Properties of Statistical Tests of Neutrality for DNA Polymorphism Data

Katy L. Simonsen,[*] Gary A. Churchill[*,†] and Charles F. Aquadro[‡]

[*]Center for Applied Math, [†]Biometrics Unit, [‡]Section of Genetics and Development, Cornell University, Ithaca, New York 14853

Manuscript received February 13, 1995
Accepted for publication June 9, 1995

ABSTRACT

A class of statistical tests based on molecular polymorphism data is studied to determine size and power properties. The class includes TAJIMA's D statistic as well as the D^* and F^* tests proposed by FU and LI. A new method of constructing critical values for these tests is described. Simulations indicate that TAJIMA's test is generally most powerful against the alternative hypotheses of selective sweep, population bottleneck, and population subdivision, among tests within this class. However, even TAJIMA's test can detect a selective sweep or bottleneck only if it has occurred within a specific interval of time in the recent past or population subdivision only when it has persisted for a very long time. For greatest power against the particular alternatives studied here, it is better to sequence more alleles than more sites.

GIVEN a set of aligned DNA sequences from a sample of n individuals of the same species, we would like to make inferences about the evolutionary history of the species. The neutral equilibrium model of sequence evolution is often considered as a null hypothesis against which specific alternative models can be compared. The neutral hypothesis is rejected if the observed data are unlikely to arise under this model. A problem of interest is to construct appropriate test statistics that will reject the neutral model with high probability when specific alternative models hold. We consider a class of test statistics that includes TAJIMA's D statistic (1989a) and the D^* and F^* tests proposed by FU and LI (1993). The power properties of these tests against specific alternative hypotheses are studied using simulated data to determine how often and under which alternatives each test is able to reject the neutral model.

Critical values (rejection regions) of statistical tests are determined by the distribution of the statistics under the null hypothesis. The distributions of the test statistics we wish to examine are not known, but we can sample from these distributions by simulating data from the neutral model. Estimating the critical values is complicated because the distributions depend on the unknown value of a parameter θ which is proportional to the product of the effective population size and the mutation rate.

Our goal is to determine which statistical tests are most powerful against different alternatives and to determine the sample sizes necessary to achieve a reasonable power. We also address the issue of larger sample sizes *vs.* greater number of sites sequenced with respect to improving statistical power.

This work was motivated in part by studies of natural populations of Drosophila. These studies have shown that levels of DNA polymorphism observed for a gene region are strongly correlated with regional rates of recombination, (*e.g.*, AGUADÉ *et al.* 1989; STEPHAN and LANGLEY 1989; BEGUN and AQUADRO 1991, 1992; BERRY *et al.* 1991; AQUADRO *et al.* 1994). One hypothesis to explain this correlation is that hitchhiking associated with the fixation of advantageous mutations leads to a reduction in linked neutral variation, (*e.g.*, KAPLAN *et al.* 1989; MAYNARD SMITH and HAIGH 1974). However, in many of the cases cited TAJIMA's D test did not reject the neutral model. This suggests the following question: is TAJIMA's D powerful enough to detect deviations from the neutral model or is its behavior indistinguishable from neutrality even when the neutral model is violated? The hitchhiking effect is simplest in the total absence of recombination when all variation at a locus is eliminated due to the fixation of a completely linked, advantageous mutation. Such a "selective sweep" event is one of the alternative hypotheses we investigate here. It must be emphasized that it is not the goal of the present paper to accept or reject the hitchhiking hypothesis for particular data sets. To do that, it would be necessary to set up hitchhiking as the null hypothesis and to explore the full range of possible parameters (including strength of selection, recombination rate, population size, dominance, neutral mutation rate, time since fixation) affecting this model. Here we vary only a few of these parameters to construct different alternative hypotheses against which the power can be estimated.

In the remainder of this section, we summarize the coalescent model of neutral evolution in order both to introduce notation and to make clear those assumptions that are violated by the alternative hypotheses. In MATERIALS AND METHODS, first we describe a class of

Corresponding author: Katy Simonsen, Center for Applied Math, Engineering and Theory Center, Cornell University, Ithaca, NY 14853. E-mail: katy@cam.cornell.edu

test statistics and a method by which critical values for statistical tests of the neutral model can be obtained; then, we describe how data was simulated under alternatives to the neutral model. The RESULTS section summarizes the outcome of these simulations, showing the effect of these alternatives on the distributions of the test statistics and their power. In the final section, we discuss the implications of these results to performing statistical tests.

The neutral model: The neutral data were generated according to the coalescent model as described by HUDSON (1990, 1993). This model is based on the standard Wright-Fisher model and makes the following assumptions: (1) a large constant diploid population size of N individuals or $2N$ alleles (where $N^2 \gg N$), (2) random mating, (3) nonoverlapping generations, (4) no recombination, and (5) an infinite-sites, constant rate neutral mutation process whereby an offspring differs from its parent allele by a Poisson-distributed number of mutations with mean μ.

Under these assumptions, the probability that two particular individuals have the same parent in the previous generation is $1/2N$. The probability that any two individuals in a sample of size j have the same parent is $p \approx \binom{j}{2}/2N$. Thus, for a sample of j individuals in the current population, the probability that the first coalescent event between any two of them occurs exactly $t + 1$ generations ago is $p(1 - p)^t$. That is, the time in generations during which there are exactly j lineages in the genealogy of the sample is geometrically distributed with mean $1/p$. It is convenient to treat time as a continuous random variable. To this end, we approximate the geometric distribution with an exponential distribution with the same mean, because $p(1 - p)^t \approx pe^{-pt}$ for small p and large t. The assumption (1) that $N^2 \gg N$ ensures that p is sufficiently small. It is also convenient to measure time in units of $2N$ generations, with the result that p is replaced by $\binom{j}{2}$. Thus the time t_j in units of $2N$ generations during which there are exactly j lineages is exponentially distributed with mean $1/\binom{j}{2}$. The total time in the tree, T_{tot}, is equal to $\Sigma_{j=2}^{n} jt_j$.

The number of mutations that occur on a lineage of length t is, by assumption (5), Poisson-distributed with mean $2N\mu t = \theta t/2$, where $\theta = 4N\mu$. The assumption of infinite sites ensures that each mutation is observed as a polymorphic or segregating site. Therefore the number S of segregating sites in a sample is Poisson-distributed with mean $\theta T_{tot}/2$.

As HUDSON (1993) has pointed out, the fact that the true value of θ for data sets is unknown presents a problem when using simulation to estimate critical values for a test. Three methods of generating data are described by HUDSON (1993): conditioning on θ, conditioning on θ and S, and conditioning on S. The first method is the one consistent with our model, but it requires knowing θ. The other two methods would require modifying the null hypothesis: instead of a neutral mutation process with rate μ, we would have to postulate a fixed number of mutations that is independent of the total time in the tree. To apply the first method, we use the information contained in S to compute a range of values for θ that are consistent with the observed data. We then use values of θ in this interval to simulate the test statistic under the neutral model, and thus obtain critical values.

MATERIALS AND METHODS

Statistical tests

From n nucleotide sequences, statistics such as S, the number of segregating sites, k, the average number of pairwise differences, and η_s, the number of singletons (mutations appearing in only one sampled allele), may be calculated. These are random variables whose distribution depends on a parameter θ whose value is unknown, and each provides an unbiased estimate of θ. Let

$$a_n = \sum_{i=1}^{n-1} \frac{1}{i}, \quad b_n = \sum_{i=1}^{n-1} \frac{1}{i^2}. \tag{1}$$

Under the neutral model, $E(S) = a_n\theta$, $E(k) = \theta$, and $E(\eta_s) = [n/(n - 1)]\theta$. Their variances are

$$\mathrm{Var}(S) = a_n\theta + b_n\theta^2 \quad (\text{WATTERSON 1975}) \tag{2}$$

$$\mathrm{Var}(k) = \frac{(n + 1)\theta}{3(n - 1)} + \frac{2(n^2 + n + 3)\theta^2}{9n(n - 1)} \quad (\text{TAJIMA 1983}) \tag{3}$$

$$\mathrm{Var}(\eta_s) = \frac{n}{n - 1}\theta + \left[\frac{2a_n}{n - 1} - \frac{1}{(n - 1)^2}\right]\theta^2$$

$$(\text{FU and LI 1993}). \tag{4}$$

Therefore, S/a_n, k, and $[(n - 1)/n]\,\eta_s$ are unbiased estimators of θ, and

$$m_1^2 = S(S - 1)/(a_n^2 + b_n) \tag{5}$$

$$m_2^2 = \frac{3nk(3(n - 1)k - n - 1)}{11n^2 - 7n + 6} \tag{6}$$

$$m_3^2 = \frac{(n - 1)\eta_s(\eta_s - 1)}{2a_n + n + 1} \tag{7}$$

are unbiased estimators of θ^2.

In the following section we define a class of test statistics that includes three previously described test statistics and six new ones.

Test statistics: From the three statistics S, k, and η_s, we can calculate test statistics such as those of TAJIMA (1989a):

$$D(k, S) = \frac{k - S/a_n}{\sqrt{u_T S + v_T S^2}} \tag{8}$$

and of FU and LI (1993):

$$D^*(S, \eta_s) = \frac{S/a_n - \eta_s\left(\dfrac{n - 1}{n}\right)}{\sqrt{u_{D^*}S + v_{D^*}S^2}} \tag{9}$$

$$F^*(k, S, \eta_s) = \frac{k - \eta_s\left(\frac{n-1}{n}\right)}{\sqrt{u_{F^*}S + v_{F^*}S^2}} \qquad (10)$$

where the coefficients u and v are given in the APPENDIX. We refer to the coefficients for TAJIMA's statistic as u_T and v_T rather than u_D and u_D to distinguish them from those for FU and LI's D statistic (1993), which is not studied in this paper because it requires an outgroup. The formula for v_{F^*} given in the APPENDIX differs slightly from that given in FU and LI (1993, unnumbered equation p. 702) due to a typographical error in their paper. Under the neutral model the three statistics D, D^*, and F^* all have expected value approximately 0 and variance approximately 1.

Each of these statistics is constructed in the same way. Under neutrality, the three quantities S/a_n, k, and $\eta_s[(n-1)/n]$ all have expected value θ. Thus the difference between any two of these statistics will have expected value 0. The variances of the differences are of the form $\gamma\theta + \epsilon\theta^2$, where γ and ϵ depend on the statistic in question. The variance of the difference is estimated using S/a_n and m_1^2 as unbiased estimates for θ and θ^2, respectively. The result is an estimated variance of the form $uS + vS^2$ where $u = \gamma/a_n - v$ and $v = \epsilon/(a_n^2 + b_n)$. Each test statistic is constructed by dividing the difference by the square root of its estimated variance. Note that because this denominator depends on the data, the expected value of the statistics is not exactly 0; simulations show that the mean is slightly negative. Because it is possible for the denominator to be 0, for purposes of this study we define D, D^*, and F^* to be 0 when $S = 0$. This has the effect of making rejection of the neutral model impossible when no variation is observed.

The statistics D, D^*, and F^* use S and m_1^2 to estimate θ and θ^2 in the variance term $\gamma\theta + \epsilon\theta^2$. It is also possible to use k and m_2^2 or η_s and m_3^2 to make this estimate. S has been used because S/a_n has a smaller variance under neutrality than the other possibilities. In nonneutral situations, however, the behavior of S, k, and η_s is more complex, so that k or η_s could make a better estimator of θ or θ^2 in some cases. We can construct six new test statistics, as follows

$$T_2(k, S) = \frac{k - S/a_n}{\sqrt{u_{T_2}k + v_{T_2}k^2}} \qquad (11)$$

$$T_3(k, S, \eta_s) = \frac{k - S/a_n}{\sqrt{u_{T_3}\eta_s + v_{T_3}\eta_s^2}} \qquad (12)$$

$$D_2^*(k, S, \eta_s) = \frac{S/a_n - \eta_s\left(\frac{n-1}{n}\right)}{\sqrt{u_{D_2^*}k + v_{D_2^*}k^2}} \qquad (13)$$

$$D_3^*(S, \eta_s) = \frac{S/a_n - \eta_s\left(\frac{n-1}{n}\right)}{\sqrt{u_{D_3^*}\eta_s + v_{D_3^*}\eta_s^2}} \qquad (14)$$

$$F_2^*(k, \eta_s) = \frac{k - \eta_s\left(\frac{n-1}{n}\right)}{\sqrt{u_{F_2^*}k + v_{F_2^*}k^2}} \qquad (15)$$

$$F_3^*(k, \eta_s) = \frac{k - \eta_s\left(\frac{n-1}{n}\right)}{\sqrt{u_{F_3^*}\eta_s + v_{F_3^*}\eta_s^2}} \qquad (16)$$

The subscript 2 or 3 indicates that the estimate of θ uses k and m_2, or η_s and m_3, respectively. The coefficients u and v are defined in the APPENDIX. The properties of these tests will be investigated along with those of the standard tests. For convenience, we again define these statistics to be 0 when their denominator is 0.

Hypothesis testing issues: Because neither S, k, nor η_s is a sufficient statistic for θ, the variance of any of the above test statistics will not be one and will vary with θ (HUDSON 1993). Thus, computed critical values for these test statistics must account for the unknown θ. Furthermore, even when θ is known, the exact distribution of the statistics under the null hypothesis is not. To perform two-sided tests of level α, the critical values required are the boundaries of a $(1 - \alpha)$ confidence interval for the test statistic. That is, for the statistic D we require D_U and D_L, independent of θ, such that the sum of the p-values $p_L = \text{Prob}_{H_0}(D \leq D_L)$ and $p_U = \text{Prob}_{H_0}(D \geq D_U)$ is less than or equal to α. We note that because S, $\binom{n}{2}k$, and η are integers, the test statistics will have a discrete distribution, and so any non-randomized test will not precisely achieve the desired level. Other authors have suggested methods to determine critical values as described below. We present an alternative method.

TAJIMA (1989a) computed critical values by assuming D to have a beta distribution with mean zero and variance one, scaled to the interval $[D_{min}, D_{max}]$. TAJIMA's justification is based on a visual comparison between beta densities and histograms of simulated data. We have found that TAJIMA's critical values are often too conservative, particularly at the upper tail of the distribution. While it is true that the probability of false rejection is not increased by using conservative critical values, it can result in a serious reduction in power. Thus this method of obtaining critical values is less than satisfactory.

FU and LI (1993) used simulated data with known values of θ and n to locate appropriate percentiles as estimates for the critical value of the statistic. Then for each value of n, they took the most extreme of these critical values over all θ's in the interval [2, 20]. The effect of this technique is to reject only when the data cannot be explained by neutral evolution for any value of θ in this interval. The interval [2, 20] for θ was chosen somewhat arbitrarily to represent "most cases of interest."

While FU and LI's approach is an improvement over TAJIMA's, there are still some problems remaining. First, the critical values are not applicable when the true value of θ is not in [2, 20], and we cannot know, for a given set of data, whether this is the case. Because θ is a per-locus value, it changes with the number of nucleotides being sequenced, as well as with the underlying mutation rate. Thus it is difficult to justify why θ would have to be confined to this range. The test may falsely reject when θ is not in this interval. Further, their technique does not take into account the information about θ inherent in the data. We will attempt to address these problems below.

The problem of the unknown parameter θ may be addressed using the technique proposed by BERGER and BOOS (1994). Ideally, we would like to reject only if the data cannot be explained by any positive value of θ. For a test of level α this would mean choosing critical values $[D_L, D_U]$ for D such that

$$\sup_{\theta\in[0,\infty)} [\text{Prob}_\theta(D \leq D_L) + \text{Prob}_\theta(D \geq D_U)] \leq \alpha \qquad (17)$$

However, we cannot perform simulations for infinitely many values of θ, nor is it reasonable to do so because extremely large values of θ are unlikely. Instead, for some small number $\beta < \alpha$, we use the data to estimate C_β, a $1 - \beta$ confidence interval for θ, and require critical values to satisfy

$$\sup_{\theta \in C_\beta}[\mathrm{Prob}_\theta(D \le D_L) + \mathrm{Prob}_\theta(D \ge D_U)] \le \alpha - \beta. \quad (18)$$

For each θ in a grid covering C_β we estimate level $(\alpha - \beta)$ critical values $[D^\theta_L, D^\theta_U]$ using the percentiles of neutral data for that θ. We take the most extreme of those critical values over all θ in C_β:

$$D_L = \min_{\theta \in C_\beta} D^\theta_L, \quad D_U = \max_{\theta \in C_\beta} D^\theta_U. \quad (19)$$

The result is a level α test, as shown by BERGER and BOOS (1994). This approach is similar to that of FU and LI (1993), except that instead of arbitrarily using the interval [2, 20] we use an interval that reflects our knowledge of θ for the data set being considered. This has the advantage of giving us a test with known level for any value of the unknown parameter θ.

To construct this $1 - \beta$ confidence interval for θ, we use the exact distribution for S given θ, as given by TAVARÉ (1984). We wish to find a two-sided interval $C_\beta = [\theta_L, \theta_U]$ such that, for a particular observation $S = s$, and for fixed n,

$$\mathrm{Prob}(S \ge s | \theta = \theta_L) = \beta/2 \quad (20)$$

$$\mathrm{Prob}(S \le s | \theta = \theta_U) = \beta/2. \quad (21)$$

The cumulative distribution function for S given θ is

$$F(s, n, \theta) = \mathrm{Prob}(S \le s | \theta) = 1 - \sum_{r=1}^{n-1} (-1)^{r-1} \binom{n-1}{r} \left(\frac{\theta}{r+\theta}\right)^{s+1}. \quad (22)$$

So, (20) and (21) may be written as

$$F(s-1, n, \theta_L) = 1 - \beta/2 \quad (23)$$

$$F(s, n, \theta_U) = \beta/2 \quad (24)$$

Thus we must solve (23) and (24) for θ_L and θ_U for the particular values of $S = s$ and n observed in the data. This is computationally intensive for large values of n and requires high precision to compute accurately in many cases. We used the variable-precision capabilities of the symbolic computation package *Maple* (CHAR *et al.* 1991) to perform the calculations. The results of these computations are given in RESULTS. Note that when $S = 0$ is observed, it is appropriate to set $\theta_L = 0$ and solve $F(0, n, \theta_U) = \beta$ for θ_U in place of (24); however, because all the test statistics are defined to be 0 in this case, the resulting critical values will be irrelevant.

In summary, there are three distinct steps to computing critical values for the test statistics in this fashion. (1) For the values of n and S required, compute C_β, a $1 - \beta$ confidence region for θ given S. (2) For a grid of θ values in C_β and for each n, simulate a large number of samples and estimate level $(\alpha - \beta)$ critical values for each test statistic from the simulated empirical distributions. (3) Take the maximum upper critical value and minimum lower critical value over all values of θ in C_β, for each value of n and S and for each test statistic. This gives critical values of α-level tests for each n and S.

Simulations

To evaluate the power of the statistical tests described above, we require data simulated under a number of different alternative models. The alternatives considered here: a selective sweep event, a population bottleneck, and a subdivided population, represent a few simple deviations from strict neutrality and are meant as examples rather than as a comprehensive study. Because balancing selection is similar to population subdivision from a coalescent perspective (HUDSON 1990), we expect the results for a subdivided population to be applicable to the corresponding balancing selection alternative as well.

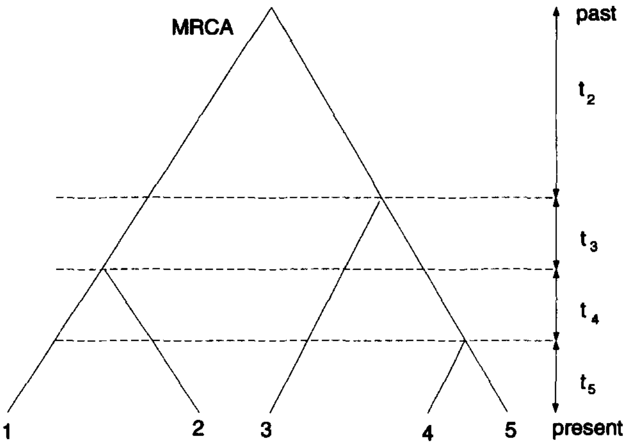

FIGURE 1.—An example of a coalescent tree for a sample of five alleles.

Neutral simulations: A sample of DNA sequences is generated by simulating a random gene genealogy according to the algorithm developed by HUDSON (1990, 1993). There are three components to this genealogy: topology, branch lengths, and mutations. First, a random tree topology is generated for the genealogy. From n individuals in the sample, two are chosen at random to be the first to coalesce. A new individual is designated as their parent, and the process is repeated on the remaining $n - 1$ individuals. The process stops when only one individual, the most recent common ancestor (MRCA) of the entire sample, remains. This gives the topology of a binary tree with n tips. Next, the branch lengths are chosen: t_j, the time (in units of $2N$ generations) during which there are exactly j lineages, is an exponentially distributed random variable with mean $1/\binom{j}{2}$ as described in the Introduction. These two steps define a tree such as that shown in Figure 1. Finally, mutations are added to the tree. The number of mutations S that have occurred during the history of the sample is generated as a Poisson-distributed random variable with mean $\theta T_{tot}/2$. For each mutation, the branch of the tree on which it occurred is chosen randomly, where the relative probability of each branch is proportional to its length. The mutation is transmitted to each offspring descended from that branch. Thus each individual is assigned a "sequence" of nucleotides designated, for example, $-+--++-$, where "$-$" indicates that the nucleotide is identical to the ancestral sequence at that site and "$+$" indicates a mutation. Under the infinite-sites model, each mutation is assumed to take place at a distinct nucleotide site, and thus each sequence generated is composed only of polymorphic or segregating sites.

Selective sweep simulations: A highly favorable mutation with selective advantage s and dominance h that occurs at a time T_s is assumed to sweep through the population and reach fixation in a deterministic fashion, such that the proportion $x(t)$ of individuals carrying the mutation at time t follows

$$\dot{x}(t) = \frac{2Nsx(1-x)[x + h(1-2x)]}{1 + s[x^2 + 2hx(1-x)]}, \quad x(T_s) = \frac{1}{2N}. \quad (25)$$

This result can be found in (MAYNARD SMITH and HAIGH 1974, equation 18). We have inserted a multiplicative factor of $2N$ to correct for the measurement of time in units of $2N$ generations. We assume that the initial frequency of the selected allele is $x(T_s) = 1/2N$, that the process is deterministic even when the frequency is low, and that the allele has been fixed in the population when the frequency reaches $1 - 1/2N$.

The selective sweep alters the coalescent process by reducing the effective population size of the parental generation at time t from $2N$ to $2Nx(t)$, because only genes carrying the selected mutation may be chosen as ancestors of the sample. The per-generation coalescent probabilities change from $\binom{j}{2}/2N$ to $\binom{j}{2}/[2Nx(t)]$. Thus the total size of the tree is reduced, and the effect of the selective sweep is to reduce variation at and around the selected locus. We are assuming no recombination between the selected and sampled loci.

To generate coalescent times under a sweep, we generate times according to the neutral model, and then scale them appropriately, as described below. This approach was suggested to us by R. R. HUDSON; also see GRIFFITHS and TAVARÉ (1994, equation 3). To convert a time from one time scale to another, we must perform a change of variables. Suppose U is a time measured in units of $2Nx(t)$ at time t. We wish to convert this time U back into the standard units of $2N$ generations. The instantaneous change of variables at time t is $2Nx(t)\,du = 2Ndt$, where dt is the interval in regular units $2N$ and du is the time interval in units $2Nx(t)$. This becomes

$$\frac{dt}{x(t)} = du, \tag{26}$$

and thus, if T represents the same time as U but in regular units, we integrate over the whole interval to obtain

$$\int_0^T \frac{dt}{x(t)} = \int_0^U du = U. \tag{27}$$

Therefore, to generate a coalescent time T under a selective sweep described by $x(t)$, we generate a time U under the neutral model and then find T which solves (27). This is done for each coalescent time in a tree, to generate a coalescent tree for a selective sweep.

If the selective sweep began quite recently, it is possible that the selected allele has not yet been fixed in the population. The length of time T_d this sweep takes to complete depends on N, s, and h; for example when $h = {}^1/_2$, the sweep lasts $T_d = 2\ln(2N - 1)/sN$. If $T_s > T_d$, then the allele has become fixed by the present time (0), and so the MRCA of the sample has to occur more recently than T_s, because it must be descended from the initial selected mutant. If the sweep began so recently that the selected allele has not yet completely reached fixation ($T_s < T_d$), then there are two types of alleles in the present population, those with, and those without, the selected mutation, in the ratio $x(0){:}1 - x(0)$. In this case we assume the number of sampled individuals having the selected mutation is binomially distributed with parameters $(n, x(0))$. The selected alleles follow the above sweep model, while the rest of the sample is drawn from a population of size $2N[1 - x(t)]$ at time t. This requires a shift in time units similar to that described above, with $x(t)$ replaced by $1 - x(t)$. Before the selected mutation occurred, ($t > T_s$) all ancestors follow a neutral model.

Our model of the selective sweep is defined in terms of four parameters: h, s, N, and its starting time T_s. For this study, we chose to fix $h = 0.5$, $N = 10^6$, and $s = 10^{-4}$, and allow T_s to vary over the range 0–2 (in units of $2N$, back in time from the present). This is relatively weak selection on a codominant allele; for comparison we also performed the simulations with $s = 10^{-2}$. For combinations of n (10, 20, 50, 100), θ (10, 20, 50), and T_s (in increments of 0.01 from 0 to 0.3, then in increments of 0.05 from 0.3 to 0.5, then in increments of 0.1 from 0.5 to 2.0), 5000 samples were generated and the proportion of rejections for each of the tests recorded.

A selective sweep is expected to reduce polymorphism at linked sites, because any observed polymorphism must be the result of mutations that have occurred since the sweep. These newly arisen mutations will at first be rare and will increase in frequency as the time since the sweep increases. Because S takes into account only the number of mutations, while k is also affected by their frequency, it is expected that S will recover more rapidly than k from the effects of a sweep. This will have the effect of reducing the expected value of TAJIMA's statistic below its neutral expectation of 0. The magnitude of this reduction has not been predicted by theory, and is one of the subjects of the present investigation.

Population bottleneck simulations: A population bottleneck is assumed to occur when the population, originally of size $2N$, is suddenly reduced to a fraction f of its former size for a length of time l, then instantaneously regains its initial size. Let T_b be the time (in units of $2N$ generations) at which the bottleneck ended, so that it began at a time $T_b + l$, which is further from the present time 0. Coalescent times under this model are obtained by scaling neutral coalescent times in the same way as the selective sweep. In this case, the changing population size $2Nx(t)$ is a step function rather than a smooth curve as it was for the sweep, so the integration in (27) is easy. We generate a time u_j under the neutral model, and then use as the coalescent time t_j given by

$$t_j = \begin{cases} u_j, & u_j < T_b \\ T_b + (u_j - T_b)f, & T_b < u_j < T_b + \dfrac{l}{f} \\ u_j - \left(\dfrac{1}{f} - 1\right)l, & T_b + \dfrac{l}{f} < u_j. \end{cases} \tag{28}$$

This is equivalent to the following probability density for the coalescent times t_j:

$$t_j \sim \begin{cases} pe^{-pt_j}, & t_j < T_b \\ \dfrac{p}{f}\, e^{-(p/f)t_j} e^{pT_b[(1/f) - 1]}, & T_b < t_j < T_b + l \\ pe^{-pt_j} e^{-pl[(1/f) - 1]}, & T_b + l < t_j \end{cases} \tag{29}$$

where $p = \binom{j}{2}$. The density can be derived by considering the per-generation probabilities of coalescence during the three stages of the bottleneck.

For purposes of this study, we kept f fixed at 0.01, l fixed at 0.1, and varied T_b, the time since the bottleneck ended, from 0 to five. These are bottlenecks of the same severity but lasting ten times the length of those considered by TAJIMA (1993). The fraction rejected out of 1000 simulations was recorded.

A population bottleneck is expected to reduce polymorphism throughout the genome, because a drastic reduction in population size is likely to eliminate many rare variants. As in the case of the selective sweep, most of the polymorphism will be a result of new mutations, which will be rare. Thus a reduction of unknown magnitude in the expectation of TAJIMA's statistic is predicted (1989b).

Subdivided population simulations: The third alternative modeled was a subdivided population with no migration. We expect the results of this model to apply to balanced polymorphism as well, because the two are similar from a coalescent perspective. We start with an ancestral population size of $2N_{AB}$. At a certain time T_m this population is assumed to split into two isolated populations A and B, of size N_A and N_B, respectively, which evolve independently from then on. Here, the sample of n alleles consists of n_A alleles of type A, and n_B of type B, with $n = n_A + n_B$.

The coalescent tree for such a subdivided population is generated in the following manner. As usual we work backward in time from the present to the time of the MRCA of the sample. Let j_A and j_B be the number of lineages of type A and B remaining at any given time. Initially, we let $j_A = n_A$ and $j_B = n_B$, and at each coalescent event, one of them is decremented. We need to know the distribution of the time back to the next coalescent event.

In the subdivided model, the coalescent probabilities before and after the population split are different. When the two populations are disjoint, the probability per generation that two A individuals coalesce is $p_A = \binom{j_A}{2}/2N_A$ while for the B population it is $p_B = \binom{j_B}{2}/2N_B$. The probability that both populations will coalesce in the same generation is negligible $[O(1/N^2)]$ compared with p_A and p_B, and so the per-generation probability of a coalescent event in either population is approximately $p_1 = p_A + p_B$ while the two populations are disjoint. When the two populations are mixed, we have a single population of size $2N_{AB}$, with $j_A + j_B$ sample lineages present. Thus the per-generation probability of coalescence for the mixed population is $p_2 = \binom{j_A+j_B}{2}/2N_{AB}$.

As before, t_i is the time during which there are exactly i lineages of any type present. Let $S_j = \Sigma_{i=j}^{n} t_i$, with $S_{n+1} = 0$. S_j keeps track of the total time generated so far. To generate the time $t_{j_A+j_B}$ to the next event, it is necessary to know the relationship between $S_{j_A+j_B+1}$ and T_m. In particular, if $S_{j_A+j_B+1} > T_m$, then we have passed the subdivision point and we may generate subsequent times $t_{j_A+j_B}$ simply as exponentially distributed random variables with parameter p_2. On the other hand, suppose $S_{j_A+j_B+1} < T_m$, say $T_m - S_{j_A+j_B+1} = M > 0$. Then the time $t_{j_A+j_B}$ generated could be less than or greater than M. The probability of coalescence after a given time $t < M$ is $(1 - p_1)^t p_1 \approx p_1 e^{-p_1 t}$. But for a time $t > M$, the probability of coalescence is $(1 - p_1)^M (1 - p_2)^{t-M} p_2 \approx p_2 e^{-p_2 t} e^{M(p_2 - p_1)}$. Thus

$$t_{j_A+j_B} \sim \begin{cases} p_1 e^{-p_1 t} I_{[t<M]} + p_2 e^{-p_2 t} e^{M(p_2-p_1)} I_{[t>M]}, & M > 0 \\ p_2 e^{-p_2 t}, & M < 0 \end{cases} \tag{30}$$

where I is an indicator function and $M = T_m - S_{j_A+j_B+1}$.

Once a time has been generated from this mixture of exponentials, two individuals must be chosen to coalesce at that time. If the total time $S_{j_A+j_B+1}$ is still less than T_m, then we must choose between group A and group B with relative probabilities p_A and p_B. If the time is greater than T_m, we have only one group. Once the group is chosen, two of the appropriate group are selected at random, and the corresponding j is decremented. The process is repeated until only one individual remains.

When a population is subdivided, the average pairwise difference k is inflated relative to the total number of mutations S, because of the large divergence between subpopulations. Thus the qualitative expectation is that D will have a positive mean in this situation. As with the selective sweep and bottleneck, we chose time since the subdivision event as the primary variable to investigate, fixing $N_A = N_B = N_{AB}/2$, $n_A = n_B = 25$, and $\theta = 20$.

RESULTS

Results of neutral simulations: Simulations of the null hypothesis were used to provide new critical values for the test statistics. Our technique uses confidence intervals for θ given S, as described in MATERIALS AND METHODS. These $1 - \beta$ confidence intervals were computed using 40 digits of accuracy to solve equations 23 and 24. Tabulation here of these confidence intervals for different values of n, S, and β would be prohibitive, so we show only a sample: the case $n = 50$, $\beta = 0.01$ in Table 1. This table shows, for example, that if $S = 23$ is observed from a sample of size $n = 50$, and the neutral model holds, then with 99% certainty θ is between 2 and 12.5. For other values of n and β, θ_L and θ_U may be closely approximated by linear functions of S, especially when S is large. For example, when $n = 20$, $C_{0.01} \approx [0.121S - 0.481, 0.709S + 2.858]$ and $C_{0.001} \approx [0.094S - 0.473, 0.904S + 4.418]$. The coefficients of these linear approximations are given in Table 2, and can be used to approximate the values corresponding to Table 1 for other values of n and β.

Table 3 shows tables of level 0.05 critical values for TAJIMA's test for a range of S values, for $n = 10, 20, 50, 100$, using $\alpha = 0.05$ and $\beta = 0.01$. For comparison, the values from the beta distribution (TAJIMA 1989a) are also shown. Corresponding values for D^* and F^* (FU and LI 1993) are given in Tables 4 and 5, along with the values that assume $\theta \in [2, 20]$ from (FU and LI 1993).

There is no simple pattern to the way in which the new critical values differ from those of the beta distribution. Generally speaking, for small n the beta distribution values are too large, while for larger n the beta distribution values are too small. The important difference is that the new values are based on a sound statistical framework that does not depend on fitting the statistic to a particular distribution, as TAJIMA did, or on the true value of θ being between 2 and 20, as FU and LI assumed.

The size of these tests (the probability of rejecting when the neutral model is true), based on the new critical values, was estimated by applying each test to 10,000 simulated neutral data sets for each value of θ. The number of false rejections was computed (data not shown). The size for most values of θ is between 3 and 4%, out of a maximum of $\alpha = 5\%$. This shortfall is attributable to three factors. First, because the statistics have discrete distributions, we cannot expect to precisely achieve the desired level with any nonrandomized test. Second, there is some error in estimating the $(\alpha - \beta)$ critical values using the empirical percentiles, because we used a finite number of simulations (10,000). This source of error could be diminished, though not eliminated, by using a larger number of simulations. Third, the Berger and Boos confidence interval procedure is conservative; using it may reduce the size of the test by as much as β. Our choice of β was arbitrary, and the effect of this choice on the size of the tests is not clear. Thus, the critical values might be improved by using a different value of β.

Results of selective sweep simulations: The effect of

TABLE 1

The 99% confidence intervals for θ

S	θ_L	θ_U	S	θ_L	θ_U	S	θ_L	θ_U	S	θ_L	θ_U
0	0.0	1.3	30	2.7	15.6	60	6.0	28.9	90	9.3	42.1
1	0.0	2.1	31	2.9	16.0	61	6.1	29.3	91	9.4	42.5
2	0.0	2.6	32	3.0	16.5	62	6.2	29.8	92	9.5	43.0
3	0.0	3.2	33	3.1	16.9	63	6.3	30.2	93	9.6	43.4
4	0.1	3.7	34	3.2	17.4	64	6.4	30.7	94	9.7	43.9
5	0.2	4.2	35	3.3	17.8	65	6.5	31.1	95	9.8	44.3
6	0.3	4.7	36	3.4	18.3	66	6.6	31.5	96	9.9	44.7
7	0.3	5.1	37	3.5	18.7	67	6.8	32.0	97	10.0	45.2
8	0.4	5.6	38	3.6	19.2	68	6.9	32.4	98	10.1	45.6
9	0.5	6.1	39	3.7	19.6	69	7.0	32.9	99	10.2	46.1
10	0.6	6.6	40	3.8	20.0	70	7.1	33.3	100	10.3	46.5
11	0.7	7.0	41	3.9	20.5	71	7.2	33.7	101	10.5	46.9
12	0.8	7.5	42	4.0	20.9	72	7.3	34.2	102	10.6	47.4
13	0.9	7.9	43	4.2	21.4	73	7.4	34.6	103	10.7	47.8
14	1.0	8.4	44	4.3	21.8	74	7.5	35.1	104	10.8	48.3
15	1.1	8.9	45	4.4	22.3	75	7.6	35.5	105	10.9	48.7
16	1.3	9.3	46	4.5	22.7	76	7.7	35.9	106	11.0	49.1
17	1.4	9.8	47	4.6	23.1	77	7.8	36.4	107	11.1	49.6
18	1.5	10.2	48	4.7	23.6	78	8.0	36.8	108	11.2	50.0
19	1.6	10.7	49	4.8	24.0	79	8.1	37.3	109	11.3	50.5
20	1.7	11.1	50	4.9	24.5	80	8.2	37.7	110	11.4	50.9
21	1.8	11.6	51	5.0	24.9	81	8.3	38.1	111	11.5	51.3
22	1.9	12.0	52	5.1	25.4	82	8.4	38.6	112	11.7	51.8
23	2.0	12.5	53	5.2	25.8	83	8.5	39.0	113	11.8	52.2
24	2.1	12.9	54	5.3	26.2	84	8.6	39.5	114	11.9	52.7
25	2.2	13.4	55	5.5	26.7	85	8.7	39.9	115	12.0	53.1
26	2.3	13.8	56	5.6	27.1	86	8.8	40.3	116	12.1	53.5
27	2.4	14.3	57	5.7	27.6	87	8.9	40.8	117	12.2	54.0
28	2.5	14.7	58	5.8	28.0	88	9.0	41.2	118	12.3	54.4
29	2.6	15.2	59	5.9	28.4	89	9.1	41.7	119	12.4	54.9

Confidence intervals for θ given S when $n = 50$, $\beta = 0.01$. $C_\beta = [\theta_L, \theta_U]$.

a selective sweep on TAJIMA's D, S/a_n and k is shown in Figure 2 for two different strengths of selection: $s = 10^{-4}$ (weak, a and c); $s = 10^{-2}$ (stronger, b and d); for 5000 simulations with $n = 50$ and $\theta = 20$. The horizontal axis is T_s, the time at which the sweep began. Because the present time is 0, this is also the amount of time *since* the sweep began. In Figure 2, a and b, the thicker curve is the mean of TAJIMA's D, while the thinner curves are the 2.5 and 97.5 percentiles. For comparison, the critical values from TAJIMA's (1989a) beta distribution are also shown (horizontal lines), though the critical values from Table 3 were used to determine rejection. The expected trend towards more negative values of D is observed. There is also a pronounced decrease in the variance of the distribution even when the sweep is very ancient. When T_s is very large (six or seven, not shown), the percentile curves eventually level off close to the critical values. Figure 2, c and d, show the values of S/a_n (solid) and k (dotted) associated with a and b respectively. The thicker lines represent the means, while the thinner lines are the 2.5 and 97.5 percentiles. For large T_s the means converge close to their neutral expectation of $\theta = 20$. It can be seen that a selective sweep affects k more strongly than S, and that k recovers more slowly both in mean and variance. Comparing Figure 2, a and c with b and d shows the effect of the strength of selection: stronger selection results in a more immediate decrease in the expected value of D, and a stronger reduction in S and k. Note, however, that the length of time T_d ($\approx 2\ \ln(2N)/Ns$ when $h = 0.5$) it takes the sweep to complete must be taken into

TABLE 2

Coefficients of linear approximations to a 1-β confidence interval for θ

β	n	b	c	q	r
0.01	10	0.133	−0.484	1.236	3.787
	20	0.121	−0.481	0.709	2.858
	50	0.108	−0.474	0.441	2.302
	100	0.101	−0.484	0.341	2.039
0.001	10	0.102	−0.483	1.782	6.304
	20	0.094	−0.473	0.904	4.418
	50	0.087	−0.468	0.519	3.408
	100	0.081	−0.856	0.389	3.420

$C_\beta = [bS + c, qS + r]$.

TABLE 3

Level 0.05 critical values of TAJIMA's *D* test

$n = 10$			$n = 20$			$n = 50$			$n = 100$		
S	D_L	D_U	S	D_L	D_U	S	D_L	D_U	S	D_L	D_U
0	−1.79	1.84	0	−1.78	1.97	0	−1.70	2.11	0	−1.58	2.21
1–26	−1.80	1.84	1–3	−1.82	1.97	1–22	−1.77	2.11	1–24	−1.70	2.21
27–41	−1.80	1.83	4–14	−1.83	1.97	23–31	−1.77	2.06	25–34	−1.70	2.15
42–48	−1.80	1.81	15–20	−1.84	1.97	32–41	−1.77	2.00	35–44	−1.70	2.07
49–63	−1.79	1.79	21–28	−1.84	1.96	42–50	−1.73	1.97	45–74	−1.70	2.04
64–71	−1.78	1.78	29–36	−1.84	1.90	51–73	−1.73	1.95	75–78	−1.68	2.01
72–135	−1.78	1.74	37–45	−1.84	1.88	74–155	−1.75	1.95	79–159	−1.69	2.01
			46–86	−1.84	1.87						
			87–144	−1.85	1.87						
			145–147	−1.85	1.82						
beta	−1.733	1.975	beta	−1.803	2.001	beta	−1.800	2.044	beta	−1.781	2.073

Values based on a 99% confidence interval for θ given S for several sample sizes n.

account; this is ≈0.3 and 0.003 for the weak and strong cases, respectively (shown inset in Figure 2, a and b). Thus in a and c, when $T_s < 0.3$, the selected allele has not yet been fixed at time 0 when we take the sample. The apparently neutral behavior when $T_s < 0.15$ is due to the low frequency of the selected allele in the population, and hence a low probability that they will be found in the sample. When $0.15 < T_s < 0.3$, we see the effect of a selective sweep in progress, while when $T_s > 0.3$, we have the result of a completed selective sweep. In Figure 2, b and d, on the other hand, the sweep is virtually instantaneous compared with the scale shown (though it takes 6000 generations), so the effect is seen immediately.

The power of TAJIMA's *D* test against the selective sweep alternative is shown in Figure 3: $\theta = 10$, $s = 10^{-4}$ (a); $\theta = 20$, $s = 10^{-4}$ (b); $\theta = 50$, $s = 10^{-4}$ (c); $\theta = 20$, $s = 10^{-2}$ (d). The horizontal axis is T_s as in Figure 2, but note that the scale is enlarged. The different curves are for different values of n as labeled on the graphs. Figure 3 shows that the sample size has a profound effect on the power to reject. While a sample of size 50 or 100 can give a substantial power, no significant result can be expected from a sample size of 10 in most cases. It appears that even with large sample sizes, it is only possible to detect selective sweeps that occurred in a specific window of time. For example, if $n = 100$ and $\theta = 20$, TAJIMA's test will reject with probability 90%

TABLE 4

Level 0.05 critical values of FU and LI's (1993) D^* test

$n = 10$			$n = 20$			$n = 50$			$n = 100$		
S	D_L^*	D_U^*	S	D_L^*	D_U^*	S	D_L^*	D_U^*	S	D_L^*	D_U^*
0	−2.06	1.41	0	−2.4	1.40	0	−2.57	1.48	0	−2.68	1.32
1–48	−2.08	1.42	1–2	−2.49	1.44	1–13	−2.58	1.51	1	−2.68	1.51
49–63	−2.08	1.40	3–7	−2.59	1.44	14–17	−2.59	1.51	2–4	−2.68	1.55
64–78	−2.06	1.36	8–13	−2.67	1.44	18–19	−2.61	1.51	5–24	−2.68	1.59
79–861	−2.06	1.35	14–41	−2.70	1.44	20–24	−2.71	1.51	25–44	−2.54	1.59
87–108	−2.06	1.34	42–45	−2.73	1.44	25–42	−2.72	1.51	45–49	−2.50	1.59
109–135	−2.06	1.32	46–53	−2.73	1.43	43–50	−2.76	1.51	50–52	−2.52	1.59
			54–61	−2.73	1.42	51–60	−2.76	1.50	53–58	−2.54	1.59
			62	−2.73	1.38	61–68	−2.80	1.50	59–64	−2.56	1.59
			63–84	−2.76	1.38	69–71	−2.80	1.45	65–74	−2.56	1.57
			85–86	−2.78	1.38	72–73	−2.84	1.45	75–103	−2.56	1.54
			87–102	−2.78	1.36	74–77	−2.92	1.45	104–123	−2.56	1.51
			103–111	−2.78	1.35	78–114	−2.92	1.41	124–143	−2.56	1.49
			112–135	−2.78	1.34	115–151	−2.92	1.39	144–146	−2.57	1.49
			136–144	−2.78	1.33	152–155	−2.92	1.35	147–159	−2.58	1.49
(1993)	−2.02	1.38	(1993)	−2.43	1.37	(1993)	−2.45	1.44	(1993)	−2.33	1.53

Values based on a 99% confidence interval for θ given S for several sample sizes n.

TABLE 5

Level 0.05 critical values of Fu and Li's F^* test

$n = 10$			$n = 20$			$n = 50$			$n = 100$		
S	F_L^*	F_U^*	S	F_L^*	F_U^*	S	F_L^*	F_U^*	S	F_L^*	F_U^*
0	−2.22	1.60	0	−2.54	1.65	0	−2.57	1.74	0	−2.52	1.67
1–48	−2.26	1.61	1–2	−2.62	1.67	1–6	−2.60	1.74	1–24	−2.52	1.83
49–63	−2.26	1.58	3–7	−2.69	1.67	7–19	−2.61	1.74	25–44	−2.47	1.83
64–71	−2.25	1.57	8–13	−2.74	1.67	20–24	−2.62	1.74	45–64	−2.42	1.83
72–101	−2.25	1.53	14–42	−2.76	1.67	25–41	−2.72	1.74	65–103	−2.40	1.83
102–108	−2.25	1.52	43–45	−2.78	1.67	42–50	−2.72	1.72	104–113	−2.40	1.82
109–135	−2.25	1.51	46–61	−2.78	1.62	51–60	−2.72	1.71	114–125	−2.40	1.81
			62	−2.78	1.58	61–68	−2.77	1.71	126–159	−2.43	1.81
			63–102	−2.81	1.58	69–73	−2.77	1.70			
			103–144	−2.81	1.56	74–133	−2.85	1.70			
			145–147	−2.81	1.55	134–155	−2.85	1.68			
(1993)	−2.21	1.59	(1993)	−2.57	1.61	(1993)	−2.43	1.66	(1993)	−2.30	1.73

Values based on a 99% confidence interval for θ given S for several sample sizes n.

only if the sweep (weak selection) began between $T_s = 0.18$ and $T_s = 0.27$, which, with $N = 10^6$, corresponds to between 360,000 and 540,000 generations ago. It must be emphasized that these results apply only to the particular model of sweep and the parameter values ($s = 10^{-4}$, $h = 0.5$) used in the simulation. For clarity, the graphs in Figure 3 are shown with T_s in the range 0–1. However, simulations were actually performed with T_s as large as 10. It can be seen in Figure 3 that the power drops well below the neutral expectation of 0.05 when T_s is close to one. In fact, a selective sweep reduces the power even when T_s is four or five. For these T_s, the sweep was long enough ago that new mutations have had a chance to reach intermediate frequency in the population, but polymorphism is still quite reduced. In other words, the difference between the expectations of k and S/a_n is fairly small, but the variance of that difference is still reduced well below one. This has the paradoxical result of making the test less likely to reject under the alternative than under the null hypothesis, when T_s takes on intermediate values. In other words, Tajima's D test is biased.

Figure 4 shows the power of all nine tests against the selective sweep alternative when $n = 50$ and $\theta = 20$; $s = 10^{-4}$ (a), $s = 10^{-2}$ (b). Among all the tests considered, Tajima's test showed the most power against the selective sweep for each value of n and θ we simulated. The tests T_2 and F_3 were almost as powerful as D, and were more powerful than Fu and Li's F^* and D^* tests. Although Tajima's test statistic does lack power in many cases, it appears to be the most powerful test of this class against the selective sweep alternative as modeled here, both for weak and strong selection. We cannot, of course, generalize this result to any alternative hypotheses that we did not simulate, such as other values of N, h, and s, or models of hitchhiking which allow recombination, but lacking evidence to the contrary it would be reasonable to assume the result holds at least for parameters within the range of those used in the simulations, such as s between 10^{-4} and 10^{-2}, and θ between 10 and 50.

Results of population bottleneck simulations: The results for the population bottleneck are summarized in Figures 5 and 6. Each figure represents a bottleneck lasting 0.1 (units $2N$ generations) and dropping to 1% of its original size. The horizontal axis in each case is the time T_b at which the bottleneck ended, and each data point is based on 1000 simulations. Figure 5 shows the mean and 2.5 and 97.5 percentiles of Tajima's statistic D, *vs.* T_b, for $n = 50$ and $\theta = 20$. Figure 6 shows the fraction rejected by Tajima's test for the cases $\theta = 10$ (a), $\theta = 20$ (b), $\theta = 50$ (c), and by Fu and Li's F^* test for the case $\theta = 20$ (d). The results are similar to those for the selective sweep. A bottleneck is only likely to be detected if it is very recent, and if the sample size is large. Again, Tajima's test performs the best of all the tests considered. The similarity of the results to those for a selective sweep is to be expected, since the effect on the coalescent process of the two situations is similar.

Results of population subdivision simulations: Population subdivision has an effect opposite to that of a selective sweep on the statistics being studied. A subdivided population results in a higher value of k than would be expected under neutrality, while the effect on S is less severe. Thus population subdivision tends to produce positive values of the test statistics D, F^*, and D^*. The more ancient the division, the greater this effect becomes. A plot of the median and 2.5 and 97.5 percentiles of Tajima's D against the time of separation T_m, for a sample size of 50 ($n_A = n_B = 25$) and $\theta = 10$, with $N_{AB} = N_A + N_B$, is shown in Figure 7. Power curves for all nine tests are shown in Figure 8. It can be seen

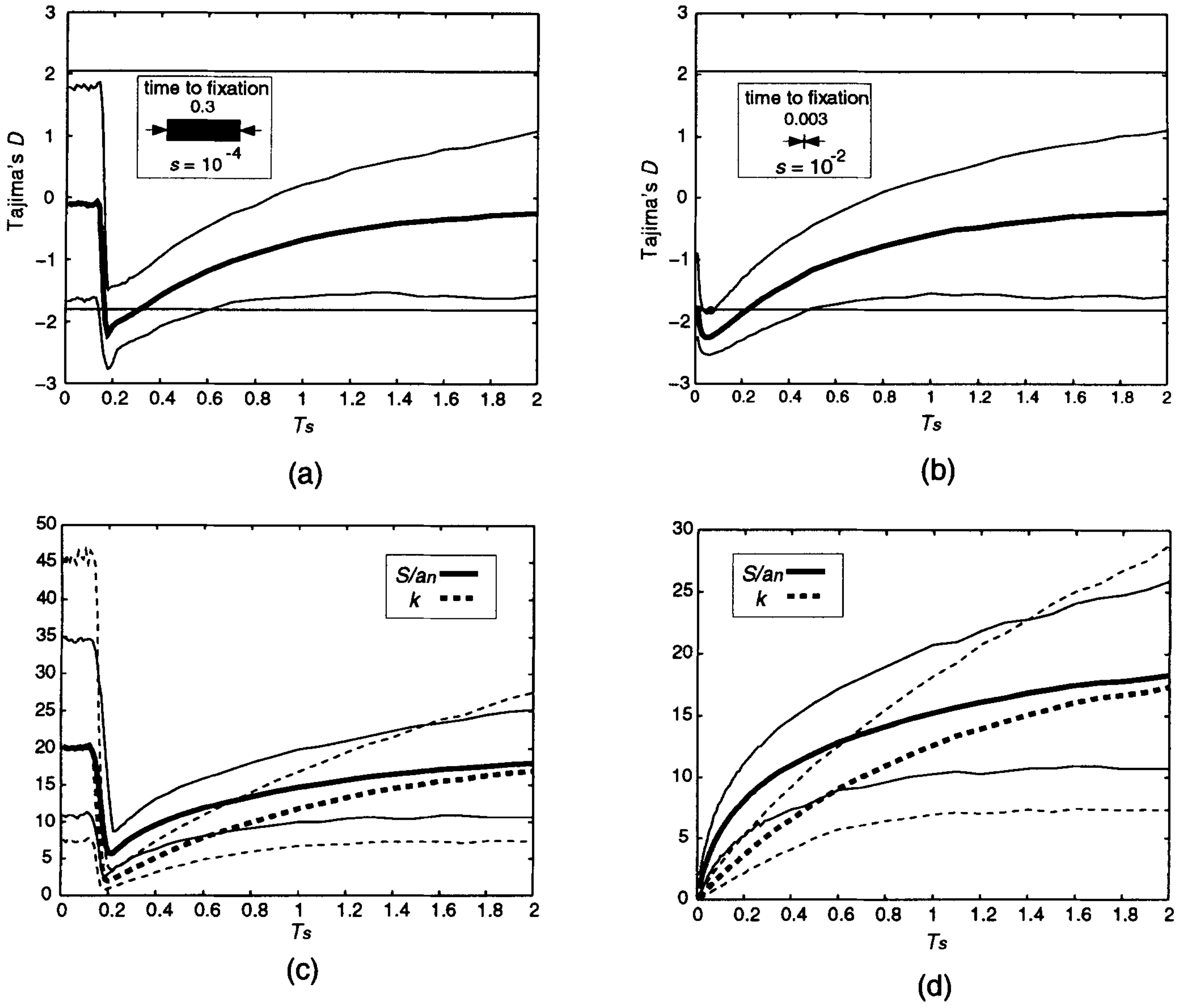

FIGURE 2.—The effect of a selective sweep on TAJIMA's D statistic, S, and k vs. the time T_s at which the sweep began. The mean and 2.5 and 97.5 percentiles of D are shown in a and b. Horizontal lines are TAJIMA's (1989a) critical values. The means and 2.5 and 97.5 percentiles of S/a_n (solid) and k (dotted) are shown in c and d. Each data point is based on 5000 simulations of a selective sweep with parameters $\theta = 20$, $n = 50$, $h = 0.5$, and $N = 10^6$: (a and c) $s = 10^{-4}$; (b and d) $s = 10^{-2}$. The length of time it takes the selected allele to reach fixation is also depicted (inset); a sweep beginning at T_s ends at the given distance to the left of T_s.

from this figure that the probability of detecting this type of population subdivision with these tests is quite small unless the division is fairly ancient. Again, TAJIMA's D is the most powerful test against this alternative, with T_2 having almost identical power to D, and FU and LI's F^* the next most powerful. The above results were given for $n_A = n_B = 25$. When we choose $n_A \neq n_B$, (e.g. $n_A = 10$, $n_B = 40$) the power is even less, with all other parameters held fixed (results not shown).

Some comments on sample size: We have shown that sampling a greater number of individuals increases the power of the test. But, sampling longer sequences (effectively increasing θ) should also increase the power. Which is better: longer sequences or more individuals? To answer this question, we must assign a relative cost to these two options. Let us assume that the cost per nucleotide sequenced is the same whether that nucleotide comes from a new individual or from extending the sequenced region. This ignores costs associated with both the acquisition and preparation of a new individual and the analysis of longer regions. Further suppose that the per-locus mutation rate is proportional to the length of the sequence, so that doubling the number of bases doubles θ. Under these assumptions, the cost is proportional to the product of n and θ. Therefore, we compare power curves where the product of n and θ is the same.

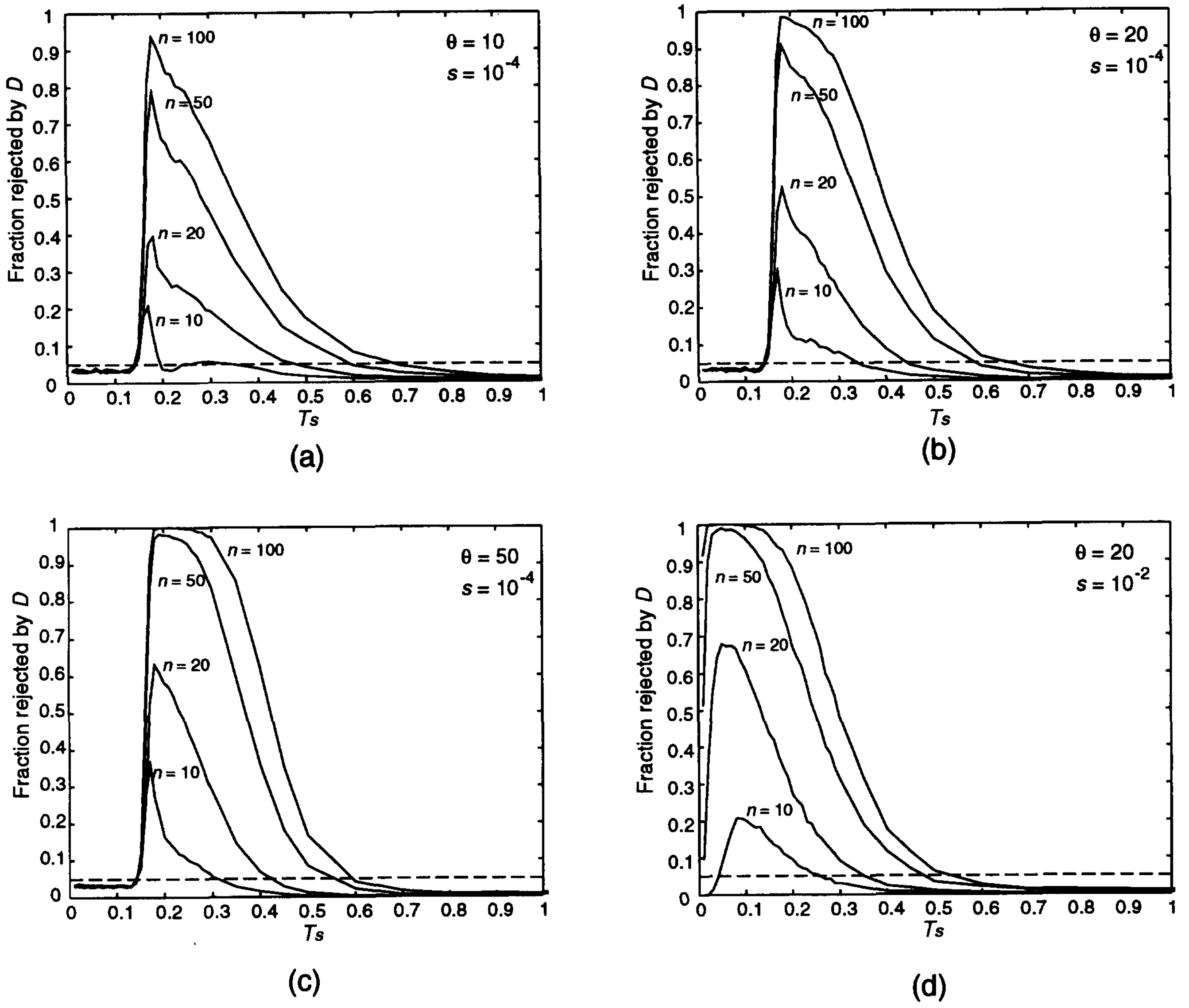

FIGURE 3.—Power of TAJIMA's D against a selective sweep *vs.* the time T_s at which the sweep began for $n = 10$, 20, 50, and 100: (a) $\theta = 10$; (b) $\theta = 20$; (c) $\theta = 50$; (d) $\theta = 20$. Each data point is based on 5000 simulations of a sweep with parameters $h = 0.5$, $s = 10^{-4}$, and $N = 10^6$, except (d), which uses $s = 10^{-2}$.

In Figure 9, we show the power of TAJIMA's test against a selective sweep (weak selection) for the product $n\theta = 200$, 500, and 1000. Note that in this context, increasing θ means increasing the size of the region examined (and thus μ) for a given N. These results show that, against this particular selective sweep alternative, it is better to sequence more individuals than more sites, so long as the number of sites is not too small. For example, against the selective sweep alternative as modeled here, TAJIMA's test is always more powerful when $n = 20$ and $\theta = 10$ than when $n = 10$ and $\theta = 20$. Similar results hold for other tests.

DISCUSSION

The new method of calculating critical values for the class of tests presented here allows us to eliminate from the null hypothesis the requirement (FU and LI 1993) that θ is between two and 20. If the true value of θ for a studied locus is indeed in that range, there is very little difference between the two methods. However, our method has the advantage that rejection cannot be explained by a too-small or too-large θ. If FU and LI's published critical values are used, it should be with the understanding that the true level of the test should have added to it the probability that θ is not in that range. For TAJIMA's test (1989a), our critical values are a clear improvement over the beta distribution method. In many cases, TAJIMA's published values are too conservative, with the result that rejection is almost impossible. Our new critical values result in a more powerful test.

As an alternative method of examining the behavior of D when θ is unknown, other authors (HUDSON 1993;

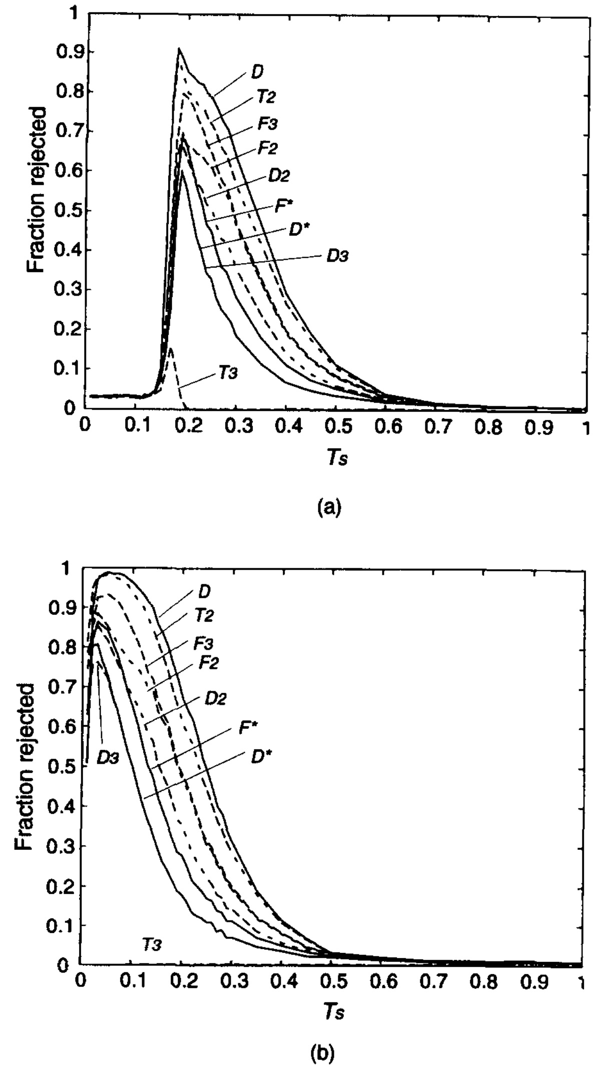

FIGURE 4.—Power of all nine statistical tests against a selective sweep *vs.* the time T_s at which the sweep began for $n = 50$ and $\theta = 20$. Each data point is based on 5000 simulations of a sweep with parameters $h = 0.5$, and $N = 10^6$: (a) $s = 10^{-4}$; (b) $s = 10^{-2}$.

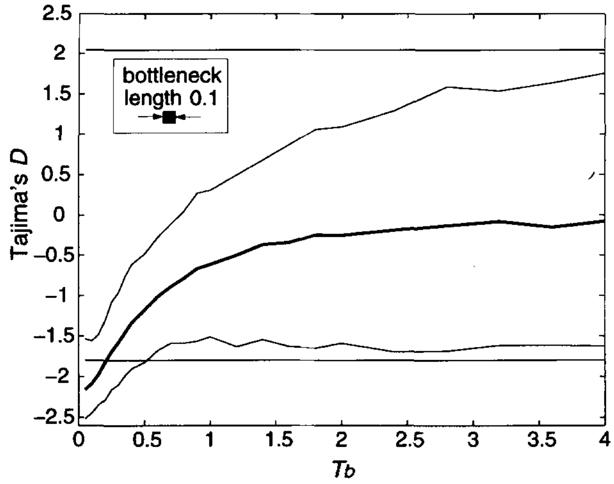

FIGURE 5.—The effect of a population bottleneck on TAJIMA's D statistic. Shown are the mean (thick line) and 2.5 and 97.5 percentiles (thinner lines) *vs.* the time T_b at which the bottleneck ended. Horizontal lines are approximate critical values for rejection. Each point is based on 1000 simulations of a population bottleneck with parameters $\theta = 10$, $n = 50$, $f = 0.01$ and $l = 0.1$.

BRAVERMAN *et al.* 1995) have suggested sampling from the conditional distribution of D given S, where S is obtained from the data set to be tested. With S fixed, D is simply a linear transformation of k and may therefore have a smaller variance under neutrality because the contribution of S to the variance is eliminated. Both methods choose a genealogy at random, but their method fixes S for all genealogies, whereas we fix θ and from this generate a value of S based on the total time in the genealogy. The contrast between the two methods of generating S is most evident when simulating data from alternative hypotheses to estimate power. The two methods represent two different views of the power of a test: as a function of the parameter θ, and as a function of the statistic S. We investigate the behavior of D after a selective sweep, with several different, but fixed, mutation rates. Their method would examine the effect of a selective sweep after which a fixed number of mutations has occurred.

Among all the tests considered, TAJIMA's test (with the new critical values) was the most powerful against the specific alternatives we simulated. Certainly we cannot extrapolate from this to say it is more powerful against all possible alternatives and parameter values. Because the chance of spurious rejection increases with the number of tests performed, we want to perform only the test with the greatest chance of rejection. In the absence of other evidence, that would appear to be TAJIMA's test. The new test statistics described above do not perform as well as TAJIMA's test, although they do have more power than FU and LI's tests in many cases. Thus we do not recommend their use. BRAVERMAN *et al.* (1995) also found TAJIMA's D (conditional on S) to be more powerful than FU and LI's D^* against the alternative of recurrent hitchhiking with recombination (KAPLAN *et al.* 1989) and very recent, strong selection.

Our results indicate that sample sizes of at least 50 are typically necessary to achieve any reasonable power. Many sample sizes for sequence data seen in the literature are much smaller than this. However, even for large sample sizes, the probability of detecting a selective sweep that is not recent is quite small.

We have shown that the expected value of TAJIMA's D is in fact negative at linked neutral sites after the selective fixation of an advantageous mutation in a model with no recombination. This agrees both with theoretical prediction and with the findings of BRAVERMAN *et al.* (1995) for a different hitchhiking model. It is also apparent that the ability to detect the selective

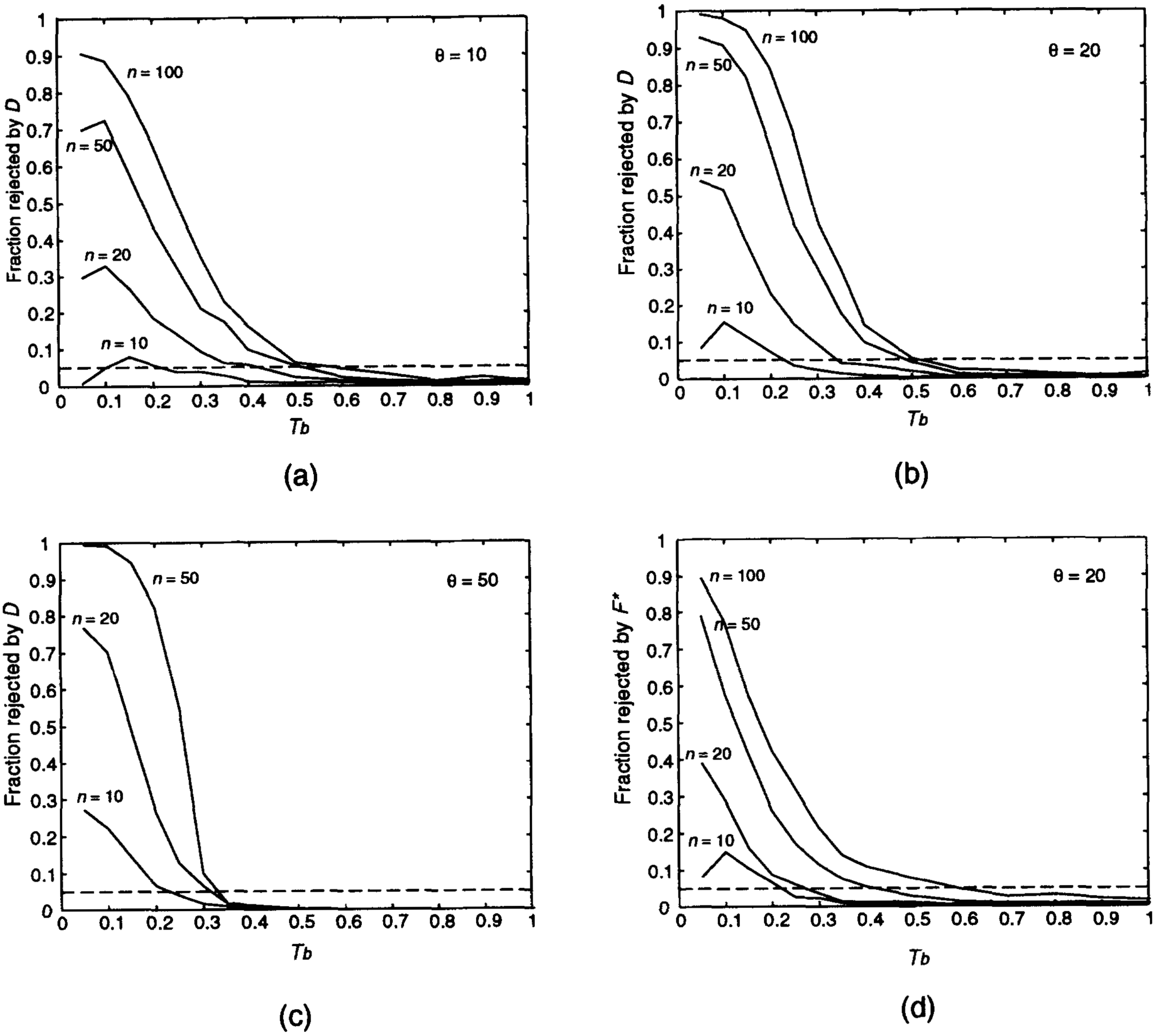

FIGURE 6.—Power of statistical tests against a population bottleneck *vs.* the time T_b at which the bottleneck ended for n = 10, 20, 50, and 100. The tests are (a) D, θ = 10; (b) D, θ = 20; (c) D, θ = 50; (d) F^*, θ = 20. Each data point is based on 1000 simulations of a population bottleneck with parameters f = 0.01 and l = 0.1.

sweep by either TAJIMA's (1989a) or FU and LI's (1993) test statistics is strongly influenced by the strength of selection and by the amount of time since the selective sweep occurred. With an effective population size of 10^6, selective sweeps of codominant mutations with a selective advantage of 10^{-4} result in distributions of variation that are unlikely to be found incompatible with a neutral model using these tests. Increasing the selective advantage 100-fold to 10^{-2} leads to a certain increase in the power of available tests. Nonetheless, there exists a defined window over which the tests have reasonable (say, >90%) statistical power to reject the neutral model. For strong selection, this window appears to be from roughly 40,000 to 280,000 generations when n = 50 and θ = 20 (Figure 3d). More recent sweeps are undetectable because there has been too little time for sufficient new variants to arise, and sweeps too distant in the past are hidden by the accumulation of new neutral variants.

These results suggest that while recent, strong, selective sweeps are likely to produce a significant TAJIMA's D if the sample size is large enough, in general the selective sweep explanation for reduced levels of variation cannot be ruled out based solely on the observation of a nonsignificant TAJIMA's test. Furthermore, nonsignificant values of TAJIMA's D are consistent with less recent sweeps and weaker selection, as well as with the neutral model. For example, in a sample of *Drosophila melanogaster* from North Carolina (AGUADÉ *et al.* 1994) with n = 50, D = −0.40 and S = 17 were observed for

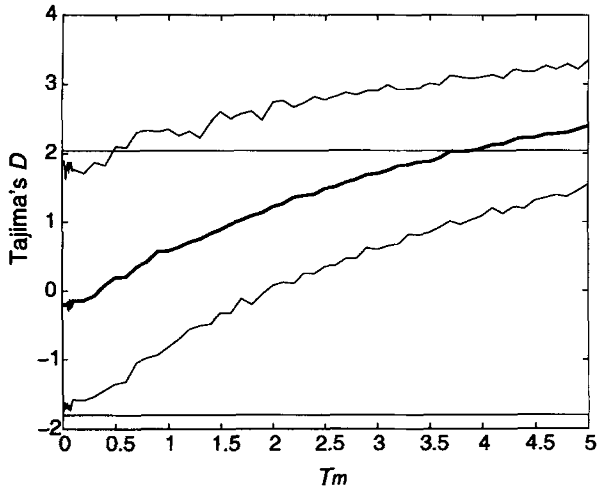

FIGURE 7.—The effect of a subdivided population on TAJIMA's D: the median (thick line) and 2.5 and 97.5 percentiles (thinner lines) *vs.* time of separation T_m. Horizontal lines are approximate critical values for rejection. Each point is based on 1000 simulations of population subdivision with parameters θ = 10, n = 50, n_A = n_B = 25.

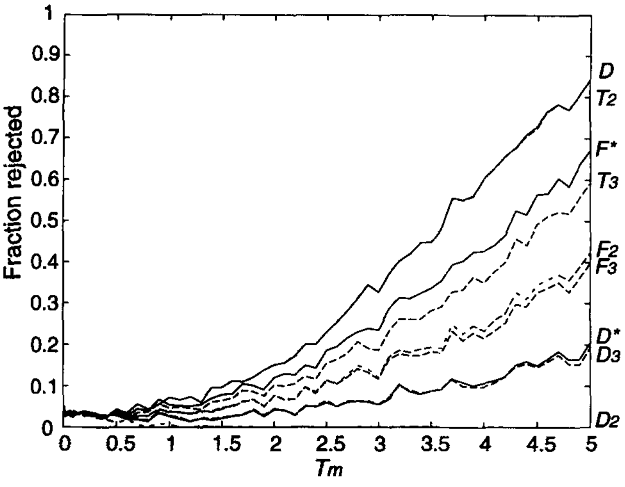

FIGURE 8.—Power of all nine tests against population subdivision. Fraction rejected *vs.* time of separation T_m. Based on 1000 simulations of population subdivision with n = 50, n_A = n_B = 25, and θ = 10.

a 1955-bp region of the $su(w^a)$ locus. If we were to take as a null hypothesis a selective sweep with parameters n = 50, s = 10^{-2}, θ = 10, N = 10^6, h = 0.5 and T_s = 0.45, our 95% confidence interval (estimated from simulated data) for D would be [−1.925, −0.289], and for S would be [16, 37]. Thus for these data we would not be able to reject that particular sweep hypothesis at the 5% level. We do not claim that those data were the result of such a sweep, only that the hypothesis cannot be ruled out on the basis of TAJIMA's D. Such a point hypothesis would not be particularly useful; in practice one would want to include intervals of the parameters involved, as we have done for θ in the neutral model. The extent to which weaker or more distant selection, or selection with recombination could result in the observed patterns of data needs further examination. Situations where negative TAJIMA's D have been observed together with reduced variation do appear consistent with a simple selective sweep model (MARTÍN-CAMPOS *et al.* 1992). The generally low level of power of the test statistics studied here indicates that other means to distinguish between selective sweeps and other hypotheses, such as background selection (CHARLESWORTH *et al.* 1993; CHARLESWORTH 1994; HUDSON and KAPLAN 1994), should be sought before firm conclusions are drawn. To do this, any test will have to take into account more information from the data than just differences between the three summary statistics k, S, and η_s. The apparent contrast between predictions for X-linked *vs.* autosomal gene variation is but one possibility (AQUADRO *et al.* 1994).

The population bottleneck simulations have shown that, for the size and length of bottleneck studied, the mean and variance of TAJIMA's D and the other statistics are decreased well below their neutral expectations. TAJIMA's D is more powerful against such a bottleneck than the other tests in this class, but sample sizes of at least 50–100 (depending on the mutation rate) are necessary for reasonable power. In these simulations we chose a bottleneck of length l = 0.1, a relatively long time (if N= 10^6 this is 200,000 generations). In a bottleneck of such length, most variation may be eliminated by the time the bottleneck ends so that subsequent variants are rare. In bottlenecks of shorter duration, results may be very different, because polymorphism that survives the bottleneck may quickly reach intermediate frequency. Thus it is important not to extrapolate these simulation results to parameter ranges we did not consider.

Simulating a subdivided population allowed us to explore the way in which TAJIMA's D and the other tests were affected by an excess of intermediate-frequency variants, which tend to make D positive. Even against this extreme no-migration model, the power of TAJIMA's D was very low unless the two subpopulations had been separated for at least 10 Ngenerations. While we do not suppose that the subdivision model used is particularly realistic over such a long time scale, the results of these simulations suggest that coalescent trees must be quite strongly skewed to produce data with a significant positive D. TAJIMA's D and the new statistic T_2 were equally powerful against this alternative and were more powerful than the other tests studied. It is therefore suggested that, of the nine tests studied in this paper, only TAJIMA's D test be performed.

This approach to estimating the power of statistical tests should prove useful in investigating many other types of alternatives and statistical tests. For example, it would be useful to know which, if any, existing tests are able to detect background selection. Tests that use

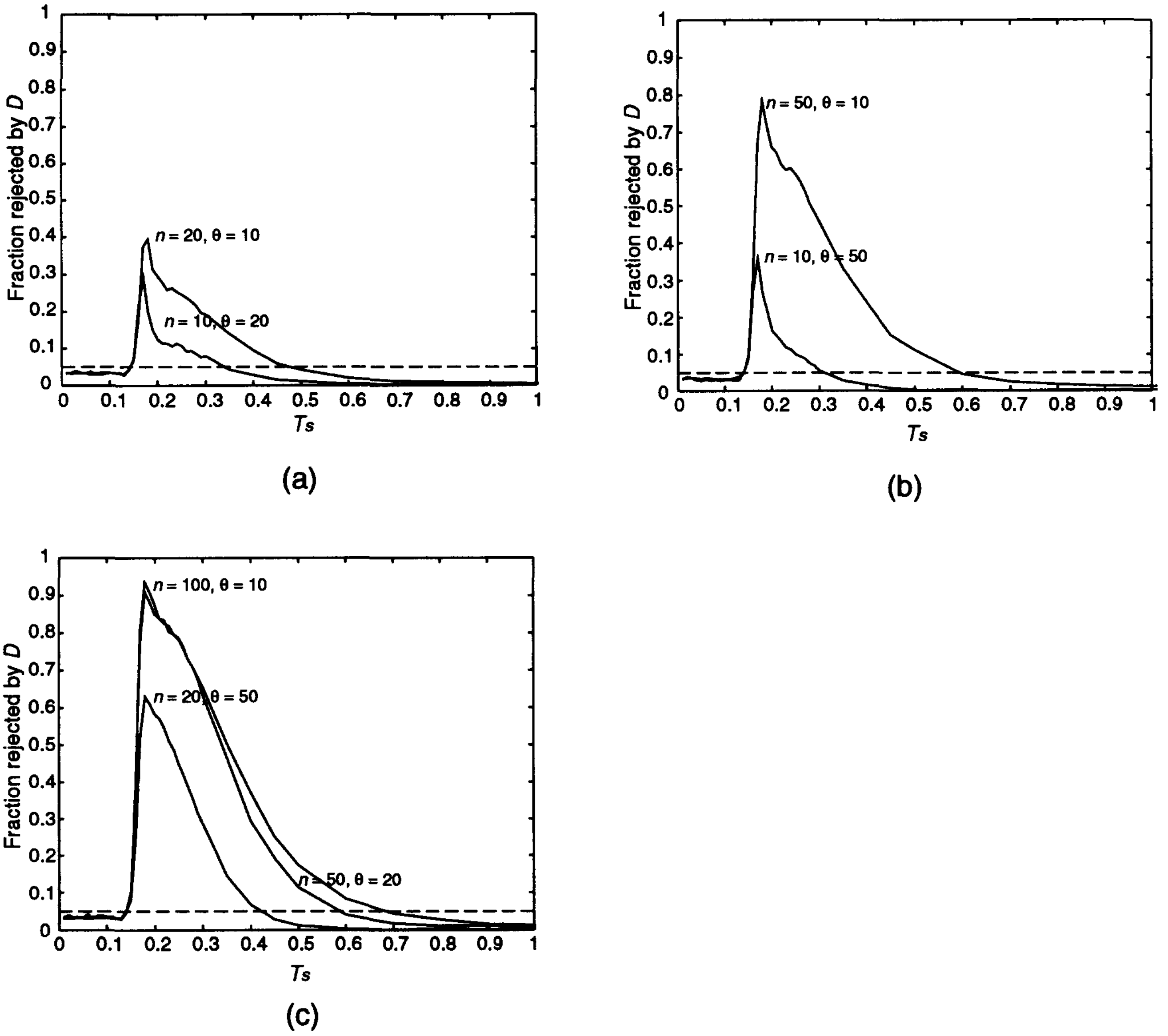

FIGURE 9.—Power of TAJIMA's test against a selective sweep *vs.* time T_s at which the sweep began. Each plot is for a constant value of the product of n and θ: (a) $n\theta = 200$; (b) $n\theta = 500$; (c) $n\theta = 1000$. Each data point is based on 1000 simulations of a selective sweep with parameters $h = 0.5$, $s = 10^{-4}$, and $N = 10^6$.

more information from the data, such as outgroups, may be more powerful than the tests studied here. We (with M. J. FORD) are currently undertaking an similar analysis of the properties of the HKA test (HUDSON *et al.* 1987).

We thank R. L. BERGER, R. R. HUDSON, one anonymous reviewer, and members of the AQUADRO Lab: D. J. BEGUN, M. J. FORD, M. T. HAMBLIN, M. W. NACHMAN and K. S. PHILLIPS, for helpful suggestions. Preliminary discussions on this project involved the above lab members as well as D. GROVE, C. MCCULLOCH, S. SCHWAGER, and J. THORNE. Some algorithms used in computer programs were based on code originally written by R. R. HUDSON. This work was supported in part by National Institute of General Medical Sciences grant GM-36431 to C.F.A., Department of Energy grant DEFG0293ER61567 to G.A.C., and National Science Foundation training grant BIR-9113307 to J. GUCKENHEIMER.

LITERATURE CITED

AGUADÉ, M., N. MIYASHITA and C. H. LANGLEY, 1989 Reduced variation in the *yellow-achaete-scute* region in natural populations of *Drosophila melanogaster*. Genetics **122:** 607–615.

AGUADÉ, M., W. MEYERS, A. D. LONG and C. H. LANGLEY, 1994 Reduced DNA sequence polymorphism in the *su(s)* and *su(w^a)* regions of *Drosophila melanogaster* as revealed by SSCP and stratified DNA sequencing. Proc. Natl. Acad. Sci. USA **91:** 4658–4662.

AQUADRO, C. F., D. J. BEGUN and E. C. KINDAHL, 1994 Selection, recombination, and DNA polymorphism in *Drosophila*, pp. 46–56 in *Non-Neutral Evolution: Theories and Molecular Data*, edited by B. GOLDING. Chapman and Hall, New York.

BEGUN, D. J., and C. F. AQUADRO, 1991 Molecular population genetics of the distal portion of the X chromosome in Drosophila: evidence for genetic hitchhiking of the *yellow-achaete* region. Genetics **129:** 1147–1158.

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

Berger, R. L., and D. D. Boos, 1994 *P*-values maximized over a confidence set for a nuisance parameter. J. Am. Stat. Assoc. **89:** 1012–1016.

Berry, A. J., J. W. Ajioka and M. Kreitman, 1991 Lack of polymorphism in the Drosophila fourth chromosome resulting from selection. Genetics **129:** 1111–1117.

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. in press.

Char, B. W., K. O. Geddes, G. H. Gonnet, B. L. Leong, M. B. Monagan *et al.*, 1991 *Maple V Library Reference Manual.* Springer-Verlag.

Charlesworth, B., 1994 The effect of background selection against deleterious mutations on weakly-selected, linked variants. Genet. Res. **63:** 213–227.

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

Fu, Y.-X., and W.-H. Li, 1993 Statistical tests of neutrality of mutations. Genetics **133:** 693–709.

Griffiths, R. C., and S. Tavaré, 1994 Ancestral inference in population genetics. Stat. Sci. **9:** 307–319.

Hudson, R. R., 1990 Gene genealogies and the coalescent process. Oxf. Surv. Evol. Biol. **7:** 1–44.

Hudson, R. R., 1993 The how and why of generating gene genealogies, pp. 23–36 in *Mechanisms of Molecular Evolution: Introduction to Molecular Paleopopulation Biology,* edited by N. Takahata and A. G. Clark. Sinauer Associates, Inc., Sunderland, MA.

Hudson, R. R., and N. L. Kaplan, 1994 Gene trees with background selection, pp. 140–153 in *Non-Neutral Evolution: Theories and Molecular Data,* edited by B. Golding. Chapman and Hall, New York.

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

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

Martín-Campos, J. M., J. M. Cameron, N. Miyashita and M. Aguadé, 1992 Intraspecific and interspecific variation at the *y-ac-sc* region of *Drosophila simulans* and *Drosophila melanogaster.* Genetics **130:** 805–816.

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

Stephan, W., and C. H. Langley, 1989 Molecular genetic variation in the centromeric region of the X chromosome in three *Drosophila ananassae* populations. I. Contrasts between the *vermillion* and *forked* loci. Genetics **121:** 89–99.

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

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

Tajima, F., 1989b The effect of change in population size on DNA polymorphism. Genetics **123:** 597–601.

Tajima, F., 1993 Measurement of DNA polymorphism, pp. 37–59, in *Mechanisms of Molecular Evolution: Introduction to Molecular Paleopopulation Biology,* edited by N. Takahata and A. G. Clark. Sinauer Associates, Inc., Sunderland, MA.

Tavaré, S., 1984 Line-of-descent and genealogical processes and their applications in population genetics models. Theor. Popul. Biol. **26:** 119–164.

Watterson, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256–276.

Communicating editor: G. B. Golding

APPENDIX

The following are the coefficients of Tajima's and Fu and Li's tests.

$$v_T = \left[\frac{2(n^2 + n + 3)}{9n(n-1)} - \frac{n+2}{a_n n} + \frac{b_n}{a_n^2}\right] \Big/ (a_n^2 + b_n) \quad \text{(A1)}$$

$$u_T = \left[\left(\frac{n+1}{3(n-1)} - \frac{1}{a_n}\right) \Big/ a_n\right] - v_T \quad \text{(A2)}$$

$$v_{D^*} = \left[\frac{b_n}{a_n^2} - \frac{2}{n}\left(1 + \frac{1}{a_n} - a_n + \frac{a_n}{n}\right) - \frac{1}{n^2}\right] \Big/ (a_n^2 + b_n) \quad \text{(A3)}$$

$$u_{D^*} = \left[\left(\frac{(n-1)}{n} - \frac{1}{a_n}\right) \Big/ a_n\right] - v_{D^*} \quad \text{(A4)}$$

$$v_{F^*} = \left[\frac{2n^3 + 110n^2 - 255n + 153}{9n^2(n-1)} + \frac{2(n-1)a_n}{n^2} - \frac{8b_n}{n}\right] \Big/ (a_n^2 + b_n) \quad \text{(A5)}$$

$$u_{F^*} = \left\{\left[\frac{4n^2 + 19n + 3 - 12(n+1)a_{n+1}}{3n(n-1)}\right] \Big/ a_n\right\} - v_{F^*} \quad \text{(A6)}$$

The following are the coefficients of the new statistical tests described in the materials and methods section.

$$v_{T_2} = \frac{\left[2(n^2 + n + 3) - \frac{9(n-1)(n+2)}{a_n} + \frac{9n(n-1)b_n}{a_n^2}\right]}{11n^2 - 7n + 6} \quad \text{(A7)}$$

$$u_{T_2} = \frac{n+1}{3(n-1)} - 1/a_n - \frac{n+1}{3(n-1)}\, v_{T_2} \quad \text{(A8)}$$

$$v_{T_3} = \frac{\left[\frac{2(n^2 + n + 3)}{9n} - \frac{(n+2)(n-1)}{na_n} + \frac{b_n(n-1)}{a_n^2}\right]}{(2a_n + n + 1)} \quad \text{(A9)}$$

$$u_{T_3} = \frac{n+1}{3n} - \frac{n-1}{na_n} - v_{T_3} \quad \text{(A10)}$$

$$v_{D_2^*} = \frac{9(n-1)}{11n^2 - 7n + 6} \times \left[\frac{nb_n}{a_n^2} - 2\left(1 + \frac{1}{a_n} - a_n + \frac{a_n}{n}\right) - \frac{1}{n}\right] \quad \text{(A11)}$$

$$u_{D_2^*} = \frac{n-1}{n} - \frac{1}{a_n} - \frac{(n+1)}{3(n-1)}\, v_{D_2^*} \quad \text{(A12)}$$

$$v_{D_3^*} = \frac{n-1}{2a_n + n + 1} \times \left[\frac{b_n}{a_n^2} - \frac{2}{n}\left(1 + \frac{1}{a_n} - a_n + \frac{a_n}{n}\right) - \frac{1}{n^2}\right] \quad \text{(A13)}$$

$$u_{D_3^*} = \frac{(n-1)^2}{n^2} - \frac{n-1}{na_n} - v_{D_3^*} \quad \text{(A14)}$$

$$v_{F_2^*} = \frac{[2n^3 + 110n^2 - 255n + 153 + 18(n-1)^2 a_n - 72n(n-1)b_n]}{n(11n^2 - 7n + 6)} \quad \text{(A15)}$$

$$u_{F_2^*} = \frac{4n^2 + 19n + 3 - 12(n+1)a_{n+1}}{3n(n-1)} - \frac{n+1}{3(n-1)} v_{F_2^*} \quad \text{(A16)}$$

$$v_{F_3^*} = \frac{\left[\frac{2n^3 + 110n^2 - 255n + 153}{9n^2} + \frac{2(n-1)^2 a_n}{n^2} - \frac{8b_n(n-1)}{n}\right]}{2a_n + n + 1} \quad \text{(A17)}$$

$$u_{F_3^*} = \frac{4n^2 + 19n + 3 - 12(n+1)a_{n+1}}{3n^2} - v_{F_3^*} \quad \text{(A18)}$$