Selective Sweeps in the Presence of Interference Among Partially Linked Loci

Yuseob Kim*,[1] and Wolfgang Stephan†

*Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York 14853 and †Department of Biology II, Section of Evolutionary Biology, University of Munich, Munich, 80333 Germany

Manuscript received July 19, 2002
Accepted for publication January 17, 2003

ABSTRACT

Recurrent directional selection on a partially recombining chromosome may cause a substantial reduction of standing genetic variation in natural populations. Previous studies of this effect, commonly called selective sweeps, assumed that at most one beneficial allele is on the way to fixation at a given time. However, for a high rate of selected substitutions and a low recombination rate, this assumption can easily be violated. We investigated this problem using full-forward simulations and analytical approximations. We found that interference between linked beneficial alleles causes a reduction of their fixation probabilities. The hitchhiking effect on linked neutral variation for a given substitution also slightly decreases due to interference. As a result, the strength of recurrent selective sweeps is weakened. However, this effect is significant only in chromosomal regions of relatively low recombination rates where the level of variation is greatly reduced. Therefore, previous results on recurrent selective sweeps although derived for a restricted parameter range are still valid. Analytical approximations are obtained for the case of complete linkage for which interference between competing beneficial alleles is maximal.

GENETIC linkage causes a correlation of ancestral histories among neighboring loci. The behavior of a neutral allele thus reflects that of a selected allele at a closely linked locus. Standing variation at neutral sites is suddenly wiped out when a rapid fixation of a strongly selected beneficial mutation occurs in this region. This "hitchhiking" effect of a beneficial mutation or "selective sweep" (Maynard Smith and Haigh 1974; Kaplan *et al.* 1989; Stephan *et al.* 1992; Barton 2000), along with "background selection" caused by recurrent purifying selection (Charlesworth *et al.* 1993), may be responsible for a substantial reduction of genetic variation in a genomic region of low recombination (Begun and Aquadro 1992). As the degree of the positive correlation between variation and recombination is determined by the strength and rate of directional and purifying selection, polymorphism data from various genomic regions can be used to estimate these parameters (Wiehe and Stephan 1993; Stephan 1995; Charlesworth 1996; Andolfatto 2001).

The action of natural selection is readily detectable as a reduction of variation in regions of low recombination rates. However, it is difficult to identify the form of natural selection responsible for this reduction because several selective forces (including those from distant chromosomal regions) may act simultaneously on variation at a given neutral site when the rate of recombination is reduced. As a consequence, interactions between selective forces such as between positive directional selection and recurrent purifying selection may occur. This problem was investigated in a previous study (Kim and Stephan 2000). The present article addresses the interaction among competing beneficial mutations.

Previous analyses of recurrent selective sweeps (Kaplan *et al.* 1989; Wiehe and Stephan 1993) are based on theories developed for rare selected substitutions. In these theories, one assumes that a neutral site is under the influence of at most one linked beneficial mutation at any given time. This assumption is satisfied when the rate of selected substitutions is low and the length of the selective phase is short. However, selected substitutions causing hitchhiking effects need to occur at least once in $2N$ generations to substantially reduce the level of variation. With such a rate of substitutions, the selective phases of different beneficial mutations will overlap with each other with high probability, if the length of the selective phase is not sufficiently short compared to $2N$ generations. Then the current theories of recurrent selective sweeps may not be applicable.

Only a few authors have analyzed the dynamics of competing beneficial alleles. Barton (1995) investigated the effect of one selected substitution on the fixation probability of a beneficial allele at a linked locus. The fixation probability is increased when the beneficial allele occurs in the genetic background of the previous beneficial allele that is on the way to fixation but decreased in the other background (repulsion phase of two beneficial alleles). The net effect, averaging over the two backgrounds, is a reduction of the fixation probability. Gerrish and Lenski (1998) showed that the trajectories of allele frequency changes are also affected

[1]*Corresponding author:* Department of Biological Statistics and Computational Biology, Warren Hall 440, Cornell University, Ithaca, NY 14853. E-mail: yk249@cornell.edu

by this interference between selected loci. We extend this work by analyzing the effect of competing beneficial alleles on neutral variation. We specifically ask whether the available theories of recurrent selective sweeps are still valid in the presence of interference.

SIMULATIONS

Examination of genetic variation by full-forward simulation: Genetic variation under complex models for which an analytical method is not readily available is usually investigated by computer simulations. Full-forward simulation (FFS), where processes of the entire population are simulated generation by generation forward in time, can accommodate any complex feature in the population. Usually a mutant allele at a neutral locus is introduced in the simulation and the frequency change of that allele is monitored in FFS. The number of generations is on the order of N_e, the effective population size, before a significant change in the allele frequency is obtained. Therefore, it requires a long simulation time for a population of a realistic size. However, it is possible to measure the level of genetic variation without introducing mutants in FFS. In coalescent simulations, the amount of variation is directly proportional to the size of the coalescent tree at the neutral locus under the infinite-site model. Similarly, identity by descent (IBD) can be measured and substituted as a measure of genetic variation in FFS. The relationship between IBD and coalescent time is well known (Slatkin 1991; Barton 1998). Assume that the number of generations is counted backward in time. Two genes randomly selected from a population find a common ancestor at generation T, where T is a random variable. Then, $g(t) = P[T \leq t]$ is the probability of IBD by generation t. According to the standard coalescent theory,

$$g(t) \approx 1 - e^{-t/2N_e}. \quad (1)$$

In FFS, $g(t)$ can be estimated in the following way. At the beginning of the simulation, all $2N$ chromosomes carry unique numbers at a neutral locus. Let us assume that we assign the "ancestral number" i to the ith gene ($i = 1, \ldots, 2N$). Forward in time (generation number is counted down from t to 0), the composition of ancestral numbers changes. Suppose that $p_i(t)$ is the frequency of the ancestral number i at present (generation 0). We define

$$I_k(t) = \sum_{i=1}^{2N} p_i(t)^k. \quad (2)$$

Then, the expectation of $I_2(t)$ is $g(t)$. Therefore, we may use $I_2(t)$ as an estimator of $g(t)$.

Using Equation 1, it follows that the effective population size, N_e, can be estimated by

$$\hat{N}_e = -\frac{t}{2\ln(1 - I_2(t))} \quad (3)$$

for a single simulation run, assuming that the coalescence of two randomly chosen genes occurs with a constant rate [defined to be $1/(2N_e)$] at each generation in the time interval $[0, t]$. Therefore N_e estimated here is the "coalescent effective" size of the population (Gillespie 2000a). Then an increase or decrease of genetic variation due to factors such as selection and population structure can be characterized by $\hat{N}_e/N$. This method requires considerably less simulation time compared to the previous ones, since the vector $\mathbf{p}(t) = \{p_1(t), \ldots, p_{2N}(t)\}$ contains more information than the frequency change of single alleles for a given time span.

Single selective sweeps: We first investigate the effect of a single selective sweep using a two-locus simulation. A diploid population of size N is simulated according to the Wright-Fisher model of reproduction. A beneficial mutation occurs at a locus linked to the neutral locus where genetic variation is measured by I_2. The recombination rate between the two loci is r. The fitness of an individual heterozygous for the beneficial allele is given by $1 + 2\eta s$ ($0 < \eta < 1$), and that of homozygous individuals by $1 + 2s$. To reduce the simulation time, 10 copies of beneficial alleles are introduced in the population at the beginning of the simulation. Chromosomes carrying these 10 copies share the same ancestral number at the neutral locus. Therefore, this has the same effect as a beneficial allele producing 10 descendants immediately, such that there is no opportunity for recombination to separate the association between the beneficial and neutral alleles. This procedure is justified since it is known that, conditional on its fixation, the initial copy number of a beneficial allele usually increases quickly by drift (Barton 1998). Furthermore, we examined the effect of initial copy number by introducing 1, 2, 5, and 10 copies of beneficial alleles ($N = 10^4$, $s = 0.05$, $\eta = 0.5$, and $r = 0.005$) and found no significant difference in the mean I_2 measured after fixation (results not shown). Even though 10 copies are given initially, the beneficial mutation may still fail to be fixed. If all beneficial alleles are lost, simulation starts again from the beginning. The observed frequency of this loss is given by l_{10}. As the early branching processes of these 10 initial copies are largely independent of each other, the fixation probability of each beneficial allele, Φ, can be obtained from $l_{10} = (1 - \Phi)^{10}$. $I_2(t)$ is measured when the fixation of the beneficial allele occurs, where t is the number of generations until fixation. $I_2(t)$ is a measure of the cumulative coalescent events during the time interval $[0, t]$. As coalescence between a pair of lineages may occur with probability $1 - \exp(-t/2N)$ at this interval without hitchhiking, the net effect of hitchhiking is measured as

$$I_{2h} = I_2(t) - (1 - e^{-t/2N}). \quad (4)$$

The expectation of I_{2h} corresponds to the probability of coalescence due to hitchhiking analyzed in previous studies (Kaplan *et al.* 1989; Stephan *et al.* 1992). Therefore we obtain

TABLE 1

Simulation of single selective sweeps

r	η	I_{2h}[a]	$E[I_{2h}]$	Φ	$1 - e^{-4\eta s}$	t_s[b]
0.5	0.5	0.000 ± 0.002	0	0.0893	0.0950	332.3
0.01	0.5	0.103 ± 0.056	0.0933	0.1000	0.0950	328.5
0.0033	0.5	0.433 ± 0.152	0.44	0.0978	0.0950	329.9
0.001	0.02	0.452 ± 0.179	0.103	0.00565	0.00399	888.6
0.001	0.1	0.582 ± 0.181	0.402	0.0176	0.0198	555.1
0.001	0.5	0.761 ± 0.141	0.777	0.0959	0.0950	332.1
0.001	0.9	0.808 ± 0.177	0.857	0.157	0.165	578.5

Results are based on 500 replicates for each parameter set. For all simulations, $N = 10^4$ and $s = 0.05$.
[a] Mean ± standard deviation.
[b] Mean length of simulation time (in generations).

$$E[I_{2h}] = 1 - \frac{2r}{s}(2Ns)^{-2r/s}\Gamma\left(-\frac{2r}{s}, \frac{1}{2Ns}\right), \qquad (5)$$

where $\Gamma(.\ ,\ .)$ is the incomplete gamma function (for $\eta = 0.5$; Stephan *et al.* 1992).

Table 1 shows the comparison between the prediction and the simulation results. For $\eta \neq 0.5$, we still used the equation above but replaced s by $2\eta s$. This may be justified if the hitchhiking effect is determined mainly at the early stage of the selective phase when the beneficial mutation is in low frequency and thus found mainly in heterozygotes (Stephan *et al.* 1992). Agreement between the simulation result and prediction is good when η is close to 0.5. However, values of I_{2h} are smaller (larger) than the prediction when η is greater (smaller) than 0.5. A particularly large discrepancy between the simulation results and the prediction for $\eta \ll 0.5$ indicates that the hitchhiking effect caused by a recessive beneficial mutation is not determined mainly at the early stage of the selective phase but at a later time when the beneficial allele is in substantially higher frequency such that homozygotes start appearing in the population. Table 1 also shows that $1 - e^{-4\eta s}$ approximates the fixation probability of the beneficial allele quite well if η does not deviate too much from 0.5. For the remainder of this article, we consider only genic selection ($\eta = 0.5$).

Two overlapping selective sweeps: The simulation scheme described above is extended to investigate the effect of two overlapping substitutions of strongly selected beneficial alleles on a linked neutral locus. Let the locus of the first beneficial mutation be *S1* and that of the following beneficial mutation be *S2*. We consider all three chromosomal arrangements of the loci: *Neu-S1-S2*, *Neu-S2-S1*, and *S1-Neu-S2*, where *Neu* represents the neutral locus. The selection coefficients for both selected loci are identical. For this three-locus model (and also the other multilocus models in this study), effects of beneficial alleles on fitness combine multiplicatively. The recombination rate between adjacent loci is r. Ten beneficial alleles are introduced in the population at each locus just as in the simulation of single selective sweeps. However, the beneficial mutation at *S2* does not occur until the allele frequency at *S1* exceeds a certain value, Q. These beneficial alleles at *S2* are initially in complete linkage with either the beneficial (background of 1) or the ancestral (background of 0) alleles of *S1*. This process is repeated until the fixations at both loci are completed. Then I_{2h} is observed as explained above. The fixation probability, Φ_2, at *S2* conditional on the fixation of the preceding beneficial allele at *S1*, is measured using the same method as in the case of single selective sweeps. The lengths of the selective phases at *S1* and *S2*, t_{S1} and t_{S2}, respectively, are also recorded.

Simulation results for *Neu-S1-S2* are shown in Table 2. The interference between substitutions causes modification of the fixation probability and the length of the selective phases. Table 2 shows that Φ_2 increases (decreases) in the beneficial (ancestral) background of *S1*. Interaction of two beneficial alleles either speeds up or slows down the course of substitution, as revealed by changes in t_{S1} and t_{S2} as a function of the genetic background. The net probability of fixation at *S2* is given by $Q\Phi_{2,1} + (1 - Q)\Phi_{2,0}$, where $\Phi_{2,1}$ and $\Phi_{2,0}$ are the fixation probabilities in the beneficial and the ancestral backgrounds, respectively. Figure 1 shows the comparison of the simulation results with the theoretical prediction obtained by numerically solving Equations 6a and 6b of Barton (1995). Interference produces the greatest effect on net fixation probability when $Q \approx 0.2$.

The effect of two overlapping selective sweeps on neutral variation can be quantified using I_{2h}. I_{2h} averaged over the two genetic backgrounds at *S1*, $\overline{I_{2h}}$, is obtained by weighting the probability of observing a substitution in either background [$Q\Phi_{2,1}$ or $(1 - Q)\Phi_{2,0}$]. $\overline{I_{2h}}$ for each of the three arrangements of loci as a function of the scaled time [$T = \ln(Q/(1 - Q))$] of the occurrence of the second beneficial mutation is shown in Figure 2. $\overline{I_{2h}}$ for *S1-S2-Neu* remained constant while Q varied from 0.03 to 0.94. However, for *Neu-S1-S2* and *S1-Neu-S2*, $\overline{I_{2h}}$ decreased with decreasing Q (for $Q < 0.5$). This dependence of $\overline{I_{2h}}$ on spatial arrangement might be explained by the following argument. If the beneficial mutation at *S2* occurs when the beneficial allele at *S1* is still in

TABLE 2

Simulation of two overlapping selective sweeps (*Neu-S1-S2*)

Q	Background	Run time[a]	I_{2h}[b]	Φ_2[c]	t_{S1}[d]	t_{S2}[e]
0.06	1	375.3	0.435	0.182	293.3	275.9
0.06	0	468.6	0.340	0.0729	353.1	368.3
0.2	1	427.6	0.390	0.143	305.4	299.0
0.2	0	495.4	0.364	0.0539	346.3	367.6
0.5	1	472.3	0.378	0.123	313.3	316.5
0.5	0	522.7	0.357	0.0382	347.7	367.0
0.8	1	505.6	0.366	0.0974	321.9	319.2
0.8	0	550.7	0.348	0.0280	357.5	365.0
0.94	1	535.4	0.366	0.0991	324.0	322.4
0.94	0	577.2	0.373	0.0262	365.4	364.2

$N = 10^4$, $s = 0.05$, $r = 0.005$. Results are based on 500 replicates for each parameter set. See text for explanations of symbols.

[a] Average total number of generations for each simulation run.

[b–e] Observed mean of I_{2h}, Φ_2, t_{S1}, and t_{S2}, respectively.

low frequency ($Q < 0.5$), it is likely to occur in the ancestral background of *S1* and thus to generate a repulsion phase between the two beneficial alleles. To achieve fixations at both loci, the two beneficial alleles must recombine onto one chromosome. Therefore, there is an excess of recombination between *S1* and *S2* conditional on the fixations at both loci. This excess recombination may reduce the hitchhiking effects of selected substitutions on *Neu*. However, this effect will occur mainly between *S2* and *Neu* rather than between *S1* and *Neu*. The hitchhiking effect of *S1* on *Neu*, which is produced mainly in the early selective phase at *S1*, is unaffected by excess recombination because the mutation at *S2* has yet to appear. On the other hand, the early selective phase at *S2* is subject to excess recombination between the selected loci, which reduces the hitchhiking effect for *S1-Neu-S2* and *Neu-S1-S2* but not for *S1-S2-Neu*. In summary, interference among beneficial mutations causes (i) a net reduction in the rate of selected substitutions and (ii) a (slight) reduction in the hitchhiking effect for a given substitution. A combination of these two effects will determine the level of genetic variation under the model of recurrent selective sweeps.

Multilocus simulation of selective sweeps: To further investigate the effect of interference on genetic variation under recurrent selective sweeps, FFS described in the previous section is extended to a multilocus model. The neutral locus under investigation is located in the middle of a chromosome. Thirty loci where beneficial mutations can occur are on each side of the neutral locus. Mutation occurs at a rate u per gene per generation if the beneficial allele is not already segregating at the same locus in the population. The recombination rate between adjacent loci is uniformly r. The first phase of the simulation, which is t_1 generations long, brings the population into an equilibrium state in which there is a constant flux of beneficial alleles reaching fixation. Then, at the beginning of the second phase, which takes t_2 generations, ancestral numbers are assigned to genes at the neutral locus. Fixation probability, Φ, of the beneficial allele is measured by counting the number of introduced and fixed alleles during the second phase of the simulation. At the end of the second phase, $I_2 = I_2(t_2)$ is recorded. The effective population size is estimated using Equation 3, where instead of t and $I_2(t)$ we use t_2 and the observed mean of $I_2(t_2)$ over replicates, respectively. Preliminary study showed that $\hat{N}_e$ obtained from Equation 3 is a decreasing function of t_2 if t_2 is small compared to the length of the single selective phase (t_s), but converges to the expected value when $t_2 > 2t_s$ (results not shown).

Using previous results on the coalescent effective pop-

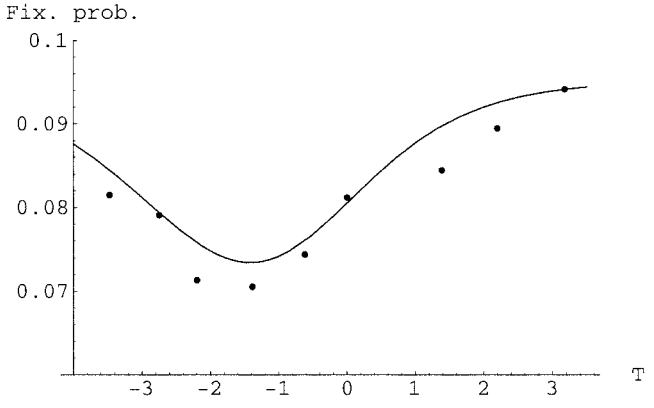

FIGURE 1.—Fixation probability at *S2* as a function of the scaled time T, where $Q = 1/(1 + \exp(-T))$. Dots represent simulation results ($N = 10^4$, $s = 0.05$ for both loci, $r = 0.005$), averaged over 1000 replicates for each Q value. The curve of the predicted fixation probability is drawn using Equations 6a and 6b of BARTON (1995).

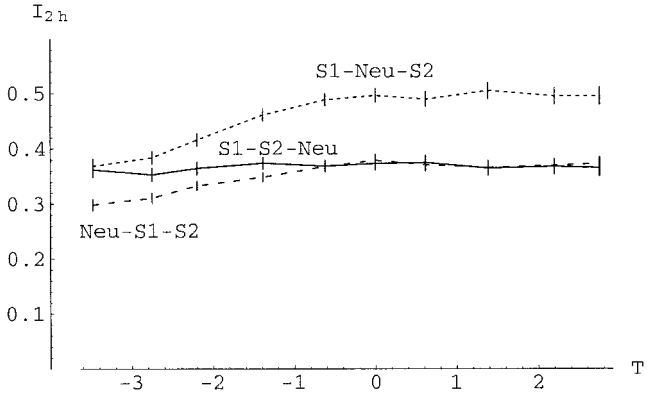

FIGURE 2.—Net effect of hitchhiking ($\overline{I_{2h}}$) conditional on the fixations of two linked beneficial alleles as a function of time T, where $Q = 1/(1 + \exp(-T))$. Simulation results were obtained from 500 replications for each parameter set ($N = 10^4$, $s = 0.05$, $r = 0.005$). Vertical lines represent twice the standard error.

ulation size under the model of selective sweeps (Wiehe and Stephan 1993; Gillespie 2000b; Kim and Stephan 2000), we predict the expectation of $\hat{N}_e$ as

$$E[\hat{N}_e] = \frac{N}{1 + 4N^2 \sum_{i=1}^{L} u_i \Phi_i (1 - h_i)}, \qquad (6)$$

where L is the number of selected loci, u_i and Φ_i are the rate and the fixation probability of the beneficial mutation at the ith site, respectively, and h_i is the reduction of expected heterozygosity at the neutral locus due to a substitution at the ith locus. The above equation assumes that fixation events occur according to a Poisson process. It is not known to what extent this assumption needs to be modified under interference among beneficial mutations. However, according to Gillespie (2000b), we expect that only a small error results from a violation of this assumption. Furthermore, assuming that a slight change in the trajectory of the beneficial allele frequency due to interference does not affect hitchhiking, we use the solution given by Equation 5 ($1 - h_i = E[I_{2h}]$). With these assumptions, we examine two expectations of $\hat{N}_e$. $E_1[\hat{N}_e]$ is given by Equation 6 using $\Phi_i = 1 - e^{-2s}$, the theoretical value without interference; thus, $E_1[\hat{N}_e]$ assumes that the effects of all selected loci combine additively. $E_2[\hat{N}_e]$ is given by Equation 6 using the observed Φ averaged over all loci.

First, we consider a uniform selection coefficient for all loci (Table 3A, $s = 0.1$, $N = 5000$). The beneficial mutation rates are high enough to cause interference among closely linked sites (see discussion). As expected, the fixation probability averaged over all sites decreased as r decreased. The average length of the selective phase, t_s, increased as r decreased. Therefore, interference slows down the course of the selected substitutions. $\hat{N}_e$ is consistently $> E_1[\hat{N}_e]$, confirming that the standing level of genetic variation is not as much decreased as expected without interference. $E_2[\hat{N}_e]$ is generally closer to $\hat{N}_e$ than is $E_1[\hat{N}_e]$. Therefore, a large part of the difference between $\hat{N}_e$ and $E_1[\hat{N}_e]$, presumably due to interference, can be explained by the reduction in the rate of substitution. The remaining discrepancy between $\hat{N}_e$ and $E_2[\hat{N}_e]$ might be explained by the reduced hitchhiking effect of a given substitution if $\hat{N}_e > E_2[\hat{N}_e]$ is observed. However, $\hat{N}_e < E_2[\hat{N}_e]$ in many cases when $r = 0$ and u is large. One possible explanation is the presence of a "leapfrog" effect, which is described below.

To mimic a more realistic situation, we also performed simulations where beneficial mutations occur with two different selection coefficients. The beneficial mutation under relatively strong directional selection occurs with selection coefficient s_s at rate u_s and the mutation under weaker directional selection with s_w at rate u_w ($> u_s$). These two different mutations occur at 30 "strong" and 30 "weak" loci, respectively, which alternate with each other along the chromosome. It should be understood that weak means only the relative strength of a mutation when compared to a strong mutation. We still consider $2Ns_w \gg 1$. Table 3B shows the result for $s_s = 2s_w$ and $u_w = 2u_s$. As expected (Barton 1995), the decline of the relative fixation probability at the weak loci is much greater than that in the case of equally strong beneficial alleles (Table 3A). However, differences among $\hat{N}_e$, $E_1[\hat{N}_e]$, and $E_2[\hat{N}_e]$ are not much greater than those among the uniform selection coefficients. For $r > 0$, the reduced rate of substitutions at weak loci might be unimportant in modifying the effective population size because the latter is determined mainly by the hitchhiking effects from strongly selected loci, the fixation rate of which does not change as much as that of weakly selected loci. However, this cannot be an explanation for the case of $r = 0$ because both strong and weak selections wipe out standing variation completely. This issue is further investigated below.

Next, the joint effect of selective sweeps and background selection was investigated by a similar simulation scheme in which half of the selected loci are now under recurrent purifying selection. A total of 48 selected loci, where loci under directional selection are alternating with those under purifying selection, were used and the neutral locus was inserted in the middle of the arrangement. Deleterious mutations with selective disadvantage s_d were introduced at a rate u_d per locus per generation. Table 4 shows the results. Φ decreased with background selection as expected (Barton 1995). $\hat{N}_e$ obtained from the simulation was in good agreement with the theoretical prediction ($E[\hat{N}_e]$ in Table 4), which was obtained by modifying Equation 6 (see Kim and Stephan 2000, Equation 6). Selective sweep and background selection act nonadditively on N_e: With low recombination, the joint effect of the two forces is almost the same as that of hitchhiking alone (compare the third, sixth, and ninth rows of Table 4). This result can be explained by the fact that the increasing effect of background selection is offset by a decreasing effect of recurrent selective sweeps, as background selection causes a reduction of the fixation probability with lower recombination rates. This multilocus simulation confirms the results obtained by a three-locus model of the joint effects of background selection and hitchhiking investigated in a previous study (Kim and Stephan 2000).

THEORY OF RECURRENT SELECTIVE SWEEPS FOR ZERO RECOMBINATION

The multilocus model described above is further investigated in the case of zero recombination for which the effect of interference is expected to be maximal. Equation 6 with $r = 0$ gives the expected level of variation under the model of recurrent selective sweeps with complete linkage. Since the substitution of any selected allele with complete linkage has the same hitchhiking effect, *i.e.*, the complete removal of genetic variation,

TABLE 3

Multilocus simulation of selective sweeps (60 loci under directional selection)

r	U	$P(\Omega)$[a]	Φ	t_s[b]	$\hat{N}_e$	$E_1[\hat{N}_e]$[c]	$E_2[\hat{N}_e]$[d]
A. Uniform selection coefficients across loci ($s = 0.1$)							
10^{-2}	2×10^{-8}	0.85	0.173	168.5	3821	3683	3730
10^{-3}	2×10^{-8}	0.40	0.165	171.1	894.2	772.2	836.0
10^{-4}	2×10^{-8}	0.65	0.169	175.2	303.0	263.2	281.0
10^{-5}	2×10^{-8}	0.68	0.166	187.5	238.7	223.9	243.8
0	5×10^{-9}	0.92	0.169	188.2	836.9	776.6	824.9
0	10^{-8}	0.83	0.166	197.2	459.3	421.0	457.5
0	2×10^{-8}	0.68	0.151	200.2	254.3	219.8	261.7
0	3.5×10^{-8}	0.51	0.139	210.5	148.5	128.0	165.8
0	5×10^{-8}	0.38	0.121	226.6	114.8	90.3	133.8

r	u_s	Φ_s	t_{ss}[b]	Φ_w	t_{sw}[b]	$\hat{N}_e$	$E_1[\hat{N}_e]$[c]	$E_2[\hat{N}_e]$[d]
B. Unequal selection coefficients and mutation rates ($s_s = 0.16$, $s_w = 0.08$, $u_w = 2u_s$)								
10^{-2}	2×10^{-8}	0.265	112.8	0.140	203.4	3161	2949	2997
10^{-3}	2×10^{-8}	0.262	114.9	0.142	208.5	533.6	468.5	486.9
10^{-4}	2×10^{-8}	0.246	120.0	0.129	220.8	177.3	168.9	189.9
10^{-5}	2×10^{-8}	0.253	128.3	0.113	235.1	155.2	144.7	171.1
0	5×10^{-9}	0.253	137.4	0.147	215.8	542.2	523.9	543.3
0	10^{-8}	0.259	126.8	0.119	236.4	292.3	276.4	314.4
0	2×10^{-8}	0.222	134.3	0.102	234.0	182.4	142.2	188.6
0	3×10^{-8}	0.233	140.6	0.091	220.4	107.6	95.7	130.5
0	4×10^{-8}	0.229	141.8	0.078	240.5	97.74	72.1	106.1

Results are based on 500 replicates for each parameter set. For all simulations, $N = 5000$, $t_1 = 1000$, and $t_2 = 1000$.

[a] Probability that two gene lineages experience only nonoverlapping selective sweeps before they find a common ancestor (Equation 15).

[b] Mean number of generations until fixation of the beneficial allele. t_{ss} and t_{sw} are the fixation times for strong and weak beneficial alleles, respectively.

[c] Expected value of $\hat{N}_e$ from Equation 6 using (A) $\Phi = 1 - e^{-2s}$ and (B) $\Phi_s = 1 - e^{-2s_s}$ and $\Phi_w = 1 - e^{-2s_w}$.

[d] Expected value of $\hat{N}_e$ from Equation 6 using the observed values of Φ.

only the rate of fixation of beneficial mutations is expected to determine the level of standing variation. The fixation probability of beneficial mutations for a nonrecombining chromosome was studied in Barton (1995) and Gerrish and Lenski (1998). However, their approximate solutions are either inaccurate or not applicable to our multilocus model. Here we present an alternative derivation of the fixation probability. The derivation assumes a haploid population of $2N$ individuals.

There are two possibilities by which the fixation prob-

TABLE 4

Multilocus simulation of selective sweeps (joint effects with background selection)

u_d	u	r	Φ[a]	t_s	$\hat{N}_e$	$E[\hat{N}_e]$
0.002	0	0.005	—	—	3733	3748
0.002	0	0.001	—	—	2540	2581
0.002	0	0.0005	—	—	2235	2271
0	5×10^{-8}	0.005	0.0954	295.7	3838	3859
0	5×10^{-8}	0.001	0.0932	296.1	1755	1646
0	5×10^{-8}	0.0005	0.1003	301.7	1136	1026
0.002	5×10^{-8}	0.005	0.0806 [0.0813]	292.6	3043	3094
0.002	5×10^{-8}	0.001	0.0675 [0.0656]	298.5	1372	1414
0.002	5×10^{-8}	0.0005	0.0611 [0.0590]	294.2	1126	1022

Results are based on 200 replicates for each parameter set. For all simulations, $N = 5000$, $s = s_d = 0.05$, $t_1 = 500$, and $t_2 = 1000$. For definition of the other symbols see Table 3.

[a] Numbers in brackets are the fixation probabilities predicted by modification of Equation 11 of Kim and Stephan (2000).

ability of a new beneficial mutation, B_1, is affected by another beneficial mutation, B_2, at a linked site. First, if B_2 is already segregating in the population when B_1 arises, the initial selective advantage of the chromosome carrying B_1 relative to others is modified depending on the frequency of B_2 and depending on which chromosome B_1 occurs. Conditional on fixation, the frequency of B_1 drifts quickly to a certain threshold above which the chance of B_1 being lost by drift is negligible. Therefore the fate of B_1, loss or fixation, is decided in a short initial period. Second, B_1 while on the way to fixation (after this short initial phase) may be displaced by another beneficial mutation that arises after B_1 (Gerrish and Lenski 1998). Therefore we approximate the fixation probability under interference as $\Phi = f_1 f_2$, where f_1 is the fixation probability that takes into account only the initial competition with preexisting alleles and f_2 is the probability that the allele that survived the initial drift is not lost in the competition with late-arising alleles.

First we consider the case where all beneficial alleles are equally advantageous with selective coefficient s (such that $Ns \gg 1$). Under no interference, the fixation probability of a beneficial allele starting at frequency x is given approximately by

$$f_0(x, s) = 1 - \exp(-4Nsx)$$

(Ewens 1979). When $2NuL$ new beneficial alleles are introduced in the population each generation, the expected number of beneficial alleles in the population found in the small frequency interval $[x, x + dx]$ (at equilibrium) is approximately

$$J(x)\,dx = 4NuL\frac{1 - \exp[-4Ns(1 - x)]}{x(1 - x)}dx \quad (0 < x < 1) \tag{7}$$

(Sawyer and Hartl 1992). We consider a small value of u so that, when a new beneficial mutation (B_1) occurs, at most one beneficial allele (B_2) exists at another locus segregating in the population. Equation 7, which was derived under the assumption of independence among sites, may be inaccurate for describing the density of allele frequency in the presence of interference. However, as we assume small values of u, the possible distortion of the frequency density may be ignored. Using (7), the probability that no segregating allele is observed can be approximated by

$$J_0 = 1 - \int_{1/4N}^{1-1/4N} J(y)\,dy. \tag{8}$$

Because of our assumption that at most one beneficial allele B_2 exists, J_0 is nonnegative.

Now suppose that B_1 appears in the population when the frequency of B_2 is x and that no additional beneficial mutation occurs after B_1. If B_1 occurs on a chromosome carrying B_2, the relative fitness of this chromosome is $\sim 1 + 2s$ while the mean relative fitness of the population is given by $1 + sx$, ignoring the contribution of the chromosome (with both B_1 and B_2) to mean fitness. Assuming that the population size remains constant each generation, the absolute fitness (*i.e.*, the mean number of its copies at the next generation) of this chromosome is thus $(1 + 2s)/(1 + sx) \approx 1 + 2s - sx$. Then, the theory of branching processes (Barton 1995) predicts that the fixation probability of B_1, f, satisfies the equation $1 - f = \exp[-(1 + 2s - sx)f]$. $f = f_0(1/2N, 2s - sx) = 1 - \exp(-4s + 2sx)$ is a good approximate solution of this equation.

On the other hand, if B_1 arises in repulsion phase with B_2, the fixation of B_1 depends on two conditions: First, since one chromosome carrying B_1 and $2Nx$ chromosomes carrying B_2 are selectively equivalent, they comprise a subpopulation of effectively identical chromosomes. This subpopulation of chromosomes increases in frequency and eventually goes to fixation with probability $f_0(x + 1/2N, s) = 1 - \exp(-4Nsx - 2s)$. Second, B_1 displaces B_2 within that subpopulation by drift with probability $1/(2Nx + 1)$. Therefore, B_1 is fixed approximately with probability

$$f_1 = f_0\left(\frac{1}{2N}, s\right)J_0 + \int_0^1 \left\{x(1 - e^{-4s+2sx}) + (1 - x)\frac{1 - e^{-4Nsx-2s}}{2Nx + 1}\right\}J(x)\,dx, \tag{9}$$

unless a beneficial mutation occurs later to compete with B_1 after B_1 survives the initial drift phase.

In the next step, we consider this latter possibility. A new beneficial mutation that arises in repulsion with B_1 can compete with B_1 for fixation. If the frequency of B_1 is y when the new mutation occurs, the fixation probability of the latter is approximately $f_0(y + 1/2N, s)/(2Ny + 1)$, following the argument above. Therefore, the probability that B_1 is not lost in the competition with a late-arising mutation is approximately

$$f_2 = \exp\{-2NuL\int_0^{t_f}(1 - y(t))\frac{1 - e^{-4Nsy(t)-2s}}{2Ny(t) + 1}dt\}, \tag{10}$$

where $y(t)$ is the frequency of B_1 as a function of time t, and $t = 0$ and $t = t_f$ mark the times of the occurrence and fixation of B_1, respectively. It is not easy to model $y(t)$ because the frequency fluctuates due to drift at the beginning and the end of the substitution. We thus approximate the course of substitution assuming that $y(t)$ immediately jumps from $1/2N$ to ε at the beginning and from $1 - \varepsilon$ to 1 at the end. Between ε and $1 - \varepsilon$, the frequency is assumed to change deterministically. Therefore $y(t) = \varepsilon/(\varepsilon + (1 - \varepsilon)\exp(-st))$ and $t_f = -(2/s)\ln(\varepsilon)$ (Stephan *et al.* 1992). We choose $\varepsilon = 1/(2Ns)$, which, in another study (Kim and Stephan 2002), led to a good approximation of the trajectory of the beneficial allele. Thus, we have an approximate formula for the fixation probability under interference as

$$\Phi = f_1 f_2, \tag{11}$$

where f_1 and f_2 are given by Equations 9 and 10, respectively. Figure 3A compares the fixation probability predicted by Equation 11 to those obtained by simulation. As expected, Equation 11 is accurate for small values of u, but discrepancies occur when u is large.

The effective population size under the model of recurrent sweeps for a nonrecombining chromosome and equally advantageous mutations is predicted from Equation 6 to be

$$N_e \approx \frac{N}{1 + 4N^2 uL\Phi}, \tag{12}$$

where Φ is given by Equation 11. Figure 4A shows the expected level of genetic variation (N_e/N) as a function of u. $\hat{N}_e$ obtained in the simulation is slightly less than the value predicted from Equation 12.

Theoretical predictions can also be derived for the multilocus model when two different selection coefficients are used (Table 3B, $r = 0$). Here we are interested in whether a great reduction of the fixation probability at the weak loci can lead to a discrepancy between the results in the presence and absence of interference that is larger than that in the case of uniformly strong beneficial mutations. As previously, we put $s_s = 2s_w$ and $u_w = 2u_s$; furthermore, we consider L_s (= 30) strong loci and L_w (= 30) weak loci. Because of the asymmetry of the effect of interference between weakly and strongly selected mutations (Barton 1995), it is assumed that the fixation probability at the strong loci is affected only by interference from other strong loci. Therefore Equation 11, with s_s and L_s instead of s and L, is used to obtain the average fixation probability, Φ_s, at the strong loci. On the other hand, the fixation probability at the weak loci, Φ_w, is determined mainly by interference due to strong beneficial mutations. If the weak mutation occurs on a chromosome carrying the strong allele, the absolute fitness of this chromosome becomes $\sim 1 + s_w + s_s(1 - x)$, where x is the frequency of the strong allele. On the other hand, if the weak allele occurs in repulsion with the strong allele, the weak allele can be fixed only if (1) the strong allele is lost and (2) the weak allele survives genetic drift. Therefore, a weak allele goes to fixation approximately with probability

$$f_1^w = (1 - e^{-2s_w}) J_0 + \int_0^1 \{x(1 - e^{-2s_w - 2s_s(1-x)}) + (1 - x)(1 - e^{-2s_w}) e^{-4Ns_s x}\} J(x)\, dx,$$

where $J(x)$ and J_0 are defined by Equations 7 and 8, respectively (but using s_s, u_s, and L_s instead of s, u, and L). Furthermore, when a weak allele that survived the initial drift increases to frequency y_w, a strong allele may appear in repulsion with the weak allele and be fixed with probability $1 - \exp(-2s_s + 2s_w y_w)$. Therefore, we approximate the probability that a weak allele on the way to fixation is not lost by a late-arising strong allele as

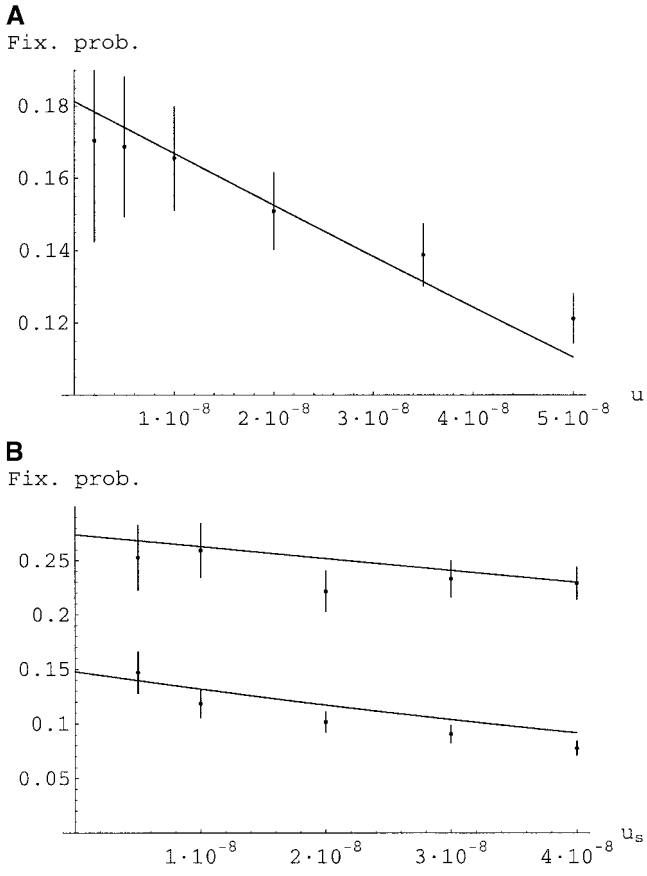

Figure 3.—Fixation probability as a function of the mutation rate under recurrent selective sweeps with no recombination. Parameter values are the same as in Table 3. (A) Uniform selection coefficients (Φ). The line is drawn using Equation 11. Mean ± 2 SE of simulation results are shown as vertical lines. (B) Unequal selection coefficients and rates. The top line (Φ_s) and bottom line (Φ_w) are drawn using Equations 11 and 13, respectively. Mean ± 2 SE of simulation results are shown as vertical lines.

$$f_2^w = \exp\left(-\int_0^{t_f} 2Nu_s L_s (1 - y_w(t))(1 - e^{-2s_s + 2s_w y_w(t)})\, dt\right),$$

where $y_w(t) = \varepsilon/(\varepsilon + (1 - \varepsilon)\exp(-s_w t))$ and $y_w(t_f) = 1 - \varepsilon$ [with $\varepsilon = 1/(2Ns_w)$]. Then the fixation probability of a weak allele is given by

$$\Phi_w = f_1^w f_2^w. \tag{13}$$

Figure 3B compares Φ_s and Φ_w obtained in this way with the simulations. Finally, N_e in this model of recurrent selective sweeps is given by

$$N_e \approx \frac{N}{1 + 4N^2(u_s L_s \Phi_s + u_w L_w \Phi_w)}. \tag{14}$$

Predictions of N_e as a function of u_s with and without interference are shown in Figure 4B. Equation 14 is in good agreement with the simulation results. It is also observed that the theoretical prediction of N_e with interference is not much different from that without interference. Therefore, despite a great reduction of the fixa-

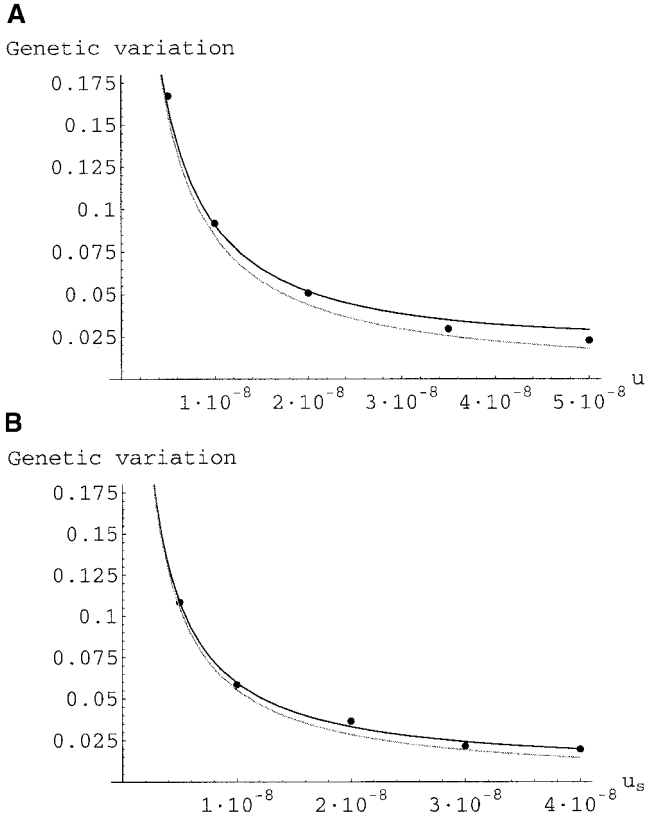

FIGURE 4.—Relative reduction of genetic variation (N_e/N) as a function of the beneficial mutation rate for zero recombination. Parameter values are the same as in Table 3. (A) Uniform selection coefficients. The solid curve is drawn using Equation 12; the shaded curve is drawn by substituting $1 - \exp(-2s)$ for Φ in Equation 12 (prediction with no interference). Dots represent simulation results. (B) Unequal selection coefficients and rates. The solid curve is drawn using Equation 14; the shaded curve is obtained by replacing Φ_s and Φ_w with $1 - \exp(-2s_s)$ and $1 - \exp(-2s_w)$, respectively, in Equation 14 (prediction with no interference). Dots represent simulation results.

tion probability at weak loci, interference failed to elevate N_e much from its expectation under no interference (compare Figure 4A with 4B).

It is consistently found in Figure 4A and Table 3, A and B ($r = 0$), that values of $\hat{N}_e$ are between the predictions of N_e with and without interference. This is unexpected because the reduction of the hitchhiking effect for a given substitution (Figure 2), which may further elevate N_e, was not considered in the prediction with interference. Therefore, another process causing an additional reduction of variation that has so far been neglected may play a role. One possible explanation is that many beneficial mutations increase in frequency substantially but fail to reach fixation due to the interference among them. This transient increase of mutations, called a "leapfrog" event (GERRISH and LENSKI 1998), may also contribute significantly to the reduction of the level of genetic variation. To confirm that leapfrog effects occur, simulations were run using the same parameter values as in Table 3. The number, $k_{0.5}$, of beneficial alleles that increase above frequency 0.5 but fail to be fixed were counted. The number, k_1, of beneficial alleles that went to fixation during the same period was also recorded. For uniform selection coefficients, $k_{0.5}/k_1$ was 0.025 and 0.043 for $u = 2 \times 10^{-8}$ and 4×10^{-8}, respectively. On the other hand, for unequal selection coefficients, $k_{0.5}/k_1$ was 0.028, 0.055, and 0.173 for $u_w = 10^{-8}$, 2×10^{-8}, and 4×10^{-8}, respectively, for the weak loci.

DISCUSSION

KAPLAN *et al.* (1989) studied recurrent selective sweeps in a restricted parameter range for which an overlap of selective phases is minimal. They assumed the rate of selected substitutions to be under a certain limit such that the probability, $P(\Omega)$, of two gene lineages experiencing only nonoverlapping selective phases before they find a common ancestor (Equation 21 of KAPLAN *et al.* 1989) remains close to 1.0. WIEHE and STEPHAN (1993) and BRAVERMAN *et al.* (1995) followed these guidelines. By applying the approach of KAPLAN *et al.* (1989) to our multilocus model (uniform selection coefficients), we obtain

$$P(\Omega) = \frac{1 + 2N\nu Hp}{1 + 2N\nu(Hp + L_{\max}(1 - p))}, \qquad (15)$$

where $L_{\max}$ is the number of selected loci defined below, $\nu = 2Nu(1 - e^{-2s})$, $H = \sum_{i=1}^{L_{\max}} (1 - h_i)$, and $p = 2N^{-2L_{\max}\nu/s}$. The maximum number of selected loci, $L_{\max}$, is determined such that the maximum of the recombination fraction between the neutral locus and the most distant selected locus [equivalent to $M/(2N)$ in KAPLAN *et al.* 1989] is $s/2$, since the hitchhiking effect is effectively negligible with $r > s/2$. Table 3A shows that $P(\Omega)$ for parameters chosen in this study may be $\ll 0.99$, the value KAPLAN *et al.* (1989) used. Although the theoretical prediction assuming no interference underestimates N_e in this parameter range, the difference is not great (Table 3A and Figure 4A). Thus, it appears that, as long as the recombination rate between loci is not too small, analytic solutions for selective sweeps developed for rare selected substitutions can still perform well for calculating the reduction of variation. Therefore, previous data analyses that used the theoretical results obtained for the case of no interference have restricted the parameter ranges unnecessarily much (WIEHE and STEPHAN 1993; STEPHAN 1995).

Little information is available to assess how widespread interference among beneficial mutations is in natural populations. However, there is some evidence that interference is common. WIEHE and STEPHAN (1993), STEPHAN (1995), and ANDOLFATTO (2001) showed that for the recurrent selective sweep model to fully account for the positive correlation between variation and recombination in *Drosophila melanogaster*, the intensity of directional selection $\alpha\nu$, where $\alpha = 2Ns$ and ν is the rate of selected substitution per nucleotide, should be

somewhere between 10^{-8} and 10^{-7}. Assuming $\alpha\nu = 10^{-8}$, $N = 10^6$, and $s = 10^{-3}$, $P(\Omega)$ calculated for a chromosomal region with a moderately low per-nucleotide recombination rate $\rho = 10^{-9}$ is 0.58 (Equation 21 of Kaplan *et al.* 1989). Therefore, as Przeworski (2002) pointed out, overlapping selective sweeps should occur with high probability if selective sweeps contribute significantly to the observed level of variation in *D. melanogaster*. Similarly, the patterns of variation in humans indicate the presence of overlapping selective sweeps (Przeworski 2002).

It is important to understand the relationship between the standing level of variation and local recombination rate in various models of selection. The level of variation determined by recurrent selective sweeps (without interference) may be much more sensitive to the change of local recombination rates over genomic regions than that determined by background selection (Kim and Stephan 2000). It was suggested that this difference may be used to distinguish selective sweeps and background selection as main contributors to the level of variation (H. Innan, personal communication). However, interference among beneficial mutations may slow down the reduction of variation in a region of very low recombination and thus make it difficult to distinguish sweeps from background selection. We examined where in the parameter space interference affects the relationship between recombination and variation. It appears that interference is important only in regions of low recombination where standing variation is highly reduced (Table 3 and Figure 4). This can be understood by examining Equation 6. A decrease of the fixation probability and the hitchhiking effect due to interference cannot significantly modify $E[\hat{N}_e]$ if the term $4N^2\Sigma u_i\Phi_i(1 - h_i)$ (in the absence of interference) is <1. However, interference becomes important when N_e/N has already been greatly reduced [$4N^2\Sigma u_i\Phi_i(1 - h_i) \gg 1$]. This is consistent with the observation that the level of neutral variation decreases more rapidly with a decreasing recombination rate than does the fixation probability of beneficial mutation (compare the changes of Φ and $\hat{N}_e$ in Table 3). Therefore interference among beneficial mutations does not happen without having a great reduction in neutral variation.

Although we have shown that the reduction of effective population size or heterozygosity by selective sweeps is not much influenced by interference, it is not known at present whether other important aspects of variation, such as the frequency spectrum or linkage disequilibrium (Fay and Wu 2000; Kim and Stephan 2002; Przeworski 2002), are also insensitive to interference among beneficial mutations. With the forward simulation proposed in this study, one may detect a change in the frequency spectrum by observing higher moments of the frequency of ancestral numbers, namely $I_k(t)$ ($k \geq 3$). A preliminary study, using $(I_3(t) - I_4(t))/(I_2(t)^2 - I_4(t))$ as a test statistic, showed that the frequency spectrum is less skewed if beneficial alleles arise on different chromosomes and thus compete with each other. More analysis of this statistic is in progress.

We thank Allen Orr for pointing out the leapfrog effect and Hideki Innan and Rasmus Nielsen for discussion. We also thank Deborah Charlesworth and two anonymous reviewers for comments that greatly improved the manuscript. This research was supported by funds from the Deutsche Forschungsgemeinschaft to W.S. and National Science Foundation grant DEB-0089487 to Rasmus Nielsen.

LITERATURE CITED

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

Barton, N. H., 1995 Linkage and the limits to natural selection. Genetics **140:** 821–884.

Barton, N. H., 1998 The effect of hitch-hiking on neutral geneologies. 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.

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., 1996 Background selection and patterns of genetic diversity in *Drosophila melanogaster*. Genet. Res. **68:** 131–149.

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

Ewens, W. J., 1979 *Mathematical Population Genetics*. Springer-Verlag, New York.

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

Gerrish, P. J., and R. E. Lenski, 1998 The fate of competing beneficial mutations in an asexual population. Genetica **102/103:** 127–144.

Gillespie, J. H., 2000a The neutral theory in an infinite population. Gene **261:** 11–18.

Gillespie, J. H., 2000b Genetic drift in an infinite population: the pseudohitchhiking model. Genetics **155:** 909–919.

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

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

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

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

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

Sawyer, S. A., and D. L. Hartl, 1992 Population genetics of polymorphism and divergence. Genetics **132:** 1161–1176.

Slatkin, M., 1991 Inbreeding coefficients and coalescence times. Genet. Res. **58:** 167–175.

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.

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.

Communicating editor: D. Charlesworth