Statistical Tests of Neutrality of Mutations Against Population Growth, Hitchhiking and Background Selection

Yun-Xin Fu

Human Genetics Center, University of Texas, Houston, Texas 77225

Manuscript received August 28, 1996
Accepted for publication June 19, 1997

ABSTRACT

The main purpose of this article is to present several new statistical tests of neutrality of mutations against a class of alternative models, under which DNA polymorphisms tend to exhibit excesses of rare alleles or young mutations. Another purpose is to study the powers of existing and newly developed tests and to examine the detailed pattern of polymorphisms under population growth, genetic hitchhiking and background selection. It is found that the polymorphic patterns in a DNA sample under logistic population growth and genetic hitchhiking are very similar and that one of the newly developed tests, F_S, is considerably more powerful than existing tests for rejecting the hypothesis of neutrality of mutations. Background selection gives rise to quite different polymorphic patterns than does logistic population growth or genetic hitchhiking, although all of them show excesses of rare alleles or young mutations. We show that Fu and Li's tests are among the most powerful tests against background selection. Implications of these results are discussed.

WHETHER the observed pattern of polymorphism in a set of DNA sequences is consistent with a neutral model of evolution is of great interest to the study of evolution. Several statistical tests (for example, WATTERSON 1977; TAJIMA 1989; FU and LI 1993; FU 1996) are available for testing, for a sample of DNA sequences from a population, whether the polymorphism can be explained by the neutral Wright-Fisher model. That is, the population evolves according to the Wright-Fisher model and all mutations are selectively neutral. These tests are often referred to as tests of neutrality (*e.g.*, TAJIMA 1989; FU and LI 1993). A statistical test is useful only if it has some chance of rejecting the neutral Wright-Fisher model when it is false. However, since there are many factors or natural forces that can play important roles in the evolution of a population, it is unlikely that one statistical test will be powerful enough to detect all kinds of evolutionary forces that may affect the pattern of polymorphism. It is therefore useful to develop a number of statistical tests, each being the most powerful one for detecting a class of departures from the neutral model.

A mutation that results in a polymorphic site can be regarded as old if it happened a long time ago, *i.e.*, at a time close to the generation in which the most recent common ancestor (MRCA) of the sequences lived, and can be regarded as young if it happened recently. It is thus convenient to classify various population genetics models into two major groups according to their tendencies of having more old or more young mutations. The first group consists of those models that, when compared with the neutral model, often exhibit excesses of old mutations or reductions of young mutations, or both. The second group consists of those models that often exhibit excesses of young mutations or reductions of old mutations, or both. Since a recent mutant is most likely to be present in a small number of individuals, a model in the latter group often results in an excess of the number of rare alleles, *i.e.*, alleles at low frequencies.

Recently, I (FU 1996) have developed several new statistical tests that are overall more powerful than existing tests for detecting the presence of the evolutionary forces described by the first group of population genetics models, which includes population subdivision, population shrinkage and over-dominance selection. In this article I will present several new statistical tests for detecting the presence of the evolutionary forces described by the second group of population genetics models, including population growth, genetic hitchhiking and background selection. We shall use simulations under these models to examine in detail the patterns of polymorphism and to study the powers of both new and existing statistical tests. We shall show that one of the statistical tests developed in this paper is considerably more powerful than existing tests for detecting population growth and genetic hitchhiking, and that FU and LI (1993)'s tests are among the most powerful tests for detecting the presence of background selection.

CONSTRUCTING STATISTICAL TESTS

All the statistical tests discussed in this paper except one are dependent on an essential parameter θ, which

Corresponding author: Yun-Xin Fu, Human Genetics Center, University of Texas at Houston, 6901 Bertner Ave., Houston, TX 77030. E-mail: fu@hgc.sph.uth.tmc.edu

is defined as $4N\mu$ for an autosomal locus, and $2N\mu$ for haploid, such as mitochondria or Y chromosome, where N is the effective population size and μ is the mutation rate per sequence per generation.

Test F_S: Let $p(k|\theta)$ be the probability of having k alleles in a sample of n sequences, given the value of θ. For a sample with k_0 alleles and the mean number of nucleotide differences between two sequences equal to $\hat{\theta}_\pi$, we define S' to be the probability of having no fewer than k_0 alleles in a random sample provided that $\theta = \pi$. Then

$$S' = p(k_0 \leq k|\theta = \hat{\theta}_\pi) = \sum_{k \geq k_0} \frac{|S_k| \hat{\theta}_\pi^k}{S_n(\hat{\theta}_\pi)},$$

where $S_n(\hat{\theta}_\pi) = \hat{\theta}_\pi(\hat{\theta}_\pi - 1) \cdots (\hat{\theta}_\pi - n + 1)$ and S_k is the coefficient of $\hat{\theta}_\pi^k$ in S_n (Ewens 1972; Karlin and McGregor 1972). It should be noted that S' is the opposite of Strobeck's statistic S (Strobeck 1987; Fu 1996), which is the probability of having k_0 or fewer alleles in a sample. In a sample with excess of recent mutations, θ estimated by $\hat{\theta}_\pi$ is likely to be smaller than that based on the number of alleles, therefore, S' can give a good indication whether there are too many recent mutations. Although S' can be used directly as a test statistic, it is not convenient to obtain its critical points because they are often too close to zero, as in the case of Strobeck's S (Fu 1996). I will instead use the logistic of S' as a test statistic, namely

$$F_S = \ln\left(\frac{S'}{1 - S'}\right). \tag{1}$$

Since F_S tends to be negative when there is an excess of recent mutations (therefore an excess of rare alleles), a large negative value of F_S will be taken as evidence against the neutrality of mutations. In other words, a one sided-test will be used.

Tests $F(r, r')$ and $F'(r, r')$: Segregating sites and mutations that result in segregating sites can be classified into a number of types. We define a segregating site as type i if the two segregating nucleotides at the site are present in i and $n - i (i \leq n - i)$ sequences, respectively, where n is the sample size, and a mutation that results in a segregating site as type i if exactly i sequences in the sample carry the mutant nucleotide (see Fu 1994b and 1995 for details).

Let $\eta_i (i \leq n - i)$ be the number of segregating sites of type i and ξ_i be the number of mutations of type i. Then the expectations of ξ_i and η_i are (Fu 1995)

$$E(\xi_i) = \alpha_i \theta$$

$$E(\eta_i) = \beta_i \theta,$$

where

$$\alpha_i = \frac{1}{i} \tag{2}$$

$$\beta_i = \begin{cases} \frac{1}{i} + \frac{1}{n-i}, & i \neq n - i \\ \frac{1}{i}. & i = n - i. \end{cases} \tag{3}$$

Since the variances and covariances of η_i and ξ_i are also known (Fu 1995), the mean and variance of any linear function of η_i's or ξ_i's can then be computed, so can the covariance between any pair of linear functions of η_i's or ξ_i's. Therefore, one can construct a statistical test from any pair of linear functions L_1 and L_2 of η_i or ξ_i as

$$\frac{L_1 - L_2}{\sqrt{\text{Var}\ (L_1 - L_2)}}. \tag{4}$$

However, it is better to impose the condition $E(L_1) = E(L_2) = \theta$ so that the expectation and variance of the test statistic are approximately zero and one, respectively under the assumption of neutrality of mutations.

Consider linear functions of the forms

$$L(r) = c_\alpha^{-1} \sum_i \alpha_i^r \xi_i' \tag{5}$$

$$L'(r) = c_\beta^{-1} \sum_i \beta_i^r \eta_i', \tag{6}$$

where α_i and β_i are given by (2) and (3), respectively, $\xi_i' = \xi_i/\alpha_i$, $c_\alpha = \Sigma_i \alpha_i^r$, $\eta_i' = \eta_i/\beta_i$ and $c_\beta = \Sigma_i \beta_i^r$. Because $E(\xi_i') = E(\eta_i') = \theta$ and $E[L(r)] = E[L'(r)] = \theta$, $L(r)$ and $L'(r)$ are estimators of θ that are weighted averages, respectively, of $n - 1$ and $[n/2]$ unbiased estimators of θ where $[n/2]$ is the largest integer that is not larger than $n/2$; the value of r determines the relative contributions of ξ_i' and η_i'.

The linear forms $L(r)$ and $L'(r)$ are generalization of several well-known quantities. For example, Watterson's estimator θ_W (Watterson 1975) of θ is given by

$$\theta_W = \left(\sum_{i=1}^{n-1} \frac{1}{i}\right)^{-1} \sum_i \eta_i = \left(\sum_{i=1}^{n-1} \frac{1}{i}\right)^{-1} \sum_i \xi_i$$

$$= L(1) = L'(1).$$

Another example is the mean number of nucleotide differences between two sequences, known as Tajima's estimator θ_π of θ (Tajima 1983) that can be written as

$$\frac{2}{n(n-1)} \sum_i i(n - i)\xi_i.$$

Therefore, it is easy to show that

$$L'(0) = \begin{cases} \theta_\pi, & n \text{ is odd,} \\ \frac{n-1}{n}\theta_\pi + \frac{1}{2}\eta_{(n/2)}, & n \text{ is even.} \end{cases}$$

Because $\alpha_1 > \alpha_{i+1} > 0$ and $\beta_i > \beta_{i+1} > 0$, it follows that the larger the value of r in $L(r)$ is, the more weight is given to ξ_i' than to ξ_{i+1}', and similarly the larger the value of r in $L'(r)$ is, the more weight is given to η_i'

than to η'_{i+1}. On the other hand, a negative r in $L(r)$ and $L'(r)$ gives more weight to ξ'_{i+1} and η'_{i+1} than to ξ'_i and η'_i. In the extreme, we have

$$\xi_1 = L(\infty)$$

$$\eta_1 = \frac{n+1}{n} L'(\infty).$$

For a pair of values of r and r' ($r < r'$), we define tests $F(r, r')$ and $F'(r, r)$ as

$$F(r, r') = \frac{L(r) - L(r')}{\sqrt{\text{Var}\,(L(r) - L(r'))}} \quad (7)$$

$$F'(r, r') = \frac{L'(r) - L'(r')}{\sqrt{\text{Var}\,(L'(r) - L'(r'))}}. \quad (8)$$

To compute the values of $F(r, r')$ and $F'(r, r')$, an estimate of θ is required for substituting the θ in Var $(L(r) - L(r'))$ and Var $(L'(r) - L(r'))$, both being of the form $a\theta + b\theta^2$. Unless stated otherwise, Watterson's estimate $\hat{\theta}_W$ is assumed to be the substitute for θ. Because the purpose of these tests is to detect departures characterized by an excess of the number of rare alleles and a reduction of the number of common alleles, which tends to give rise negative values for these tests, large negative values are taken as evidence against the neutrality model. That is, one-sited test will be used.

It follows from the above analysis that the tests D, D^* and F^* by Fu and Li (1993) are equivalent to

$$D = F(1, \infty), \quad (9)$$

$$D^* = F'(1, \infty), \quad (10)$$

$$F^* = F'(0, \infty). \quad (11)$$

The test F by Fu and Li (1993) can be written as

$$F = \frac{L'(0) - L(\infty)}{\sqrt{\text{Var}\,(L'(0) - L(\infty)}}. \quad (12)$$

Tajima's test is for all practical purposes equivalent to

$$T = F'(0, 1). \quad (13)$$

Watterson's test: In addition to the three types of tests presented above, we will also include Watterson's (1978) homozygosity test for comparisons. Watterson's test W is defined as

$$W = n^{-2} \sum_i f_i^2, \quad (14)$$

where f_i is the number of allele i in a sample of size n. Since given the number of alleles in a sample, the frequencies of allele of various types are independent of θ (Ewens 1972; Karlin and McGregor 1972), Watterson's test W is thus independent of θ when conditioning on the observed number of alleles in the sample.

THE CRITICAL POINTS OF THE TESTS

The critical points (or value) of each test described earlier (except for Watterson's test W) can be obtained by the Monte-Carlo method used by Fu (1996). The process of finding the critical values of a test for a sample of size n essentially consists of two steps. The first step is to obtain an estimate $\hat{\theta}$ of θ from the sample for computing the value of the test statistics and later for use in the second step. For test F_S, $\hat{\theta}$ is Tajima's estimate $\hat{\theta}_\pi$ and for other tests, $\hat{\theta}$ is Watterson's estimate $\hat{\theta}_W$. The second step is to obtain the critical points of the test from simulated samples from a random-mating population with θ equal to $\hat{\theta}$. In other words, we first generate a large number of simulated samples of size n from a random-mating population with $\theta = \hat{\theta}$ and for each simulated sample, the value of the test statistic is computed, and after all the samples have been examined obtain an empirical distribution of the test statistic, from which we obtain the critical points. After determining the critical points of a test, and if the value of the test statistic is less than the critical point, the null hypothesis of neutrality of mutations is then rejected. Although one should try to simulate as large number of samples as possible to avoid random errors, it is usually unnecessary to simulate >10,000 samples.

To illustrate the procedure described above, consider Fu and Li's test D and a hypothetic sample of size 50. Suppose we obtain from the sample that Watterson's estimate $\hat{\theta}_W = 5$ and $D = -1.95$. To determine whether this result is statistically significant, we generate 10,000 samples of size 50 from the neutral Wright-Fisher model with $\theta = 5$. Since for each simulated sample, we have a value of D, we thus have 10,000 D's. If in 5% percent of the samples D is smaller than -1.83 (this is taken to be the 5% cutoff value of the test, *i.e.*, the critical value of the test at 5% significance level), then we can conclude that the test is significant at 5% level since the observed D from the real DNA sample is smaller than -1.83.

Extensive simulations were carried out to study the critical points of the tests examined in this article. It was found that the critical point at 5% significance level for each of the tests except for F_S is the point corresponding to the lower fifth percentile of its empirical distribution. The critical point for test F_S is however the value corresponding to the lower second percentile of its empirical distribution. Therefore, if the value corresponding to the lower fifth percentile of the empirical distribution of F_S is used, the probability of rejecting the neutral model when it is true will be larger than 5%. Although why F_S behaves differently from the other tests is not fully understood, it appears partly due to the form of the test statistic and partly due to be the large sampling variance of Tajima's estimate of θ.

The above procedure for determining the critical points in a test is adequate when we have only one or a few DNA samples, but it is too time-consuming to use for investigating the power of a statistical test because many hundreds or thousands of samples need to be tested as in this study. To reduce the amount of compu-

tations in our simulations, we obtained the critical points of a test for only a number of values of θ and used a linear interpolation to obtain the critical points for a given value of θ as follows. Suppose θ_i ($i = 1, \ldots, m$) are the values of θ examined and the corresponding critical values are $c(\theta_i)$. Then the critical point for $\theta_i < \theta < \theta_{i+1}$ is determined by

$$c(\theta) = \frac{c(\theta_{i+1}) - c(\theta_i)}{\theta_{i+1} - \theta_i}(\theta - \theta_i) + c(\theta_i).$$

Similar interpolation of critical points for sample sizes can also be made. This method is very effective and sufficiently accurate. In fact, one can go even further by finding a regression equation to summarize the critical points so that for a given sample size and θ within certain ranges they can be computed from the regression equation, as in FU (1996) for several statistical tests.

The critical points of Watterson's test W in this study are also determined by Monte-Carlo simulation, using the algorithm by STEWART (1977). Although the critical points of this test are available in the literature (*e.g.*, WATTERSON 1978; EWENS 1979) for some combinations of n and the number of alleles, it is simpler for our purpose to regenerate all the critical points. Our critical points agree well with those in EWENS (1979) for comparable combinations of n and the number of alleles.

POLYMORPHISM PATTERNS AND THE POWERS OF TESTS

In this section, we will examine the pattern of polymorphisms and the powers of various tests described in the early sections under three models: population growth, genetic hitchhiking and background selection. These will be accomplished by using simulated samples under the three models. Since there are simply too many tests of types $F(r, r')$ and $F'(r, r')$, we will focus on T, D^*, F^* and a few others that appear promising and sufficiently different from T, D, F, D^* and F^*. It should be pointed out that in our simulation studies, recombinations are not considered, so one should be cautious when applying the tests described in this article to DNA samples containing recombinations. The effects of recombinations are different for different tests and they will be discussed in DISCUSSION section.

Population growth: Let N_t be the effective size of the population at generation t. The generation 0 ($t = 0$) is the reference time point and somewhat arbitrary. Consider the logistic model of population growth

$$N_t = N_{\min} + \frac{N_{\max} - N_{\min}}{1 + e^{-r(t-c)}}, \qquad (15)$$

where $N_{\min}$ and $N_{\max}$ are the minimum and maximum effective population sizes, r and c are both nonnegative, and one unit of time corresponds to $2N_{\max}$ generations. The parameter r in the logistic equation determines the speed of growth and parameter c is the reflection point of the growth curve. Because the logistic model of population growth has four parameters, it is a very general model of population growth and is much more flexible than the exponential model of population growth. Figure 1 plots $N_t/N_{\max}$ for several values of r when $N_{\min} = 1000$ and $N_{\max} = 20{,}000$.

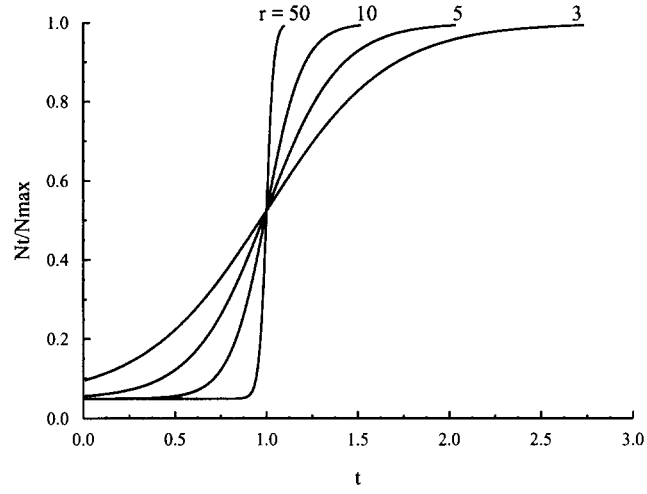

FIGURE 1.— $N_t/N_{\max}$ of the logistic model with $N_{\min} = 1000$, $N_{\max} = 20{,}000$ and $c = 1$. One unit of t corresponds to $2N_{\max}$ generations.

We are interested in the effects of sampling at different times on the pattern of DNA sequence polymorphism and the powers of some statistical tests. Suppose a sample of size n is taken at time T_s. Select a new time scheme so that generation 0, 1, 2, . . . represent, respectively, the generation at T_S, 1, 2, . . . generations before the generation at T_S. That is, time is counted backward starting at the generation represented by time T_S. Then the effective population size N_t' at the new time t is

$$N_t' = N_{\min} + \frac{N_{\max} - N_{\min}}{1 + e^{-r(T_s - t - c)}}. \qquad (16)$$

The coalescent theory for a deterministic change in population size such as (16) was developed by GRIFFITHS and TAVARÉ (1994). Let t_k be the kth coalescent time (one unit corresponds to $2N_{\max}$ generations), its density function $f(t_k)$ conditional on $t_{k+1}, \ldots, t_n$ is then given by

$$f(t_k) = \frac{k(k-1)}{2} v^{-1}(s_{k+1} + t_k) \times \exp\left[-\frac{k(k-1)}{2}\int_{s_{k+1}}^{s_{k+1}+t_k} v^{-1}(t)\,dt\right], \qquad (17)$$

where $s_{n+1} = 0$, $s_{k+1} = t_n + \cdots + t_{k+1}$, $v(t) = N_t'/N_{\max}$ and thus $v^{-1}(t) = N_{\max}/N_t'$. A random value of t_k conditional on s_{k+1} can be generated by solving t_k for the equation

$$\int_{s_{k+1}}^{s_{k+1}+t_k} v^{-1}(t)\,dt = \frac{-2}{k(k-1)} \log(U), \qquad (18)$$

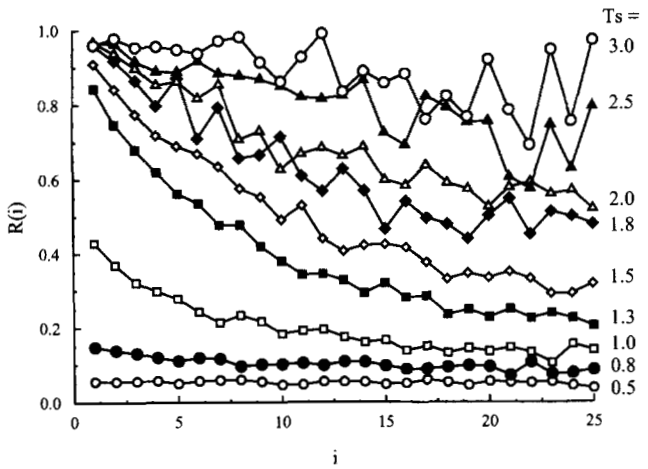

FIGURE 2.—$R(i) = E(\eta_i')/E(\eta_i)$ of logistic population growth with $N_{min} = 1000$, $N_{max} = 20{,}000$, $r = 10$ and $c = 1.0$. Each curve is based on 20,000 independent samples and $R(i)$ $(i = 1, \ldots, 25)$ for each T_s are connected by line segments for clarity.

where U is a random value of a uniform random variable over $(0, 1)$. The solution to the equation can be obtained by a numerical integration. Therefore, we can generate the coalescent times for a sample of size n sequentially by generating t_n first, then t_{n-1} and so on until obtaining t_2.

One way to measure the effect of population growth on the pattern of polymorphism is to examine the ratio $R(i) = E(\eta_i')/E(\eta_i)$, where $E(\eta_i')$ and $E(\eta_i)$ are, respectively, the expected numbers of segregating sites of type i under the logistic model of population growth and under the neutral model, or similarly one can examine the ratio $E(\xi_i')/E(\xi_i)$. For example, if $R(i)$ $(i = 1, \cdots)$ are roughly constant, then the effect of population growth is about the same on each type of segregating sites. We expect to observe this pattern when T_s is either small or very large, because when T_s is small, the population size at the time of sampling is only marginally larger than N_{min}; while when T_s is very large, the population size has already been close to N_{max} for some times, thus coalescent to the common ancestor often occurs before population size decreases significantly. Figure 2 shows the effects of sampling at different times on $R(i)$ for a sample of 50 sequences with $r = 10$ and $c = 1$.

Figure 3 shows the powers of several tests for sampling at different times with two different values of $\theta = 4N_{max}\mu$. As indicated by the above analysis of $R(i)$, it is indeed true that all these tests have little power when the sampling time T_s is either too small or too large. The peak of the power for each test lies between $T_s = 1$ and 1.5, which happens to correspond to the period in which the population size differs significantly from the initial size but before it reaches a steady size when $r = 10$ and $c = 1$. Similar patterns were observed for different values of r and c, suggesting that in general, sampling at a time when the population size has grown

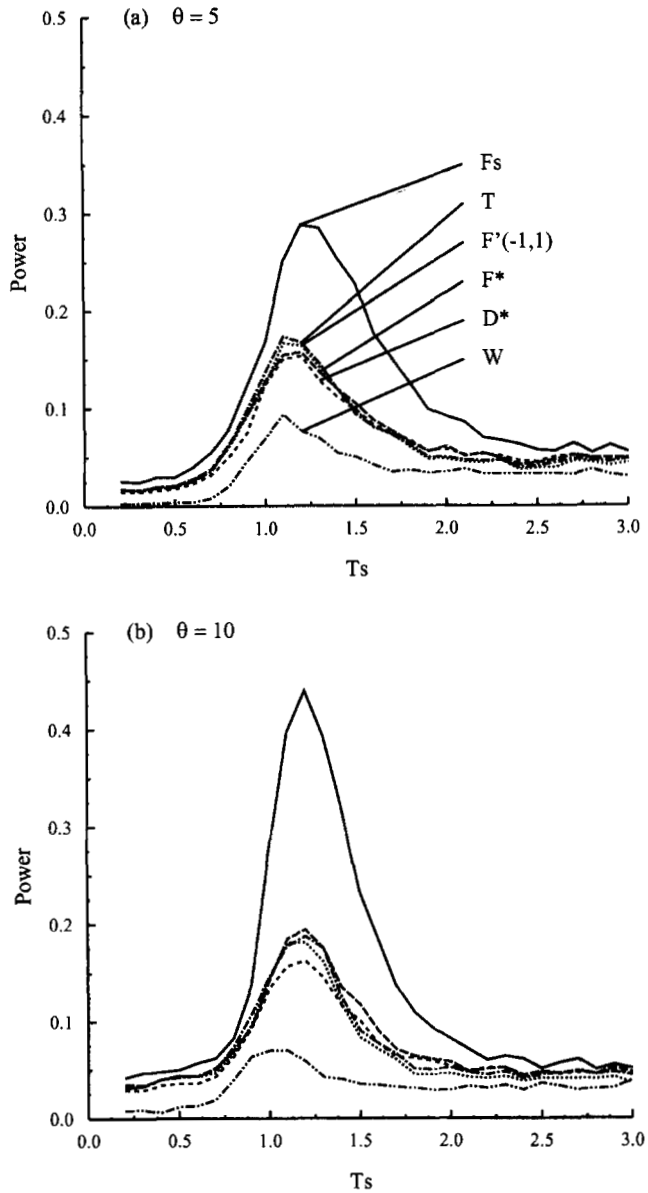

FIGURE 3.—Powers of tests when $n = 50$ against logistic population growth with $N_{min} = 1000$, $N_{max} = 20{,}000$, $r = 10$ and $c = 1.0$. The same line pattern is used in both panels for each test.

substantially but before it reaches a steady size provides the best opportunity to detect a population growth.

Among the tests considered, the new test F_S is clearly the most powerful one; in fact, it is often more than twice as powerful as any other test examined. On the other hand, Watterson's test W is the least powerful test. In between are Tajima's test T, Fu and Li's tests D^* and F^* and the new test $F'(-1, 1)$. These four tests do not differ significantly in their powers. We also examined several other tests, including Fu and Li's test D and F, tests $F(-0.5, 1.5)$, $F'(-0.5, 1)$ and $F'(0, 2)$, and found that their powers are all similar to those of T, D^*, F^* and $F'(-1, 1)$.

Genetic hitchhiking: Consider a neutral locus that is linked to a locus under natural selection. When a

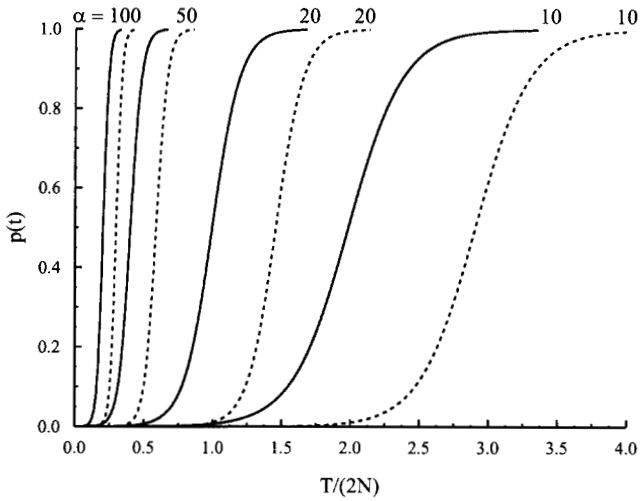

FIGURE 4.—Frequency of selected allele with respect to times. Solid lines are with $N = 10^4$ and dotted lines are with $N = 10^6$. In all the cases $h = {}^1/_2$.

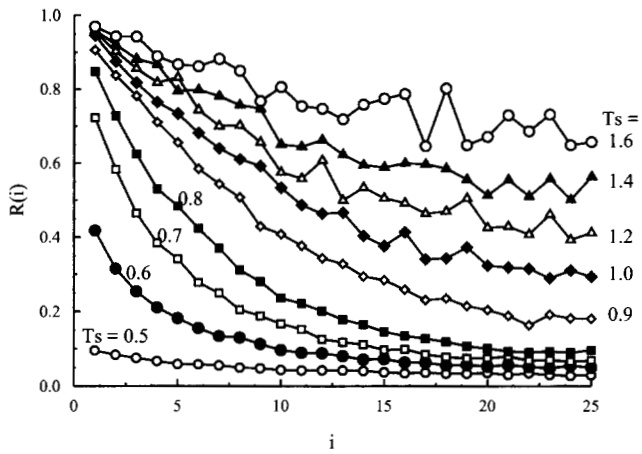

FIGURE 5.— $R(i) = E(\eta_i')/E(\eta_i)$ under hitchhiking with $N = 10^6$, $2Ns = 50$ and $h = {}^1/_2$. Each curve is based on 20,000 independent samples and $R(i)$ $(i = 1, \ldots, 25)$ for each T_s are connected by line segments for clarity.

favorable mutant at the locus under selection sweeps the whole population, it drags along the neutral locus and therefore the pattern of polymorphism at the neutral locus can be strongly affected by the linkage to the selected locus. Suppose there are two alleles at the selected locus and the fitness of genotypes are

$$\begin{matrix} AA & Aa & aa \\ 1+s & 1+hs & 1 \end{matrix},$$

where allele A is a mutant favored by natural selection, and $s(s > 0)$ and $h(1 \geq h > 0)$ are the selection coefficient and the dominance coefficient, respectively. Assuming the initial frequency of the allele A is $1/(2N)$ and neglecting the effect of random drift, MAYNARD SMITH and HAIGH (1974) showed that the frequency of allele A at $n+1$-th generation is given by

$$p_{n+1} = p_n + \frac{sp_n(1-p_n)[h + p_n(1-2h)]}{1 + sp_n[2h + p_n(1-2h)]}. \quad (19)$$

The speed at which the mutant allele A reaches fixation is largely determined by selection intensity defined as $\alpha = 2Ns$. Equation 19, however, is inconvenient to use because a huge number of iteration is usually needed to compute the allele frequency. Instead of using one generation as an unit of time, we can define $\delta(2N)$ generations as one unit of time where $\delta > (2N)^{-1}$, then an approximation to (19) is

$$p(t+1) = p(t) + (\delta 2N) \times \frac{sp(t)(1-p(t))[h + p(t)(1-2h)]}{1 + sp(t)[2h + p(t)(1-2h)]}. \quad (20)$$

Figure 4 shows $p(t)$ for several values of the selection intensity α.

An algorithm for simulating a sample under the hitchhiking model used here was developed by KAPLAN *et al.* (1987). For simplicity, we assume that the sequences in a sample are randomly drawn from those carrying allele A at the selected locus. With this simplification, the simulation of the genealogy of a sample is similar to the algorithm for population growth discussed in the previous section. Let T_s be the time at which a sample is taken. Start at the generation and look backward in time. Then the frequency $v(t)$ of allele A at time t is $p(T_s - t)$, namely

$$v(t) = p(t') + [\delta(2N)] \times \frac{sp(t')[1-p(t')][h + p(t')(1-2h)]}{1 + sp(t')[2h + p(t')(1-2h)]}, \quad (21)$$

where $t' = T_s - t - 1$. Substituting the $v(t)$ in (17) by the above $v(t)$ gives the density function of the kth coalescent time under the hitchhiking model, and thus the kth coalescent time can be generated by solving (18). For the purposes of studying the powers of tests, we found that setting $\delta = 10^{-3}$ gives sufficiently accurate results, which was also the increment value used by BRAVEMAN *et al.* (1995).

Similar to the case of population growth, we can examine the ratio $R(i) = E(\eta_i')/E(\eta_i)$, where $E(\eta_i')$ and $E(\eta_i)$ are, respectively, the expected numbers of segregating sites of type i under the hitchhiking model and under the neutral model. Figure 5 shows how sampling at different times affects the value of $R(i)$ for a sample of 50 sequences. Comparing the pattern of $R(i)$ to that in Figure 2, it is clear that they are overall very similar. A closer examination shows that $R(i)$ $(i = 1, \ldots, 25)$ decreases more deeply under the hitchhiking model than under the population growth model. For example, the value of $R(1)$ for $T_s = 0.8$ under hitchhiking is about the same as that for $T_s = 1.3$ under population growth, but the values of $R(25)$ under hitchhiking and population growth are, respectively, 0.11 and 0.23. Similar patterns were also observed for a number of different parameter sets.

Figure 6 shows the powers of several tests under dif-

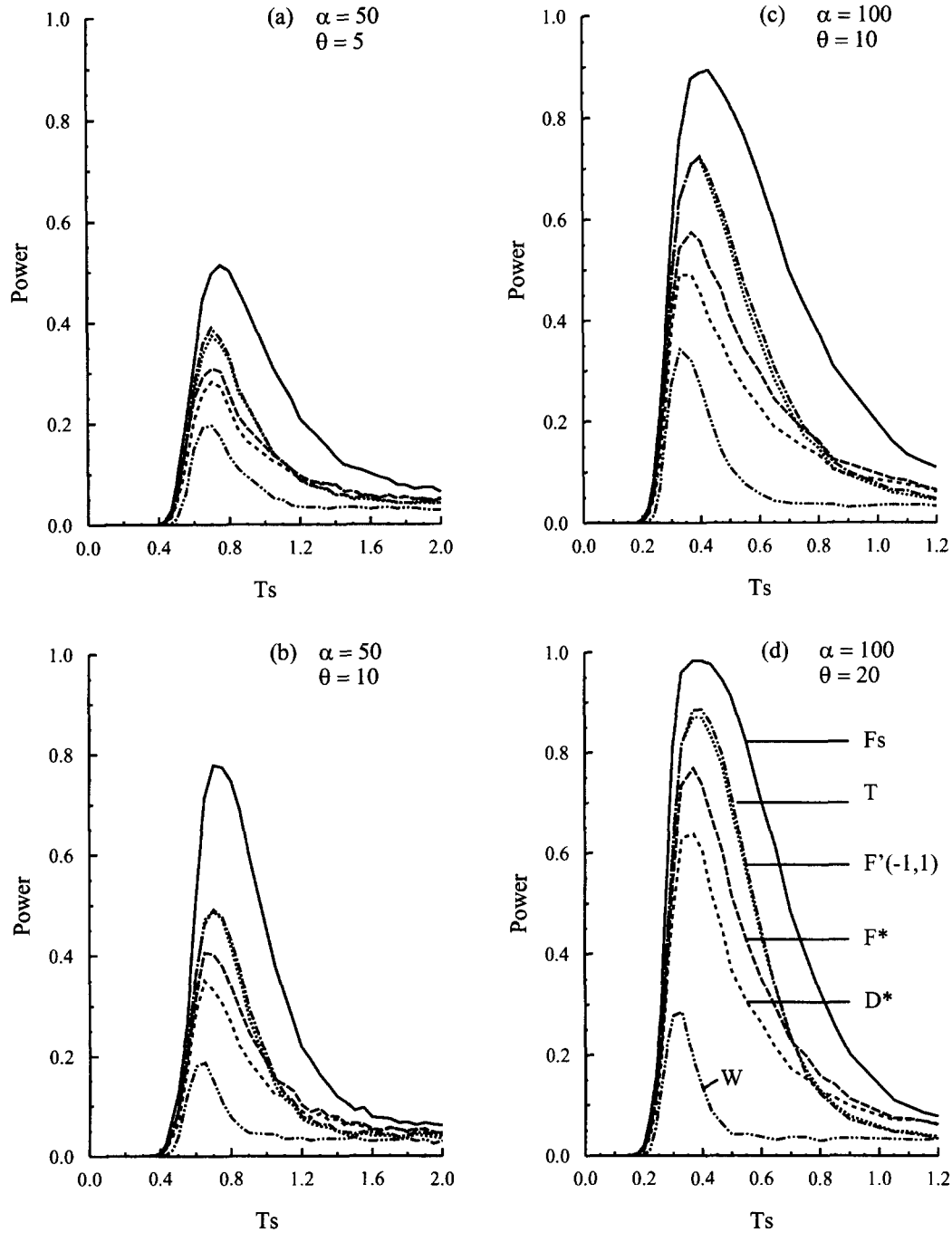

FIGURE 6.—Powers of tests when $n = 50$ for hitchhiking with $N = 10^6$, $h = 0.5$ and $\alpha = 2Ns = 50$ [(a) $\theta = 5$ and (b) $\theta = 10$] and 100 [(c) $\theta = 10$ and (d) $\theta = 20$]. (The power of $F'(-1, 1)$ is almost identical to that of $F'(-1/2, 3/2)$.) The same line pattern is used in all the panels for each test.

ferent conditions with $N = 10^6$. One can see from Figure 4 that for $N = 10^6$ and $\alpha = 50$ the frequency of allele A starts to increase significantly when $T_s = 0.45$, reaches 0.50 when $T_s = 0.60$ and becomes 0.99 when $T_s = 0.77$; while for $\alpha = 100$, the frequency of allele A starts to increase significantly when $T_s = 0.20$, reaches 0.50 when $T_s = 30$ and reaches nearly fixation (*i.e.*, 0.99) when $T_s = 0.39$. Figure 6 shows that there is a sharp increase in the power of each test when allele A climbs from low frequency to fixation, and afterward the power gradually declines.

It should be noted that we assume that the sequences are a random sample from those carrying allele A at the selected locus. When allele A is fixed or nearly fixed in the population, our sample is not different from a random sample from the entire population. However, when the frequency of allele A is not close to 1, a random sample from the population may contain some sequences carrying allele a at the selected locus, and because it takes longer to coalesce one sequence carrying alleles A and one sequence carrying allele a than to coalesce two sequences under the neutral Wright-

Fisher model, the excess of recent mutations in such a sample is less severe than in a sample of sequences carrying only allele A. Consequently when the frequency of allele A is not close to fixation, the power of a test for a random sample from the entire population will be less than that shown in Figure 6. In other words, the power of a test for a random sample of sequences will start to climb later than indicated in Figure 6 and increase more rapidly, as observed by SIMONSEN *et al.* (1995). On the other hand, if a random sample is taken and the allelic status at the selected locus is known for each sequence, it is more powerful to use only those sequences carrying the advantageous allele.

It is clear that F_S is the most powerful test among the six tests showed in Figure 6: Watterson's test W is the least powerful one and in between are tests T, F^*, D^* and $F'(-1, 1)$, which is similar to the situation of population growth. It is also true that the powers of tests T and $F'(-1, 1)$ [and $F(-0.5, 1.5)$, result not shown] are very similar, but unlike the situation of population growth, test T and $F'(-1, 1)$ are now considerably more powerful than tests D^* and F^* (and D and F, results not shown). These results probably reflect the similarity and difference between the patterns in Figures 2 and 5, and they also agree with those by SIMONSEN *et al.* (1995).

Figure 6 also shows that the value of the selection intensity α is a key factor determining the power of a test. Comparing b and c shows that the larger the value of α is, the more powerful a test becomes, which is naturally expected. It is also obvious that a larger value of θ results in more powers in all these tests.

If there are recombinations between the neutral and selected loci, the effect of genetic hitchhiking will be reduced and so will the power of a test (BRAVEMAN *et al.* 1995). However, we expect that test F_S continues to be a powerful test in such a situation.

Background selection: Consider a neutral locus that is linked to a number of loci subject to the natural selection that eliminates gametes carrying too many deleterious mutations. Such type of selection is known as background selection (*e.g.*, CHARLESWORTH *et al.* 1993). We consider a simple model of fitness in which a gamete carrying j deleterious mutation has fitness $w_i = (1 - sh)^j$ where s and h are the selection and dominance coefficients. Assume that the number of new mutations per individual that arise each generation is a Poisson variable with mean U. Then under the above fitness model, the frequency of gametes carrying i mutations will reach the equilibrium frequency (KIMURA and MARUYAMA 1966; CROW 1970)

$$f_i = \frac{e^{-U/(2sh)}[U/(2sh)]^i}{i!}. \qquad (22)$$

Our simulation algorithm is a slight modification of the algorithm by CHARLESWORTH *et al.* (1995). To simulate the genealogy of a sample of DNA sequences from the neutral locus, one first generates the number n_i of gametes with i mutations from the equilibrium distribution (22). At each generation, the first step of the algorithm is to determine the number of mutations in the parent gamete of each gamete. Given a gamete carries i mutations, the probability that its parent has j mutations is

$$Q_{ij} = \frac{f_j w_j m_{i-j}}{\sum f_j w_j m_{i-j}},$$

where m_{i-j} is the probability that a gamete experiences $i - j$ new mutations in one generation. Therefore, $m_{i-j} = e^{-U}U^{i-j}/(i-j)!$. For each gamete with i deleterious mutations, we generate a random number and determine from Q_{ij}, $(j = 0, \ldots, i)$ the number of deleterious mutations in its parent gamete. Note that CHARLESWORTH *et al.* (1995) determined this number by using Poisson variable with mean Q_{ij} that is economic in computation but is less accurate than using Q_{ij} directly. After the number of mutation in the parent gamete of each gamete has been found, we determine the coalescent events. Coalescence can occur only between alleles with the same number of deleterious mutations. Let n_i' be the number of sequences with i deleterious mutations in the parent generation. Then the probability of a coalescent event within this group of alleles is

$$\prod_{k=0}^{n_i'-1}\left(1 - \frac{k}{2Nf_i}\right) \approx 1 - \frac{n_i'(n_i' - 1)}{4Nf_i}.$$

Note that multiple coalescences in different groups of alleles can occur. This process continues until there is only one ancestral gamete left. Once the genealogy is obtained, we superimpose neutral mutations onto the genealogy.

As in the previous sections, we can examine the ratio $R(i) = E(\eta_i')/E(\eta_i)$, where $E(\eta_i')$ and $E(\eta_i)$ are the expected number of segregating sites of type i under the background selection and the neutral model, respectively. Figure 7 shows how $R(i)$ are affected by background selection. In comparison with the effects of population growth and genetic hitchhiking (Figures 2 and 5), background selection shows strikingly different pattern: the frequencies of segregating sites of various type, except for that of singletons, are reduced by about the same proportion from the neutrality, the frequency of singletons on the other hand is much closer to that under the neutral model. This pattern largely explains the observation by CHARLESWORTH *et al.* (1995) that Fu and Li's test D is more powerful than Tajima's test T, because D is a contrast between singleton and non-singleton. However, when computing the value of D and T, CHARLESWORTH *et al.* (1995) used fixed values of θ to substitute the unknown θ in these two statistics, instead of estimating θ by Watterson's estimate $\hat{\theta}_W$ of θ as proposed, thus it is not clear whether their conclusions still hold when these two tests are used as they are in practice.

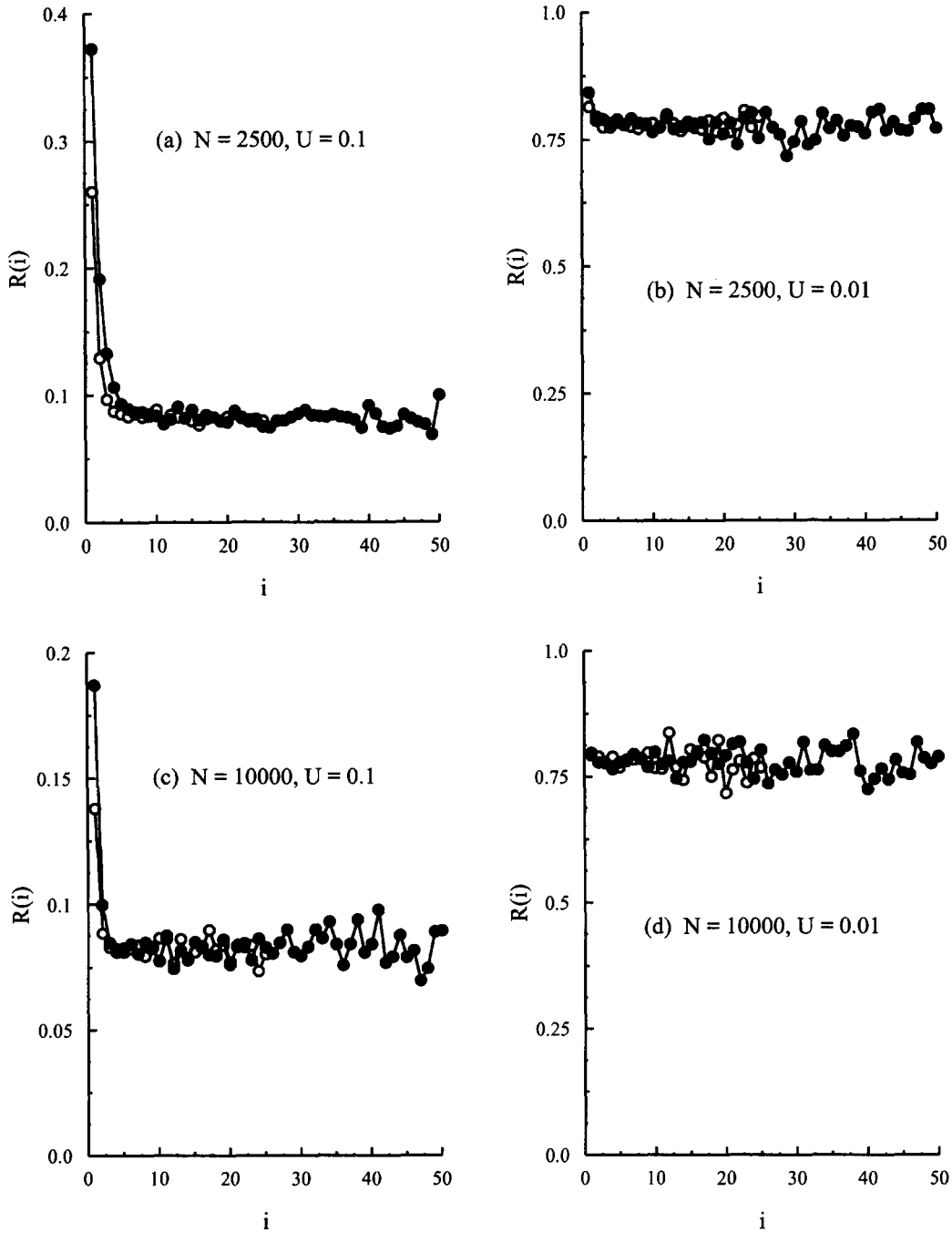

FIGURE 7.— $R(i) = E(\eta_i')/E(\eta_i)$ under background selection with $h = 0.1$ and $s = 0.2$. ○, $n = 50$; ●, $n = 100$. Each curve is based on 20,000 independent samples.

Table 1 gives the powers of several tests for detecting background selection under several parameter sets and we summarize the results as follows:

- The value of U has a substantial effect on the amount of polymorphism and the power of a test. The larger the value of U is, the less the polymorphism and larger the chance of detecting background selection.
- It is more effective to increase sample size than to increase sequence length for detecting background selection, but a sufficient amount of polymorphism is necessary.
- Among the tests considered, the four tests by FU and LI (1993) are the most powerful tests and the powers do not differ much among them, but they are often more than twice as powerful as Tajima's test T. Interestingly the powers of tests $F'(1, r)$ $(r > 2)$ are all similar to that of test $D^* = F'(1, \infty)$.
- Watterson's homozygosity test is the least powerful test among all the tests examined, similar to what was observed in the cases of population growth and genetic hitchhiking. Overall the power of test F_S is between those of Fu and Li's (1993) tests and Tajima's test T.

TABLE 1

Power of tests against background selections

Parameters			Estimates of θ			Powers of tests							
N	n	θ	π	θ_W	ξ_1	W	T	F_S	D	F	D^*	F^*	$F'(1,3)$
$U = 0.01$													
25000	100	100	78.6	78.4	78.7	0.04	0.05	0.03	0.05	0.05	0.05	0.05	0.05
10000	100	100	77.7	78.1	80.2	0.05	0.05	0.04	0.06	0.06	0.06	0.06	0.06
5000	100	100	78.6	79.2	81.8	0.04	0.05	0.04	0.06	0.06	0.05	0.06	0.06
2500	100	100	78.5	79.5	83.5	0.04	0.06	0.05	0.07	0.07	0.07	0.07	0.07
$U = 0.1$													
25000	50	10	0.8	0.9	1.1	0.02	0.06	0.06	0.06	0.07	0.06	0.07	0.07
		50	4.2	4.4	5.3	0.04	0.07	0.08	0.09	0.09	0.09	0.08	0.08
		100	8.3	8.8	10.7	0.04	0.07	0.11	0.11	0.10	0.09	0.09	0.09
	100	10	0.8	0.9	1.3	0.05	0.09	0.08	0.08	0.09	0.08	0.09	0.09
		50	4.2	4.6	6.4	0.05	0.09	0.12	0.16	0.16	0.15	0.15	0.15
		100	8.3	9.2	12.8	0.06	0.09	0.15	0.21	0.19	0.20	0.18	0.19
5000	50	10	0.9	1.1	1.8	0.04	0.14	0.25	0.15	0.17	0.15	0.17	0.16
		50	4.4	5.4	9.3	0.09	0.17	0.31	0.33	0.31	0.28	0.28	0.27
		100	8.8	10.8	18.5	0.11	0.18	0.39	0.40	0.37	0.32	0.32	0.32
	100	10	0.9	1.2	2.6	0.13	0.22	0.20	0.29	0.30	0.30	0.29	0.30
		50	4.4	6.2	13.1	0.14	0.30	0.46	0.66	0.63	0.64	0.61	0.64
		100	8.7	12.3	26.2	0.20	0.33	0.60	0.81	0.75	0.79	0.74	0.78
$U = 0.2$													
5000	50	10	0.1	0.1	0.2	0.00	0.02	0.02	0.02	0.02	0.02	0.02	0.02
		50	0.4	0.5	1.1	0.04	0.15	0.17	0.15	0.18	0.15	0.17	0.17
		100	0.7	1.0	2.2	0.10	0.21	0.21	0.25	0.29	0.26	0.27	0.27
	100	10	0.1	0.1	0.4	0.02	0.05	0.05	0.04	0.04	0.04	0.04	0.04
		50	0.4	0.6	1.8	0.16	0.30	0.28	0.30	0.34	0.30	0.33	0.33
		100	0.8	1.3	3.6	0.22	0.35	0.33	0.54	0.53	0.54	0.52	0.54
10000	50	10	0.1	0.2	0.7	0.02	0.13	0.15	0.12	0.13	0.12	0.13	0.13
		50	0.5	1.1	3.4	0.23	0.51	0.52	0.47	0.63	0.58	0.61	0.62
		100	1.0	2.3	6.8	0.30	0.64	0.69	0.83	0.83	0.81	0.81	0.81
	100	10	0.1	0.3	1.1	0.13	0.30	0.30	0.23	0.27	023	0.27	0.26
		50	0.5	1.6	5.6	0.39	0.72	0.72	0.84	0.86	0.84	0.86	0.86
		100	1.0	3.2	11.2	0.48	0.86	0.91	0.98	0.98	0.97	0.98	0.98

10,000 samples were simulated for each parameter set.

Our simulation results on the power of Tajima's test T and Fu and Li (1993) test D agree with those by Charlesworth *et al.* (1995) in the case $N = 25000$, $U = 0.1$, but the powers of the two tests in our simulation are both less powerful than found in Charlesworth *et al.* (1995). One reason for this is that we used all the samples simulated regardless of the amount of polymorphism, while Charlesworth *et al.* (1995) used only those samples with polymorphism, since samples without polymorphism do not result in rejecting the neutral model, the power of a test in our study should be less. Another difference between this study and that by Charlesworth *et al.* (1995) is that our tests are performed in the way they are used (or should be used) in practice, while Charlesworth *et al.* (1995) used fixed values of θ to substitute the unknown θ.

However, our simulation results in the case $N = 25000$, $U = 0.01$ are quite different from those of Charlesworth *et al.* (1995). For example, when $n = 100$ and $\theta = 10$, Charlesworth *et al.* (1995) found that tests T and D have powers 0.116 and 0.289 (their Table 4), respectively, at 5% significance level; while in our simulation, we found that none of the tests has power significantly larger than the nominal level (0.05) (see Table 1). This appears to be a result of the difference in applying these tests and not a result of the way samples are selected because almost all the samples are polymorphic in this situation (see Table 1).

DISCUSSION

The statistical properties of tests for detecting an excess of the number of rare alleles are more complex than those of tests for detecting an excess of the number of common alleles. We developed in this paper the new test F_S and several new tests of types $F(r, r')$ and $F'(r, r')$. Although it is unlikely that the resource for developing new and hopefully more powerful statistical tests is exhausted, it appears that F_S is a very promising test for detecting population growth and genetic hitchhiking while Fu and Li's (1993) tests are among the best for detecting background selection. There are a number of other statistical tests examined in this study. Their results are not presented because they are either

less powerful or are not significantly better than those presented. For example, instead of using Watterson's estimate of θ, one can use the estimate ($\hat{\theta}_{\eta_1} = \eta_1/[1 + 1/(n-1)]$) based on only the number of singleton-segregating sites to substitute θ in $F(r, r')$ and $F'(r, r')$, doing so results in slightly better tests in most cases. We also examined the three tests W, G_η and G_ξ by Fu (1996) and found that W is less powerful than F_S while G_η and G_ξ have little power against an excess of the number of rare alleles.

As in my previous study (Fu 1996), I used the infinite-sites model to generate critical values of each test. Therefore, when multiple hits at some sites are evident for a given sample of DNA sequences, some corrections should be taken before applying these tests. One effective way to minimize the effect of multiple hits is to compute the values of statistics in a test from a parsimony tree of the sample. For example, instead of assigning the number of segregating sites in a sample to K in Tajima's test T, one should use the number of mutations inferred by the parsimony analysis [also see the discussion in Fu (1996)].

This study also assumes no recombination within the locus from which DNA sequences are obtained. When there are recombination events, the number of alleles is usually inflated, while the means of π and θ_W are not affected. Therefore, test F_S may be sensitive to recombinations, so one should be cautious when applying F_S to a sample if there is evidence of recombination. If future studies show that F_S is indeed sensitive to recombination, it may be a good statistic for testing the presence of recombination. Since F_S appears to be a very powerful test against population growth and hitchhiking, it will be very useful to explore in future study whether it can be modified to allow recombination.

Statistical tests of type $F(r, r')$ and $F'(r, r')$ should be less sensitive to the existence of recombination because the expectations of the estimates of θ used in these tests are the same with or without recombination. Therefore, tests of type $F(r, r')$ and $F'(r, r')$ can be used when there is recombination. However, since recombination reduces variances of the estimates of θ, these tests may be conservative when there is recombination. Therefore, there is also a need to expand these tests to allow recombination without significant loss of power.

The observation that Fu and Li's (1993) tests are considerably more powerful than Tajima's test and F_S in the case of background selection, and the reverse for population growth and genetic hitchhiking, has an interesting implication: these tests can indicate the likely mechanism that is responsible for the observed polymorphism. For example, if only Fu and Li's tests are significant, this suggests that background selection is the more likely cause. On the other hand, if only F_S is significant, it is more likely to be due to population growth or hitchhiking (or perhaps recombination). In my previous study on statistical tests for detecting an excess of common alleles (Fu 1996), it was found that the relative powers of tests are consistent over different population genetic models that all result in an excess of common alleles, although only a few alternative models were examined.

Computer programs to perform the statistical tests discussed in this article will be available at the web page: http://hgc.sph.uth.tmc.edu/fu

I thank two reviewers for their comments. This research was supported by National Institutes of Health grant R29 GM-50428.

LITERATURE CITED

Braverman, J. M., R. R. Hudson, C. H. Kaplan, N. L. Langley and W. Stephan, 1995 The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics **140:** 783–796.

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

Charlesworth, D., B. Charlesworth and M. T. Morgan, 1995 The pattern of neutral molecular variation under the background selection model. Genetics **141:** 1619–1632.

Crow, J. F., 1970 Genetic loads and the cost of natural selection, pp. 1–35 in *Mathematical Topics in Population Genetics*, edited by K. I. Kojima. Springer-Verlag, Berlin.

Ewens, W. J., 1972 The sampling theory of selectively neutral alleles. Theoret. Popul. Biol. **3:** 87–112.

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

Fu, Y. X., 1994 Estimating effective population size or mutation rate using the frequencies of mutations of various classes in a sample of DNA sequences. Genetics **138:** 1375–1386.

Fu, Y. X., 1995 Statistical properties of segregating sites. Theoret. Popul. Biol. **48:** 172–197.

Fu, Y. X., 1996 New statistical tests of neutrality for DNA samples from a population. Genetics **143:** 557–570.

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 Sampling theory for neutral alleles in a varying environment. Phil. Trans. R. Soc. Lond. B **344:** 403–410.

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, London.

Karlin, S., and J. L. McGregor, 1972 Addendum to a paper of W. Ewens. Theoret. Popul. Biol. **5:** 95–105.

Kimura, M., and T. Maruyama, 1966 Mutational load with epistatic gene interactions in fitness. Genetics **54:** 1337–1351.

Kingman, J. F. C., 1982a The coalescent. Stoch. Proc. Appl. **13:** 235–248.

Kingman, J. F. C., 1982b On the genealogy of large populations. J. Appl. Probab. **19A:** 27–43.

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

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

Stewart, F. M., 1977 Appendix to P. A. Fuest, R. Chakraborty and M. Nei, Statistical studies on protein polymorphism in natural populations. I. Distribution of sigle-locus heterozygosity. Genetics **86:** 455–483.

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

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

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

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

Watterson, G. A., 1978 The homozygosity test of neutrality. Genetics **88:** 405–417.

Communicating editor: D. Charlesworth