Soft Sweeps II—Molecular Population Genetics of Adaptation from Recurrent Mutation or Migration

Pleuni S. Pennings and Joachim Hermisson

Section of Evolutionary Biology, Department of Biology II, Ludwig-Maximilians-University Munich, Planegg-Martinsried, Germany

In the classical model of molecular adaptation, a favored allele derives from a single mutational origin. This ignores that beneficial alleles can enter a population recurrently, either by mutation or migration, during the selective phase. In this case, descendants of several of these independent origins may contribute to the fixation. As a consequence, all ancestral haplotypes that are linked to any of these copies will be retained in the population, affecting the pattern of a selective sweep on linked neutral variation. In this study, we use analytical calculations based on coalescent theory and computer simulations to analyze molecular adaptation from recurrent mutation or migration. Under the assumption of complete linkage, we derive a robust analytical approximation for the number of ancestral haplotypes and their distribution in a sample from the population. We find that so-called "soft sweeps," where multiple ancestral haplotypes appear in a sample, are likely for biologically realistic values of mutation or migration rates.

Introduction

When a beneficial allele rises to fixation in a population, it erases genetic variation in a stretch of DNA that is linked to it. This phenomenon is called "genetic hitchhiking" or a "selective sweep" and was first described by Maynard Smith and Haigh (1974). In the classical scenario for such an adaptive substitution, the beneficial allele arises in the population as a single new mutation and then increases to fixation under a constant selection pressure. Under this scenario, genetic variation in parts of the genome that are tightly linked to the selected site is lost and will only be recovered by new mutation. Ancestral variation, that is, genetic variation that has been present in the population prior to the selective phase, is only maintained if recombination during the selective phase breaks the association between the study locus and the selected site. The resulting pattern of a selective sweep, a valley of reduced variation around the target of selection, has been described in some detail and is well understood (e.g., Kaplan, Hudson, and Langley 1989; Stephan, Wiehe, and Lenz 1992; Barton 1995; Durett and Schweinsberg 2004; Etheridge, Pfaffelhuber, and Wakolbinger 2005).

There is, however, a second scenario as to how ancestral variation can be maintained in the face of positive selection, namely, if an adaptive substitution involves multiple copies of the same beneficial allele. This can happen in the following two ways. If adaptation occurs from the standing genetic variation, a large number of copies of the beneficial allele may be initially present. Fixation of the allele may then involve descendants of more than one of these copies. Alternatively, a beneficial allele can enter the population recurrently by mutation or migration during the selective phase. Again, descendants of several of these independent origins may contribute to the fixation of the allele. In both cases, ancestral haplotypes that are linked to any of these copies will be retained in the population. Clearly, this would affect the pattern of a selective sweep on linked DNA variation. We call selective sweeps that involve (descendants of) more than one copy of the selected allele, "soft sweeps." They are distinguished from the classical "hard sweeps" where ancestral variation is maintained only through recombination.

Selective sweeps from the standing genetic variation have been described in three recent publications. Hermisson and Pennings (2005) derive the probability for a soft sweep for adaptation from the standing genetic variation. Innan and Kim (2004) and Przeworski, Coop, and Wall (2005) describe the effect of an adaptive substitution from the standing variation on summary statistics for DNA variation, assuming that the allele had been neutral prior to the onset of positive selection. There is then the chance that ancestral variation—due to mutation during this first time period—is retained in the population even without recombination. However, as long as there is only a single origin of the beneficial allele (as assumed by Innan and Kim [2004] and Przeworski, Coop, and Wall [2005]), the effect is necessarily limited. Other than in the case of recombination, the surviving ancestral haplotypes are not independent but identical by descent.

In this study, we focus on selective sweeps from a beneficial allele that enters the population recurrently by mutation or migration. We derive the probability for a soft sweep, given the mutation/migration rate and the selection coefficient of the beneficial allele. More generally, we determine the expected number of independent ancestral haplotypes and their frequency distribution in a sample from a locus that is tightly linked to the selected site. Our results show that soft sweeps are likely under biologically realistic conditions.

Model and Methods

Model and Definitions

We study a single locus under selection in a haploid population of effective size N_e. For most of this study, only two alleles (or classes of alleles) at this locus are considered, an ancestral allele b and a new beneficial variant B with fitness advantage s. In general, we will allow s to depend on time and/or on the frequency of the beneficial allele. The B allele enters the population through either recurrent mutation at rate u or migration at rate m (where m is the per generation probability for an individual to be replaced by a migrant). We consider mutation and migration

Key words: adaptive evolution, selective sweeps, hitchhiking, migration.

E-mail: hermisson@zi.biologie.uni-muenchen.de.

Mol. Biol. Evol. 23(5):1076–1084. 2006
doi:10.1093/molbev/msj117
Advance Access publication March 6, 2006

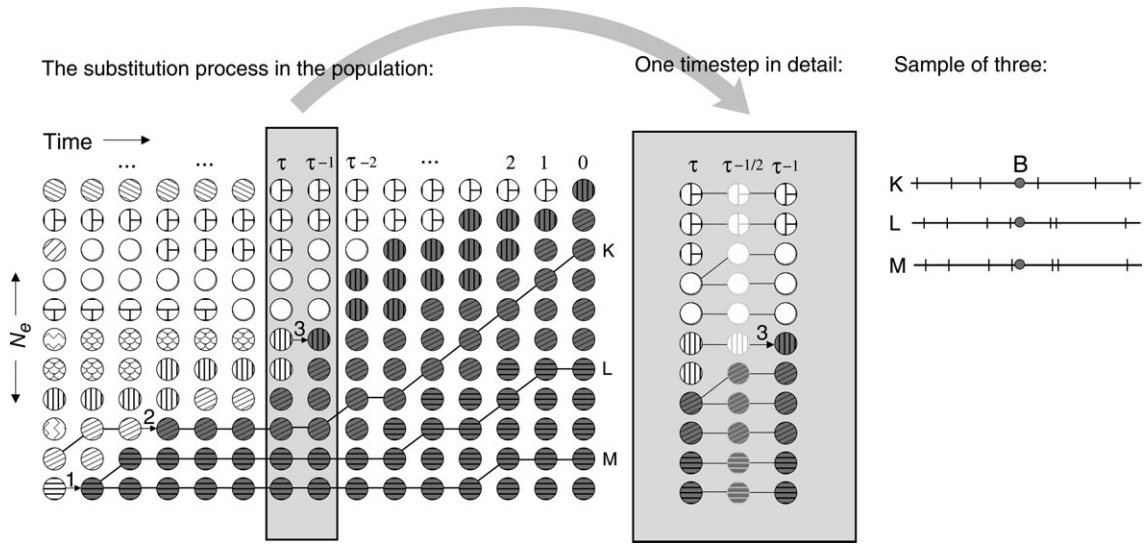

FIG. 1.—Soft selective sweep from recurrent mutation in a schematic Wright-Fisher model. Circles represent individuals, the different patterns indicate independent ancestral haplotypes. The beneficial allele *B* (dark gray individuals) substitutes the ancestral *b* allele (white). The *B* allele arises three times by independent mutation; individuals then change their color from white to gray but keep their haplotype pattern. The "zoom" into a single time step shows how reproduction and mutation are separated. Directly after fixation (time 0), we take a sample of size three (K, L, M) that contains descendants from the first (L, M) and the second (K) mutational origins of *B*. The right panel shows DNA fragments of the sampled individuals. The vertical ticks represent neutral polymorphisms. Individuals L and M share a recent ancestor and are identical in this region of the genome. Individual K carries a different ancestral haplotype.

separately. Back mutation or migration is ignored. We define population-level parameters for selection, mutation, and migration as $\alpha = 2N_e s$, $\Theta = 2N_e u$, and $M = 2N_e m$. Every generation consists of reproduction (including fertility selection), followed by mutation or migration, see figure 1.

Assume that the population is originally monomorphic for the ancestral allele *b*. After successful substitution, all individuals carry the *B* allele. Because mutation or migration is recurrent, this substitution may involve several copies of the *B* allele with independent origins in the sense that they do not trace back to a single ancestor in the study population. Independent copies are linked to independent genetic backgrounds that are randomly drawn either from the study population prior to the substitution or from the source population of migrants. We call these independent genetic backgrounds at the selected locus "independent ancestral haplotypes," or "ancestral haplotypes" for short. Note that with this definition differences due to new mutations or recombination events (i.e., events after the first beneficial mutation or migration event) are not considered. Note also that "independent" does not necessarily mean "different" because it includes the possibility that the same haplotype is drawn multiple times.

Suppose that we take a sample from the selected locus or from a tightly linked fragment (so that no recombination has taken place between this fragment and the selected locus) some time after fixation of the *B* allele. If there is more than one ancestral haplotype in the sample, we call this a soft selective sweep from recurrent mutation or migration. The opposite case (only one ancestral haplotype) is called a hard sweep. Note that soft and hard sweeps can be defined either with respect to a sample or with respect to the population. A soft sweep in a population means that there are several ancestral halpotypes at the selected locus in the population. In this paper, we usually consider samples.

Simulations

We checked all analytical results by forward-in-time computer simulations. For this, a Wright-Fisher model with $N_e = 500{,}000$ haploid individuals is simulated. Each run starts with a population that is monomorphic for the ancestral *b* allele. Reproduction is simulated by fitness-weighted multinomial sampling. After reproduction, every *b* individual has probability *u* to mutate to *B*. In the migration model, every individual, independent of its genotype, is replaced by a migrant with probability *m*. Descendants of mutants and migrants are followed separately; at the observation time their frequencies are determined in a population sample. Data points are averages over 100,000 runs (10,000 runs for $\alpha = 100$). The code is available on request.

Results

This section is organized in four parts. The first two consider recurrent mutation. We start with a detailed derivation for a sample of two, which is the simplest case. We then use intuitive arguments to motivate our main result, which is the frequency distribution of ancestral haplotypes for a sample of size *n*. All formal derivations for this general case are given in the Supplementary Material online. In the third part, we show how these results apply to the recurrent migration case. Finally, we briefly discuss several generalizations of the model.

Soft Sweeps from Recurrent Mutation in a Sample of Size Two

Consider a sample of size two that is taken from a population at some time t_{obs}, measured from the time of fixation of the beneficial *B* allele. Initially, we will assume that sampling occurs directly at fixation, that is, $t_{obs} = 0$. We

want to derive the probability $P_{\text{soft},2}$ that the two copies of the B allele in the sample are not identical by descent, that is, the probability of a soft selective sweep. We use a coalescent framework and define τ as the time in the past before the sample was taken, that is, $\tau = 0$ for $t = t_{\text{obs}}$ and if $\tau_2 > \tau_1$, then τ_2 is further back in the past than τ_1.

Let x_τ be the fraction of the population that carries the beneficial allele B at time τ. We follow the fate of the two lineages backward in time until they either coalesce or one of the two mutates. $P_{\text{soft},2}$ is the probability that mutation happens before coalescence; we denote the alternative possibility that the lines coalesce before one of the two mutates as $P_{\text{hard},2} = 1 - P_{\text{soft},2}$.

Let $P_{\text{coal},2}(\tau)$ be the coalescence probability in generation τ and $P_{\text{mut},2}(\tau)$ the probability that one of the two lineages has mutated and had a b ancestor in generation τ. We can then express $P_{\text{hard},2}$ as

$$P_{\text{hard},2} = \left\langle \sum_{\tau=1}^{\infty} \left(P_{\text{coal},2}(\tau) \prod_{i=1}^{\tau-1} (1 - P_{\text{mut},2}(i) - P_{\text{coal},2}(i)) \right) \right\rangle_x, \tag{1}$$

where the empty product, $\prod_{i=1}^{0}$, is defined to be 1. The product is the probability that neither mutation nor coalescence has happened until generation τ. $\langle \ldots \rangle_x$ denotes the expectation over the stochastic path $\{x_\tau\}_\tau$ of the frequency x_τ of B.

To calculate $P_{\text{mut},2}$, it is convenient to separate reproduction (and therefore coalescence) from mutation by introducing an artificial intermediate generation after reproduction but before mutation: Using the backward-in-time notation, individuals of generation τ reproduce to form generation $\tau - \frac{1}{2}$, and the individuals in this intermediate generation can mutate or not to form $\tau - 1$ (see fig. 1). Ignoring back mutation from B to b, the number of B alleles in the $(\tau - 1)$th generation is given by

$$x_{\tau-1} = \left(1 - x_{\tau-\frac{1}{2}}\right) u + x_{\tau-\frac{1}{2}}. \tag{2}$$

in which the first term on the right-hand side is the new mutants and the second term is the B's that were already there. For a single B lineage, the probability that it is a mutant is

$$P_{\text{mut},1}(\tau) = \frac{\left(1 - x_{\tau-\frac{1}{2}}\right) u}{\left(1 - x_{\tau-\frac{1}{2}}\right) u + x_{\tau-\frac{1}{2}}} = \frac{(1 - x_{\tau-1}) u}{(1 - u) x_{\tau-1}}, \tag{3}$$

where $x_{\tau-\frac{1}{2}} = \frac{x_{\tau-1} - u}{1 - u}$ (from eq. 2). Thus, the probability for (at least) one mutation in a sample of two is $P_{\text{mut},2} = 2P_{\text{mut},1} - P_{\text{mut},1}^2$. If no mutation has happened, coalescence happens with rate $1/(N_e x_\tau)$. The exact coalescence probability in generation τ therefore is

$$P_{\text{coal},2}(\tau) = \frac{1 - P_{\text{mut},2}(\tau)}{N_e x_\tau}. \tag{4}$$

In a sufficiently large population, and for small values of u, we can safely ignore the occurrence of several events in a single generation. Formally, this is done by ignoring terms of order $u/(N_e x)$ and u^2. If we can also ignore terms of order $s/(N_e x)$, we can further set $x_{\tau-1} \approx x_\tau$. We then obtain

$$P_{\text{mut},2}(\tau) \approx \frac{\Theta(1 - x_\tau)}{N_e x_\tau}, \quad P_{\text{coal},2}(\tau) \approx \frac{1}{N_e x_\tau} \tag{5}$$

for the probability of mutation and coalescence. Using equation (5) in equation (1), $P_{\text{hard},2}$ can now be expressed as

$$\begin{aligned} P_{\text{hard},2} &= \left\langle \sum_{\tau=1}^{\infty} \left\{ \frac{1}{N_e x_\tau} \prod_{i=1}^{\tau-1} \left(1 - \frac{1 + \Theta(1 - x_i)}{N_e x_i}\right) \right\} \right\rangle_x \\ &= \frac{1}{1 + \Theta} \left[\left\langle \sum_{\tau=1}^{\infty} \left\{ \frac{1 + \Theta(1 - x_\tau)}{N_e x_\tau} \right.\right.\right. \\ &\quad \left.\left.\left. \times \prod_{i=1}^{\tau-1} \left(1 - \frac{1 + \Theta(1 - x_i)}{N_e x_i}\right) \right\} \right\rangle_x \right. \\ &\quad \left. + \left\langle \sum_{\tau=1}^{\infty} \left\{ \frac{\Theta x_\tau}{N_e x_\tau} \prod_{i=1}^{\tau-1} \left(1 - \frac{1 + \Theta(1 - x_i)}{N_e x_i}\right) \right\} \right\rangle_x \right] \\ &= \frac{1}{1 + \Theta} \left[1 + \frac{\Theta}{N_e} \left\langle \sum_{\tau=1}^{\infty} \prod_{i=1}^{\tau-1} \left(1 - \frac{1 + \Theta(1 - x_i)}{N_e x_i}\right) \right\rangle_x \right], \end{aligned} \tag{6}$$

where the sum in the second line is the probability that the two lineages eventually either coalesce or mutate, which is 1 for every realization of the path $\{x_\tau\}_\tau$. The expectation in the last line of equation (6) has a simple interpretation. It is the average time, T_1, in generations until either coalescence or mutation happens. T_1 certainly lies between 0 and T_{fix}, the average fixation time for the beneficial allele in the population. This gives an upper and lower bound for $P_{\text{hard},2}$ as

$$\frac{1}{1 + \Theta} \le P_{\text{hard},2} \le \frac{1}{1 + \Theta} \left(1 + \frac{\Theta T_{\text{fix}}}{N_e}\right). \tag{7}$$

Equivalently,

$$\frac{\Theta}{1 + \Theta} \ge P_{\text{soft},2} \ge \frac{\Theta}{1 + \Theta} \left(1 - \frac{T_{\text{fix}}}{N_e}\right). \tag{8}$$

This result has several important implications. First, none of the details of the stochastic process that underlies the path $\{x_\tau\}_\tau$ enters into the estimate for $P_{\text{hard},2}$ or $P_{\text{soft},2}$. In fact, the value of T_{fix} in one of the bounds is the only quantity that depends on this process—and thus on the selection coefficient. Second, we see that the estimate gets very precise (upper and lower bounds converge) if $T_{\text{fix}}/N_e \ll 1$. This is easily fulfilled for strong selection. In this case, $P_{\text{hard},2}$ and $P_{\text{soft},2}$ depend only on Θ but are entirely independent of all selection parameters.

Finally, one should note that the derivation does not depend on the assumption that the sample is taken directly after fixation. Assume, instead, that the population is sampled some time t_{obs} after fixation. In that case T_{fix} in equations (7) and (8) has to be replaced by the expected age of

the oldest B allele that is found in the population at the time of observation. The approximation will be good as long as $(t_{obs} + T_{fix})/N_e \ll 1$. If the sample is taken before full fixation, the above estimates (7) and (8) hold if we condition on a sample that is monomorphic for B.

To assess the quality of the bounds for $P_{soft,2}$ in equation (8), we need an estimate of T_{fix}. For a single copy of a beneficial allele that rises to fixation under a constant selection pressure $\alpha = 2N_e s$, a precise estimate of the fixation time is $T_{fix}/N_e \approx 4 \log(\alpha)/\alpha$ (Hermisson and Pennings 2005). For $\Theta \ll \alpha$, the same approximation holds also for fixation under recurrent mutation. With this estimate for T_{fix}, both bounds deviate by $<5\%$ for $\alpha > 500$. Figure 2 confirms that simulation data fall between the predicted bounds. Only for extremely strong selection ($s \approx 1$), some deviations appear (data not shown). The reason is that the approximation $x_{\tau-1} \approx x_\tau$ that we have used in the derivation is no longer accurate in this case.

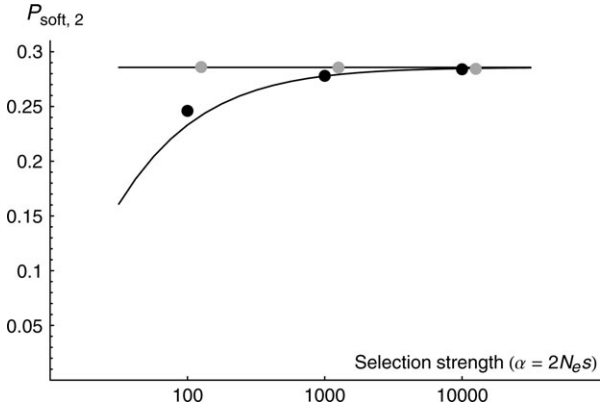

FIG. 2.—The probability of a soft selective sweep in a sample of size two, taken directly after fixation. The horizontal line represents the first-order approximation (upper bound, eq. 8) and the curved line the second-order approximation (lower bound, eq. 8). Dots are simulation results; black dots are for mutation ($\Theta = 0.4$) and the gray dots are for migration ($M = 0.4$).

Soft Sweeps from Recurrent Mutation in Larger Samples

Consider now a sample of size n taken from the population at some time t_{obs}. If sampling occurs before fixation, we condition on samples that are monomorphic for the B allele. We are interested in the number and the frequency distribution of ancestral haplotypes in the sample.

If there are k ancestors of the sample that are associated with a B allele at time τ, the probability for mutation and coalescence at this time is approximately

$$P_{mut,k} \approx \frac{k\Theta(1-x_\tau)}{2N_e x_\tau}, \quad P_{coal,k} \approx \frac{k(k-1)}{2N_e x_\tau}, \tag{9}$$

where x_τ is the frequency of the beneficial allele. Using these relations, we can extend the above approach and calculate upper and lower bounds for the probability of a soft selective sweep. These derivations are given in the Supplementary Material online. Below, we focus on just one of the bounds where a more intuitive derivation is possible.

We need two steps for our argument. First, note that the leading order approximation for a sample of size two (i.e., the lower bound in eq. 7 and the upper bound in eq. 8) is equivalent to an approximation of the mutation probability $P_{mut,2}$. In fact, equation (6) reduces to $1/(1 + \Theta)$ if we ignore the factor $(1 - x_\tau)$ in the numerator of $P_{mut,2}$ in equation (5). We can apply the same approximation to $P_{mut,k}$ in equation (9) and justify this step as follows: the denominator of $P_{mut,k}$ guarantees that mutation is only likely if x_τ is small. In this case, however, $(1 - x_\tau) \approx 1$.

Secondly, without the $(1 - x_\tau)$ term, we see that the coalescence and mutation rates in equation (9) are both proportional to $(1/x_\tau)$. If we are only interested in the order of events in the genealogy of the sample (and not in the exact times at which coalescence and mutation happen), only the relative rates matter and we can ignore the x_τ dependence altogether (see the Supplementary Material online for a formal derivation). The result is that the problem is now equivalent to a standard neutral coalescent in a population of constant size where lines are stopped at mutations (also called "coalescent with killings," Durett 2002). This problem is long known and can be exactly solved (e.g., Ewens 2004, p. 335ff). In particular, the expected number of haplotypes and their frequency distribution are given by the Ewens sampling formula: given the mutation rate Θ for the B allele, the probability to find k haplotypes, occurring $n_1, \ldots, n_k$ times in a sample of size $n = \sum_i n_i$ is

$$\Pr(n_1 \ldots n_k | n, \Theta) = \frac{n!}{k! n_1 \cdots n_k} \frac{\Theta^k}{\Theta(\Theta+1)\cdots(\Theta+n-1)}. \tag{10}$$

Using this result for $k = 1$ and $n_1 = n$, we obtain an upper bound for the probability of a soft sweep as

$$P_{soft,n} \leq 1 - \Pr(n|n,\Theta) = 1 - \prod_{i=1}^{n-1} \frac{i}{i+\Theta} = a_{n-1}\Theta + \mathcal{O}(\Theta^2), \tag{11}$$

where $a_n = \frac{1}{1} + \frac{1}{2} + \ldots + \frac{1}{n}$. Equation (11) reduces to (8) in the case of $n = 2$. The "$\leq$" expresses the fact that we have overestimated the mutation probability by ignoring the factor $(1 - x_\tau)$ in $P_{mut,k}$. The marginal distributions for the number of haplotypes k and the distribution for fixed k can also be given

$$\Pr(k|n,\Theta) = \frac{\Theta^k S_n^{(k)}}{\Theta(\Theta+1)\cdots(\Theta+n-1)}, \tag{12}$$

where $S_n^{(k)}$ is Stirling's number of the first kind and

$$\Pr(n_1 \ldots n_k | k, n, \Theta) = \frac{n!}{k! n_1 \cdots n_k S_n^{(k)}}. \tag{13}$$

In figures 3–5 we compare the estimates from equations (11)–(13) with simulation data for samples that are drawn at the time of fixation of the B allele. As can be seen from figures 3 and 5, the predictions are good for strong selection. For $\alpha = 100$, the simulation data deviate more strongly. The same effect is seen if the sample is taken a long time after fixation. The reason is the same as for a sample of size two: if the time from the first origin of the allele to

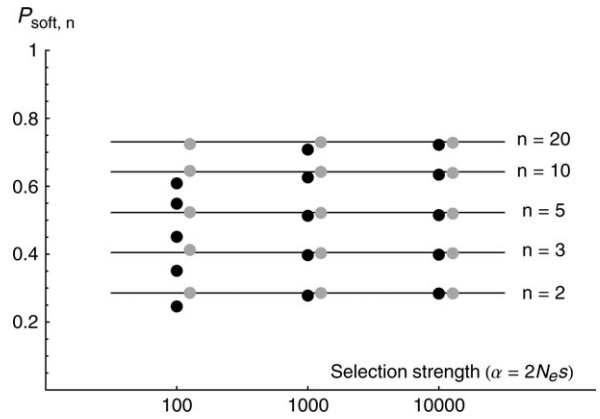

FIG. 3.—The probability of a soft sweep in samples of varying size n, taken directly after fixation. The horizontal lines represent the first-order approximation (upper bound, eq. 11). The dots are simulation results. Black dots are for mutation ($\Theta = 0.4$) and gray for migration ($M = 0.4$).

the observation of the sample is very long, the small error that we have made by ignoring the factor $(1 - x_\tau)$ in the mutation probability accumulates over many generations. In a time-forward picture, this corresponds to the fact that ancestral haplotypes with a low frequency will slowly drift out of the population. Figure 5 shows that the distribution of the remaining haplotypes then becomes more uniform, as is predicted by Kimura (1955). Figure 3 also shows that the approximation works best for small samples sizes (see also the Supplementary Material online).

Equation (11) and figure 4 show that the probability of a soft sweep depends strongly on Θ, the recurrent mutation rate of the beneficial allele on the population level. For low $\Theta < 0.01$, soft selective sweeps from recurrent mutation are rare. For Θ between 0.01 and 0.02 (depending on sample size), they will appear in about 5% of all cases. $0.01 < \Theta < 1$ is the transitional range where both soft and hard sweeps will be found. For high mutation rates with $\Theta > 1$, almost all selective sweeps will be soft (see fig. 4). Equation (11)

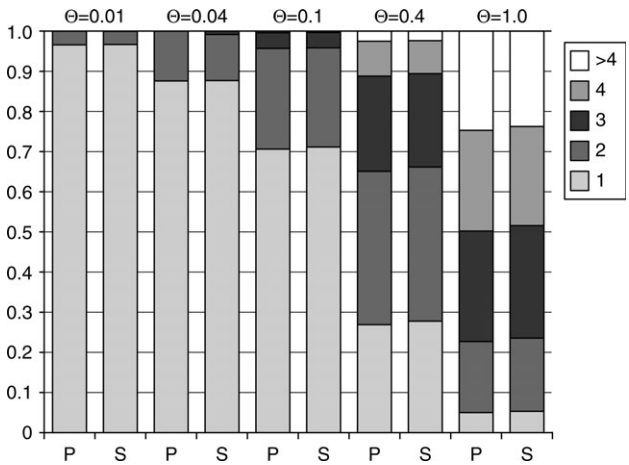

FIG. 4.—The probability of finding 1, 2, 3, 4, or > 4 ancestral haplotypes (different mutational origins of the B allele) in a sample of 20 for different Θ values. For each Θ value, we show the simulation results on the right (marked S) and the prediction left (according to eq. 12, marked P). The simulations use $\alpha = 10{,}000$, the population is sampled directly after fixation.

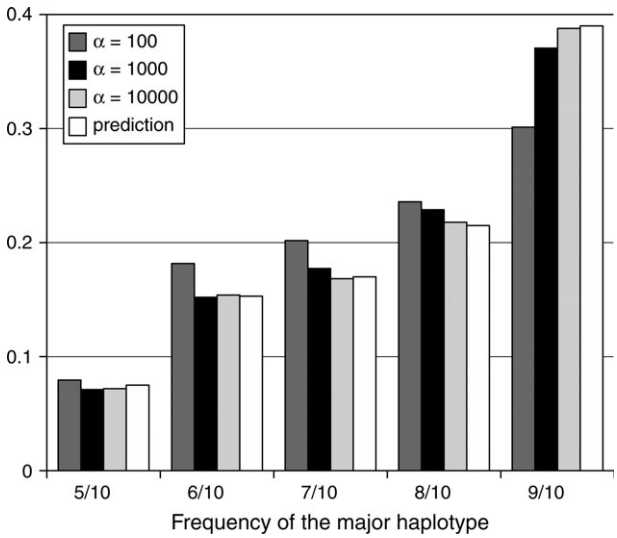

FIG. 5.—Haplotype frequency spectrum: the probability for a major ancestral haplotype with frequency 5 out of 10, 6 out of 10, etc., given that there are two haplotypes in the sample of 10. We show from left to right: $\alpha = 100$; $\alpha = 1{,}000$; $\alpha = 10{,}000$; prediction according to equation (13). $\Theta = 0.1$

shows that there is a logarithmic dependence on the sample size: $P_{\text{soft},n} \approx a_{n-1}\Theta \approx (\gamma + \log(n-1))$ (with Euler's $\gamma = 0.577\ldots$), which can also be seen in figure 3.

To leading order, the probability of a soft selective sweep is independent of the selection strength. However, to second order, and as can be seen in figure 3, $P_{\text{soft},n}$ increases with selection strength. In other words, the tendency to maintain ancestral genetic variation in the face of positive selection increases with stronger selection. This is in strong contrast to the maintenance of variation due to recombination. As explained above, the reason for this effect is the longer fixation time of weakly beneficial alleles. If we sample at a fixed time after the start of the substitution process, the increase disappears (see also the Supplementary Material online).

Finally, we note that the results are slightly different when we consider the entire population instead of a sample. As we reported previously, the probability for a soft sweep on the whole population level increases with selection strength (see Hermisson and Pennings 2005, fig. 6). This holds true even if the sample is taken at a fixed time after the start of the substitution process (results not shown). This indicates that under strong selection more alleles are maintained in a population, which afterward could be picked up by a new selection pressure. Note that our analytical results cannot be extended to the entire population because the approach depends on the assumption that multiple events in a single generation and coalescent events with multiple mergers can be ignored.

Migration

Instead of new mutation, a beneficial allele can also enter a population through recurrent migration. We consider the following scenario. A population is split into two subpopulations. At the B locus, the subpopulation in the first deme is fixed for the B allele since a long time ago; the second subpopulation is initially fixed for the

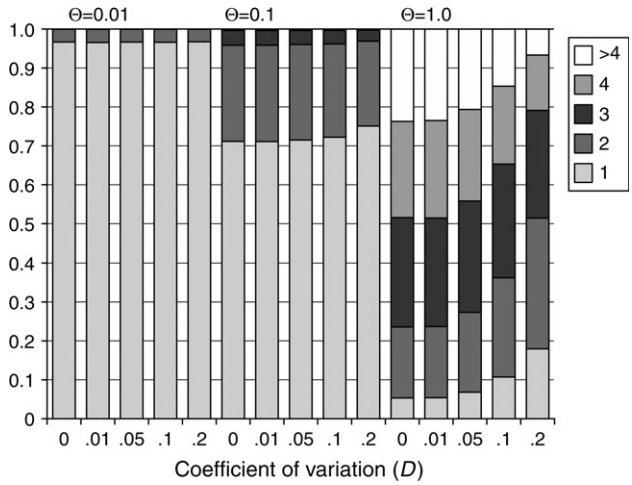

FIG. 6.—Effect of variance in the fitness of B alleles on the number of ancestral haplotypes in a sample. Beneficial mutation produces two kinds of B alleles, $B_{\pm}$, with equal probability. The scaled selection coefficients are $\alpha_{\pm}=(1\pm D)\bar{\alpha}$, where the mean selection strength is $\bar{\alpha}=10,000$. A sample of size 20 is taken at fixation of the $B_{\pm}$ alleles (i.e., when the ancestral b allele is lost). Simulation results are shown for three different Θ values and values of D ranging from 0 (homogeneous fitness) to 0.2 (corresponding to a 50% larger α_+ relative to α_-).

b allele. We assume that gene flow at the B locus into the second subpopulation was inhibited for a long time, either because of geographical isolation or because of selection against the B allele in the second deme. Now, however, both populations are linked through weak migration, and the B allele is beneficial in both demes. We assume that a mutational origin of the B allele in the second subpopulation is unlikely and can be ignored. Thus, adaptation in the second deme will only occur from migrants.

Migration is modeled by a fixed probability m for every individual to be replaced by a migrant. For a (fixed) population size of N_e in the second deme, $N_e m$ is then the average number of successful migrants that arrive in that deme per generation. We ignore the possibility that over the relevant time scale (i.e., the typical fixation time of B), a lineage sampled in the second deme migrates to the first deme and back to the second deme. As a consequence, we can entirely focus on the evolutionary process in the second deme and treat the first deme as a reservoir of independent B haplotypes that enter the second deme at a constant rate.

We are interested in the expected number of different B haplotypes and their frequency distribution in a sample from the second deme. As in the mutation model, we separate the two stages that produce a new generation and introduce an intermediate step after reproduction but before migration. Using the backward-in-time notation, individuals of generation τ reproduce to form generation $\tau-\frac{1}{2}$ and the individuals in this intermediate generation can be replaced or not by migrants to form $\tau-1$. A migrant replaces a random resident individual, independent of the resident's genotype. We can thus write $x_{\tau-1}$ in terms of $x_{\tau-\frac{1}{2}}$ as

$$x_{\tau-1}=m+x_{\tau-\frac{1}{2}}(1-m), \tag{14}$$

where the last term represents the resident B's that are not replaced by migrants. For a B lineage, the probability that it has migrated and has an ancestor in deme one in generation τ is

$$P_{\mathrm{mig},1}(\tau)=\frac{m}{m+x_{\tau-\frac{1}{2}}(1-m)}=\frac{m}{x_{\tau-1}}. \tag{15}$$

Ignoring the probability that multiple events happen in one generation, and using $x_{\tau-1}\approx x_\tau$, the probability for migration, backward in time, for k ancestors at time τ is

$$P_{\mathrm{mig},k}\approx\frac{kM}{2N_e x_\tau}, \tag{16}$$

where $M=2N_e m$. The probability for migration lacks the $(1-x_\tau)$ factor of the mutation probability (eq. 9). While only mutations from b to B introduce a new ancestral haplotype associated with the B allele, every migration, replacing either a b individual or a B individual in the subpopulation in deme two, will add a new B haplotype.

We thus see that the migration and coalescence probabilities are strictly proportional; their relative rates do not depend on the frequency x_τ of the B allele. We can therefore directly map the coalescent to a neutral coalescent. The problem is fully solved by the Ewens sampling formula (eqs. 10–13), with Θ replaced by M, for arbitrary values of the selection coefficient. The simulation data in figures 2 and 3 show that this estimate is highly accurate. Our results can also be applied if the origin of the B allele in the first deme is more recent (less than $2N_e$ generations ago). In this case, however, there is a higher chance that different B haplotypes have a common origin and are thus similar or even identical.

Generalizations of the Model

We have derived our results under a number of simplifying assumptions mainly for the clarity of the presentation. As it turns out, several of these assumptions can be significantly relaxed without changing our results. In this section, we show that the sampling distribution of ancestral haplotypes follows the Ewens sampling scheme as a good first-order approximation under a wide range of biological scenarios.

Back Mutations

Inclusion of back mutations at rate v into the model brings three small changes. First, there is a small additional term proportional to uv added to the mutation probability $P_{\mathrm{mut},n}$ in the coalescent. Second, there is a slight chance that multiple mutations from B to b and back occur on a single line of descent. On the time scales considered here, both these effects can be safely ignored even for high back-mutation rates. Third, back mutation also changes the expected frequency x_τ of the beneficial allele B. In particular, with high v, B may never reach full fixation. However, as long as we condition on samples that are monomorphic for B, this does not affect our results, which do not depend on the stochastic path $\{x_\tau\}_\tau$.

Changing Population Size

In the migration model, we can allow arbitrary changes in the effective population size N_e of the population

in the second deme. To maintain our results, we only need to keep the average number of successful immigrants, $N_e m$ (and thus $M = 2N_e m$), fixed. In generations with small N_e, this is compensated by a higher probability m for each individual to be replaced by a migrant. (For recurrent mutation, a similar assumption of a constant Θ despite changing N_e does not seem to be meaningful.) An important limiting case is that the second deme is initially altogether empty and only colonized by descendants of immigrants that appear at a constant rate. We stress that this is a purely demographic scenario without any positive selection that leads to the same expected pattern of ancestral haplotypes at the study locus. In contrast to selection, however, the pattern should be genome wide in this case.

Mutation and Migration

Without any additional problem, we can combine mutation and migration into a single model. To leading order, the sampling distribution of ancestral haplotypes is then still given by the Ewens equations (10)–(13), with Θ replaced by $\Theta + M$. The leading correction terms are the same as above and depend on Θ only.

Adaptation from Standing Genetic Variation

Because our approximations do not depend on the path, $\{x_\tau\}_\tau$, they are not affected by changes of the selection pressure s as a function of time or of frequency, as long as the fixation time does not become too long. In particular, s may also change its sign during the course of the substitution process. This will be the case if the allele adapts from the standing genetic variation. As in the purely beneficial case, the Ewens approximation will be accurate as long as $(T_{\text{fix}} + t_{\text{obs}}) \ll N_e$. This is always the case if selection (either positive or negative) is strong enough. Note that the sampling distribution counts the numbers of independent haplotypes (independent origins in Hermisson and Pennings 2005). It does not count the number of descendants of all B copies that segregated in the population at the start of positive selection because the latter may still be identical by descent.

Diploidy and Dominance

Formally, our derivations above apply for a haploid model or for a diploid model with complete dominance. In these two cases, every B allele in a parent generation has the same expected number of offspring. In a diploid model with dominance coefficient $h < 1$, the expected number of offspring of a B allele depends on whether it comes from a homozygote BB or a heterozygote Bb individual. This increases the variance in offspring number relative to the haploid case and therefore also the coalescence rate. As shown in the Supplementary Material online, however, the effect is very small, of the order s^2, and can usually be ignored. Dominance further changes the expected frequency path $\{x_\tau\}_\tau$ of the beneficial allele. Because this does not affect our results, they also apply to randomly mating diploids with an arbitrary level of dominance.

Variance in the Fitness Effects

Until now, we have assumed that all beneficial B alleles are of a single type and have the same fitness advantage. If B corresponds to a class of (more of less) physiologically equivalent alleles rather than to a unique molecular genotype, this may not be realistic. It is therefore important to check the stability of our results under variations in fitness among the beneficial alleles. With this aim, we ran additional simulations where we split the B alleles into two classes B_+ and B_-. New mutations are assigned with equal probability to either of these classes. $B_\pm$ alleles have scaled selection coefficients $\alpha_\pm = \bar{\alpha}(1 \pm D)$. With this definition, D is the coefficient of variation of the distribution of α values.

Figure 6 shows that for low Θ, there is no visible difference in the number of ancestral haplotypes relative to the homogeneous case ($D = 0$), even for a large variance among the selection coefficients. For higher Θ, the probability of a soft sweep is significantly reduced if D gets large. Note, however, that soft sweeps are very likely in this parameter range anyway. For the frequency spectrum, the predictions from the homogeneous case are even more stable. We find no visible deviation from the values predicted by equation (13) even for $D = 0.2$ (figure S1 in the Supplementary Material online).

Discussion

How much genetic variation can be maintained in a population in the face of positive selection? Ever since the work of Maynard Smith and Haigh (1974), we know that positive selection removes genetic variation from a population. This has important consequences. First, the characteristic valleys of reduced variation around a selected site can be used to detect loci that underlie adaptation (e.g., Harr, Kauer, and Schlötterer 2002; Storz, Payseur, and Nachman 2004; Haddrill et al. 2005; Ometto et al. 2005). Second, if positive selection acts recurrently along the chromosome, it may be selection rather than genetic drift that controls the level of genetic variation in a population. This was formalized in the theory of genetic draft by Gillespie (1991). Positive selection and linkage may also limit the rate of the future adaptive process (Barton 1995).

The classical view is that selection erases all ancestral variation (variation that existed before the onset of selection) unless recombination during the substitution process breaks the linkage between the selected site and its genetic background. The point of this paper is that ancestral variation can also be retained if the favorable allele occurs recurrently and if several independent origins contribute to the adaptive substitution. Positive selection then results in what we call a soft selective sweep. Because every beneficial mutation is eventually recurrent, the crucial question is for which mutation rate will recurrent mutation result in soft sweeps and thus affect the standard results of genetic hitchhiking? From our results, we can answer this question as follows. If $\Theta = 2N_e u$ is the population-level mutation rate of the beneficial allele (or allelic class), then

- For $\Theta < 0.01$, soft sweeps are rare (less than 5%) even in a large sample. In this parameter range, the classical results on hitchhiking and selective sweeps hold as a good approximation.
- For $\Theta > 0.01$, soft sweeps start to play a role and will be observable for recent substitutions. In a transitional

range, $0.01 < \Theta < 1$, soft sweeps coexist with classical hard sweeps. For $\Theta > 1$, almost all adaptive substitutions will result in soft sweeps.
- Analogous results hold if beneficial alleles are introduced by recurrent migration instead of mutation. Other parameters such as selection strength, dominance, etc. play only a minor role.
- Our results show much more than the probability of a soft sweep: for a given Θ, the expected number and distribution of ancestral haplotypes in a sample follow approximately the Ewens sampling formula.

The relatively low values for Θ that are necessary to obtain soft sweeps and the independence of the selection strength may come as a surprise. After all, if selection is strong adaptation is fast and the time for recurrent mutation limited. In fact, input of neutral mutations during the selective phase can often be neglected, even if their combined mutation rate on a DNA fragment is high ($\Theta \approx 10$ typical for *Drosophila* species). So why is the same not true for beneficial mutations that are much rarer? Here, it is important to note that the neutral mutations can be ignored because they are unlikely to be seen in a sample, not because they are unlikely to happen in the population during the selective phase. Also for beneficial mutations, multiple origins during the substitution process are likely, even for quite low values of Θ. And because of their positive fitness, they have a much higher probability to survive stochastic loss and to make it into the sample.

In a forward-in-time picture, this can be estimated as follows. For a beneficial allele with selective advantage $\alpha = 2N_e s$, the average fixation time is $T_{\text{fix}} \approx 4N_e \log(\alpha)/\alpha$. The average number of mutations that occur in this time is $2\Theta N_e \log(\alpha)/\alpha$. To get an idea of this quantity, if $\Theta = 0.01$, $N_e = 2 \times 10^6$, and $\alpha = 1{,}000$, then T_{fix} is about 55,000 generations, and the mutation will occur about 276 times during the fixation process of the first mutation. For neutral mutations, the probability for a given mutation to occur in a sample of size n is about n/N_e (for a star-like phylogeny). We thus obtain a probability of $2n\Theta \log(\alpha)/\alpha$ for recurrent neutral mutations to enter the sample, which strongly decreases with α. In contrast, the probability for beneficial mutations to escape stochastic loss and to appear in the sample is proportional to the selection coefficient s (approximately $2s(1 - x)$ if x is the frequency of beneficial alleles that already segregate in the population). As a result, the dependence on s of the probability $P_{\text{soft},n}$ to observe recurrent beneficial mutations in a sample will largely cancel.

The fact that $P_{\text{soft},n}$ and, more generally, the number and distribution of ancestral haplotypes are independent of α is only one aspect of the remarkable robustness of these estimates. Under the sole assumption that the substitution was relatively fast and recent, the approximations are independent of most details of the adaptive process. They are valid whether beneficial mutations arise through mutation or migration or both, in haploids or diploids, for arbitrary patterns of time-dependent or frequency-dependent selection, any level of dominance, and even for moderate variance in the selection coefficient among the beneficial alleles. Because of this generality, there should be a realistic chance that patterns associated with soft sweeps can be found in data.

Where should we expect soft selective sweeps due to multiple origins of the beneficial allele in nature? Two factors contribute to Θ, which is the crucial parameter: soft selective sweeps should be expected if either the effective population size N_e or the allelic mutation rate u is high. For example, in African *Drosophila melanogaster* with an estimated haploid size $N_e \approx 2 \times 10^6$, Watterson's estimator for Θ per site was measured to be $\Theta_W \approx 0.013$ (Ometto et al. 2005). This translates into a $P_{\text{soft},n}$ of ~5%, if only mutation at a single site produces the beneficial allele. One should note, however, that Watterson's estimator is strongly affected by past demographic events. If the population has experienced recent strong growth, this estimator will severely underestimate the real Θ (which depends on the inbreeding effective population size at the time of the adaptation rather than on the variance effective size). For humans, in particular, it is questionable whether the often-cited low values for $\Theta_W \approx 0.001$ (or $N_e \approx 10{,}000$) are relevant for recent adaptations (e.g., to agriculture or diseases). A second scenario where soft selective sweeps from recurrent mutation are likely are adaptations with a high allelic mutation rate, such as adaptive loss-of-function mutations. Finally, situations where beneficial alleles may have been introduced into a population by recurrent migration at a low, but steady rate are easy to imagine.

In human population genetic data, quite a few alleles are known that have risen in frequency due to positive selection and are associated with different haplotypes. These could be cases of soft sweeps from independent mutational origin. Some of these alleles are indeed produced by loss-of-function mutations (e.g., the FY-0 allele at the Duffy locus, Hamblin and Rienzo [2000]; α and β thalassemia mutations, Flint et al. [1993]) but others are not (e.g., HbS, which causes sickle cell anemia, Flint et al. [1993]; HbE, which causes a mild variant of β thalassemia, Antonarakis, Orkin, and Kazazian [1982]).

Schlenke and Begun (2005) found three immunity genes in *Drosophila simulans* that show clear signs of soft sweeps. The genes have extreme linkage disequilibrium values, in each case caused by two distinct haplotypes at intermediate frequencies that have not recombined. In one case, there is also a third invariant haplotype at low frequency. Each of the haplotypes has little or no polymorphism, ruling out the possibility of long-term balanced polymorphisms. The authors also did simulations to rule out the possibility that the patterns are caused by purely demographic scenarios such as bottlenecks. However, the pattern that is found in these three genes is perfectly compatible with soft sweeps.

Pathogens can have extremely high population sizes. It may, therefore, not be surprising that evidence for soft sweeps also comes from a recent study of *Plasmodium falciparum*, with an estimated population size of 10^{10}–10^{12} per infected person (Roper et al. 2004). In this study, microsatellite variation in both pyrimethamine-resistant and sensitive parasites was studied. The haplotype structure in the data clearly suggests that the double-mutant *dhfr* allele (with longer clearance times than the sensitive parasites) in Africa has three independent mutational origins. The triple-mutant allele (making the parasite almost resistant) seems, however, to have only one origin (Roper et al. 2003). In some cases, for example in viruses, Θ values

may be so high that selective sweeps, at least for single mutants, can never be detected. All sweeps would involve alleles of many different origins, and there will be no visible signature of selection.

An obvious next step to be taken is to add recombination to the model and study how soft sweeps affect patterns of nucleotide variation at linked neutral loci. Also, more realistic demographic scenarios still remain to be investigated. Aspects that we have not addressed in this paper include changes in population size for the mutation case or more complex population structures. In general, population structure should make soft sweeps more likely. This is easy to see from the extreme case, where subpopulations are linked by very weak migration. If M is lower than Θ, it is more likely that adaptation in each population will be from its own mutational origin of the beneficial allele. On the meta-population level this would result in a soft sweep.

Supplementary Material

Supplementary figure S1 and other supplementary materials are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Peter Pfaffelhuber for fruitful discussions and Sascha Glinka, Andreas Gros, John Parsch, Marcy Uyenoyama, and the referees for helpful comments on the manuscript. This work was supported by an Emmy Noether grant by the Deutsche Forschungsgemeinschaft to J.H.

Literature Cited

Antonarakis, S. E., S. H. Orkin, and H. H. Kazazian, Jr. 1982. Evidence for multiple origins of the (E)-globin gene in Southeast Asia. Proc. Natl. Acad. Sci. USA **79**:9–117.

Barton, N. H. 1995. Linkage and the limits to natural selection. Genetics **140**:821–841.

Durett, R. 2002. Probability models for DNA sequence evolution. Springer, New York.

Durett, R., and J. Schweinsberg. 2004. Approximating selective sweeps. Theor. Popul. Biol. **66**:129–138.

Etheridge, A., P. Pfaffelhuber, and A. Wakolbinger. 2005. An approximate sampling formula under genetic hitchhiking. Ann. Appl. Probab. (in press).

Ewens, W. J. 2004. Mathematical population genetics. 2nd edition. Springer, Berlin.

Flint, J., R. M. Harding, J. B. Clegg, and A. J. Boyce. 1993. Why are some genetic diseases common? Distinguishing selection from other processes by molecular analysis of globin gene variants. Hum. Genet. **91**:91–117.

Gillespie, J. 1991. The causes of molecular evolution. Oxford University Press, New York.

Haddrill, P. R., K. R. Thornton, B. Charlesworth, and P. Andolfatto. 2005. Multilocus patterns of nucleotide variability and selection history of *Drosophila melanogaster* populations. Genome Res. **15**:790–799.

Hamblin, M. T., and A. D. Rienzo. 2000. Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. Am. J. Hum. Genet. **66**:1669–1679.

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

Hermisson, J., and P. S. Pennings. 2005. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics **169**:2335–2352.

Innan, H., and Y. Kim. 2004. Pattern of polymorphism after strong artificial selection in a domestication event. Proc. Natl. Acad. Sci. USA **101**:10667–10672.

Kaplan, N. L., R. R. Hudson, and C. H. Langley. 1989. The "Hitchhiking Effect" Revisited. Genetics **123**:887–899.

Kimura, M. 1955. Solution of a process of random genetic drift with a continuous model. Proc. Natl. Acad. Sci. USA **41**:144–150.

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

Ometto, L., S. Glinka, D. D. Lorenzo, and W. Stephan. 2005. Inferring the effects of demography and selection on Drosophila melanogaster populations from a chromosome-wide scan of DNA variation. Mol. Biol. Evol. **22**:2119–2130.

Przeworski, M., G. Coop, and J. D. Wall. 2005. The signature of positive selection on standing genetic variation. Evolution **59**:2312–2323.

Roper, C., R. Pearce, B. Bredenkamp, J. Gumede, C. Drakeley, F. Mosha, D. Chandramohan, and B. Sharp. 2003. Antifolate antimalarial resistance in southeast Africa: a population-based analysis. Lancet **361**:1174–1181.

Roper, C., R. Pearce, S. Nair, B. Sharp, F. Nosten, and T. Anderson. 2004. Intercontinental spread of pyrimethamine-resistant malaria. Science **305**:1124.

Schlenke, T. A., and D. J. Begun. 2005. Linkage disequilibrium and recent selection at three immunity receptor loci in *Drosophila simulans*. Genetics **169**:2013–2022.

Stephan, W., T. 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.

Storz, J. F., B. A. Payseur, and M. W. Nachman. 2004. Genome scans of DNA variability in humans reveal evidence for selective sweeps outside Africa. Mol. Biol. Evol. **21**:1800–1811.

Marcy Uyenoyama, Associate Editor

Accepted March 1, 2006