived alleles, revealed by a difference between $\hat{\theta}_\pi$ and $\hat{\theta}_H$ (Fay and Wu 2000), was observed. Different values of l (8–14) produced almost identical results (data not shown). If the spatial overlap of the lowest K and the largest difference between $\hat{\theta}_\pi$ and $\hat{\theta}_H$ is not coincidental, it may indicate that the excess of LD and the excess of high-frequency-derived alleles share the same underlying cause. In fact, a simple genealogical model of a selective sweep given in the discussion supports this argument. At this point, it is worthwhile to summarize the main conclusion in that section: The following three patterns of LD are predicted from a genealogical model of a selective sweep.

1. A high level of LD is expected in regions close, but not immediately adjacent, to the site where the fixation of the beneficial allele occurred.
2. When the chromosome is divided by the location of the beneficial mutation, the high level of LD is expected within each side but not across two sides.
3. The probability of observing a high frequency of derived alleles in the sample is greater in regions where LD is strong.

The first and third patterns are evident in Figure 1. It is not straightforward to verify the second prediction from Figure 1. However, a moderate level of LD in a

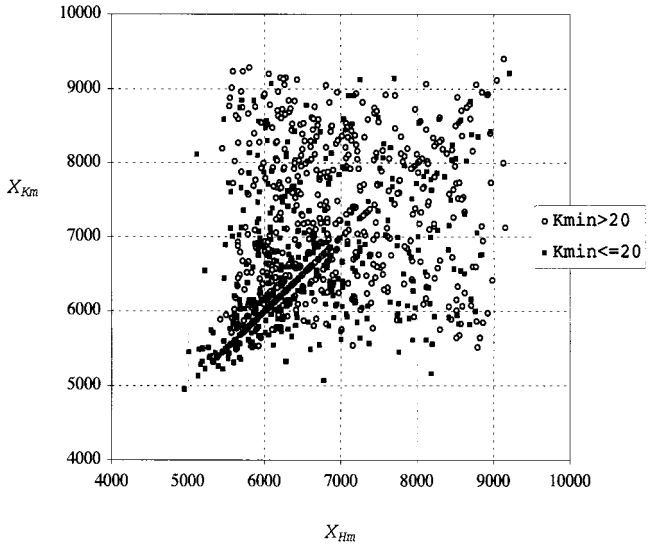

FIGURE 2.—Joint distribution of X_{K_m} and X_{H_m} from 1000 replicates of selective sweep simulation (n = 40, θ = 0.002, τ = 0.001, R = 500, α = 500). Beneficial mutation occurs at the position 4 kb on a 10-kb-long sequence. The sliding window of 20 consecutive polymorphic sites moves on the right side of the beneficial mutation (position 4–10 kb). Numbers on the axes are the distance in base pairs to the beneficial mutation.

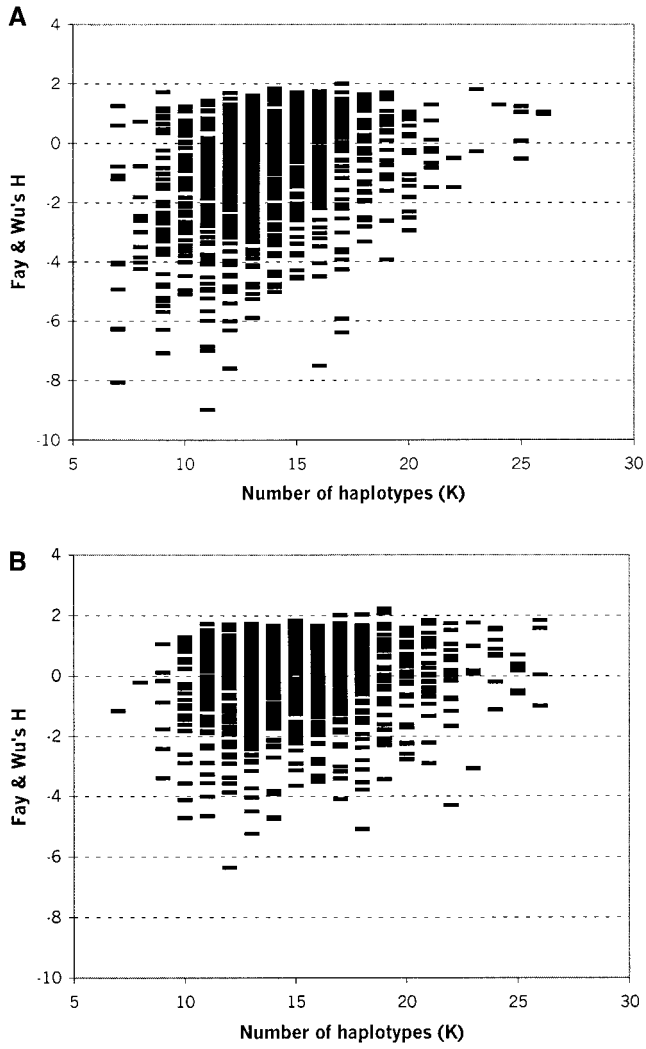

FIGURE 3.—Joint distribution of the number of distinct haplotypes (K) and Fay and Wu's H. Ten consecutive polymorphic sites were chosen randomly from each replicate of selective sweep (A) and neutral (B) simulations (1000 replicates each). Selective sweep simulations are those described in Figure 2. Neutral simulations use the same set of parameters except α = 0.

window of consecutive polymorphic sites that is centered on the beneficial mutation is compatible with the second prediction. Each of three expected patterns of LD is further examined below.

To investigate the statistical properties of K, more simulations were performed (n = 40, R = 500, θ = 0.002, τ = 0.001, and α = 500). The fixation of the beneficial mutation occurs at position 4 kb of a 10-kb-long sequence. For each simulation run, the smallest value of K, $K_{\min}$, in sliding windows of l consecutive polymorphic sites was found. As the patterns generated on both sides of the beneficial mutation are symmetrical, we analyzed only the sequence on one side of the beneficial mutation, between positions 4 and 10 kb. Let X_{K_m} be the position of the sliding window where $K_{\min}$ is obtained, where the position of a window is defined as the arithmetic mean of the locations of polymorphic sites within it. We also calculated Fay and Wu's H ($\hat{\theta}_\pi - \hat{\theta}_H$) for each sliding window and obtained the position X_{H_m}, where the minimum value of H is observed. Figure 2 shows the joint distribution of X_{H_m} and X_{K_m} (l = 20). In about half of the simulation runs, X_{H_m} and X_{K_m} are close to each other ($|X_{H_m} - X_{K_m}| < 0.5$ kb). This correlation becomes weaker with l = 15, mainly due to the occurrence of multiple windows of K_m (in this case the window closest to the site of selection is chosen). Increasing l above 20 produces too few sliding windows covering the region. In Figure 2, it is shown that the level of correlation depends on LD. Within the subset of data with strong LD, producing $K_{\min} \leq 20$, the correlation coefficient (r) is 0.412 ($P < 0.0001$). However, it decreases to 0.173 (P = 0.0002) with $K_{\min} > 20$. X_{H_m} and X_{K_m} are closer to the site of beneficial mutation with $K_{\min} \leq 20$ than with $K_{\min} > 20$ (Figure 2). Therefore, when a selective sweep generates a significant excess of LD near the site of directional selection it is also likely to generate a strong skew in the frequency spectrum at the same location. We suspect that this correlation, presumably due to common underlying genealogies, exists even under neutral evolution. To examine this, a window of 10 consecutive polymorphic sites (still restricted between positions 4 and 10 kb) was chosen randomly from each replicate of simulations above. Then K and Fay and Wu's H were calculated for each window. For a comparison, the same analysis was conducted for data sets simulated under the neutral model. Figure 3 shows the joint distribution of K and H. A

significantly positive correlation was observed in the selective sweep simulations (using linear regression, $r^2 = 0.087$, $P < 0.0001$). From neutral simulations a weaker but still significant correlation was observed ($r^2 = 0.009$, $P = 0.0022$). If it is generally true, as indicated by this result and Figure 2, that strong LD is seen together with a skewed frequency spectrum, there may not be a large advantage to including LD in an analysis that already takes the frequency spectrum into account. However, the correlation between LD and the skew of frequency spectrum (as measured by H) shown here is not strong enough to suggest that these two signatures of selective sweeps are completely redundant. For a given level of frequency spectrum change, selective sweeps generate stronger LD than expected under the neutral model. For example, with H between -3 and -2, the simulations of selective sweeps yield significantly lower K [$K = 12.8 \pm 2.6$ (mean $\pm$ SD)] than the neutral simulations do ($K = 14.9 \pm 3.1$; t-test with unequal variances, $P < 10^{-10}$). This implies that there is room for improving the power to detect selective sweeps by adding LD into frequency spectrum.

For the selective sweep simulations above, the 2.5 and 97.5 percentiles of $K_{\min}$ ($l = 20$) were 12 and 23, respectively, whereas those from neutral simulations were 18 and 27 (sliding windows move along the entire sequence). It should be noted that the neutral and selective sweep simulations used the same parameter values of sequence length, θ and R. There are thus more polymorphic sites, which produce more sliding windows, in neutral data sets. The difference in the range of $K_{\min}$ between neutral and selective sweep simulations would have been even larger, and the power of this test greater, had it been obtained for a fixed number of polymorphic sites. This suggests that high LD is a strong signature of selective sweeps and that $K_{\min}$ could be used as a test statistic to detect a selective sweep in a sample of DNA sequences. In the case above ($l = 20$), the power of rejecting the null hypothesis (neutral equilibrium) was 47.1%, where the criteria for rejection are $K_{\min}$ less than the 2.5th percentile from neutral data sets. When we use $\overline{r^2}$ [r^2 averaged over all pairs of polymorphic sites, or Z_{nS} of Kelly (1997)] as a test statistic, a similar power to rejecting the neutrality is obtained (0.479).

Detecting selective sweeps using $K_{\min}$ or $\overline{r^2}$, however, requires that the null distributions of these statistics be obtained from neutral simulations using correct values of θ and R. A correct and independent estimate of the local recombination rate is usually unavailable. In many cases, estimates of recombination rates after a selective sweep are lower than the correct value. If a recombination rate that is too low ($R = 250$ instead of 500) is used to simulate neutral data sets, the power to detect selective sweeps using $K_{\min}$ and $\overline{r^2}$ decreases to 0.128 and 0.096, respectively, for the case above. Therefore it is practically difficult to detect a change in the absolute level of LD caused by a selective sweep without knowing the correct value of R. Not having the correct value of θ for the simulation may also seriously affect the power of the test. However, as suggested by Figures 1 and 2, a selective sweep not only increases the average level of LD across a region, but also generates a specific spatial pattern of LD around the site of selection, which cannot be simply summarized by $K_{\min}$ or $\overline{r^2}$. As mentioned above, an examination of the genealogy shaped by a selective sweep predicts that the excess of LD is produced within each of the two regions flanking the site of selection but does not extend across the two regions (see below). We reason that this spatial pattern of LD should not depend critically on θ or R. To detect this pattern, the following test statistic was designed. If there are S polymorphic sites in the data set, we divide them into two groups: one from the first to the lth polymorphic site from the left and the other from the $(l + 1)$th to the last site ($l = 2, \ldots, S - 2$). Then we define

$$\omega = \frac{\left(\binom{l}{2} + \binom{S-l}{2}\right)^{-1}\left(\sum_{i,j\in L} r_{ij}^2 + \sum_{i,j\in R} r_{ij}^2\right)}{(1/l(S-l))\sum_{i\in L, j\in R} r_{ij}^2}, \qquad (1)$$

where L and R represent the left and right set of polymorphic sites, respectively, and r_{ij}^2 is r^2 between the ith and the jth sites. Obviously ω increases with increasing LD within each group but with decreasing LD between groups. For a given data set, a value of l that maximizes ω is found. We use this maximum value of ω, $\omega_{\max}$, as the test statistic. Applying this to the simulated data sets shown above (using polymorphic sites on both sides of the beneficial mutation), the power of rejection is 0.597 using the null distribution with the correct recombination rate ($R = 500$). When $R = 250$ is used instead, the power decreases to 0.375 (37.2% reduction). However, this reduction of power is modest compared to those of $K_{\min}$ and $\overline{r^2}$ tests (a reduction of 79.6 and 73.3%, respectively). Therefore, as expected, the detection of selective sweeps is more robust to assumptions regarding the recombination rate by targeting the specific spatial pattern of LD than by observing the average level of LD across the region. As $\omega_{\max}$ was designed to detect the disruption of LD across the two flanking regions, the power to detect a sweep will be lost when the data do not include the selected locus. Indeed, the power of the $\omega_{\max}$ test was only 0.269 when it was applied to simulated data in which the beneficial mutation occurs at position 0 of the 10-kb-long sequence (other parameters are identical to the sweep simulation above, using the correct $R = 500$).

LIKELIHOOD METHOD FOR DETECTING SELECTIVE SWEEP WITH LD

So far we quantified the excess of LD caused by selective sweep by $K_{\min}$, r^2, or ω. But the use of these summary statistics may result in the loss of a considerable amount of information. Furthermore, it is not obvious how to

incorporate the information from these statistics into a likelihood analysis that uses other features of genetic variation. We solve these problems using an approach analogous to that of Hudson (2001). Briefly, this study determined the distribution of allelic configurations for a pair of polymorphic sites in a standard neutral model using two-locus coalescent simulations. These two-locus sampling probabilities were obtained for various genetic distances between loci. Then a composite-likelihood function was obtained by multiplying joint sampling for all pairs of polymorphic sites. This composite likelihood was used to assess the level of LD in data and estimate the recombination rate. While Hudson (2001) considers sampling probabilities under the neutral model, we try to obtain the equivalent quantities for the selective sweep model. Therefore, in this study the sampling probabilities are obtained from coalescent simulations in which joint genealogies are constructed for two neutral loci under the hitchhiking effect from a third locus.

The two neutral sites are denoted by M1 and M2, and the site under directional selection by Sel. M1 is defined to be the neutral (marker) locus closer to Sel than M2 is. The relative positions of these three loci are specified by the continuous variables R_1 and R_2, where R_1 (>0) is the scaled recombination rate between M1 and Sel and $|R_2|$ ($\geq R_1$) is the scaled recombination rate between Sel and M2. If R_2 is positive (negative), it indicates that M2 is on the same (opposite) side as M1 relative to Sel. Therefore, if R_1 and R_2 are not large, the recombination rate between the two marker loci is approximately $|R_2 - R_1|$. The sample configuration at the neutral loci is denoted by $\mathbf{n} = (n_{00}, n_{01}, n_{10}, n_{11})$, where n_{ij} is the number of chromosomes carrying allele i at M1 and allele j at M2. At each locus the ancestral allele is denoted 0 and the derived allele is denoted 1. Let the scaled mutation rate per locus be θ ($= 4N\mu$). For small values of θ, the sampling probability $Q(\mathbf{n}; R_1, R_2, \alpha, \theta)$, for two variable sites, can be approximated by $(\theta/2)^2 h(\mathbf{n}; R_1, R_2, \alpha)$, where $h(\mathbf{n}; R_1, R_2, \alpha)$ is equivalent to the "scaled likelihood" of Hudson (2001).

To estimate $h(\mathbf{n}; R_1, R_2, \alpha)$, simulations are performed using the method of Kim and Stephan (2002) with $\tau = 0$ (although here we describe a three-locus simulation, in practice we first construct a multilocus ancestral recombination graph under the selective sweep model and then extract marginal trees corresponding to M1 and M2). Let G_{1i} be the marginal tree for the M1 locus obtained from the ith simulation run and G_{2i} be that for M2. Then the two-locus genealogy from this particular simulation replicate is denoted by $\mathbf{G}_i = (G_{1i}, G_{2i})$. $\mathbf{G}_i$ is determined by three parameters: R_1, R_2, and the strength of selection on the beneficial allele, $\alpha = 2Ns$. After m runs of coalescent simulations, $h(\mathbf{n}; R_1, R_2, \alpha)$ is estimated by

$$\hat{h}(\mathbf{n}; R_1, R_2, \alpha) = \frac{1}{m} \sum_{i=1}^{m} \sum_{j,k} I(\mathbf{G}_i, \mathbf{n}, j, k)\, a_j(i)\, b_k(i) \qquad (2)$$

(Nielsen 2000; Hudson 2001). In this equation, $I(\mathbf{G}_i, \mathbf{n}, j, k)$ is an indicator function equal to one if a mutation on branch j of G_{1i} and a mutation on branch k of G_{2i} would generate sample configuration $\mathbf{n}$ and zero otherwise, when branches of each marginal tree are arbitrarily labeled from 1 to $2n - 2$. $a_j(i)$ is the length of the jth branch of G_{1i} and $b_k(i)$ is the length of the kth branch of G_{2i}. The branch lengths are in units of $2N$generations.

Our aim is to obtain a table of sampling probabilities for many different cases of selective sweeps, *i.e.*, for many different values of R_1, R_2, and α. However, it is computationally very difficult to obtain accurate estimates of $h(\mathbf{n}; R_1, R_2, \alpha)$ on a dense three-dimensional grid of parameter values. Therefore, we use a rescaling of the parameters R_1, R_2, and α that will reduce the number of parameters from three to two.

In the Appendix of Kim and Stephan (2002), the final frequency of the descendants of a neutral allele that was originally linked to a beneficial allele sweeping through the population is given by $\varepsilon^{r/s}$, where r is the recombination rate between the selected and the neutral sites and s is the selection coefficient, and ε is the starting frequency of the beneficial mutation. We use $\varepsilon = 1/(4Ns) = 1/(2\alpha)$ instead of $1/(2N)$ because, conditional on fixation, the early increase in the frequency of the beneficial allele is accelerated by a factor $1/(2s)$ relative to the deterministic increase from $1/(2N)$ (Maynard Smith 1971; Barton 1998). At the end of the selective phase, the final frequency of the descendants of the neutral allele originally linked to the beneficial allele is then given approximately by $(2\alpha)^{-R/(2\alpha)}$. Therefore, the scaled genetic distance, defined as the probability that a neutral lineage recombines away from the beneficial allele during the selective phase, is $\sim C = 1 - (2\alpha)^{-R/2\alpha}$. We then rescale R_1 and R_2 in the model above into C_1 and C_2, where $C_1 = 1 - (2\alpha)^{-R_1/2\alpha}$ and $|C_2| = 1 - (2\alpha)^{-|R_2|/2\alpha}$ (this definition ensures that R_2 and C_2 have the same sign).

While this approximation has been shown to be adequate to describe the reduction of expected heterozygosity and the skew in the frequency spectrum caused by a selective sweep (Barton 1998; Kim and Stephan 2002), it is not known how suitable it is for describing the pattern of LD or the distribution of sample configurations ($\mathbf{n}$). We examined this problem using simulations. Table 1 compares the scaled likelihoods of sample configurations $[\hat{h}(\mathbf{n}; R_1, R_2, \alpha)]$ estimated from the simulations ($n = 15$) using $\alpha = 1000$ and 5000. R_1 and R_2 were adjusted to give fixed values of C_1 and C_2. For $C_1 = 0.2$ and $|C_2| = 0.3$, we obtained a close agreement of sample configurations between two cases. However, for $C_1 = 0.05$ and $|C_2| = 0.06$ the agreement is not as good. In general, the discrepancy is largest when the scaled distance between markers, $|C_2 - C_1|$, is small (data not shown). Nonetheless, for computational reasons we pursue an approach using the scaled genetic distances instead of the full parametric model. It should be noted

TABLE 1

Examples of $\hat{h}(\mathbf{n}; C_1, C_2)$ ($n = 15$)

$\min(n_{00}, n_{11})$		$\alpha = 1000$			$\alpha = 5000$		
$n_{1.}$	$n_{.1}$	0	1	2	0	1	2
			$C_1 = 0.05$, $C_2 = 0.06$				
1	1	0.20018	0.36915		0.12717	0.36203	
1	14	0.31931	0.07021		0.30964	0.05253	
2	2	0.00456	0.00079	0.08081	0.00205	0.00154	0.04850
3	2	0.00145	0.00010	0.00414	0.00075	0.00023	0.00346
3	13	0.00200	0.00012	0.00062	0.00179	0.00018	0.00041
14	14	0.00859	0.30202		0.01959	0.29064	
			$C_1 = 0.05$, $C_2 = -0.06$				
1	1	0.25227	0.02026		0.19500	0.01469	
1	14	0.01144	0.14045		0.01084	0.14231	
2	2	0.00678	0.00067	0.00164	0.00460	0.00111	0.00050
3	2	0.00172	0.00038	0.00069	0.00134	0.00038	0.00014
3	13	0.00049	0.00041	0.00158	0.00011	0.00058	0.00155
14	14	0.07670	0.00621		0.10562	0.00785	
			$C_1 = 0.2$, $C_2 = 0.3$				
1	1	1.04039	0.41801		1.05324	0.42335	
1	14	0.18071	0.27949		0.16717	0.28869	
2	2	0.07649	0.05063	0.07663	0.06990	0.07151	0.05155
3	2	0.02719	0.01890	0.03071	0.02223	0.02550	0.02878
3	13	0.01388	0.01080	0.01087	0.01403	0.01378	0.00880
14	14	0.11038	0.12327		0.13184	0.10781	
			$C_1 = 0.2$, $C_2 = -0.3$				
1	1	1.21589	0.09724		1.27142	0.09276	
1	14	0.03389	0.40149		0.03073	0.42568	
2	2	0.11127	0.02939	0.00521	0.11723	0.03834	0.00281
3	2	0.04072	0.01770	0.00423	0.04021	0.01956	0.00301
3	13	0.00232	0.01320	0.02779	0.00186	0.01432	0.02869
14	14	0.17824	0.01643		0.20158	0.01410	

$n_{1.}$ and $n_{.1}$ are the number of derived alleles observed at M1 and M2, respectively. $\hat{h}(\mathbf{n}; C_1, C_2)$ for each parameter set was estimated from 10^5 runs of selective sweep simulations.

that any inadequacy of the approximation will not result in an anticonservative test because critical values of the test are obtained using simulations based on the full model.

Table 1 also supports the argument that the excess of LD is produced within each of the two regions flanking the site of selection but does not extend across the two regions (see DISCUSSION). For example, examine the case of $n_{1.} = n_{.1} = 1$. With $C_1 = 0.05$ and $C_2 = -0.06$, $\Pr[n_{11} = 1]$ [probability of two mutant alleles on the same chromosome; proportional to $\hat{h}(\mathbf{n} = (14, 0, 0, 1); C_1, C_2, \alpha)$] is much smaller than $\Pr[n_{11} = 0]$ [$\propto \hat{h}(\mathbf{n} = (13, 1, 1, 0); C_1, C_2, \alpha)$]. Indeed, $\Pr[n_{11} = 1]/\Pr[n_{11} = 0]$ is 0.080 for $\alpha = 1000$ and 0.075 for $\alpha = 5000$, which is only slightly higher than the expectation under random association of alleles ($= 1/14$). However, we obtain much higher values of $\Pr[n_{11} = 1]/\Pr[n_{11} = 0]$ with $C_1 = 0.2$ and $C_2 = 0.3$, even though the two sites are in a similar genetic distance ($|C_1 - C_2|$) and farther away from the site of selection. Therefore, high LD is found only between two sites located on the same side of the beneficial mutation. Other examples in Table 1 support this conclusion.

To generate a complete table of $\hat{h}(\mathbf{n}; C_1, C_2)$, all possible combinations of C_1 and C_2, in which $C_2 = C_1 + c$ or $C_2 = -C_1 - c$ (C_1 or $c = 0.001$, 0.01, 0.02, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9) and $|C_2| < 1$, were simulated using $\alpha = 1000$. For each combination of C_1 and C_2, 10^5 replicates of two-locus genealogies were generated. We also obtained the mean total tree lengths of marginal trees for positions corresponding to scaled distances $C = 0.001j$ ($j = 1, \ldots, 999$). Using this table, a composite likelihood under the model of selective sweep was obtained. Let $P_i(k) = P_i(k; X, \alpha, \theta, R_n)$ be the probability of observing k derived alleles at the ith site ($k = 0, \ldots, n - 1$), where X is the position of the beneficial mutation and $R_n = 4N\rho$ is the scaled recombination rate per nucleotide. In the limit of $\theta \rightarrow$

0, $P_i(0)$ can be approximated by $\hat{P}_i(0) = 1 - (\theta/2)t_i$, where t_i is the mean total tree length of the marginal tree closest to the ith site. Also, let $\hat{Q}_{jk}(\mathbf{n}) = (\theta/2)^2\hat{h}(\mathbf{n}; C_1, C_2)$ be the simulation estimate of $Q_{jk}(\mathbf{n})$, where $\hat{h}(\mathbf{n}; C_1, C_2)$ is interpolated from the table. Then $\hat{Q}_{jk}(\mathbf{n})/\Sigma_{\mathbf{m}\in\Psi}\hat{Q}_{jk}(\mathbf{m})$, where Ψ is the set of all possible configurations in which both sites are polymorphic, provides an estimate of the sampling probability of configuration $\mathbf{n}$, from two variable sites, j and k, conditional on variability in these sites. We then propose a composite likelihood,

$$L_1 = \prod_{i\in A}\hat{P}_i(0)\prod_{j\in B}[1 - \hat{P}_j(0)]\left(\prod_{j,k\in B}\frac{\hat{Q}_{jk}(\mathbf{n})}{\Sigma_{\mathbf{m}\in\Psi}\hat{Q}_{jk}(\mathbf{m})}\right)^{1/(S-1)}, \quad (3)$$

where A (B) denotes the set of the locations of the monomorphic (polymorphic) sites in the data and S is the number of polymorphic sites. Therefore, L_1 depends on the spatial distribution of the polymorphic sites as well as on the frequency spectrum and the LD. The exponent $1/(S-1)$ comes from the fact that a given polymorphic site is counted $S-1$ times in the multiplication of two-locus sampling probabilities. To further assess the contribution of LD, we define the second composite likelihood,

$$L_2 = \left(\prod_{j,k\in B}\frac{\hat{Q}_{jk}(\mathbf{n}')}{\Sigma_{\mathbf{m}'}\hat{Q}_{jk}(\mathbf{m}')}\right)^{1/(S-1)}, \quad (4)$$

where $\mathbf{n}'$ and $\mathbf{m}'$ represent the "folded" sample configurations that do not distinguish the derived from the ancestral allele at each site (see Equation 6 of Hudson 2001). This maximizes the relative contribution of LD by eliminating the effect of high-frequency-derived alleles. However, the excess of rare alleles, which is the component of frequency spectrum that causes negative Tajima's D (Tajima 1989), can still be detected by L_2. We also calculate

$$L_3 = \prod_{i\in A\cup B}P_i(k_i) \quad (k_i = 0, \ldots, n-1) \quad (5)$$

and

$$L_4 = \prod_{i\in B}\frac{P_i(k_i)}{1 - P_i(0)} \quad (k_i = 1, \ldots, n-1), \quad (6)$$

where k_i is the number of the derived allele observed at the ith site. $P_i(k)$ under the selective sweep model is calculated from Kim and Stephan (2002) using $\varepsilon = 1/(2\alpha)$. Therefore, L_3 is equivalent to the composite likelihood proposed in Kim and Stephan (2002), which depends on the spatial distribution and the frequency spectrum of polymorphic sites. On the other hand, L_4 depends only on the frequency spectrum. It should be noted that L_2 and L_4 are not functions of θ. For each definition from L_1 to L_4, it is possible to obtain the corresponding quantities under the neutral equilibrium model. To calculate L_1 and L_2 under the neutral model, we use the table of the scaled likelihoods of two-locus sampling configurations from Hudson (2001; available at http://home.uchicago.edu/rhudson1/source/twolocus.html). Then the log-likelihood ratio is defined as $LR_k = \log(L_k^1/L_k^0)$, where L_k^1 (L_k^0) is L_k ($k = 1, 2, 3$, and 4) maximized under the model of selective sweep (neutral equilibrium).

TABLE 2

Composite-likelihood analysis of simulated data under selective sweep [$\mathrm{Log}_{10}(2Ns) = 2699$, $X = 4$ (kb)]

	Power (0.975)	$\mathrm{Log}_{10}(2Ns)$	X (kb)
		$n = 15$	
Test 1 (θ)	NA	2.72 ± 0.219	3.97 ± 1.01
Test 1 ($\hat{\theta}_W$)	0.675	2.52 ± 0.230	3.99 ± 1.11
Test 2	0.244	2.66 ± 0.306	4.41 ± 2.11
Test 3 (θ)	NA	2.78 ± 0.244	3.96 ± 1.05
Test 3 ($\hat{\theta}_W$)	0.608	2.62 ± 0.260	3.96 ± 1.20
Test 4	0.435	2.51 ± 0.467	4.22 ± 2.06
		$n = 40$	
Test 1 (θ)	NA	2.70 ± 0.198	4.01 ± 0.905
Test 1 ($\hat{\theta}_W$)	0.774	2.54 ± 0.215	4.01 ± 0.920
Test 2	0.399	2.68 ± 0.227	4.24 ± 1.72
Test 3 (θ)	NA	2.71 ± 0.238	4.01 ± 1.02
Test 3 ($\hat{\theta}_W$)	0.728	2.61 ± 0.270	4.05 ± 1.15
Test 4	0.595	2.54 ± 0.381	4.18 ± 1.64

θ and $\hat{\theta}_W$ in parentheses indicate that the likelihood was calculated using the true value of θ and Watterson's estimate of θ, respectively. NA, not applicable.

The statistical test for detecting a selective sweep using LR_k as a test statistic is referred to as test k. For each test, the empirical null distributions of the likelihood ratio come from a large number of data sets simulated under the neutral model (Kim and Stephan 2002). The composite likelihood under the selective sweep model is maximized by changing values of X and α. It is assumed that a correct estimate of R_n for the data set under investigation is available from a source other than the data analyzed. Test 1 and test 3 also require an independent estimate of θ to calculate likelihoods. We use either the true value of θ (used in the simulation) or Watterson's θ, $\hat{\theta}_W$ (Watterson 1975), estimated directly from each data set. Only the latter scheme is appropriate for detecting the signature of selective sweeps from the pattern of variation rather than from the absolute level of variation. [If a correct θ is used, very large LR_1 and LR_3 will be obtained because of the reduction of average heterozygosity due to selective sweep (Kim and Stephan 2002). This is not much different from the Hudson-Kreitman-Aguadé test (Hudson *et al.* 1987). However, a correct value of θ is not usually known.] Null distributions of LR_1 and LR_3 obtained using $\hat{\theta}_W$ are independent of the number of polymorphic sites (data not shown). However, we use the true value of θ to test the accuracies

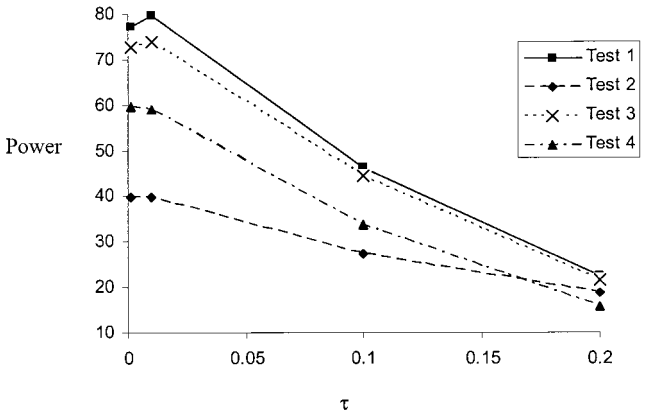

FIGURE 4.—Powers of composite-likelihood tests for various times since the last hitchhiking event (τ). Each test was applied to 1000 replicates of selective sweep simulations ($n = 40$, $\theta = 0.002$, $R = 500$, $\alpha = 500$, position of beneficial mutation, 4 kb).

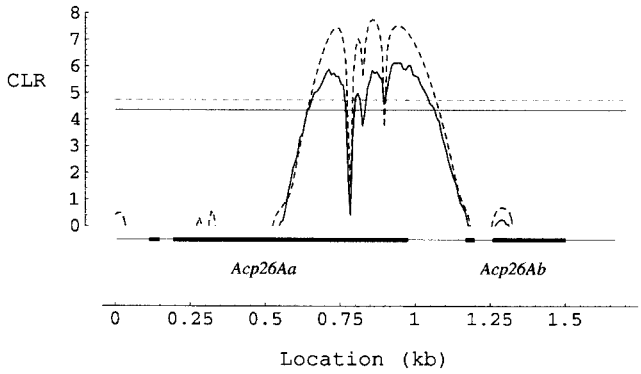

FIGURE 5.—Plots of LR_1 (solid curve) and LR_2 (dashed curve) for varying locations of the candidate target of selection over the *Acp26A* locus. Only LR_1 and LR_2 that are greater than zero are shown. Solid and dashed horizontal lines indicate the 95th percentile of LR_1 and LR_2, respectively, obtained from neutral simulations. Locations of two exons for both *Acp26Aa* and *Acp26Ab* are shown as solid boxes below the plots.

of likelihood estimates of X and α. The performance of each statistical test was evaluated against the data sets simulated under the selective sweep model ($L = 10$ kb, $R = 500$, $\theta = 0.002$, $\tau = 0.001$, $\alpha = 500$, and the beneficial mutation at position 4 kb). Table 2 shows the power to detect a selective sweep using two different sample sizes ($n = 15$ and 40). From the comparison of test 1 and test 3, it is shown that including two-locus sampling probabilities into the composite likelihood improves both the statistical power and the accuracy of the parameter estimation. However, the degree of improvement is rather small. This suggests that information contained in LD is to some extent redundant with that contained in the other aspects of genetic variation. With $n = 40$, test 2 (which depends heavily on the pattern of LD) and test 4 (which depends mainly on the frequency spectrum) rejected the null hypothesis for 39.9 and 59.5% of simulated data sets, respectively (Table 2). For 33.5% of the same data sets, both test 2 and test 4 rejected the null hypothesis. Therefore, in only 6.4% (39.9 − 33.5) of the data LD contributed exclusively to detecting selective sweeps. This is in agreement with the observation of the positive correlation between Fay and Wu's H and K observed earlier (Figures 2 and 3), although it was also shown that the correlation is not strong enough to make two signatures of selective sweeps completely redundant.

Finally, as the likelihoods were calculated under the assumption that DNA sequences are sampled immediately after the fixation of beneficial alleles, the power of these tests is expected to decrease with increasing time since the last fixation of the beneficial mutation. Figure 4 shows that the power to detect a selective sweep decreases substantially when $\tau > 0.1$ ($0.2N$ generations after the sweep). Test 2 exhibited a relatively slow decline of the power presumably because it is not dependent on the excess of high-frequency-derived alleles, which disappears quickly with increasing τ (KIM and STEPHAN 2002; PRZEWORSKI 2002).

APPLICATION TO DATA

The composite-likelihood method developed here is applied to two available data sets in Drosophila. First, we analyze polymorphism in *Acp26A* genes from a North Carolina population of *Drosophila melanogaster* (AGUADÉ *et al.* 1992). The data set is composed of 10 sequences, 1.6 kb long. FAY and WU (2000) detected a significant excess of high-frequency-derived alleles at this locus as evidence of a selective sweep. Furthermore, they observed a local reduction of variation in the middle of the *Acp26Aa* gene, flanked by narrow regions of high-frequency-derived alleles. This pattern is consistent with the recent fixation of a beneficial alleles in the middle of *Acp26Aa*. However, as shown by KIM and STEPHAN (2002), stochastic fluctuation of variation along the sequence may create such a pattern by chance. Therefore, a proper statistical test is needed to test the significance of the spatial pattern observed. We applied tests 1, 2, and 3 to the data set. The scaled recombination rate is 33.2, assuming $N = 10^6$ and $r = 5 \times 10^{-8}$ (FAY and WU 2000). The ancestral/derived status of alleles was determined by comparison with outgroup sequences (AGUADÉ *et al.* 1992). For those sites where the derived allele could not be identified, we arbitrarily assigned the derived status to rare alleles to make the test conservative. Both test 1 and test 3 rejected neutrality in favor of selective sweep [test 1, $LR_1 = 5.22$ ($P = 0.011$), $\hat{\alpha} = 13.7$, $\hat{X} = 946$; test 3, $LR_3 = 7.54$ ($P = 0.010$), $\hat{\alpha} = 15.1$, $\hat{X} = 943$]. However, test 2, the test based only on two-locus sampling configurations, could not reject neutrality [$LR_2 = 1.01$ ($P = 0.423$), $\hat{\alpha} = 19.9$, $\hat{X} = 751$]. The

signature of the selective sweep becomes stronger with a larger population size [$N = 5 \times 10^6$; test 1, $LR_1 = 6.20$ ($P = 0.007$), $\hat{\alpha} = 110.2$, $\hat{X} = 943$; test 3, $LR_3 = 7.76$ ($P = 0.001$), $\hat{\alpha} = 100.9$, $\hat{X} = 857$]. Figure 5 shows the composite-likelihood ratios obtained for varying X ($N = 5 \times 10^6$). Profiles of LR_1 and LR_3 are very similar to each other, indicating that the incorporation of LD into the test did not change the estimated range for the candidate site of selection. This result is expected since no significant contribution of LD was detected [by test 2, $LR_2 = 2.71$ ($P = 0.171$)].

The next data set, on the other hand, contains a very strong pattern of LD as a signature of selective sweep. Schlenke and Begun (2004) have analyzed DNA sequence variation from many loci across chromosome arm 2R in *D. simulans*. Using eight chromosomes sampled from a California population, they surveyed polymorphism at 28 short sequence segments (900 bp long on average; here denoted Sgm1 to Sgm28) that are scattered over 3.2 Mb. Surprisingly, eight segments (Sgm10 to Sgm17) that span a 100-kb region contain no polymorphic sites. The probability of observing no variation in those segments is extremely low considering the amount of variation usually observed in *D. simulans* populations (Schlenke and Begun 2004). Within each of the four segments that are located adjacent to this region of no variation (Sgm8, -9, -18, and -19), segregating alleles are in complete linkage disequilibrium ($\overline{r^2} = 1.0$). Thereafter, moving away in both directions, $\overline{r^2}$ decays and haplotype diversity increases gradually. Thus, from several aspects of the data, a recent fixation of a very strong beneficial allele is overwhelmingly supported. The spatial pattern of LD is in complete agreement with the prediction from the genealogical modeling of a selective sweep (see DISCUSSION). We obtained composite-likelihood ratios for this data set. (Since the scale of data is very large, the test of significance using neutral simulations was not conducted. However, we argue that these signatures of selective sweep are significant beyond doubt.) Scaled recombination rate per nucleotide was assumed to be 0.04 ($N = 10^6$ and $r = 10^{-8}$). From Tests 1–3, the maximum composite-likelihood locations of the target of selection were all found within the 100-kb span of no variation, as expected. The profiles of LR_1 and LR_3 over X are similar to each other (Figure 6) even though there is a clear contribution of LD. The profile of LR_2 is quite different from the others. While LR_1 and LR_3 are maximized in the middle of the 100-kb span, highest LR_2 is found around position 1820 kb. This is an interesting observation since the candidate target of selection, *Cyp6p1*, which appears to be associated with an insecticide resistance in the California population (Schlenke and Begun 2004), is located near Sgm16 (Figure 6). Therefore the fixation of a strongly selected beneficial mutation on the *Cyp6p1* locus is consistent with the pattern of LD, but not as well supported by other features of the data.

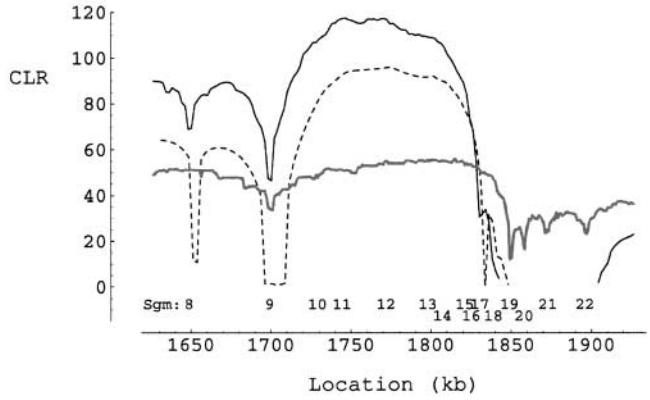

Figure 6.—Plots of composite-likelihood ratios (CLRs) obtained for varying locations of the target of selection around the 100-kb span of zero heterozygosity in the data of Schlenke and Begun (2004). Three CLRs that are greater than zero are shown: LR_1 (solid curve), LR_2 (shaded curve), and LR_3 (dashed curve). The positions of Sgm8 to Sgm22 are shown by numbers from 8 to 22 below the plots. The relative position on the *x*-axis follows that used in Table 1 of Schlenke and Begun (2004).

DISCUSSION

Previous studies using simulations have shown that the level of LD increases after a selective sweep (Kim and Stephan 2002; Przeworski 2002). Table 2 shows that the power to detect a selective sweep using LD (test 2) is not so small compared to that using frequency spectrum alone (test 4). Therefore, LD is an important signature of selective sweeps. However, it is surprising to find that the addition of LD into the composite likelihood only slightly increases the power to detect a selective sweep (comparing test 1 and test 3). This implies that LD and the other features of sequence variation are correlated to some extent. To understand this, we first need to understand why and how LD increases due to selective sweep. While a full mathematical analysis on this phenomenon is difficult, a rather simple argument in the framework of the coalescent theory can explain why LD increases due to hitchhiking. Recent studies demonstrate that the causes of LD can be clearly understood by studying genealogy (McVean 2002; Nordborg and Tavaré 2002).

In our model, only recombination due to meiotic crossing over, but not gene conversion, is considered. Assume that DNA sequences are sampled immediately after the fixation of a beneficial allele, *B*. Also assume that the selective phase (the period when the beneficial mutation is on the way to fixation) is very short. The genealogy of the segment completely linked with *B* can be described by a star-like tree (tree 1 of Figure 7).

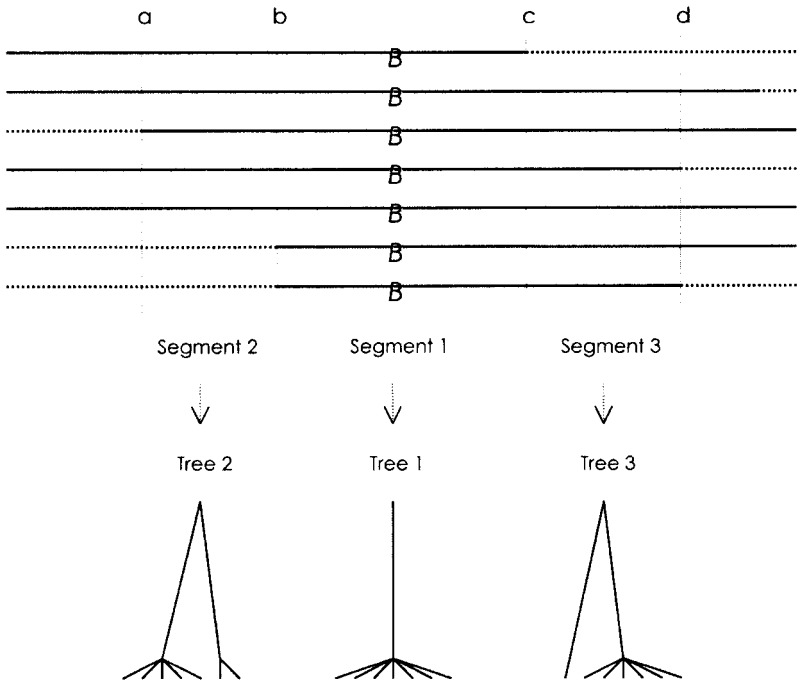

FIGURE 7.—An example of a genealogical structure of DNA sequences produced by a recent selective sweep. Horizontal lines represent DNA sequences sampled immediately after the fixation of the beneficial allele *B*, which is located in the middle of the sequence. Solid lines represent sequences that were initially linked with *B*. Dashed lines represent "recombinant" sequences that were linked with *b* when the mutation to *B* occurred in the population but recombined with *B* during the selective phase. Recombination breakpoints that mark the beginnings of recombinant sequences are labeled a, b, c, and d. Segments between breakpoints are defined as segments 1, 2, and 3 as shown above. Below each segment is a coalescent tree that represents genealogical structure of each segment.

Looking backward in time, coalescence of gene lineages proceeds with increasing rate as the frequency of *B* in the population decreases until one lineage is left as a common ancestor. This describes the ancestral history for segment 1 in Figure 7. During the selective phase, a recombination event may occur, allowing an ancestral sequence originally linked with *B* to recombine with another sequence that was previously linked with *b*, the allele ancestral to *B*. As a result, (descendants of) "recombinant" sequences appear in the sample, which are shown in Figure 7 as dotted lines. They begin at recombination breakpoints and extend distal to *B*. As one moves away from the site of selection, more recombinant sequences on other chromosomes appear as more recombination breakpoints are encountered. The ancestral history of the region between the first and the second breakpoint to the left of *B*, labeled segment 2 in Figure 7, is described by coalescent trees with the structure of tree 2 in Figure 7. A recombinant sequence does not coalesce with sequences carrying *B* during the selective phase. Therefore, tree 2 has two long inner branches whose length is approximately exponentially distributed with the mean of $2N$ generations. Similarly, the ancestral history of segment 3 can be represented by tree 3 in Figure 7.

Recombination events producing breakpoints as shown in Figure 7 occur mainly when the frequency of *B* is low in the population (KAPLAN *et al.* 1989; STEPHAN *et al.* 1992). Since the fixation of *B* occurs very quickly, there is a very limited opportunity for recombination events to occur in a given interval of sequence during the selective phase. Therefore, the space between the breakpoints becomes larger as the length of the selective phase becomes shorter. When mutations are mapped on coalescent trees for segments 2 and 3, most will map onto the long inner branches. Then rare alleles in the sample will be found along the chromosomes containing the recombined sequences. If segments 2 and 3 stretch long enough, strings of consecutive polymorphic sites with complete allelic association will be observed. By simulating coalescence under both selective sweep and neutral models, KIM and STEPHAN (2002) showed that the appearance of such consecutive polymorphic sites in complete LD is a unique signature of a recent selective sweep.

The ancestral history of the region to the left of segment 2 or to the right of segment 3 in Figure 7 can be represented by a group of coalescent trees in which more than one gene lineage escapes coalescence at the end of the selective phase. In this region, complete association of segregating alleles does not necessarily occur because recombination events involving inner branches disrupt association of alleles descended onto recombined segments. However, a considerable level of LD is still expected because rare alleles are confined to only a few chromosomes in the sample. Moving farther away from the site of selection, LD quickly disappears as more recombination breakpoints are encountered and genealogies approach those under neutrality. Therefore, the largest excess of LD is expected to occur in a region corresponding to either segment 2 or 3.

From the above discussion, one may predict two important properties of LD generated by selective sweeps. First, when the location of the beneficial mutation di-

vides the region into two, we expect to find a lack of LD between the two sides. As recombination breakpoints are expected to occur independently across the two sides, recombinant segments where rare alleles accumulate should also occur independently. Therefore, allelic associations should be observed within each side but not across the two sides. This prediction is well supported by simulation results (Table 1). It should be noted, however, that this property is expected under the model assuming only meiotic crossing over but not gene conversion.

Second, LD will be generated in regions where an excess of high-frequency-derived alleles is also generated. The occurrence of high-frequency-derived alleles is caused by trees with a small number of long inner branches, such as tree 2 or tree 3 in Figure 7 (Fay and Wu 2000; Przeworski 2002). Therefore, these two signatures of selective sweep have a common underlying genealogical structure. This explains the correlated occurrence of LD and high-frequency-derived alleles, shown in Figures 1–3.

The small improvement of test 1 relative to test 3 in Table 2 might also be explained by the correlation of LD and the skew of frequency spectrum. However, the degree of correlation shown in Figure 3 was not as strong as to suggest such a small increase of power in detecting selection. Therefore, there could be other reasons for this small improvement of test 1 over test 3. It might be argued that the excess of high-frequency-derived alleles or a large value of Fay and Wu's H is not a complete summarization of the change in the frequency spectrum caused by selective sweeps. A summary statistic capturing all information regarding frequency spectrum might have shown a stronger correlation with LD. Alternatively, there might be a systematic limitation of extracting information regarding LD in our new composite-likelihood method using two-locus sampling probabilities. The full information may not be obtained unless a higher-order LD, the association of more than two consecutive polymorphic sites, is taken directly into account in the calculation of the likelihood.

So far, in the model shown in Figure 7, it was assumed that DNA sequences are sampled shortly after the fixation and that the length of the selective phase is small enough to be ignored. Let t_o be the length of the period between the time of sampling and the completion of coalescence among B alleles. Unless directional selection is very strong and/or sample size is very small, the sum of outer branch lengths, approximately equal to nt_o for tree 1 of Figure 7, can be substantial. Then, mutations along these outer branches cannot be ignored. These mutations, mainly singletons randomly distributed over chromosomes in the sample, will obscure the pattern of LD produced by a selective sweep as explained above. LD further decays by recombination events involving outer branches as t_o increases. Therefore, as time since the fixation of a beneficial mutation increases it becomes difficult to detect a selective sweep using LD (Figure 4).

We thank anonymous reviewers for useful comments. This work was supported by funds from National Science Foundation grant DEB-0089487 to R.N.

LITERATURE CITED

Aguadé, M., N. Miyashita and C. H. Langley, 1992 Polymorphism and divergence in the *Mst26A* male accessory gland gene region in *Drosophila*. Genetics **132:** 755–770.

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

Bustamante, C. D., R. Nielsen, S. A. Sawyer, K. M. Olsen, M. D. Purugganan *et al.*, 2002 The cost of inbreeding in *Arabidopsis*. Nature **416:** 531–534.

Depaulis, F., and M. Veuille, 1998 Neutrality tests based on the distribution of haplotypes under an infinite-site model. Mol. Biol. Evol. **15:** 1788–1790.

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

Fay, J. C., G. J. Wyckoff and C.-I Wu, 2001 Positive and negative selection on the human genome. Genetics **158:** 1227–1234.

Harr, B., M. Kauer and C. Schlötterer, 2002 Hitchhiking mapping: a population-based fine-mapping strategy for adaptive mutations in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **99:** 12949–12954.

Hill, W. G., and A. Robertson, 1968 Linkage disequilibrium in finite populations. Theor. Appl. Genet. **38:** 473–485.

Hudson, R. R., 2001 Two-locus sampling distributions and their application. Genetics **159:** 1805–1817.

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

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

Kelly, J. K., 1997 A test of neutrality based on interlocus associations. Genetics **146:** 1197–1206.

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

McVean, G. A. T., 2002 A genealogical interpretation of linkage disequilibrium. Genetics **162:** 987–991.

Maynard Smith, J., 1971 What use is sex? J. Theor. Biol. **30:** 319–335.

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

Nielsen, R., 2000 Estimation of population parameters and recombination rates using single nucleotide polymorphisms. Genetics **154:** 931–942.

Nielsen, R., and Z. Yang, 2003 Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA. Mol. Biol. Evol. **20:** 1231–1239.

Nordborg, M., and S. Tavaré, 2002 Linkage disequilibrium: what history has to tell us. Trends Genet. **18:** 83–90.

Parsch, J., C. D. Meiklejohn and D. L. 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.

Piganeau, G., and A. Eyre-Walker, 2003 Estimating the distribution of fitness effects from DNA sequence data: implications for the molecular clock. Proc. Natl. Acad. Sci. USA **100:** 10335–10340.

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

Przeworski, M., 2003 Estimating the time since the fixation of a beneficial allele. Genetics **164:** 1667–1676.

Quesada, H., U. E. M. Ramírez, J. Rozas and M. Aguadé, 2003 Large-scale adaptive hitchhiking upon high recombination in *Drosophila simulans*. Genetics **165:** 895–900.

Sabeti, P. C., D. E. Reich, J. M. Higgins, H. C. Levine, D. J. Richter *et al.*, 2002 Detecting recent positive selection in the human genome from haplotype structure. Nature **419:** 832–837.

Schlenke, T. A., and D. J. Begun, 2004 Strong selective sweep associ-

ated with a transposon insertion in *Drosophila simulans*. Proc. Natl. Acad. Sci. USA **101:** 1626–1631.

Smith, N. G. C., and A. Eyre-Walker, 2002 Adaptive protein evolution in *Drosophila*. Nature **415:** 1022–1024.

Stephan, W., 1995 An improved method for estimating the rate of fixation of favorable mutations based on DNA polymorphism data. Mol. Biol. Evol. **12:** 959–962.

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.

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

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

Vigouroux, Y., M. McMullen, C. T. Hittinger, K. Houchins, L. Schulz *et al.*, 2002 Identifying genes of agronomic importance in maize by screening microsatellites for evidence of selection during domestication. Proc. Natl. Acad. Sci. USA **99:** 9650–9655.

Watterson, G. A., 1975 On the number of segregating sites. Theor. Popul. Biol. **7:** 256–276.

Wootton, J. C., X. Feng, M. T. Ferdig, R. A. Cooper, J. Mu *et al.*, 2002 Genetic diversity and chloroquine selective sweeps in *Plasmodium falciparum*. Nature **418:** 320–323.

Communicating editor: J. B. Walsh