Estimating the Time Since the Fixation of a Beneficial Allele

Molly Przeworski[1]

Max Planck Institute for Evolutionary Anthropology, 04103 Leipzig, Germany

Manuscript received January 15, 2003
Accepted for publication April 21, 2003

ABSTRACT

The fixation of a beneficial allele in a population leaves a well-characterized signature in patterns of nucleotide variation at linked sites. This signature can be used to estimate the time since fixation from patterns of polymorphism in extant individuals. I introduce a method to assess the support in polymorphism data for a recent episode of directional positive selection and to estimate the time since fixation. I summarize the polymorphism data by three statistics that carry information about levels of diversity, the allele frequency spectrum, and the extent of allelic associations. Simulations are then used to obtain a sample from the posterior distribution of the time since fixation, conditional on the observed summaries. I test the performance of the approach on simulated data and apply it to the gene *tb1* in maize. The data support the recent fixation of a favored allele, consistent with what is known about the importance of *tb1* in the domestication process of maize.

PATTERNS of nucleotide variation are shaped by the evolutionary history of the genomic region and, in particular, by adaptation. Polymorphism data are therefore informative about the nature and timing of positive selection, including what proportion of loci show evidence of selective changes or when the selective pressure was exercised. In many contexts, the time since a beneficial substitution is of interest. For example, one might want to date adaptations to temperate habitats in species of Drosophila whose range was originally restricted to the tropics (David and Capy 1988). In domesticated species, one might be interested in the number of genes (or genes in particular pathways) for which the timing of beneficial substitutions coincides with the advent of agriculture (*e.g.*, Whitt *et al.* 2002).

Similarly, one might hope to date the genetic changes that led to the emergence of anatomically modern humans, in particular those underlying speech and language (*e.g.*, Enard *et al.* 2002). A common theory of human evolution contends that anatomically modern humans were the first to possess current cognitive and linguistic abilities (Klein 1995; Mellars 1998). In this view, the emergence of language is manifested by a burst of evidence for "symbolic thought" in the fossil record, centered mainly on art and artifacts found 30–50 thousand years ago (KYA; Mellars 1998). Although the extent to which these novel cognitive abilities have a genetic basis is unclear, it seems likely that some genetic changes were involved in the acquisition of these traits. Since there is evidence for the colonization of Australia by modern humans as far back as 40–60 KYA and modern humans are thought to have emerged in the past 150 KYA (Stringer 2002), it follows that these adaptations would have had to occur ~40–150 KYA.

If many favored substitutions occurred as "selective sweeps," in which a rare allele arises and rapidly increases in frequency until fixation in the population, their timing can be estimated from patterns of polymorphism in extant individuals. Indeed, under simplifying assumptions, a favorable substitution leaves a well-characterized signature on linked neutral variability (Maynard Smith and Haigh 1974; Kaplan *et al.* 1989; Braverman *et al.* 1995; Kim and Stephan 2002), detectable in humans for >8000 generations or ~200,000 years (Przeworski 2002). This signature can be used to identify regions that have experienced recent directional selection, as well as to estimate parameters of interest. Luckily, this time frame covers the period of relevance for the emergence of modern human-specific traits. Similarly, theoretical investigations suggest that the effects of adaptation to temperate habitats should still be visible in the genomes of model Drosophila species (Kim and Stephan 2002; Przeworski 2002), given what is known about their history (*e.g.*, Lachaise *et al.* 1988).

The most common approach to identifying adaptive genetic changes from polymorphism data is to assume a neutral null model and then assess whether the value of some summary of the data is unexpected under this model. A poor fit is interpreted as evidence for the action of natural selection. Among these "tests of neutrality" are allele frequency spectrum-based tests such as D (Tajima 1989) and H (Fay and Wu 2000) as well as various tests based on the extent of allelic associations (*e.g.*, Hudson *et al.* 1994; Sabeti *et al.* 2002). However, even when such tests indicate that the neutral null model is a poor fit to the data, a selective sweep may be no more likely an explanation. Furthermore, these

[1]*Address for correspondence:* Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany. E-mail: przewors@eva.mpg.de

tests do not allow one to estimate parameters of interest, such as the timing of the selective sweep.

A number of alternative approaches to estimating parameters of a selective sweep model exist. In particular, Kim and Stephan (2002) introduced a "composite-likelihood" estimator that allows one to estimate the strength of selection and the location of the favored substitution, assuming that other parameter values are known. In their framework, time is measured backward, so that the time since the fixation of the beneficial allele, T, is 0 when the selective sweep has just ended. Under simple demographic assumptions, the likelihood of an allelic configuration at a given site is given explicitly when $T = 0$ (Fay and Wu 2000). By multiplying the likelihoods at different sites, thereby ignoring the dependence between sites, the authors obtain a so-called composite likelihood (Hudson 2001). The composite likelihood of $T = 0$ is compared to the one for no selection to assess the support for a very recent selective sweep. Approximate confidence intervals can be obtained by simulation. However, the time T cannot be estimated, since it is assumed to be 0 in the selective sweep model.

To estimate T for the case of no recombination, Perlitz and Stephan (1997) proposed a method of moments estimator based on observed diversity levels at linked neutral sites. More recently, Jensen *et al.* (2002) used an acceptance-rejection algorithm to find the joint maximum-likelihood estimates of T and the population mutation rate from two summaries of diversity levels. Enard *et al.* (2002) applied a similar rejection-sampling method to estimate T from data for a recombining locus, FOXP2, a gene involved in speech and language in humans. Their estimate of T was also based on two diversity statistics; it assumed that all nuisance parameters (besides the recombination rate) were measured without error.

I introduce a method to assess the support for a recent beneficial substitution in polymorphism data from linked neutral sites and to estimate T. The approach applies to the situation where researchers are interested in a fixed difference between species or populations that they consider a candidate for a recent selective sweep. The idea is to draw a sample from the posterior distribution of T, conditional on summaries of the polymorphism data. If a recent adaptation occurred at a closely linked site, most of the posterior probability should be on recent T values. Conversely, if most of the posterior probability is on more distant times, this suggests that the selective sweep did not occur recently, or did not occur at all, in the genealogical history of the sample. This method improves on existing ones by using more summaries of the data, including a measure of linkage disequilibrium (LD) as well as a summary of the frequency spectrum and of levels of diversity. An attractive feature of the approach is that the parameters not of interest are integrated out, so one need not assume that their values are known exactly (*cf.* Beaumont *et al.* 2002).

METHODS

The model: I consider the following model of a selective sweep: a neutrally evolving region is linked to a site where a favorable allele reached fixation in the population at some time, T, measured into the past. The neutral locus is assumed to evolve according to the infinite-sites model. The population mutation rate for the neutral locus is $\theta = 4N\mu L$, where N is the diploid effective population size of the species, μ is the mutation rate per base pair per generation, and L is the length of the locus in base pairs. Similarly, the population recombination rate for the neutral locus is $\rho = 4NrL$, where r is the rate of recombination per base pair per generation. Recombination occurs at a constant rate per base pair throughout the region and all recombination events are crossovers without gene conversion. The population recombination rate between the neutral and selected loci is $C = 4NrK$, where K is the physical distance (in base pairs) between the closest edge of the neutral locus and the selected site. The population is random mating and of constant size.

Time is scaled in units of $4N$ generations, so $T = 1$ is equivalent to fixation of the favored allele $4N$ generations ago. Selection is additive. The increase in frequency of the favored allele is modeled deterministically, from introduction to fixation [*i.e.*, frequency $1 - 1/(2N)$]. The sojourn time of the favored allele in the population is then $\sim 2\ln(2N)/s$, where s is the selection coefficient of the favored allele (Stephan *et al.* 1992). In reality, the increase in frequency is governed by genetic drift as well as selection; however, modeling the trajectory as stochastic rather than deterministic makes little difference so long as $4Ns$ is large (results not shown).

Simulations of selective sweep: The model is implemented in a coalescent framework. There are two phases: a neutral phase, in which the coalescent is the standard coalescent with recombination (*cf.* Hudson 1990), and a selected phase. During the latter, there are two allelic classes at the neutral locus: lineages carrying the favored allele and those carrying the unfavored one. These allelic classes can be modeled analogously to subpopulations, with recombination acting as migration; the sizes of the subpopulations change over time with the frequency of the favored allele (Barton 1998). The details of the implementations are described elsewhere (Przeworski 2002).

Summaries of the polymorphism data: At present, it is not computationally feasible to use all the polymorphism data to estimate parameters, even for models simpler than the one considered here (Fearnhead and Donnelly 2002). I therefore summarized the data by three statistics: the number of segregating sites, S; a

summary of the allele frequency spectrum, Tajima's D (Tajima 1989); and the number of distinct haplotypes in the sample, H, which is a measure of linkage disequilibrium (Strobeck 1987; Wall 2000). The choice of summaries is discussed in results.

Simulating a sample from the posterior distribution: In addition to the time since the fixation of the beneficial allele, T, there are four parameters in this selective sweep model whose values cannot be measured exactly: N, s, μ, and r. (While in many population genetic contexts, the parameter N does not appear on its own, here, it specifies the frequency at which the beneficial allele first appears [*i.e.*, $1/(2N)$].) I assume that researchers are interested in a particular site where a substitution may have been selected, so K is known. To model the uncertainty in the other parameters, I use a distribution of prior values, rather than assuming a fixed value (see below).

To obtain a sample from the posterior distributions of the parameters conditional on the summaries of the data, I follow the rejection-algorithm 2 outlined in Tavare *et al.* (1997). The idea of the procedure is to accept parameter values chosen from prior distributions with probability proportional to their likelihood. Specifically, let S, D, and H be the observed summaries in polymorphism data from a number of chromosomes sequenced at a locus of length L bp. For each independent replicate i:

1. Pick N_i, s_i, μ_i, r_i, and T_i values independently from the prior distributions given below.
2. To model the history of the sample at the neutral locus, simulate a recombination graph (including coalescent events and recombination events, but no mutation) under a selective sweep model with parameter values K, N_i, s_i, μ_i, r_i, and T_i.
3. Because of recombination, different segments of the neutral locus will have distinct genealogies. Suppose that there are n segments, where each segment is a set of consecutive base pairs with the same genealogy; let L_j be the length of segment j and τ_j be the length of the genealogy for segment j. Accept the graph simulated in step 2 with probability $u = \mathrm{Po}(S, 4N\mu\tau)/\mathrm{Po}(S, S)$ where Po denotes the probabilities of the Poisson distribution and $\tau = \sum_{j=1}^{n} (\tau_j L_j/L)$. If the graph is rejected, return to step 1.
4. Place S segregating sites on the graph.
5. Tabulate two summaries of the simulated data: D_i and H_i.
6. If $|D_i - D| \leq \varepsilon$ and $H_i = H$, record values N_i, s_i, μ_i, r_i, and T_i.

Run this algorithm until there are M_ε sets of recorded values. The list of sets of recorded values is a sample of size M_ε from the joint posterior distribution of the parameters, conditional on $S_i = S$, $|D_i - D| \leq \varepsilon$, and $H_i = H$. A sample from the posterior distribution of one or a subset of the parameters can be obtained by considering only the values of the parameters of interest. A sample from the posterior distribution of $T_{\mathrm{gen}} = 4NT$, the time in generations since the fixation of the beneficial allele, is given by the appropriate product of T and N in this list.

In step 1, the recombination rate r is chosen from a gamma distribution with parameters $(z \times 10, 10^9)$; the mean is then $z \times 10^{-8}$. Assuming little error in the physical map and homogeneity of local recombination rates, the sampling error associated with large-scale estimates of genetic distance provides some sense of the accuracy of r estimates. I assume that the estimate obtained from a comparison of genetic and physical maps would be used as the mean of the gamma distribution and choose the parameters to be in rough accordance with the estimates of sampling error provided for the latest genetic map in humans (Kong *et al.* 2002; Weber 2002). The mutation rate μ is also drawn from a gamma distribution, with parameters $(10 \times m, 10^9)$. The effective population size N is chosen from a gamma distribution with parameters $(5, 5/Y)$. As an illustration, for $Y = 10^4$, 95% of this distribution covers the approximate interval (3200, 20,500) of N values. When estimates of the parameters exist, they can inform the choice of m, z, and Y. The selection coefficient s is drawn from a uniform on $(50/N, 0.05)$ (if $N < 1000$, s is set to 0.05). When $s < 50/N$, the deterministic approximation becomes inaccurate; furthermore, the rise in allele frequency of the favored allele (even when modeled as a stochastic process) is not rapid enough to cause a severe sweep at nearby neutral sites (results not shown).

The aim is to assess the support for a beneficial substitution that happened recently in the genealogical history of the sample. The time depth of that history depends on N: as an illustration, a substitution that occurred 40,000 generations ago is recent for *Drosophila melanogaster*, where N is on the order of 10^6, but not for humans, where N is estimated to be $\sim 10^4$. I therefore place the prior on the scaled time since the fixation of the beneficial allele, T, rather than on the true time in generations, $4NT$. In the absence of a selective sweep at a nearby site, the scaled time to the most recent common ancestor of a sample is on average ~ 1 (for a sample size greater than two; *cf.* Hudson 1990). Thus, if $T > 1$, a selected substitution has little detectable effect on the mean number of haplotypes or on mean summaries of the allele frequency spectrum at linked neutral sites (Simonsen *et al.* 1995; Przeworski 2002), while $T \gg 1$ is equivalent to no sweep.

I use two different priors on T to test the robustness of the approach to the choice of distribution. In one implementation, T is exponential with parameter 1.2 [referred to as Exp(1.2)]; the $\Pr(T < 1)$ is then ~ 0.70. This distribution arises if beneficial mutations occur infrequently and independently of one another (Kaplan *et al.* 1989). I also use a uniform prior on (0, 1) [referred to as $U(0, 1)$]. The posterior distribution of

T is then proportional to the likelihood of T given the data, for values of $T < 1$. In RESULTS, I arbitrarily consider $T \leq 0.2$ to be a recent selective sweep. The $\Pr(T \leq 0.2)$ is very similar for both distributions: if T is $U(0, 1)$, it is obviously 0.2, while if T is $\text{Exp}(1.2)$, it is $\sim$0.21.

Test of performance: I generate 100 simulated data sets under the selective sweep model with fixed values of the parameters $N_x = 10^4$, $s_x = 0.01$, $\mu_x = 10^{-8}$/bp/generation, and $r_x = 1$ cM/Mb. These parameters are chosen to be plausible descriptions of a region of average recombination in humans. The neutral locus is 10^4 bp in length and the sample size is 50 chromosomes. The physical distance between selected and neutral loci, K, is 10^3 bp. I consider three cases: $T = 0$, $T = 0.10$, and $T = 100$ (which is equivalent to no selective sweep in the genealogical history of the sample). For each simulated data set, I obtain a sample from the posterior distribution of T, according to the method outlined above, with $\varepsilon = 0.1$ and $M_\varepsilon = 2 \times 10^3$. Parameters m, z, and Y are chosen such that the mean $N = N_x$, the mean $r = r_x$, and the mean $\mu = \mu_x$. I assume that the physical distance to the selected site is known.

Application to *tb1*: As an "empirical test," I apply the method to the gene *tb1* in maize, a locus known to have experienced selective pressure during the domestication process that occurred over the past 5–10 KYA (*cf.* WANG *et al.* 1999). The initial screen of the *tb1* gene in maize and its wild progenitor teosinte failed to identify a substitution differentiating the two species (WANG *et al.* 1999). The authors estimated that the selected site lies within $\sim$1 kb upstream of the 5′ nontranscribed region that they sequenced; I use a value of 500 bp in my simulation. Parameterization of the prior distributions is as follows: m, z, and Y are chosen such that the mean $N = 5 \times 10^5$ (EYRE-WALKER *et al.* 1998; TENAILLON *et al.* 2001), the mean $\mu = 10^{-8}$, and the mean $r = 1.35 \times 10^{-8}$ (this estimate was kindly provided by L. Zhang and B. Gaut). I use $\varepsilon = 0.1$ and $M_\varepsilon = 4 \times 10^4$. For the results reported in Figure 3, the prior distribution of T is $\text{U}(0, 1)$; results are very similar if instead the prior on T is $\text{Exp}(1.2)$ (results not shown).

Code: The C program used to simulate a sample from the posterior distribution of the parameters is available from http://email.eva.mpg.de/~przewors. The sample size from the posterior distribution, M_ε, is specified by the user, as is the tolerance ε. For a data set reported in Table 1, for which $M_\varepsilon = 2000$, it took anywhere from 1 min to 3 days on a 1667-Mhz AMD Athlon single processor. A second version of the program is available upon request. In this implementation, step 3 of the algorithm is eliminated. Instead, parameter values are recorded whenever $|D_i - D| \leq \varepsilon$ and $H_i = H$ and each set of recorded values is weighted by the probability u. The algorithm is repeated until an estimate of the effective sample size (based on the mean and sample variance of the importance weights, u) exceeds a value specified by the user. This "importance sampling" procedure is known to be more efficient than the acceptance-rejection algorithm described above (*cf.* Appendix of DONNELLY *et al.* 2001).

RESULTS

To assess the support for a recent selective sweep, I follow the approach developed by PRITCHARD *et al.* (1999) to estimate the time since the onset of growth in humans. Specifically, I summarize the polymorphism data and obtain a sample of the posterior distribution of the parameters conditional on the summaries being close to (*i.e.*, within a prespecified neighborhood) or equal to the observed value. Similar rejection-sampling methods have been used in other contexts, including estimating the effective population size (BACHTROG and CHARLESWORTH 2002), population parameters (TAVARE *et al.* 1997; WALL 2000; FEARNHEAD and DONNELLY 2002), and the age of an allele (TISHKOFF *et al.* 2001), as well as demographic inference (WEISS and VON HAESELER 1998; BEAUMONT *et al.* 2002). For a discussion of the differences between implementations, see BEAUMONT *et al.* (2002).

Choice of summaries: Short of being able to use all the information in the data, one would like to use summaries that are sensitive to the parameter of interest (here, the time since the fixation of the beneficial allele, T) and capture different facets of the data. I focus on three statistics: the number of segregating sites (S), a summary of the allele frequency spectrum (D), and a summary of linkage disequilibrium (H). Previous studies have shown two of the statistics, S and D, to be sensitive to T. Specifically, selective sweeps are expected to reduce diversity, thus reducing S, and skew the frequency spectrum toward rare alleles, leading to negative D values (MAYNARD SMITH and HAIGH 1974; BRAVERMAN *et al.* 1995; SIMONSEN *et al.* 1995). I chose D among various frequency-spectrum summaries because when it is used as a test statistic it, more than other statistics, retains power to reject a neutral model when data are generated under a selective sweep model for larger T values (KIM and STEPHAN 2002; PRZEWORSKI 2002). D is the (approximately) normalized difference between a measure of diversity based on S and π, the mean pairwise difference in the sample (TAJIMA 1989). Thus, specifying S and D determines π as well.

The number of haplotypes, H, also carries information about T. Its behavior depends on the strength of selection and the recombination rate. If recombination occurs between selected and neutral loci during the selective sweep, then at $T = 0$, $H/(S + 1)$ will be lower on average than it would be in the absence of selection (PRZEWORSKI 2002). In other words, allelic associations will tend to be stronger than they would be in the absence of selection. As T increases, new alleles will arise by mutation. These rare alleles will create new haplotypes, such that $H/(S + 1)$ will rapidly exceed the neutral

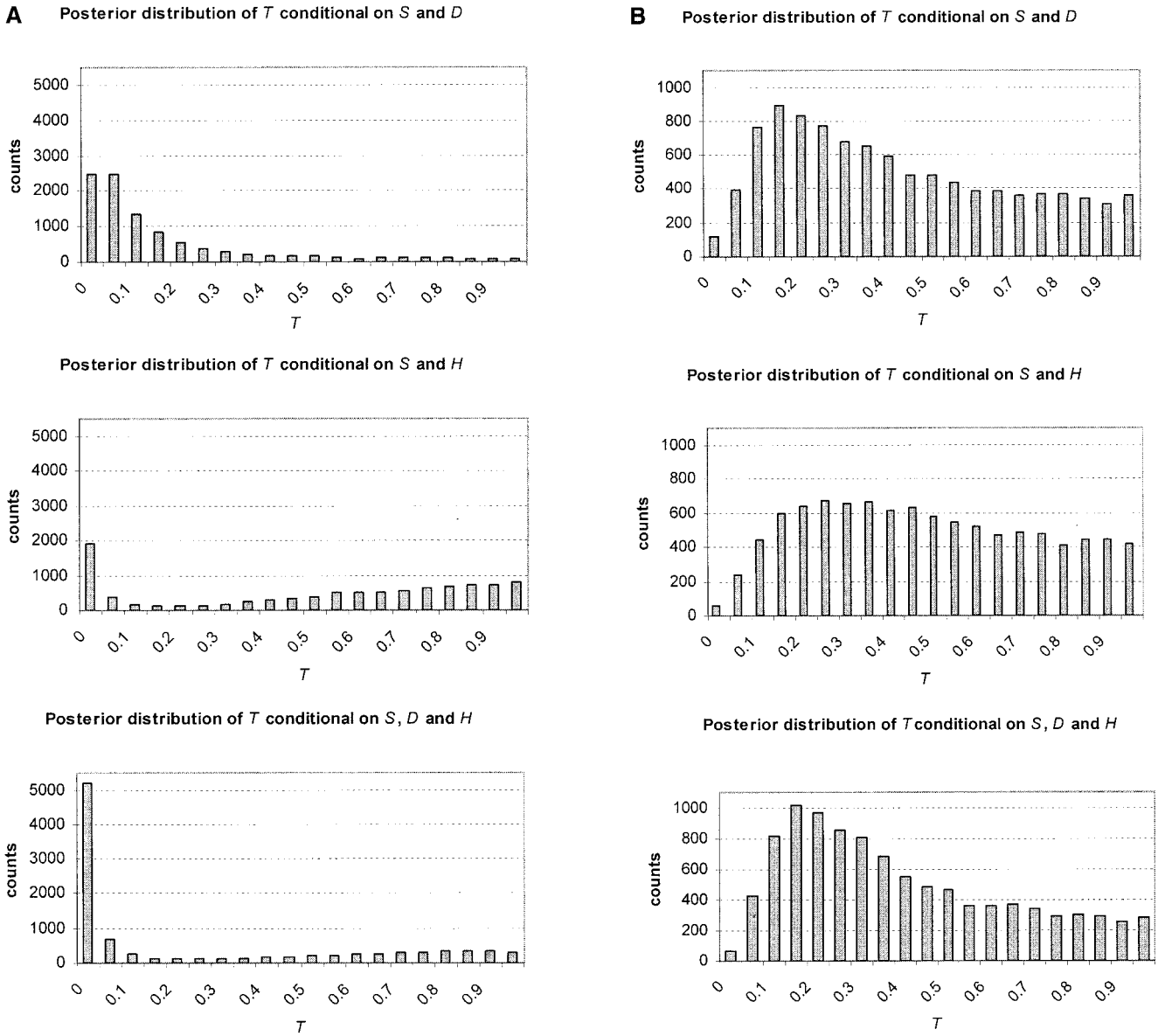

FIGURE 1.—(A) A sample from the posterior distribution of T for a simulated data set, when the true time $T_0 = 0$. Other parameters used to generate the data set are the same as in Table 1 (with a uniform prior on T) with ε set to 0.1 and $M_e = 10^4$. In this example, $S = 7$, $D = -1.78$, and $H = 4$. (B) A sample from the posterior distribution of T for a simulated data set, when the true time $T_0 = 0.2$. Other parameters are as in A. In this example, $S = 8$, $D = -0.91$, and $H = 10$.

expectation. The ratio will subsequently decrease (at $T \gg 0.1$), as the alleles gradually increase in frequency and recombine onto other backgrounds (PRZEWORSKI 2002). If there is no recombination between selected and neutral loci during the selective sweep, most of the alleles will be rare, and $H/(S + 1)$ will be larger than expected under neutrality, with its largest value attained for $T > 0.1$ (PRZEWORSKI 2002).

In summary, while to a rough approximation, S and D are expected to increase monotonically with T, $H/(S + 1)$ tends to have a maximum value at some intermediate T. This suggests that using H as an additional statistic may help to distinguish between recent T values and therefore to refine the estimates of T obtained using D and S alone. I illustrate this in Figure 1 by plotting the posterior distribution of T for two simulated data sets, conditional on S and D (first row); S and H (second row); and S, D, and H (last row). As can be seen, conditioning on all three summaries leads to a tighter distribution around the true value than does the use of only two; this finding is confirmed by more extensive simulations (results not shown).

The extent to which the three summaries are informative about T depends on prior knowledge about the parameters. In particular, detecting a reduction in diversity requires some knowledge of what levels of diversity are expected to be in the absence of selection. Thus, if one has accurate prior knowledge about the population

TABLE 1

Performance of the method on simulated data

	Proportion of 100 runs where $\Pr(T \leq 0.2)$ is greater in posterior than in prior	Proportion of 100 runs where posterior $\Pr(T \leq 0.2) \geq 0.50$
Prior $T \sim U(0, 1)$		
$T_o = 0$	0.96	0.83
$T_o = 0.1$	0.72	0.36
$T_o = 100$	0.06	0.04
Prior $T \sim \text{Exp}(1.2)$		
$T_o = 0$	0.91	0.85
$T_o = 0.1$	0.79	0.41
$T_o = 100$	0.04	0.00

T_o is the true time since the fixation of the beneficial allele, in units of $4N$ generations. $T_o = 100$ is equivalent to no selective sweep in the genealogical history of the sample. Data are generated by coalescent simulations (see METHODS), with the following parameters: a sample size of 50 chromosomes; a neutral locus of length 10^4 bp; a mutation rate, μ, of 10^{-8}/bp/generation; a recombination rate, r, of 1 cM/Mb; a selection coefficient, s, of 0.01; and a diploid effective population size, N, of 10^4. The physical distance between selected and neutral loci, K, is 10^3 bp. To obtain a sample from the posterior distribution, the tolerance, ε, is set to 0.1 and $M_\varepsilon = 2 \times 10^3$ (see METHODS).

mutation rate θ ($= 4N\mu$), the decrease in the number of segregating sites and in the number of haplotypes can be highly informative about the time since the selective sweep (SIMONSEN *et al.* 1995). When less is known about θ, most of the information about the time since the selective sweep will come from the observed value of D and the value of H given S.

An additional benefit of using distinct aspects of the data is that the approach may be less sensitive to misspecification of the prior distributions. For example, methods that estimate T on the basis of diversity levels alone are highly sensitive to the estimate of θ. If θ is estimated to be higher than it is in reality but no selection has occurred, levels of variation will appear reduced. On this basis, methods may spuriously suggest the recent fixation of a beneficial allele. However, if the data are generated under a neutral model with an elevated mutation rate, the values of D and H will tend to be less likely under a recent selective sweep than under neutrality. Thus, the use of all three summaries may result in less support for a recent T. To examine this, I ran 20 simulations with no selection in which the mean of the prior distribution of θ was twofold larger than the value used to generate the data (parameters as in Table 1 for a uniform prior on T). For none of the simulated data sets was there strong support for a recent selective sweep (results not shown).

Performance of approach: One concern is that the posterior probabilities may not be well estimated. In that respect, an advantage of this method over more efficient ones such as Monte Carlo Markov chain is that it provides independent samples from the posterior distribution, so one can easily assess the accuracy of estimates of the posterior probabilities. In particular, if the sample from the posterior is of size M_ε, the sampling error associated with B_j, the observed number of counts in interval j, is binomial (CARLIN and LOUIS 1998) and can be estimated using parameters $(B_j/M_\varepsilon, M_\varepsilon)$. This indicates that probabilities on the order of $1/M_\varepsilon$ are poorly estimated, while those $\gg 1/M_\varepsilon$ are fairly precisely estimated. In the simulations presented here, $M_\varepsilon = 2000$ and probabilities of interest are $\gg 5 \times 10^{-4}$.

A second question is whether data generated under a selection sweep model with realistic parameters carry much information about the parameter T. Let T_o be the true time since the fixation of the beneficial allele. If the data are informative, the support for recent times should be stronger in the posterior distribution than in the prior if T_o is 0 or 0.10, while there should be weaker support for recent times in the absence of selection. As can be seen in Table 1, this is true of almost all simulated runs, whether the prior distribution of T is Exp(1.2) or U(0, 1). To measure the proportion of data sets with strong support for a "recent" selective sweep, I tabulate the proportion of 100 simulated data sets where the posterior $\Pr(T \leq 0.2) > 0.50$. For $T_o = 0$, the proportion is very high. It decreases with T_o, but even when the beneficial substitution occurred some time ago ($T_o = 0.10$), over one-third of simulated data sets strongly support a selective sweep (Table 1). In contrast, one rarely (in <5% of the runs) finds strong support for a recent selective sweep when none has occurred. In summary, the simulated data sets appear to be informative about recent genetic adaptations. In humans, the parameters chosen for the simulations correspond loosely to 25 individuals sequenced for 10 kb in a region of average recombination, comparable to what is currently collected for studies of putatively selected loci (*e.g.*, ENARD *et al.* 2002; HAMBLIN *et al.* 2002).

Note further that these tests of performance were carried out for two quite different prior distributions for T. The results are very similar in both implementations, suggesting that the method is robust to the choice of a prior distribution. This is reassuring, as little or nothing is usually known about T—in contrast to other parameters, where there is often independent knowledge to guide the specification of the prior.

In some contexts, one is interested in an estimate of the unscaled time (in generations) since the fixation of the beneficial allele, $T_{gen} = 4NT$. As a point estimate, one might consider the mode of the sample from the posterior distribution of T_{gen}. In Figure 2, I plot the distribution of modes (*i.e.*, the bin with the largest number of counts) for 100 simulated data sets. When the data are generated under a no-selection model, few

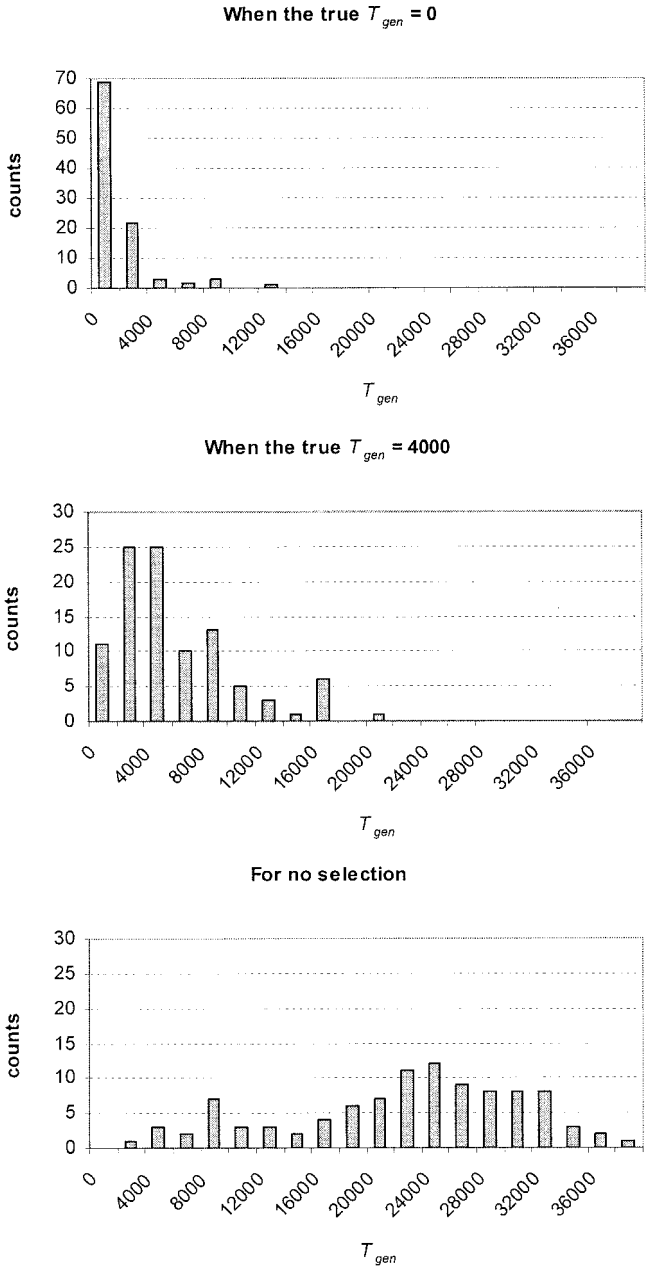

FIGURE 2.—The modes of posterior distribution samples of $T_{gen} = 4NT$ for 100 simulated data sets. For each data set, the values of T_{gen} are binned in increments of 2000 from 0 to 80,000; the mode refers to the bin with the highest number of counts. The simulated data sets are the same as summarized in Table 1 for a uniform prior on T; results are similar if the prior is instead Exp(1.2) (results not shown).

modes are at recent times (*e.g.*, 6 are at $T_{gen} \leq 8000$ generations). In contrast, when the data are generated under a recent selective sweep model, most modes are close to the true time. For example, if the true T_{gen} is 0, 94 of the modes are at $T_{gen} \leq 4000$ generations. As the true T_{gen} increases, the precision of the estimate decreases: thus, if the true T_{gen} is 4000 generations, only 60 of the modes are within a factor of two of the true value.

Application to *tb1*: The *tb1* locus is responsible for the short branches that distinguish maize from its wild progenitor, teosinte. This trait is thought to have fixed during the domestication process, 5–10 KYA (*cf.* WANG *et al.* 1999). I use polymorphism data collected for 2740 bp of the maize *tb1* by TENAILLON *et al.* (2001; available from http://bgbox.bio.uci.edu/data/maud1asd.html). I focus on the 14 landraces among the 23 lines that were sequenced; results are similar if the 9 additional inbred lines are included (results not shown). For these data, $S = 39$, $H = 14$, and $D = -2.25$ [statistics were calculated using DNAsp (ROZAS and ROZAS 1999)]. A sample from the posterior distribution of T is presented in Figure 3A. Over 99.99% of the support is on $T \leq 0.2$. Thus, consistent with what is known about the role of *tb1* in the domestication of maize, polymorphism data strongly suggest the recent fixation of a beneficial allele.

I also present a sample from the joint posterior distribution of s, the selection coefficient of the favored allele, and T_{gen}, the time in generations since the fixation of the beneficial allele (Figure 3B). The results suggest a large selection coefficient, in accordance with evidence that the trait was under artificial selection. However, most of the support is on older than expected times from the archeological record (assuming approximately one generation per year for maize). This discrepancy may be due to chance, since few estimates of T_{gen} will be on the true value even under ideal conditions (see Figure 2). Alternatively, it may reflect an incorrect assumption about the location of the selected site or a salient aspect of the history of maize not captured by the demographic or selective model (see below).

DISCUSSION

Advantages: This rejection-sampling method is computationally feasible yet uses enough summaries of the data to capture information about the time since the fixation of the beneficial allele. Indeed, the limited simulations in Table 1 and Figure 2 suggest that the summaries can provide strong support for a selective sweep when it occurred recently, as well as fairly accurate point estimates of the time since fixation. Further, the precision of the posterior density estimates depends on the sample size from the posterior distribution. This is under the investigator's control, at least to some extent, as the number of replicates can easily be increased.

As can be seen in Table 1, the proportion of data sets with strong support for a selective sweep decreases with the time since fixation. This is consistent with the finding that summary-statistic-based "tests of neutrality" and the composite-likelihood method of KIM and STEPHAN (2002) lose power to reject the null model in favor of a selective sweep as time increases (SIMONSEN *et al.* 1995; KIM and STEPHAN 2002; PRZEWORSKI 2002). As new mutations arise, and recombination breaks down allelic associations, the signature of a selective sweep dissipates.

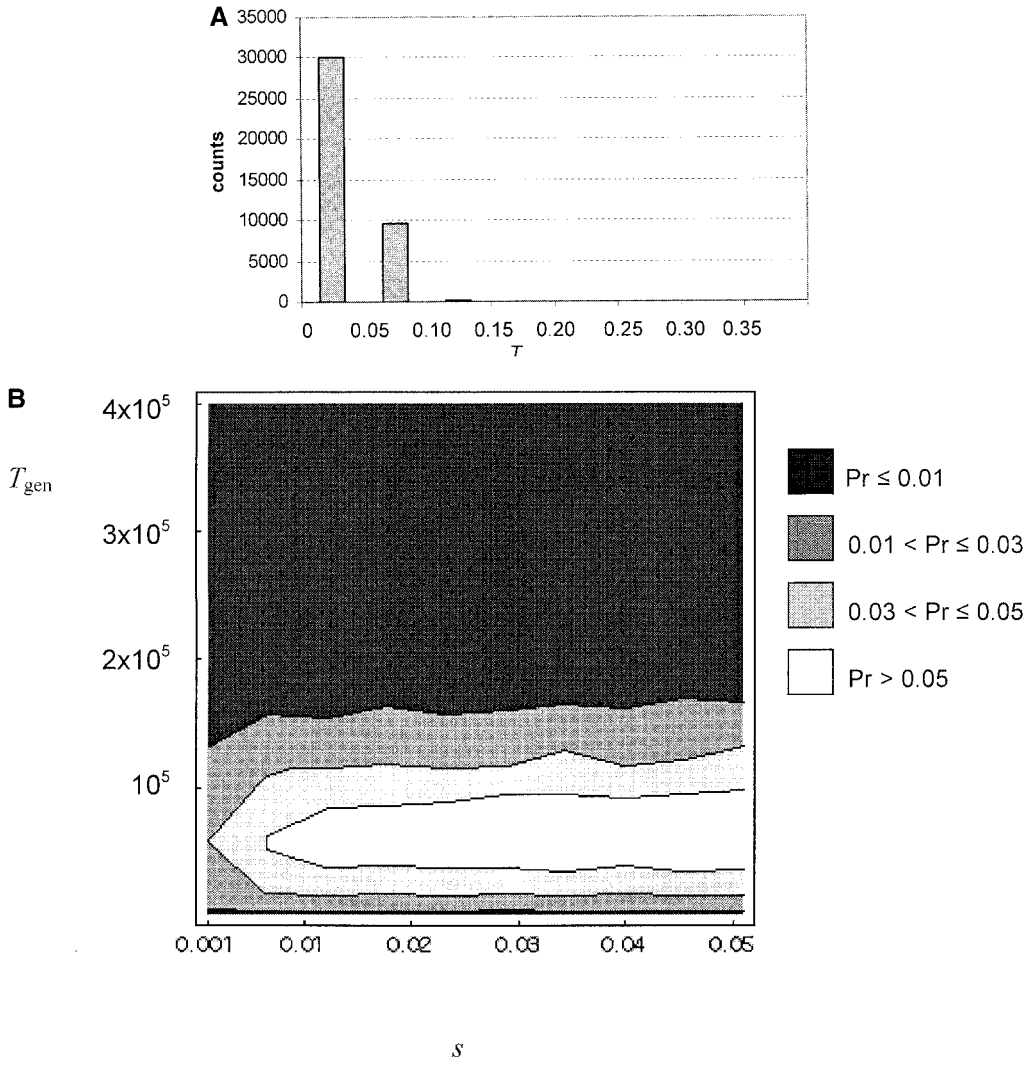

FIGURE 3.—(A) Sample from the posterior distribution of T for polymorphism data from the *tb1* gene in maize. For details of the implementation, see METHODS. (B) Sample from the joint posterior distribution of s, the selection coefficient of the favored allele, and T_{gen} (= $4NT$), the time in generations since the fixation of the beneficial allele.

As a result, this method and others cannot distinguish a selective sweep that occurred a long time ago from no selective sweep in the history of the sample.

This said, the power to detect a *recent* selective sweep varies between methods, depending on the approach and the choice of summaries of the data. In that light, it is worth noting that existing estimates of the power of tests of neutrality or the composite-likelihood approach tend to assume that the values of nuisance parameters are known (*e.g.*, KIM and STEPHAN 2002; PRZEWORSKI 2002). In practice, however, estimates are inaccurate or unavailable. In particular, the population recombination rate can rarely be estimated with any precision. Assuming that the true values of the parameters are known will overestimate the true power (WALL 1999; KIM and STEPHAN 2002; PRZEWORSKI 2002). In contrast, this type of approach models uncertainty in the large number of nuisance parameters and integrates it out in the simulation process (see the discussion of this point in BEAUMONT *et al.* 2002).

Like the composite-likelihood approach, this method allows one to estimate parameters as well as to assess support for a recent selective sweep model. An advantage of this framework is that one can also quantify the uncertainty associated with these estimates. While for composite likelihood it is not obvious how to obtain confidence intervals for the point estimates (*e.g.*, FRISSE *et al.* 2001), in this framework, probability intervals are directly available.

Possible improvements: Although this approach can provide strong support for a recent selective sweep, the limited number of summaries entails a considerable loss of information. It is therefore important to choose summary statistics judiciously. Simulations suggest that the statistics used here capture distinct aspects of the data (Figure 1; results not shown). However, it would be useful to have a more rigorous way to compare sets of statistics than visual inspection. It may also be possible to substantially increase the number of statistics by using more sophisticated implementations of rejection-sampling methods (BEAUMONT *et al.* 2002).

This approach makes a number of assumptions that can fairly easily be relaxed, for example, that sequences are contiguous or that haplotypes are known. The latter

is likely to be true for Drosophila or maize, but not for human autosomes. One possibility would be to consider a summary of linkage disequilibrium that can be calculated on genotypic data, such as an estimate of the population recombination rate (Hudson 2001), instead of the number of haplotypes. Another assumption is that local recombination rates are constant per base pair; this is unlikely to be true for humans or plants (*e.g.*, Lichten and Goldman 1995; Jeffreys *et al.* 2001). The approach can be modified accordingly, once a good model of rate variation is available.

In addition, the method is designed for the situation where researchers have a candidate site for positive selection and wish to assess support for this hypothesis (*e.g.*, Enard *et al.* 2002), so it assumes that the physical distance to the selected site is known. More common may be the situation where polymorphism data seem to support a selective sweep at a nearby site, but the exact location is unknown. To model this, the distance to the selected site could be chosen from a prior distribution rather than being fixed. Whether there is enough information in a single locus to estimate an additional parameter is unclear. A possible solution is to combine information from a large number of independently evolving loci. These loci would share demographic parameters, but have different selection parameters (*e.g.*, Bustamante *et al.* 2002). Even in this approach, however, there would be three parameters specific to each locus: the selection coefficient, the location of the favored allele, and the time since fixation. Thus, it may be possible to obtain accurate estimates of a subset of the parameters only when independent information is available about others.

A further shortcoming of this method is that it makes a number of demographic assumptions that are unrealistic for humans (Cavalli-Sforza *et al.* 1994), Drosophila (Andolfatto 2001), and maize (Matsuoka *et al.* 2002) as well as for many other species. This can have two effects: first, a departure from demographic assumptions, such as a population size increase, could result in apparent support for a recent selective sweep. This problem might be addressed by contrasting patterns of variability at the locus of interest with data from a large number of independently evolving loci (*e.g.*, Hamblin *et al.* 2002). More problematic is that a selective sweep may not have the same effect if it occurs in other demographic settings (Slatkin and Wiehe 1998). This shortcoming is not specific to this method; any inference about natural selection depends on demographic assumptions. In fact, an attractive feature of this method is that it will be fairly easy to relax demographic assumptions during the neutral phase, once we have a better idea of what modifications are relevant. For instance, once we have a good estimate of the onset of population growth in humans, a change in population size can easily be added. This said, it is not currently possible to model natural selection *and* complicated demographic histories within a coalescent context. Thus, much less computationally efficient, forward simulations may be required if the selective episode occurred in a complex demographic setting.

Thanks go to P. Andolfatto, B. Gaut, P. Donnelly, Y. Gilad, R. Hudson, S. Ptak, M. Tenaillon, J. Wakeley, J. Wall, and L. Zhang for helpful discussions and/or comments on the manuscript.

LITERATURE CITED

Andolfatto, P., 2001 Contrasting patterns of X-linked and autosomal nucleotide variation in *Drosophila melanogaster* and *Drosophila simulans*. Mol. Biol. Evol. **18:** 279–290.

Bachtrog, D., and B. Charlesworth, 2002 Reduced adaptation of a non-recombining neo-Y chromosome. Nature **416:** 323–326.

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

Beaumont, M. A., W. Zhang and D. J. Balding, 2002 Approximate Bayesian computation in population genetics. Genetics **162:** 2025–2035.

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.

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

Carlin, B. P., and T. A. Louis, 1998 *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman & Hall, Boca Raton, FL.

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

David, J. R., and P. Capy, 1988 Genetic variation of *Drosophila melanogaster* natural populations. Trends Genet. **4:** 106–111.

Donnelly, P., M. Nordborg and P. Joyce, 2001 Likelihoods and simulation methods for a class of nonneutral population genetics models. Genetics **159:** 853–867.

Enard, W., M. Przeworski, S. E. Fisher, C. S. Lai, V. Wiebe *et al.*, 2002 Molecular evolution of FOXP2, a gene involved in speech and language. Nature **418:** 869–872.

Eyre-Walker, A., R. L. Gaut, H. Hilton, D. L. Feldman and B. S. Gaut, 1998 Investigation of the bottleneck leading to the domestication of maize. Proc. Natl. Acad. Sci. USA **95:** 4441–4446.

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

Fearnhead, P., and P. Donnelly, 2002 Approximate likelihood methods for estimating local recombination rates. J. R. Stat. Soc. Ser. B **64:** 657–680.

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

Hamblin, M. T., E. E. Thompson and A. Di Rienzo, 2002 Complex signatures of natural selection at the Duffy blood group locus. Am. J. Hum. Genet. **70:** 369–383.

Hudson, R. R., 1990 Gene genealogies and the coalescent process, pp. 1–14 in *Oxford Surveys in Evolutionary Biology*, edited by D. Futuyma and J. Antonovics. Oxford University Press, Oxford.

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

Hudson, R. R., K. Bailey, D. Skarecky, J. Kwiatowski and F. J. Ayala, 1994 Evidence for positive selection in the superoxide dismutase (Sod) region of *Drosophila melanogaster*. Genetics **136:** 1329–1340.

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

Jensen, M. A., B. Charlesworth and M. Kreitman, 2002 Patterns of genetic variation at a chromosome 4 locus of *Drosophila melanogaster* and *D. simulans*. Genetics **160:** 493–507.

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

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

Klein, R. G., 1995 Anatomy, behavior and modern human origins. J. World Prehist. **9:** 167–198.

Kong, A., D. F. Gudbjartsson, J. Sainz, G. M. Jonsdottir, S. A. Gudjonsson *et al.*, 2002 A high-resolution recombination map of the human genome. Nat. Genet. **31:** 241–247.

Lachaise, D., M. L. Cariou, J. R. David, F. Lemeunier, L. Tsacas *et al.*, 1988 Historical biogeography of the Drosophila-melanogaster species subgroup. Evol. Biol. **22:** 159–225.

Lichten, M., and A. S. H. Goldman, 1995 Meiotic recombination hotspots. Annu. Rev. Genet. **29:** 423–444.

Matsuoka, Y., Y. Vigouroux, M. M. Goodman, G. J. Sanchez, E. Buckler *et al.*, 2002 A single domestication for maize shown by multilocus microsatellite genotyping. Proc. Natl. Acad. Sci. USA **99:** 6080–6084.

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

Mellars, P., 1998 Neanderthals, modern humans and the archeological evidence for language, pp. 89–115 in *The Origin and Diversification of Language*, edited by N. G. Jablonski and L. C. Aiello. California Academy of Sciences, San Francisco.

Perlitz, M., and W. Stephan, 1997 The mean and variance of the number of segregating sites since the last hitchhiking event. J. Math. Biol. **36:** 1–23.

Pritchard, J. K., M. T. Seielstad, A. Perez-Lezaun and M. W. Feldman, 1999 Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Mol. Biol. Evol. **16:** 1791–1798.

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

Rozas, J., and R. Rozas, 1999 DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics **15:** 174–175.

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

Simonsen, K. L., G. A. Churchill and C. F. Aquadro, 1995 Properties of statistical tests of neutrality for DNA polymorphism data. Genetics **141:** 413–429.

Slatkin, M., and T. Wiehe, 1998 Genetic hitch-hiking in a subdivided population. Genet. Res. **71:** 155–160.

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

Stringer, C., 2002 Modern human origins: progress and prospects. Philos. Trans. R. Soc. Lond. B Biol. Sci. **357:** 563–579.

Strobeck, C., 1987 Average number of nucleotide differences in a sample from a single subpopulation—a test for population subdivision. Genetics **117:** 149–153.

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

Tavare, S., D. J. Balding, R. C. Griffiths and P. Donnelly, 1997 Inferring coalescence times from DNA sequence data. Genetics **145:** 505–518.

Tenaillon, M. I., M. C. Sawkins, A. D. Long, R. L. Gaut, J. F. Doebley *et al.*, 2001 Patterns of DNA sequence polymorphism along chromosome 1 of maize (Zea mays ssp. mays L.). Proc. Natl. Acad. Sci. USA **98:** 9161–9166.

Tishkoff, S. A., R. Varkonyi, N. Cahinhinan, S. Abbes, G. Argyropoulos *et al.*, 2001 Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. Science **293:** 455–462.

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

Wall, J. D., 2000 A comparison of estimators of the population recombination rate. Mol. Biol. Evol. **17:** 156–163.

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

Weber, J. L., 2002 The Iceland map. Nat. Genet. **31:** 225–226.

Weiss, G., and A. von Haeseler, 1998 Inference of population history using a likelihood approach. Genetics **149:** 1539–1546.

Whitt, S. R., L. M. Wilson, M. I. Tenaillon, B. S. Gaut and E. S. T. Buckler, 2002 Genetic diversity and selection in the maize starch pathway. Proc. Natl. Acad. Sci. USA **99:** 12959–12962.

Communicating editor: N. Takahata