Detecting a Local Signature of Genetic Hitchhiking Along a Recombining Chromosome

Yuseob Kim[*,†] **and Wolfgang Stephan**[*,1]

[*]*Department of Evolutionary Biology, University of Munich, 80333 Munich, Germany and* [†]*Department of Biology, University of Rochester, Rochester, New York 14627*

Manuscript received May 10, 2001
Accepted for publication November 19, 2001

ABSTRACT

The theory of genetic hitchhiking predicts that the level of genetic variation is greatly reduced at the site of strong directional selection and increases as the recombinational distance from the site of selection increases. This characteristic pattern can be used to detect recent directional selection on the basis of DNA polymorphism data. However, the large variance of nucleotide diversity in samples of moderate size imposes difficulties in detecting such patterns. We investigated the patterns of genetic variation along a recombining chromosome by constructing ancestral recombination graphs that are modified to incorporate the effect of genetic hitchhiking. A statistical method is proposed to test the significance of a local reduction of variation and a skew of the frequency spectrum caused by a hitchhiking event. This method also allows us to estimate the strength and the location of directional selection from DNA sequence data.

THE level of genetic variation at a neutral locus can be influenced by natural selection at linked loci. The substitution of a strongly selected beneficial mutation produces a "hitchhiking" effect on the frequency of neutral alleles at linked loci (Maynard Smith and Haigh 1974; Kaplan *et al.* 1989; Stephan *et al.* 1992). Neutral variants are either lost or fixed along with the ancestral or beneficial allele at the selected locus unless recombination breaks down the association between neutral and selected alleles during the substitution process. As a result, genetic variation around the site of directional selection is greatly reduced by this hitchhiking event or "selective sweep." Selection against recurrent deleterious mutations also reduces variation at linked loci (Charlesworth *et al.* 1993). This mechanism, known as "background selection," causes the continuous removal of linked sequences along with deleterious mutations, resulting in a reduced effective population size.

To elucidate the relative contributions of selective sweeps and background selection in shaping the positive correlation between genetic variation and recombination (Begun and Aquadro 1992), recent investigations focused on the features of genetic variation where the two mechanisms make different predictions. A hitchhiking event produces an excess of rare alleles (Braverman *et al.* 1995; Fu 1997) and high-frequency-derived alleles (Fay and Wu 2000) in a sample of DNA sequences. It can also greatly reduce differentiation among subdivided populations (Stephan *et al.* 1998). These mechanisms are more efficient when the level of recombination between neutral and selected loci is reduced. Therefore, efforts to discover such phenomena were undertaken mainly for polymorphism data from regions of low recombination.

Another unique feature of genetic hitchhiking is the expected pattern of genetic variation along a recombining chromosome, *i.e.*, in regions of intermediate to high recombination rates. The reduction of genetic variation is greatest at the site of directional selection, but not as great at distant sites due to recombination. Therefore, it produces a "valley" of expected heterozygosity along the sequence. This pattern was used to demonstrate recent episodes of directional selection in populations (Benassi *et al.* 1999; Wang *et al.* 1999; Fullerton *et al.* 2000; Nurminsky *et al.* 2001).

Although the expected spatial pattern of variation along a chromosome caused by hitchhiking is straightforward, it is not certain whether it can be detected in a sample of DNA sequences. The size of the area affected by a single hitchhiking event can be very large if selection is strong or recombination rate is low. On the other hand, for relatively weak selection and high recombination rates, the size of the area might be sufficiently small to be detected in a survey of a gene of moderate length. However, the large variance of nucleotide diversity in a DNA sample makes it difficult to distinguish the pattern caused by a weak hitchhiking effect from a similar pattern generated randomly under neutral evolution with recombination. In the presence of recombination, different regions on a sequence have different genealogies whose sizes can differ considerably. Therefore, a local reduction of variation in a certain region of a recombining chromosome can happen by chance without hitchhiking events.

[1] *Corresponding author:* Department of Evolutionary Biology, University of Munich, Luisentstr. 14, 80333 Munich, Germany. E-mail: stephan@zi.biologie.uni-muenchen.de

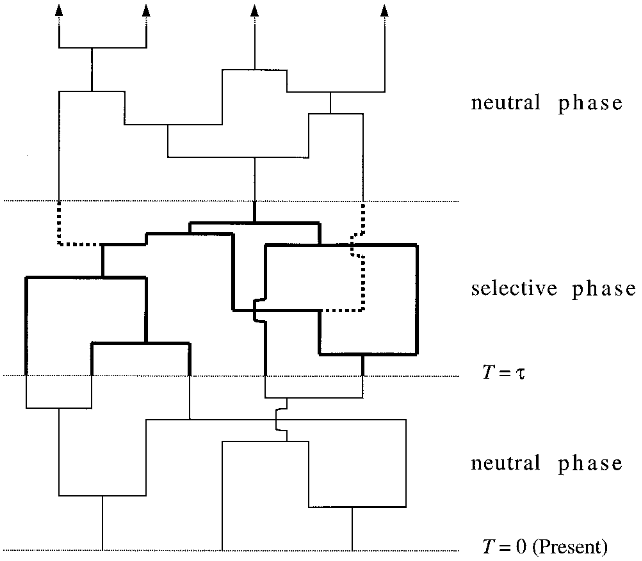

Figure 1.—Ancestral recombination graph with genetic hitchhiking for a sample of $n = 3$. The graph is constructed from the bottom ($T = 0$) to top ($T > 0$). Vertical edges represent gene lineages that contain the "ancestral material" (Wiuf and Hein 1999). Nodes where two edges join into one define coalescences. Nodes where one edge splits into two define recombinations. At each recombination node, the recombinational break point (U) is specified. At the beginning of the selective phase ($T = \tau$), all edges change into B edges (thick solid lines). Some recombination nodes produce b edges (dashed lines). There is only one B edge at the end of the selective phase, after which the distinction between B and b edges is erased. The construction of the ARG continues until $T = T_{\text{limit}}$.

In this article, we investigate the pattern of genetic variation resulting from a single hitchhiking event on a recombining chromosome. A likelihood-based statistical test is developed to evaluate the significance of a local reduction of variation. It is also examined if the strength and location of directional selection can be estimated from DNA sequence data.

COALESCENT SIMULATION

This study requires a coalescent simulation in which both intragenic recombination and directional selection take place during the ancestry of a DNA sample. The ancestral recombination graph (ARG) described by Griffiths and Marjoram (1997) allows the realization of intragenic recombination for any length of a DNA sequence. We modified the ARG to incorporate the effect of genetic hitchhiking. Figure 1 illustrates an ARG with hitchhiking for a sample of three sequences. The ancestral history of the sample is divided into neutral phases and a selective phase.

During a neutral phase the ARG is constructed as described by Griffiths and Marjoram (1997), with the following modifications. Edges of the ARG represent ancestral sequences at a given time (T) in the past measured in units of $2N$ generations, where N is the effective number of individuals in a diploid population. If there are k edges at a given time ($k = n$ at $T = 0$), either coalescent or recombination events can occur with rates $k(k - 1)/2$ and $kR/2$, respectively. R is $4N$ times the recombination rate between both ends of the sequence. Therefore, if the sequence is L nucleotides long, $R = 4NL\rho$, where ρ is the per-nucleotide recombination rate. Each of the k edges is labeled by a pair of integers (I_i, J_i) $(i = 1, \ldots, k)$. This pair of integers delimits the region within which sequences ancestral to sample sequences are found. Therefore, recombination outside this region can be ignored. At $T = 0$, $(I_i, J_i) = (1, L)$ for all n edges. At a coalescence event, two randomly chosen edges (for example, the lth and mth edges) join to a new edge, which is then labeled by $(\text{Min}(I_l, I_m), \text{Max}(J_l, J_m))$. At a recombination event, an edge is chosen randomly and a random uniform integer, U, is drawn between 1 and $L - 1$. If an edge labeled by (I, J) was chosen, it joins to two parental edges only if $I \leq U < J$. Then, the two parental edges are labeled as (I, U) and $(U + 1, J)$. If $U < I$ or $> J$, no change is made at the edge. This procedure is necessary to minimize k in the simulation.

The selective phase is the period when a substitution of a beneficial mutation that causes a hitchhiking effect takes place. The beneficial allele B has a genic selective advantage s over the parent allele b. This substitution occurs at a site M nucleotides away from the left end of the sequence and the fixation of B is completed at $T = \tau$. The allele frequency of B, x, is assumed to change deterministically from $1 - \xi$ to ξ. Therefore, x at $T = \tau + t$ is given by

$$x(t) = \frac{\xi}{\xi + (1 - \xi)e^{\alpha(t - t_s)}} \quad (0 \leq t \leq t_s) \qquad (1)$$

(Stephan *et al.* 1992), where $\alpha = 2Ns$ and $t_S = -(2/\alpha)\log(\xi)$, which is the length of the selective phase. The choice of ξ does not change the resulting genealogy significantly (Braverman *et al.* 1995). We use $\xi = 100/(2N)$ for the simulations. During the selective phase, B and b edges exist, indicating whether an ancestral sequence includes the beneficial allele or not. Therefore, all edges are B edges at the beginning of the selective phase ($T = \tau$). The system of labeling edges is also changed: (I, J) at the end of the neutral phase at $T = \tau$ changes to $(\text{Min}(I, M), \text{Max}(J, M))$. This change means that recombination between the site of directional selection and the ancestral sequence should be followed during the selective phase. There are four possible events during the selective phase: (1) coalescence between B edges; (2) coalescence between b edges; (3) recombination in a B edge; (4) recombination in a b edge. The probability of these four events during the time interval $[t, t + \Delta t]$ is given by

$$\lambda_1(t)\Delta t = \frac{k_B(k_B - 1)}{2x(t)}\Delta t, \tag{2a}$$

$$\lambda_2(t)\Delta t = \frac{k_b(k_b - 1)}{2(1 - x(t))}\Delta t, \tag{2b}$$

$$\lambda_3(t)\Delta t = \frac{k_B R}{2}\Delta t, \tag{2c}$$

$$\lambda_4(t)\Delta t = \frac{k_b R}{2}\Delta t, \tag{2d}$$

respectively, where k_B and k_b are the numbers of B and b edges at $T = \tau + t$, respectively. The waiting time, Δt, between events is randomly drawn from an exponential distribution with parameter $\lambda_T(t) = \lambda_1(t) + \lambda_2(t) + \lambda_3(t) + \lambda_4(t)$. Then, one of the four events is allowed to occur according to its probability. This method should be used only when waiting time, Δt, is short ($\ll 1/\alpha$) such that the change of $x(t)$ between events is negligible. In this study, due to large values of R, values of Δt are sufficiently small. With lower values of R, a rejection method such as the one by Braverman *et al.* (1995) should be used. When a recombination event occurs in a B edge with (I, J), it joins to two parental edges only if a random integer U is between I and J. Then, one parental edge is labeled by (Min(I, M), Max(U, M)) and the other one by (Min$(U + 1, M)$, Max(J, M)). If $M \leq U$, the former parental edge must become a B edge, since the beneficial allele has descended from the ancestral sequence in this edge. The other parental edge, however, becomes either a B edge with probability $x(t)$ or a b edge with probability $1 - x(t)$. Likewise, if $M > U$, the parental edge with ($U + 1$, Max(J, M)) becomes a B edge, with the other parental edge becoming either B or b. The same principle is applied to a recombination event in a b edge. The selective phase, which ends when $x(t) \leq \xi$ or the combined number of B and b edges becomes 1, is followed by another neutral phase where the distinction between B and b edges is erased.

The coalescent for each nucleotide site (or the "marginal tree") is embedded in the ARG. The marginal tree is extracted as described in Griffiths and Marjoram (1997). The number of edges, k, at a given time changes stochastically during the construction of the ARG. Theoretically, k eventually hits 1, and the ARG is completed. However, if $R \gg 10$, k fluctuates at high numbers, and waiting until $k = 1$ in the simulation is then practically impossible. We therefore stop the construction of the ARG at an arbitrary point, T_{limit}, which is chosen to be a large number relative to the mean time to the most recent common ancestor (MRCA) for a marginal tree. If a marginal tree cannot find the MRCA until T_{limit}, the remaining branches of the tree are forced to coalesce at T_{limit}. This procedure has a negligible effect on our results since most nucleotide sites find their MRCAs before T_{limit}. $T_{\text{limit}} = 7.0$ was used in this study. Source codes written in C for the simulation of ARGs and other procedures described in this article are available upon request.

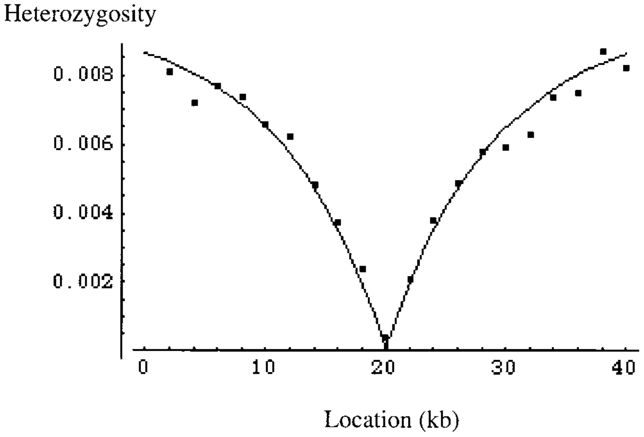

Figure 2.—Average nucleotide diversity along a recombining chromosome under the model of genetic hitchhiking, with $n = 5$, $\rho = 10^{-8}$, $N = 2 \times 10^5$, $\theta = 0.01$, and $T_{\text{limit}} = 7.0$. Squares represent average heterozygosity at single nucleotide sites averaged over 50,000 replicates of the simulations. The expected π value (continuous curve) was calculated using Equation 13 of Kim and Stephan (2000), with $r = \rho\ |i - 20{,}000|$ as the recombination rate between a nucleotide site i and the site of selection. Directional selection occurs at position 20 kb with $s = 0.001$ and $\tau = 0.005$.

PATTERNS OF GENETIC VARIATION ALONG A CHROMOSOME WITH HITCHHIKING

The simulated patterns of sequence polymorphism are obtained by introducing mutations into the marginal tree for each nucleotide site. To verify that the simulation procedure generates the correct ancestral genealogy expected under the model of hitchhiking, nucleotide diversities at many fixed sites along the sequence were summarized over 50,000 replicates of the ARG for a set of parameters (Figure 2). The simulation results agreed well with the expectation on the basis of the analytic solutions by Stephan *et al.* (1992) and Kim and Stephan (2000). Without the selective phase, the mean and variance of coalescent times of marginal trees agreed well with the expectation under the standard neutral model (data not shown). The number of shifts from one MRCA to another along the sequence was also observed and compared to the prediction (Equation 5 of Wiuf and Hein 1999). With $R = 20$ ($2N = 10^5$, $L = 10^4$, and $\rho = 10^{-8}$), the observed numbers of shifts for sample sizes 2 and 10, each averaged over 200 replicates, were 13.72 and 19.74, respectively. The corresponding expectations are 13.33 and 19.64, respectively.

We assume that the derived allele can be distinguished from the ancestral allele, which is defined to

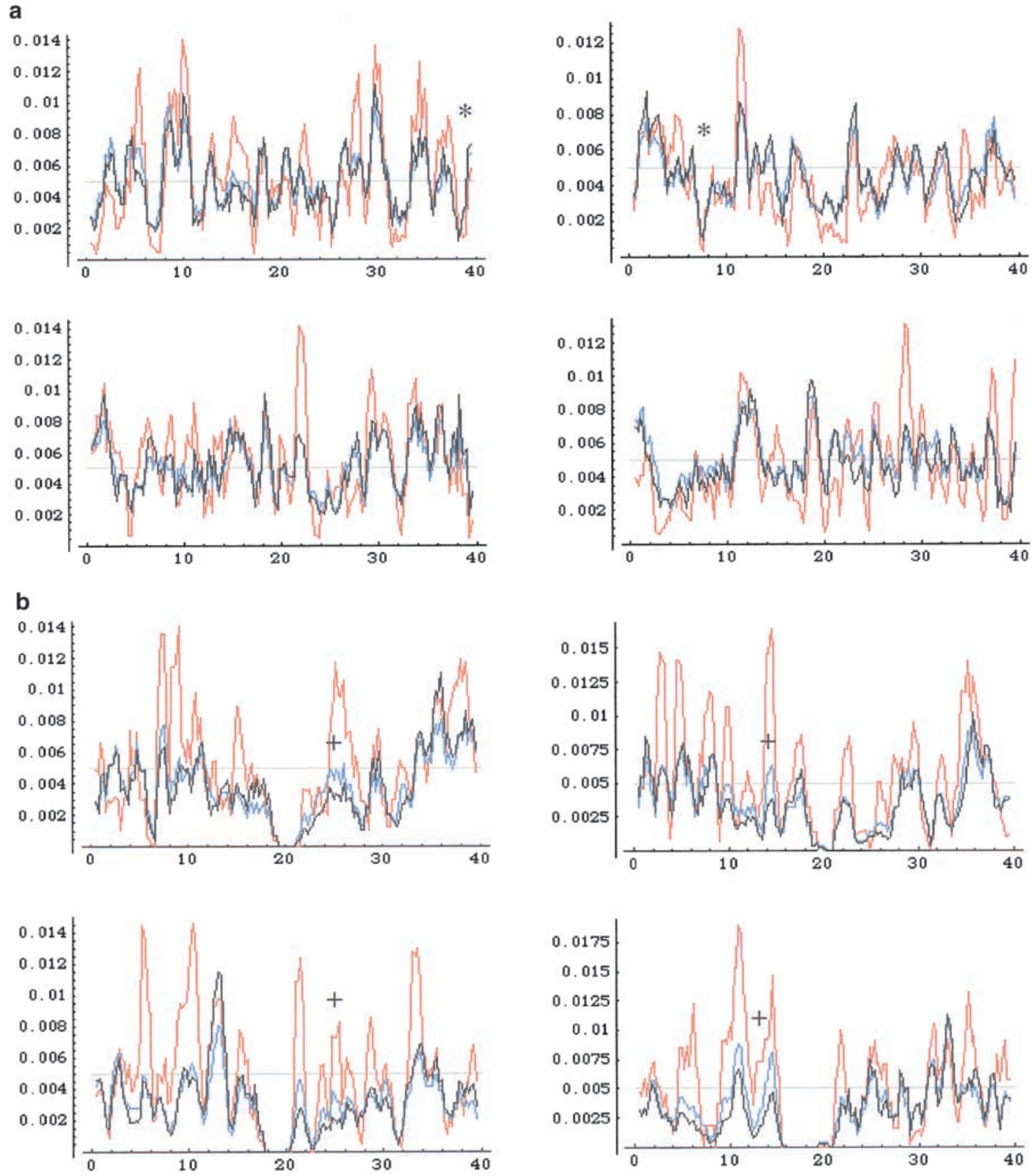

Figure 3.—Patterns of genetic variation along a recombining chromosome. Simulated data were generated with (a) a neutral model with $N = 5 \times 10^5$ ($R = 800$); hitchhiking models with (b) $N = 5 \times 10^5$, $s = 0.001$, $R = 800$, $\alpha = 1000$, and $\tau = 0.001$; (c) $N = 5 \times 10^5$, $s = 0.001$, $R = 800$, $\alpha = 1000$, and $\tau = 0.2$; and (d) $N = 5 \times 10^4$, $s = 0.001$, $R = 80$, $\alpha = 100$, and $\tau = 0.001$. The values of the other parameters are $n = 10$, $\rho = 10^{-8}$, $\theta = 0.005$, and $T_{\text{limit}} = 7.0$. Selection occurs at position 20 kb. For each model, four replicates are shown. $\hat{\theta}_\pi$ (black), $\hat{\theta}_W$ (blue), and $\hat{\theta}_H$ (red) were calculated along the sequence with window size 1 kb and step size 0.25 kb. Positions on the sequence are shown in units of 1 kb. *, ↓, and + indicate the positions of specific features described in the text.

be the allele at the root of the marginal tree. If more than one mutant is segregating at one site, all mutant alleles are classified as the derived allele and not distinguished from each other. To examine the pattern of variation, three different estimators [$\hat{\theta}_\pi$ (Tajima 1983), $\hat{\theta}_W$ (Watterson 1975), and $\hat{\theta}_H$ (Fay and Wu 2000)] of $\theta = 4N\mu$ were calculated for the simulated sequences. Differences among the three estimators reveal deviations from neutrality (Tajima 1989; Fay and Wu 2000). Figure 3 shows the values of the three estimators for sliding windows along the sequence. Four replicates are shown for each set of parameter values, where each replicate was produced from an ARG of a 40-kb sequence. The ARGs were generated using the neutral

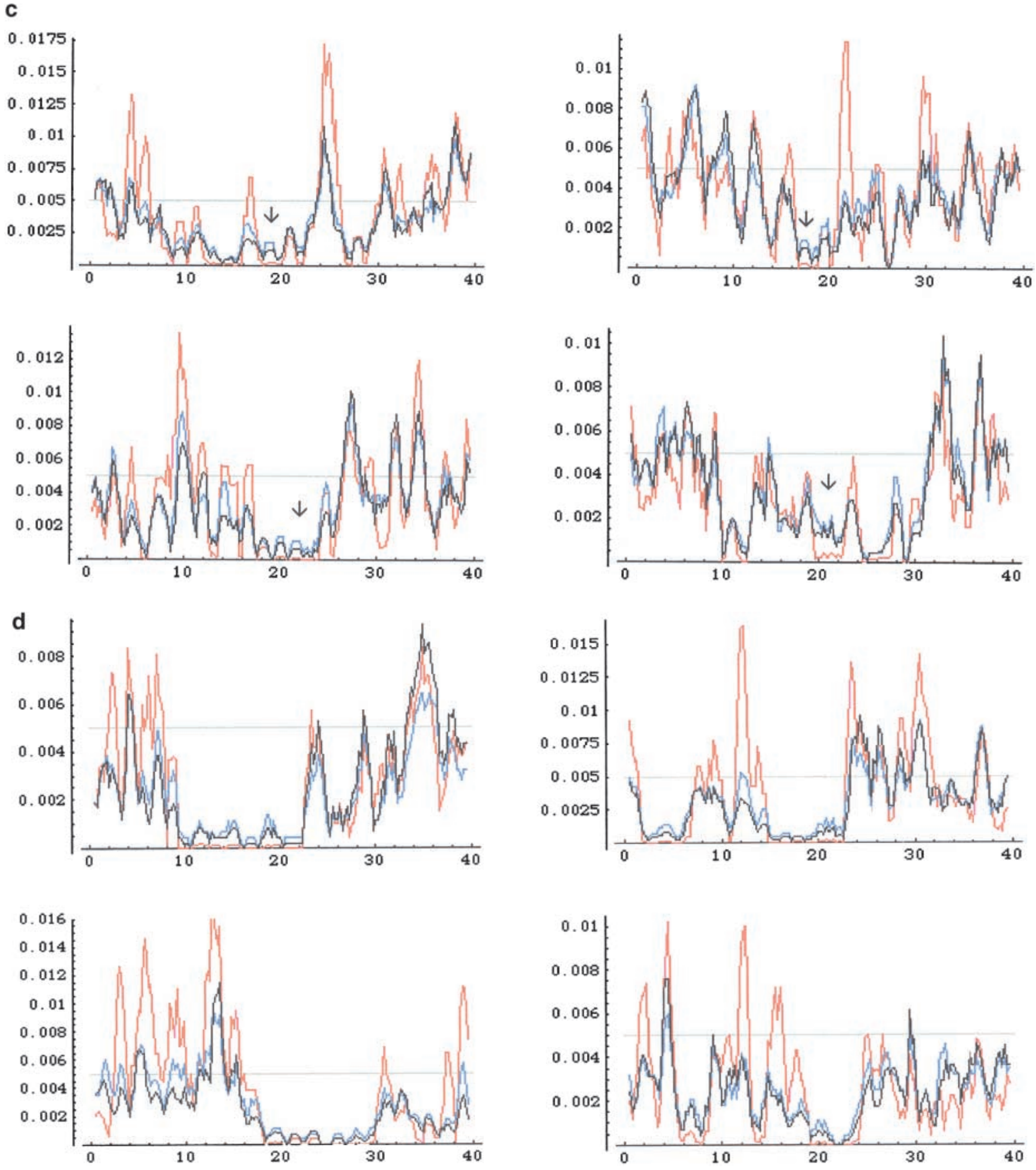

Figure 3.—*Continued.*

model and the hitchhiking model with $s = 0.001$ ($\alpha = 100$ or 1000) and $\tau = 0.001$ or 0.2. As only four examples randomly chosen from the simulations are shown for each model, one may not be allowed to draw a general conclusion from these figures. However, some features of hitchhiking effects on sequence variation could be consistently identified from these examples. We use these examples mainly to illustrate these features.

A local reduction or valley of heterozygosity ($\hat{\theta}_\pi$) along the sequence is the most important pattern expected under the model of genetic hitchhiking. Under neutral evolution (Figure 3a), the stochastic change of $\hat{\theta}_\pi$ along the sequence occasionally generates deep valleys of variation (for example, regions indicated by *). However, in this case valleys are usually narrow compared to those under hitchhiking (Figure 3, b–d). The stochastic spatial pattern of variation is influenced by R. Using the same parameter values as in Figure 3a but smaller N, deeper and wider valleys were frequently observed (data not shown). With hitchhiking (Figure 3, b–d), a deep valley always appears at or around the site of directional selection. However, the "shape" of the valleys varies considerably among realizations for a given value of s. Valleys are rather asymmetrical around the site of the beneficial mutation, which implies that the shape of the valley may provide imprecise information about the location of the target of selection (see below). The asymmetry gets larger as N decreases. Figure 3d uses the same values of s, τ, and ρ as Figure 3b but a 10 times smaller N. As a result, the stochastic noise in the spatial pattern has been dramatically increased.

Compared to neutrality (Figure 3a), the relative level of $\hat{\theta}_H$ versus $\hat{\theta}_\pi$ increased immediately after the hitchhiking event (Figure 3b), as expected by Fay and Wu (2000). However, at $\tau = 0.2$ (Figure 3c), *i.e.*, $0.4N$ generations after the hitchhiking event, a higher relative level of $\hat{\theta}_H$ as shown in Figure 3b is not observed. Especially, $\hat{\theta}_H$ is distinctively lower than $\hat{\theta}_\pi$ around the site of selection where the level of nucleotide diversity has only partially recovered since the selective sweep (regions labeled by ↓). This is consistent with the observation that the excess of high frequency variants appears suddenly after a hitchhiking event but soon disappears through the fixation of these alleles, reversing the excess of high frequency mutants (Kim and Stephan 2000). $\hat{\theta}_W$ is expected to become larger than $\hat{\theta}_\pi$ due to hitchhiking (Braverman *et al.* 1995). In Figure 3, the increase of $\hat{\theta}_W$ is not as obvious as in the case of $\hat{\theta}_H$. However, we could identify regions where $\hat{\theta}_W$ became distinctively larger than $\hat{\theta}_\pi$ in Figure 3b (labeled by +). The generality of these observations drawn from the examples of Figure 3 is further investigated by a statistical test applied to larger simulated datasets (see below).

The stronger the hitchhiking effect, the larger is the region that is expected to be affected. To find the relationship between the mean length of the region of reduced variation and the parameter values of the hitchhiking model, we generated 200 simulated datasets for a fixed combination of N, ρ, s, τ, and θ. A 1-kb-long window moves from the left end of the 40-kb-long sequence with an increment of one nucleotide and calculates $\hat{\theta}_\pi$ at each position. Regions of reduced variation are defined by the centers of windows for which $\hat{\theta}_\pi < \theta/2$. Therefore, a segment of the affected area is delimited by the centers of the two windows that mark the beginning and the end of the stretch of nucleotides with $\hat{\theta}_\pi < \theta/2$. The length of the longest of such segments found on the 40-kb sequence is defined as $W_{\theta/2}$. The mean and standard deviation of $W_{\theta/2}$ over 200 replicates are shown in Table 1. The proportion, p_{within}, of the simulated datasets in which the site of the beneficial mutation is included in the largest segment that defines $W_{\theta/2}$ is also recorded (Table 1). Examples 1–3, 7, and 8 show that an increase of the mutation rate per nucleotide, given by θ, does not lead to a proportional decrease in $W_{\theta/2}$. Therefore, the mutation rates used in Table 1 are high enough to "saturate" and reveal the underlying stochastic patterns of coalescent times along the sequence. As expected, $W_{\theta/2}$ with hitchhiking is significantly larger than that without. The mean of $W_{\theta/2}$ is roughly proportional to s/ρ, as expected from the solutions of the hitchhiking effect (Maynard Smith and Haigh 1974; Stephan *et al.* 1992). As predicted by these solutions, population size N has also an effect on $W_{\theta/2}$; $W_{\theta/2}$ decreases with increasing N (examples 6 and 12). However, this effect is not as large as that determined by the parameter s/ρ, in particular for large values of N (Equation 19 in Stephan *et al.* 1992). Barton (2000) offered the following explanation of this effect of N. A beneficial mutation in a large population takes longer to reach a high frequency, by which time the lineages originally associated with it have become separated by recombination. This is equivalent to the statement that the "effective" length of the selective phase, $-(2/s)\log(\varepsilon)$ generations, is longer in large populations, where ε ($\approx 1/(2Ns)$) is the frequency of the beneficial mutation when it starts increasing deterministically. Looking backward in time, the rate of coalescence for two gene lineages during the selective phase gets sufficiently high only when the product of population size and beneficial allele frequency becomes low. Therefore, the waiting time (in generations) until the coalescence event is longer in large populations. However, the recombination rate per generation is independent of the population size. Therefore, the probability of the recombination event being the first event is higher in a larger population. This explains the smaller effect of a single hitchhiking event in a larger population as shown in Table 1.

Examples 6 and 7 show that almost identical $W_{\theta/2}$'s are obtained with the same values of $N\rho$ and α, but with different ρ and s values. Therefore, for a given τ, $N\rho$ and α are the two principal parameters governing the pattern of variation caused by a hitchhiking event. We compared the mean $W_{\theta/2}$ to the theoretical prediction, $E[W_{\theta/2}]$ (Table 1), which is based on the expectation of $\hat{\theta}_\pi$ along the sequence from Equation 13 of Kim and Stephan (2000). Mean $W_{\theta/2}$ is consistently smaller than $E[W_{\theta/2}]$. This discrepancy occurs partly because the calculation of $E[W_{\theta/2}]$ assumes that the duration of the selective phase is negligible on the timescale of $2N$ generations. However, the lengths of the selective phase, $t_S = -(2/\alpha)\log(\xi)$, are 0.076 and 0.017 for examples 6 and 12, respectively, which are not much smaller than $\tau = 0.05$. In the early part of the selective phase (when the frequency of the beneficial mutation is high), the behavior of the genealogy is similar to that in the neutral phase. Therefore, the length of the first neutral phase, *i.e.*, the time since the last hitchhiking event, is effectively longer than τ. It should be noted that the standard deviation of $W_{\theta/2}$ is considerably large, as suggested by Figure 3. Furthermore, in >20 of 200 realizations, the site of the beneficial mutation is not included in the largest segment of reduced variation (see p_{within} in Table 1) even with strong hitchhiking (examples 11 and 12). These results again indicate a large amount of stochas-

TABLE 1

Features of genetic variation with and without hitchhiking

Examples	N	$N\rho$	α	τ	θ	$W_{\theta/2}$ (kb)[a]	$E[W_{\theta/2}]$[b]	p_{within}[c]	$\overline{r^2}$[d]	$S_{\max}$[e]
1	10^5	10^{-3}	0	—	0.002	1.91 ± 0.89	—	—	0.139 ± 0.013	5.04 ± 1.93
2	10^5	10^{-3}	0	—	0.005	1.53 ± 0.80	—	—	0.141 ± 0.015	8.60 ± 3.28
3	10^5	10^{-3}	0	—	0.01	1.48 ± 1.09	—	—	0.140 ± 0.015	11.8 ± 3.9
4	10^5	10^{-3}	40	0.05	0.005	3.46 ± 1.81	4.33	0.58	0.141 ± 0.014	9.06 ± 3.30
5	10^5	10^{-3}	100	0.05	0.005	7.04 ± 3.78	8.24	0.84	0.146 ± 0.014	9.94 ± 3.39
6	10^5	10^{-3}	200	0.05	0.005	10.4 ± 4.9	14.0	0.885	0.162 ± 0.022	12.5 ± 5.3
7	10^6	10^{-3}	200	0.05	0.005	10.2 ± 4.7	14.0	0.845	0.161 ± 0.018	11.7 ± 4.7
8	10^6	10^{-3}	200	0.05	0.01	11.3 ± 5.6	14.0	0.875	0.157 ± 0.022	15.2 ± 5.2
9	5×10^5	5×10^{-3}	0	—	0.005	0.95 ± 0.46	—	—	0.120 ± 0.006	5.12 ± 1.51
10	5×10^5	5×10^{-3}	500	0.05	0.005	4.96 ± 2.36	5.80	0.835	0.123 ± 0.006	7.20 ± 2.72
11	5×10^5	5×10^{-3}	1000	10^{-3}	0.005	8.27 ± 3.51	11.1	0.91	0.147 ± 0.015	12.2 ± 6.4
12	5×10^5	5×10^{-3}	1000	0.05	0.005	7.79 ± 3.07	10.3	0.88	0.127 ± 0.007	8.52 ± 3.40
13	5×10^5	5×10^{-3}	1000	0.2	0.005	6.23 ± 2.92	7.87	0.835	0.120 ± 0.005	5.82 ± 1.80

Results are based on 100 replicates for each parameter set. For all simulations, $L = 40$ kb, $n = 10$, $T_{\text{limit}} = 7.0$, and $\xi = 50/N$.

[a] Means ± standard deviations of $W_{\theta/2}$. See text for definitions.

[b] Prediction of $W_{\theta/2}$ using Equation 13 of Kim and Stephan (2000).

[c] Proportion of simulated data in which the site of the beneficial mutation is included in the region of low variation that defines $W_{\theta/2}$.

[d] Means ± standard deviations of $\overline{r^2}$.

[e] Means ± standard deviations of $S_{\max}$.

ticity in the pattern of variation shaped by hitchhiking effects.

After a selective sweep, the level of genetic variation is slowly restored due to new neutral mutations. Therefore, with given values of $N\rho$ and α, $W_{\theta/2}$ should become smaller with increasing τ, as is indeed observed in examples 11–13, for which $\tau = 0.001$, 0.05, and 0.2, respectively. The level of variation around the site of selection, which is zero immediately after the sweep, should be characterized by τ (Wiehe and Stephan 1993). As the site of selection is expected to be found at the center of the largest segment of reduced variation (defined above), we examined the average nucleotide diversity, $\hat{\theta}^*_\pi$, for the middle one-third of this segment. The average values (±standard deviation) of $\hat{\theta}^*_\pi$ for examples 11–13 are 3.88×10^{-4} ($\pm 4.09 \times 10^{-4}$), 5.48×10^{-4} ($\pm\ 3.85 \times 10^{-4}$), and 9.85×10^{-4} ($\pm 3.86 \times 10^{-4}$), respectively. The corresponding theoretical values obtained by numerical integration of Equation 13 of Kim and Stephan (2000) are 5.45×10^{-4}, 7.24×10^{-4}, and 1.23×10^{-3}, respectively. Therefore, the "depth" of the valley of reduced variation contains information about the time of the last hitchhiking event. However, the large standard deviations of $\hat{\theta}^*_\pi$ suggest that a correct estimation of τ from polymorphism data will be very difficult.

So far the patterns of genetic variation based on the segregation at single sites were examined. Another important aspect of sequence variation is the association of polymorphisms between neighboring loci. We calculated r^2 (Hill and Robertson 1968) for all pairs of segregating sites in each simulated set and obtained the average, $\overline{r^2}$, over the entire 40-kb region. Hitchhiking caused an increase in $\overline{r^2}$ (Table 1). It might be possible that this increase in $\overline{r^2}$ was caused by the excess of rare alleles (*i.e.*, singletons) generated by the hitchhiking effect, because r^2 frequently becomes large by chance when the allele frequencies at both loci are extreme. Unfortunately, we could not exclude singletons from the analysis since not many segregating sites are left if singletons are removed from the data generated under the hitchhiking model. However, a visual inspection of the raw hitchhiking data reveals that there are several extensive haplotype structures that are not likely to be created by chance alone. Large stretches of polymorphic sites share an identical pattern of segregation; *i.e.*, there are only two haplotypes observed in such a stretch. We recorded the maximum number, $S_{\max}$, of such consecutive sites found in each dataset. Table 1 shows that $S_{\max}$ increases with hitchhiking. With strong selection, the increase of $S_{\max}$ is very large (for instance, compare examples 5 and 11). The increase of $\overline{r^2}$ and $S_{\max}$ by hitchhiking can be explained by a coalescent argument. Ancestral histories of two neutral loci become identical if no recombination event occurs before the MRCA for both loci is found (Griffiths and Marjoram 1997; Wiuf and Hein 1999). During the selective phase the rate of coalescence is increased while that of recombination remains the same. Therefore, hitchhiking removes opportunities for recombination between two linked loci. This leads to an increase in the correlation of gene genealogies between segregating sites around the site of directional selection. However, the buildup of linkage disequilibrium due to hitchhiking quickly decays as τ increases (examples 11–13), because recombination events during that neutral phase break up associations created in the selective phase.

STATISTICAL TEST OF A LOCAL SIGNATURE CAUSED BY GENETIC HITCHHIKING

In the following, a maximum-likelihood method is developed to examine the significance of a local reduction of genetic variation and to estimate the strength of directional selection. The probability of observing a certain frequency of derived alleles at a site after a recent hitchhiking event can be obtained by previously used analytic approximations. Under neutrality, the expected number of sites where the derived variant is in the frequency interval $[p, p + dp]$ in the population is given by

$$\phi_0(p)\,dp = \frac{\theta}{p}dp \tag{3}$$

(Kimura 1971). Immediately after a hitchhiking event, this distribution is transformed approximately to

$$\phi_1(p) = \begin{cases} \dfrac{\theta}{p} - \dfrac{\theta}{C} & \text{for } 0 < p < C \\ \dfrac{\theta}{C} & \text{for } 1 - C < p < 1 \end{cases} \tag{4}$$

(Fay and Wu 2000), where C is given approximately by $1 - \varepsilon^{r/s}$ (appendix). Here, r is the recombination fraction between the neutral locus and the selected locus and ε is the frequency of the beneficial allele when it begins to increase deterministically. It should be noted that Equation 4 is obtained by assuming deterministic changes of allele frequencies during the selective phase. The probability of observing a site where k derived alleles are found in a sample of size n is given by

$$P_{n,k} = \int_0^1 \binom{n}{k} p^k (1 - p)^{n-k} \phi(p)\,dp \quad (k = 1, \ldots, n - 1) \tag{5}$$

and

$$P_{n,0} = 1 - (P_{n,1} + \ldots + P_{n,n-1}),$$

where $\phi(\cdot) = \phi_0(\cdot)$ under the neutral model, and $\phi(\cdot) = \phi_1(\cdot)$ under the hitchhiking model. $P_{n,k}$ was found to be sensitive to the choice of ε. We used $\varepsilon = 1/\alpha$, which gave the best fit to the simulation results (Figure 4). The likelihood of all data under the model of genetic hitchhiking is obtained by multiplying the probabilities for all nucleotide sites under consideration. This is a composite likelihood because there is a correlation of $P_{n,k}$ between sites due to shared ancestral histories. Therefore, it should be distinguished from the conventional likelihood-ratio test that is based on exact likelihoods. A statistical test in this analysis thus depends on an empirical distribution of the test statistic obtained by simulation. Composite likelihood is frequently used when the derivation of exact likelihoods is difficult (*e.g.*, Rannala and Slatkin 2000). It should also be noted that higher-order structures in the polymorphism data, such as linkage disequilibria, are neglected in this analysis.

As it is currently unrealistic to have polymorphic data from a reasonably large sample of long continuous sequences, we apply the test to a region for which only short segments are sequenced, interspaced with larger nucleotide stretches for which no data are available. That is, we consider a survey in which 11 1-kb-long segments distributed over a 40-kb region are sequenced. The distances between segments are uniformly 2.9 kb. The sample size is 10 for all segments. Simulated data used for Table 1 (examples 9 and 11–13) were reused, but only sites from the 11 segments were included. The maximum composite likelihood under the neutral model (L_0) and that under the hitchhiking model (L_1) were obtained for each simulated dataset. Then, the likelihood ratio is given by L_1/L_0. L_0 is a function of θ and L_1 is a function of N, μ, ρ, s, and the location of the selected locus, X. X is allowed to vary in the middle 10-kb region of the sequence (15 kb $< X <$ 25 kb); *i.e.*, we consider a situation where a candidate region for the site of selection has already been inferred. It is difficult practically to allow all these parameters to vary freely until a unique combination that maximizes L_1 is found. Therefore, we chose only s and X as free variables and assumed that separate estimates of N, μ, and ρ are available. Thus, in one test (test A), the same values of N, μ, and ρ specified in the simulation are used in the calculation of L_0 and L_1. In the other test (test B), to be conservative, we let the mutation rate be inferred from the data by using the average heterozygosity ($\hat{\theta}_\pi$) over all 11 segments of the sequence as the fixed prior estimate of θ. Therefore, the standing level of variation is simply the level observed in the data. But we still used the true value of N for the calculation of $\alpha = 2Ns$. We also assumed either that the derived neutral allele is distinguished from the ancestral allele at each site (option 1) or that they are not distinguished (option 2). For the latter, there are only five ratios of segregating variants in the sample of 10 sequences. Let $Q_{k,n}$ ($k = 1, \ldots, n/2$) be the probability of observing a $[k(n - k)$ segregation ratio. Then the likelihood ratios are calculated simply by using $Q_{n,k} = P_{n,k} + P_{n,n-k}$.

The null distribution of likelihood ratios was obtained by applying tests to datasets generated under the neutral model (200 replicates corresponding to example 9 of Table 1). To reduce the problem of local optima, eight different initial guesses of X between 15 and 25 kb, with $s = 0.01$, were used to start the maximization procedure using Powell's method (Press *et al.* 1992; program writ-

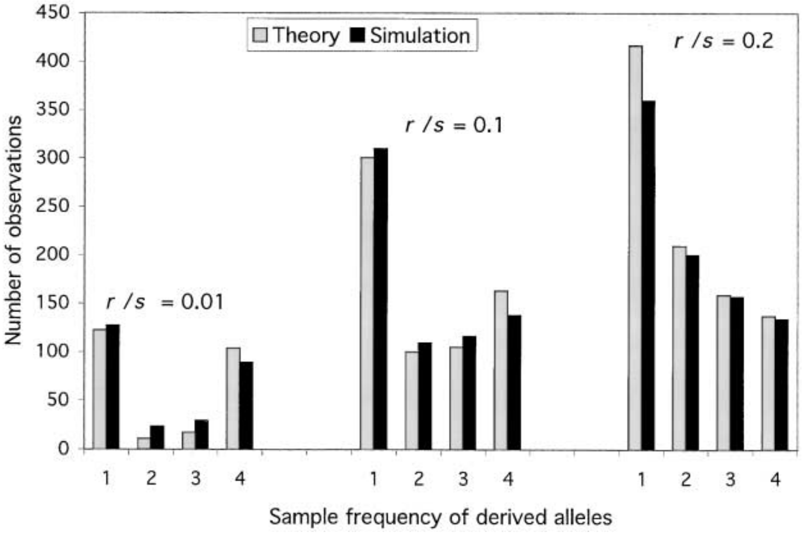

FIGURE 4.—Frequency of derived alleles at individual nucleotide sites in a sample of five sequences under genetic hitchhiking. Three nucleotide sites were chosen such that $r/s = 0.01$, 0.1, and 0.2. For each site, the numbers of occurrences of one to four derived alleles were recorded out of 50,000 replicates. The values of the other parameters are $\theta = 0.01$, $N = 2 \times 10^5$, and $s = 0.001$. Theoretical expectations were obtained using Equation 5, with $\varepsilon = 1/\alpha$.

ten by D. L. Swofford and kindly provided by J. P. Huelsenbeck). The null distribution of $\log(L_1/L_0)$ for test A/option 1 is shown in Figure 5. All test methods (tests A and B and options 1 and 2) produced almost identical null distributions (data not shown). About 30% of the $\log(L_1/L_0)$ values are negative. This may suggest a failure of obtaining the global maximum of L_1 for some null datasets, since L_1, with two more free variables, should be larger than L_0. However, increasing the number of initial guesses did not improve the likelihood ratios (data not shown). The negative values are more likely due to the restricted range of s and the assumption of $\tau = 0$. As α should be large enough to generate a hitchhiking effect (also $\varepsilon = 1/\alpha$ should remain small), s was not allowed to be $<6 \times 10^{-5}$ ($\alpha = 60$). Only when s is very small, the hitchhiking model with $\tau = 0$ can fit neutral data in which a local reduction of variation is not found. Therefore, there was a limit in maximizing L_1.

Table 2 summarizes the power of the test and the point estimates of s and X for each hitchhiking model. Power is the proportion of replicates that produce $\log(L_1/L_0)$ values greater than the 95th percentile of the corresponding null distribution. Test A yielded very high power of rejecting neutral evolution, even for larger values of τ. The main reason for obtaining large likelihood ratios from test A is that it uses the "true" standing level of variation (θ) for the calculation of L_1 and L_0. As the average heterozygosity has been reduced below θ due to selective sweep, the neutral model based on the true value of θ cannot fit the data. The negligible differences in the power between options 1 and 2 (except at $\tau = 0.2$) indicate that the additional information obtained by distinguishing between ancestral and derived alleles contributed little in test A, whereas a significant reduction of heterozygosity played a major role in increasing the likelihood ratio.

On the other hand, a reduction of heterozygosity is not a major factor for increasing the likelihood ratio in test B. To obtain a higher likelihood ratio in this test for a given number of segregating sites, the spatial distribution of those sites along the sequence and the allele

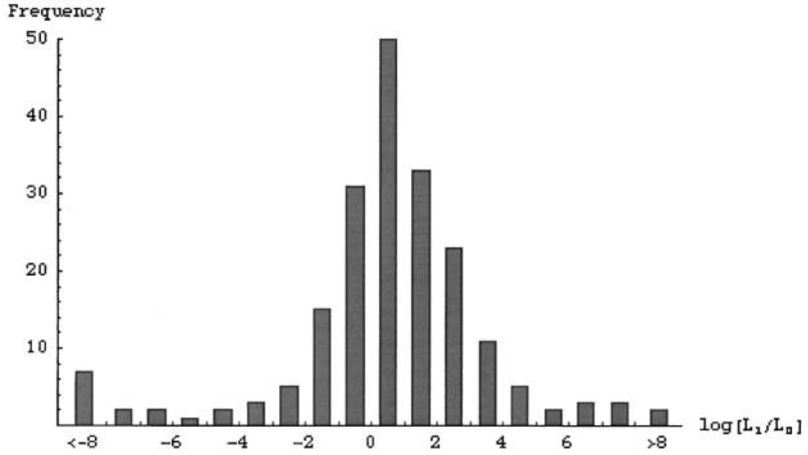

FIGURE 5.—Distribution of $\log(L_1/L_0)$ obtained by applying test A/option 1 to neutral datasets ($R = 800$).

TABLE 2

Results of likelihood-ratio tests (11 × 1-kb segments)

		Test A			Test B		
Simulation model	Option	Power	$\hat{s}$ ($\times 10^{-3}$)[a]	$\hat{X}$ (kb)[a]	Power	$\hat{s}$ ($\times 10^{-3}$)	$\hat{X}$ (kb)
Neutral	1	—	0.105 ± 0.056	19.75 ± 2.29	—	0.099 ± 0.045	19.77 ± 2.33
	2	—	0.139 ± 0.095	19.88 ± 2.34	—	0.118 ± 0.068	20.04 ± 2.37
$s = 10^{-3}$, $\tau = 10^{-3}$	1	0.995	1.16 ± 0.61	20.09 ± 2.07	0.97	0.726 ± 0.326	20.14 ± 2.05
	2	0.975	1.23 ± 0.67	20.07 ± 2.00	0.925	0.604 ± 0.275	20.02 ± 1.86
$s = 10^{-3}$, $\tau = 0.05$	1	0.96	1.05 ± 0.50	19.80 ± 1.87	0.905	0.621 ± 0.274	19.70 ± 1.87
	2	0.955	1.14 ± 0.56	19.80 ± 1.82	0.845	0.562 ± 0.240	19.83 ± 1.81
$s = 10^{-3}$, $\tau = 0.2$	1	0.855	0.709 ± 0.410	19.75 ± 2.05	0.5	0.329 ± 0.191	19.69 ± 2.21
	2	0.915	0.877 ± 0.490	19.86 ± 2.02	0.68	0.396 ± 0.194	19.70 ± 1.86

Tests A and B, options 1 and 2, and power are defined in the text.

[a] The means and standard deviations of maximum-likelihood estimates of s and X are shown. The true value of X is 20 kb. Other parameter values are $N = 5 \times 10^5$, $\rho = 10^{-8}$, $\theta = 0.005$, and $n = 10$.

frequency spectrum should be close to the expectation under the hitchhiking model. As expected, the power of test B is smaller than that of test A (Table 2). However, it is still high (84–97%) for small values of τ (0.001 and 0.05). Power declines as τ increases, since the spatial pattern and the frequency spectrum of segregating sites approach those under neutrality as time passes after the selective sweep. Tests using option 1 yield higher power than those using option 2 for $\tau = 0.001$ and 0.05, which means that the skew toward high-frequency-derived alleles at segregating sites is observed as described by $P_{n,k}$ (Figure 4). For $\tau = 0.2$, however, both tests A and B had higher power with option 2. This is obvious from the fact that, at $\tau = 0.2$, the proportion of high-frequency-derived alleles is lowered below its level under neutrality (Figure 3c). Therefore, distinguishing the derived allele from the ancestral allele in these tests has an advantage for detecting very recent hitchhiking events only. However, it should be noted that our analytic prediction of the frequency spectrum is based on the assumption of $\tau = 0$. Complete solutions for $P_{n,k}$ for any value of τ may make option 1 still useful for detecting more distant hitchhiking events.

Maximum (composite)-likelihood estimates of s and X were also obtained. Test A with option 1 produced the most unbiased estimates of s, although the accuracy is quite low for all combinations of the test methods (Table 2). Joint estimates of s and X using test A with option 1 for datasets generated under neutrality and under hitchhiking with $s = \tau = 0.001$ are shown in Figure 6. From the neutral data, joint estimates were clustered in the parameter space of small s (close to the lower limit) and X between the "sequenced" segments. This can be expected since the hypothesized valley due to hitchhiking should be sufficiently narrow to fit between the sequenced segments where the level of polymorphism is high. On the other hand, the joint estimates from the hitchhiking datasets were centered around the true value ($s = 0.001$, $X = 20$ kb). It is also shown that estimates of X tend to cluster on the sequenced segments in this case.

To further investigate the performance of the composite likelihood-ratio test, we produced additional but shorter (10-kb) sequences by simulation. Unlike in the previous analysis, polymorphism data are assumed to be obtained from the entire continuous region and X is also allowed to vary over the entire region. Only option 1 is used. To make results comparable to the previous

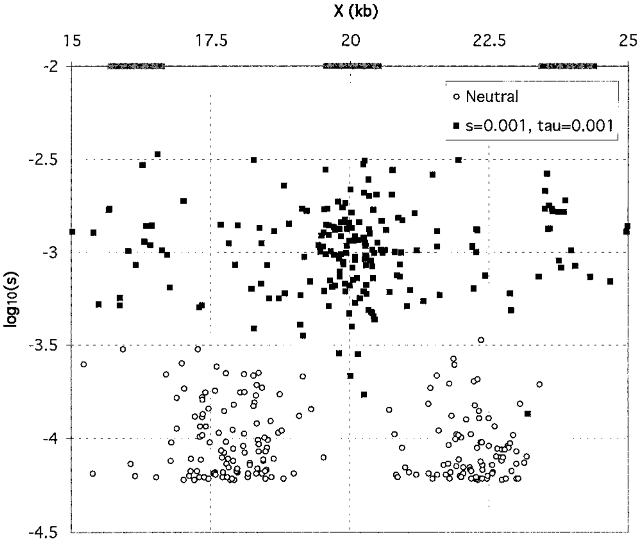

Figure 6.—Joint estimates of s and X using test A and option 1 for simulated neutral and hitchhiking ($s = 0.001$, *i.e.*, $\alpha = 1000$; $\tau = 0.001$) datasets. The substitution of the beneficial allele occurs at position 20 kb. Solid squares represent estimates obtained from hitchhiking datasets and open circles represent those from neutral datasets. A total of 200 replicates were used for each model. Shaded segments over the y-axis represent regions from which polymorphism data were obtained.

TABLE 3

Results of likelihood-ratio tests (continuous 10-kb sequence)

Simulation model	Test A			Test B		
	Power	$\hat{s}$ ($\times 10^{-3}$)[a]	$\hat{X}$ (kb)[a]	Power	$\hat{s}$ ($\times 10^{-3}$)	$\hat{X}$ (kb)
Neutral	—	0.094 ± 0.088	5.05 ± 2.11	—	0.086 ± 0.084	5.10 ± 2.13
$s = 0.0001$	0.255	0.148 ± 0.132	5.08 ± 1.60	0.13	0.122 ± 0.101	4.97 ± 1.55
$s = 0.0002$	0.6	0.232 ± 0.159	4.94 ± 1.09	0.435	0.189 ± 0.129	5.10 ± 1.16
$s = 0.0005$	0.925	0.627 ± 0.328	5.04 ± 0.69	0.855	0.437 ± 0.217	5.03 ± 0.73
$s = 0.001$	0.97	1.15 ± 0.62	4.97 ± 0.83	0.915	0.713 ± 0.368	4.98 ± 0.83
$s = 0.002$	0.995	2.66 ± 2.47	5.16 ± 0.98	0.98	1.30 ± 0.69	5.13 ± 0.96
$s = 0.005$	0.96	9.58 ± 3.50	4.97 ± 1.68	0.865	2.24 ± 1.52	4.89 ± 1.61

Tests A and B and power are defined in the text.

[a] The means and standard deviations of maximum-likelihood estimates of s and X are shown. The true value of X is 5 kb. Other parameter values are $N = 5 \times 10^5$, $\rho = 4 \times 10^{-8}$, $\tau = 0.001$, $\theta = 0.005$, and $n = 10$.

cases, R was adjusted to be 800 by setting $\rho = 4 \times 10^{-8}$. In the simulation of selective sweeps, selection occurs at position 5 kb with $\tau = 0.001$, but with various s values (Table 3). With $s = 0.001$ ($\alpha = 1000$), the powers of tests A and B were 0.97 and 0.915, respectively. These can be compared with 0.995 and 0.97 from the previous analysis (Table 2, $s = 0.001$ and $\tau = 0.001$). Considering a slight reduction of the surveyed region (11–10 kb) where informative segregating sites were observed, the power of the tests with this new scheme appears to remain as high as in the previous one using discontinuous regions. With decreasing s, the power of detecting the hitchhiking event and the accuracy of the parameter estimates decrease, as expected (Table 3). However, power declined also when s increased from 0.002 to 0.005. Examination of simulated data showed that, with $s = 0.005$, the number of segregating sites is highly reduced and those sites are frequently found clustered on one side of the site of selection. This produced very low likelihood ratios in both tests A and B. This effect should disappear if a wider region of the chromosome is surveyed.

We also conducted a few tests to assess the effect of uncertainty in prior estimates of N and ρ. Tests A and B were performed for a dataset described above (Table 3, $s = 0.001$) but with a prior estimate of $N = 10^5$, fivefold lower than the true value. New null distributions of the likelihood ratio were obtained accordingly. The power of tests A and B decreased to 0.935 and 0.865, respectively, from 0.97 and 0.915 (Table 3) due to the incorrect assumption of N. Average $\hat{s}$ decreased slightly (8.5×10^{-4} from 1.15×10^{-3} for test A and 5.1×10^{-4} from 7.1×10^{-4} for test B). Next, we used the correct value of N but a fivefold lower prior estimate of ρ (8×10^{-9}). Average $\hat{s}$ decreased about fivefold (1.78×10^{-4} and 1.19×10^{-4} for tests A and B, respectively) as expected. Power decreased to 0.855 and 0.78 for tests A and B, respectively.

DISCUSSION

In this study, coalescent simulations using the ancestral recombination graph (Hudson 1983; Griffiths and Marjoram 1997) were used to investigate patterns of nucleotide sequence polymorphism under the models of neutrality and genetic hitchhiking. The modification of the ARG to incorporate the effect of hitchhiking is analogous to the two-locus coalescent model of Braverman *et al.* (1995). In fact, our ARG with hitchhiking reduces to their model if the recombination break point is fixed rather than uniformly distributed between a neutral locus and the site under selection. Many simulation and theoretical studies based on this two-locus model generated valuable knowledge about genetic hitchhiking (Kaplan *et al.* 1989; Braverman *et al.* 1995; Barton 1998; Fay and Wu 2000). However, the study of variation at a single neutral site under hitchhiking cannot provide information about spatial patterns of variation, which is shaped by genealogical correlation between many consecutive sites. The ARG allows such a study by generating coalescent trees for all sites simultaneously.

Local reduction of genetic variation without the corresponding reduction of interspecific divergence was used as evidence of past directional selection in maize (Wang *et al.* 1999), Drosophila (Benassi *et al.* 1999; Nurminsky *et al.* 2001), and human (Nachman and Crowell 2000; Fullerton *et al.* 2000). However, these studies did not address the possibility of observing such patterns due to stochastic change of genetic variation along recombining chromosomes under neutrality. Table 1 shows that, in a chromosome with $N\rho = 10^{-3}$, segments of reduced variation that are >1 kb can be frequently found under neutrality. The average length of segments sharing the same ancestral history becomes longer as $N\rho$ becomes smaller (Wiuf and Hein 1999). For species with relatively small effective population sizes, such as human, $N\rho$ can be $\ll 10^{-3}$. In such a case, a very cautious interpretation of the data is warranted when a sudden drop of heterozygosity over a few kilobases is observed in these species.

To address this problem, we developed a composite likelihood-ratio test to detect the local signature of genetic hitchhiking along a recombining chromosome, where the null distribution of variation is obtained by neutral coalescent simulations with recombination. The composite likelihood under the hitchhiking model is based on the probability of observing a certain ratio of segregating variants, $P_{n,k}$, for each site. $P_{n,k}$ is a function of N, μ, ρ, s, and X. In test A, we assumed that the actual values of N, μ, and ρ are already known. The recombination rate per nucleotide can be determined independently for some species for which both physical and genetic maps are available. The effective population size and the mutation rate might be obtained from polymorphism and divergence data from adjacent chromosomal regions, if the standing levels of diversity and divergence are uniform in those regions and hitchhiking events do not occur frequently; *i.e.*, the standing level is determined mainly by neutrality or background selection (Kim and Stephan 2000). Therefore, test A can be used only for species such as Drosophila and human where a considerable amount of population genetic information has been obtained. In the sense that it requires information from the other loci, test A is not much different from the Hudson-Kreitman-Aguadé (HKA) test (Hudson *et al.* 1987). However, unlike the HKA test, test A uses information contained in the frequency spectrum at segregating sites and allows the parameters of the hitchhiking model to be estimated.

Test B relaxes the assumption that θ, the level of variation in the region immediately before the hitchhiking event happened, is known. In the most conservative treatment, θ is given as the average nucleotide diversity observed in the data to be tested. Therefore, test B is the method of choice when information on other loci is not available. However, due to the incorrectness of θ, the estimation of s is poorer in test B than in test A.

A similar approach to detect the signature of hitchhiking has recently been proposed by Galtier *et al.* (2000). Their method detects diversity-reducing events such as hitchhiking or bottleneck and estimates the time and the strength of those events from DNA sequence polymorphism. However, they assumed no recombination within the region to be tested and needed several independent regions to distinguish selective sweeps from bottlenecks. Our approach is different in that it detects a characteristic spatial pattern shaped by recombination around the site of directional selection. It is similar, however, in that it also detects skews in the frequency spectrum of segregating sites, as Galtier *et al.* (2000) detect "distortions" in the shape of gene genealogies. Unfortunately, a relationship between the force that distorts the genealogy and the strength of directional selection was not given in their study.

Knowledge about the strength and the rate of directional selection in natural populations is fundamental in evolutionary biology. Previously, Wiehe and Stephan (1993) and Stephan (1995) obtained a rough estimate of $\alpha\nu$, where ν is the rate of strongly selected substitutions per nucleotide, using the positive correlation of variation and recombination in *Drosophila melanogaster*. Separate estimation of α and ν might be achieved by surveying large areas of a genome for signatures of hitchhiking events, using the method proposed in this article. According to the assumption that standing variation in the region tested is not influenced by other hitchhiking events, this method is expected to be most useful in regions of high recombination rates (Kim and Stephan 2000). It should be noted, however, that the inference of s and X may be accompanied by a considerable amount of error (Figure 6). Figure 3 and Table 1 show that, for a given set of parameter values, the size of the valley of reduced nucleotide diversity varies significantly and the center of the valley may drift away from the site of selection. This stochasticity is especially serious for populations with small effective sizes (Figure 3d). This might explain the observation by Wang *et al.* (1999) that within a 1.1-kb region of highly reduced variation no fixed differences were found between maize and teosinte.

We thank two reviewers for valuable suggestions. This work was supported by funds from the University of Munich and the Deutsche Forschungsgemeinschaft (STE 325/4-1).

LITERATURE CITED

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

Barton, N. H., 2000 Genetic hitchhiking. Philos. Trans. R. Soc. Lond. B **355:** 1553–1562.

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

Benassi, V., F. Depaulis, G. K. Meghlaoui and M. Veuille, 1999 Partial sweeping of variation at the *Fbp2* locus in a west African population of *Drosophila melanogaster*. Mol. Biol. Evol. **16:** 347–353.

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.

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

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

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

Fullerton, S. M., A. G. Clark, K. M. Weiss, D. A. Nickerson, S. L. Taylor *et al.*, 2000 Apolipoprotein E variation at the sequence haplotype level: implications for the origin and maintenance of a major human polymorphism. Am. J. Hum. Genet. **67:** 881–900.

Galtier, N., F. Depaulis and N. H. Barton, 2000 Detecting bottlenecks and selective sweeps from DNA sequence polymophism. Genetics **155:** 981–987.

Griffiths, R. C., and P. Marjoram, 1997 An ancestral recombination graph, pp. 257–270, in *Progress in Population Genetics and Human Evolution*, edited by P. Donnelly and S. Tavaré. Springer-Verlag, New York.

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

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

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.

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

Kimura, M., 1971 Theoretical foundation of population genetics at the molecular level. Theor. Popul. Biol. **2:** 174–208.

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

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.

Nurminsky, D., D. De Aguiar, C. D. Bustamante and D. L. Hartl, 2001 Chromosomal effects of rapid gene evolution in *Drosophila melanogaster*. Science **291:** 128–130.

Press, W. H., S. A. Teukolsky, W. T. Vetterling and B. P. Flannery, 1992 *Numerical recipes in C.* Cambridge University Press, Cambridge, UK.

Rannala, B., and M. Slatkin, 2000 Methods for multipoint disease mapping using linkage disequilibrium. Genet. Epidemiol. **19:** S71–S77.

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.

Stephan, W., L. Xing, D. A. Kirby and J. M. Braverman, 1998 A test of the background selection hypothesis based on nucleotide data from *Drosophila ananassae*. Proc. Natl. Acad. Sci. USA **95:** 5649–5654.

Tajima, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics **123:** 437–460.

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

Wang, R. L., A. Stec, J. Hey, L. Lukens and J. Doebley, 1999 The limits of selection during maize domestication. Nature **398:** 236–239.

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

Wiehe, T. H. E., and W. Stephan, 1993 Analysis of a genetic hitchhiking model, and its application to DNA polymorphism data from *Drosophila melanogaster*. Mol. Biol. Evol. **10:** 842–854.

Wiuf, C., and J. Hein, 1999 The ancestry of a sample of sequences subject to recombination. Genetics **151:** 1217–1228.

Communicating editor: Y.-X. Fu

APPENDIX

Consider a neutral locus where a mutant, A, is segregating. The substitution of a beneficial allele, B, for the wild-type allele, b, is assumed to occur at a linked locus at recombination fraction r away from the neutral locus. p_1 is defined as the frequency of A among chromosomes carrying the B allele when the frequency of B increased to the value of near fixation, $1 - \varepsilon$. Then,

$$E(p_1) = p_{1\varepsilon} - r(p_{1\varepsilon} - p_{2\varepsilon})\int_0^T \frac{(1-\varepsilon)e^{-(s+r)t}}{\varepsilon + (1-\varepsilon)e^{-st}}dt \qquad \text{(A1)}$$

(Stephan *et al.* 1992), where $p_{1\varepsilon}$ and $p_{2\varepsilon}$ are the frequencies of A among chromosomes carrying B and b alleles, respectively, when the frequency of B is ε and $T = -2 \log(\varepsilon)/s$. Using the approximation given by Stephan *et al.* (1992), (A1) can be simplified to

$$E(p_1) \approx p_{1\varepsilon} - C(p_{1\varepsilon} - p_{2\varepsilon}), \qquad \text{(A2)}$$

where $C = 1 - \varepsilon^{r/s}$. Let p_0 be the frequency of A when one copy of the B allele first appeared in the population. Assuming that the initial linkage disequilibrium between the two loci does not break down until the frequency of B increases to ε and $p_{2\varepsilon} \approx p_0$, the expectation of p_1 is Cp_0 if B is initially linked with a and $1 - C(1 - p_0)$ if B is initially linked with A. The former event occurs with probability $1 - p_0$ and the latter with p_0. This leads to the transformation of $\phi_0(\cdot)$ into $\phi_1(\cdot)$.