New Statistical Tests of Neutrality for DNA Samples From a Population

Yun-Xin Fu

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

Manuscript received October 16, 1995
Accepted for publication January 29, 1996

ABSTRACT

The purpose of this paper is to develop statistical tests of the neutral model of evolution against a class of alternative models with the common characteristic of having an excess of mutations that occurred a long time ago or a reduction of recent mutations compared to the neutral model. This class of population genetics models include models for structured populations, models with decreasing effective population size and models of selection and mutation balance. Four statistical tests were proposed in this paper for DNA samples from a population. Two of these tests, one new and another a modification of an existing test, are based on EWENS' sampling formula, and the other two new tests make use of the frequencies of mutations of various classes. Using simulated samples and regression analyses, the critical values of these tests can be computed from regression equations. This approach for computing the critical values of a test was found to be appropriate and quite effective. We examined the powers of these four tests using simulated samples from structured populations, populations with linearly decreasing sizes and models of selection and mutation balance and found that they are more powerful than existing statistical tests of the neutral model of evolution.

DNA polymorphisms are powerful sources of information for studying the evolution of a population. Whether a locus or region from which a DNA sample has been taken evolves neutrally or under natural selection is of considerable interest in evolutionary study and can be examined using a statistical test designed for DNA polymorphisms. A popular statistical test proposed by TAJIMA (1989) is

$$T = \frac{\pi - K/a_n}{\sqrt{\mathrm{Var}(\pi - K/a_n)}},$$

where π is the mean number of nucleotide differences between two sequences, K is the number of segregating sites, n is sample size and

$$a_n = 1 + \frac{1}{2} + \cdots + \frac{1}{n-1}. \qquad (1)$$

An essential parameter in the theory of neutral evolution is $\theta = 4N\mu$, where N is the effective population size and μ is the mutation rate per sequence per generation. Almost all summary statistics of DNA polymorphisms are related to this parameter. For example, the expectations of π and K/a_n are both equal to θ under the neutral model, which assumes that the population evolves according to the WRIGHT-FISHER model, that all mutations are selectively neutral and that there is no recombination within the locus. It follows that T is expected to be zero, and its variance is approximately equal to one when the neutral model of evolution holds. Although natural selection affects the values of both π and K/a_n, the magnitudes of the effect are different and T can thus be quite different from zero. Therefore T can be used as a statistical test of the neutral model. The several tests proposed by FU and LI (1993) are based on the same idea but utilizing the information in a sample differently.

Statistical tests of the neutral model, including TAJIMA's tests and FU and LI's tests, are often referred to as tests for the presence of natural selection, but when the neutral model is rejected, the presence of natural selection is only one of potentially many possible causes. It would be ideal if not only can one reject the neutral model when it is not true, but also can one identify the causal evolutionary force(s). It is difficult to achieve the latter goal by using statistical tests alone, but it is possible to distinguish two types of departures from the neutral model. For a finite population, all the sequences in a sample are descendants of a common ancestral sequence and the sequence polymorphisms are due to mutations that occurred in the branches of the sample genealogy. Mutations that occurred at generations close to the generation at which the most recent common ancestor lived are relatively old in age and mutations that occurred at recent generations are relatively young in age. When the neutral model is violated, the numbers of old and young mutations are often different from those under the neutral model. For example, when the effective size of a population is decreasing over generations, one would expect to observe more old mutations and less young mutations than expected under the neutral model, which assumes a constant effective population size. On the other hand, when

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

most mutations at a locus are deleterious, the majority of mutations in a sample are expected to be relatively young because deleterious mutations are unlikely to survive very long.

The purpose of this paper is to investigate statistical tests for detecting departures from the neutral model that are characterized by an excess of old mutations or a reduction of young mutations or both. This class of models includes models for structured populations, models of mutation and selection balance as well as models with decreasing effective population size. The reason for focusing on detecting one class of models is that our experience suggests that no single test of the neutral model is most powerful for detecting all alternative models, but when a test is powerful for detecting one particular model, it is usually also powerful for detecting other models that yield similar patterns of polymorphism. When there is an excess of old mutations or a reduction of young mutations, TAJIMA's T and the several tests by FU and LI (1993) all tend to be positive.

Three new statistical tests will be proposed in this paper and a test due to STROBECK (1987) will be modified. A simulation and regression approach will be used to compute the critical values of these tests. Powers of these tests will then be investigated using simulated samples and will be compared to the powers of TAJIMA's test and tests by FU and LI (1993). We shall show that the new tests are in general more powerful than both TAJIMA's test and FU and LI's tests for detecting departures from the neutral model due to an excess of old mutations or a reduction of young mutations or both.

CONSTRUCTING STATISTICAL TESTS

Tests based on EWENS' sampling distribution: EWENS (1972) and KARLIN and MCGREGER (1972) showed that under the infinite-alleles model the probability of k haplotypes (alleles) in a sample of size n is

$$\Pr(k|\theta) = \frac{|S_n^k|\theta^k}{S_n(\theta)}, \qquad (2)$$

where $\theta = 4N\mu$, N is the effective size of the population, μ is the mutation rate per sequence per generation and

$$S_n(\theta) = \theta(\theta + 1)\cdots(\theta + n - 1) = \sum_{k=0}^{n} S_n^k \theta^k.$$

STROBECK (1987) proposed a statistical test for detecting population structure based on EWENS' sampling formula (2). His test statistic is defined as

$$S = \frac{1}{S_n(\hat{\theta}_\pi)} \sum_{i \le k} |S_n^i| \hat{\theta}_\pi^i, \qquad (3)$$

where $\hat{\theta}_\pi = \pi$, *i.e.*, the mean number of nucleotide differences between two sequences. STROBECK (1987) suggested that the null hypothesis of a panminctic population (the neutral model) is rejected at α (say 0.05) significance level if $S < \alpha$. As we shall show later α is not an appropriate critical value for this test.

STROBECK's test was motivated by several theoretical studies (*e.g.*, LI 1976; SLATKIN 1982; STROBECK 1987; GOLDING and STROBECK 1983) showing that the expectation of $\hat{\theta}_\pi$ is independent of the migration rate among subpopulations and that the number of haplotypes in a sample is on average smaller than that predicted by the EWENS' sampling formula when θ is assumed to be $\hat{\theta}_\pi$. Therefore, the value of S, which is the probability of having no more than the observed number of alleles, can indicate whether the population under study is subdivided.

EWENS' sampling distribution provides an interesting basis for constructing statistical tests of the neutral model. However, to use this formula, one must substitute an estimate for the parameter θ. For a sample of DNA sequences there exist a number of estimators of θ, and for each estimator of θ, one can construct a test similar to STROBECK's test. Besides the sequence diversity $\hat{\theta}_\pi$, two other widely used estimators of θ are WATTERSON's estimator $\hat{\theta}_W$ and the heterozygosity estimator $\hat{\theta}_h$. WATTERSON's estimator $\hat{\theta}_W$ (WATTERSON 1975) is given by

$$\hat{\theta}_W = K/a_n, \qquad (4)$$

where K is the number of segregating sites and a_n is given by (1). The heterozygosity estimator $\hat{\theta}_h$ (ZOUROS 1979; CHAKRABORTY and WEISS 1991) is the solution θ for the equation

$$\frac{h}{1-h} = \theta\left[1 + \frac{2(1+\theta)}{(2+\theta)(3+\theta)}\right], \qquad (5)$$

where h is the allelic heterozygosity of a sample. Recently, several new estimators of θ have been proposed (FU 1994a,b; KUHNER *et al.* 1995). Although these new estimators are more accurate than the estimators $\hat{\theta}_\pi$, $\hat{\theta}_W$ and $\hat{\theta}_h$, their calculations are time consuming, which makes it difficult to investigate the properties of tests using these estimates as substitute of θ. Among these new estimators, however, the minimum variance estimators $\hat{\theta}_\eta$ and $\hat{\theta}_\xi$ (FU 1994a), which are based on the frequencies of mutations of various classes (see the next section for their definitions), are relatively easy to compute. Therefore, in addition to the three estimators $\hat{\theta}_\pi$, $\hat{\theta}_W$ and $\hat{\theta}_h$, we include $\hat{\theta}_\eta$ and $\hat{\theta}_\xi$ as candidate estimators of θ for constructing tests from EWENS' sampling distribution.

It should be obvious for the purpose of constructing a statistical test that the best substitute for θ in (2) is an estimator that differs mostly from EWENS' estimator $\hat{\theta}_E$ of θ when the neutral model does not hold because $\hat{\theta}_E$ is derived from the formula (2). The value of $\hat{\theta}_E$ is the solution θ for the equation

$$k = 1 + \frac{\theta}{1+\theta} + \cdots + \frac{\theta}{n-1+\theta}.$$

TABLE 1

Means and variances of several estimators for samples from a panminctic and an island population

Case	θ	Mean						Variance					
		$\hat{\theta}_E$	$\hat{\theta}_W$	θ_π	$\hat{\theta}_h$	$\hat{\theta}_\eta$	$\hat{\theta}_\xi$	$\hat{\theta}_E$	$\hat{\theta}_W$	$\hat{\theta}_\pi$	$\hat{\theta}_h$	$\hat{\theta}_\eta$	$\hat{\theta}_\xi$
a	1	1.0	1.0	1.0	1.0	1.0	1.0	0.3	0.4	0.6	0.6	0.3	0.3
b	10	10.4	9.9	10.1	8.1	10.1	10.0	11.4	10.6	24.1	7.4	9.0	6.2
c	10	5.6	6.6	7.7	4.5	6.2	5.9	6.0	8.1	21.6	3.1	6.3	3.9
d	10	2.0	7.7	10.6	2.2	6.5	4.8	0.6	22.9	91.4	1.3	20.0	4.6

Cases a and b: panminctic population and sample size $n = 30$. Case c: Two-islands model with $Nm = 0.125$, where N is the size of the whole population, and a sample of 15 sequences from each island is taken. Case d: ten-islands model with $Nm = 0.5$, and a sample of 100 sequences is taken from only one of the 10 islands. Ten thousand independent samples were simulated for each case.

Examining the means and variances of candidate estimators of θ under the neutral model and an alternative model, such as WRIGHT's finite-islands model, should provide a clue on which estimators are worth of further investigation. Suppose a population is divided into k local populations (islands). Then WRIGHT's finite-islands model assumes that each individual in an island has probability $m/(k-1)$ migrating to one of other $k-1$ islands at each generation, where m is the overall migration rate. Table 1 gives several examples of the means and variances of six estimators of θ for samples from a panminctic populations and island populations.

It is clear from Table 1 that the heterozygosity estimator $\hat{\theta}_h$ is not a good choice because its mean differs too little from the mean of $\hat{\theta}_E$ and also because it is biased even under the neutral model (case b). However, CHAKRABORTY and WEISS (1991) found that $\hat{\theta}_h$ is a good substitute for θ for detecting population growth. It is also clear from Table 1 that TAJIMA's estimator $\hat{\theta}_\pi$ cannot be a bad choice because its mean differs mostly from that of $\hat{\theta}_E$ among the five candidate estimators. It is interesting that although the mean of WATTERSON's estimator $\hat{\theta}_W$ is slightly smaller than that of $\hat{\theta}_\pi$, its variance is much smaller than that of $\hat{\theta}_\pi$. It is the small variance of WATTERSON's estimator $\hat{\theta}_W$ that makes it an excellent candidate to use with EWENS' sampling distribution. We thus propose the following new test statistic:

$$W = \sum_{i \le k} \Pr(k|\hat{\theta}_W) = \frac{1}{S_n(\hat{\theta}_W)} \sum_{i \le k} |S_n^i| \hat{\theta}_W^i. \qquad (6)$$

Table 1 also suggests that both $\hat{\theta}_\eta$ and $\hat{\theta}_\xi$ should be adequate substitutes for θ. Since the mean and variance of $\hat{\theta}_\eta$ are close to those of $\hat{\theta}_W$, the test with $\hat{\theta}_\eta$ as substitute of θ is expected to be similar to the test W. In the case of $\hat{\theta}_\xi$, it is more difficult to judge, because although the mean of $\hat{\theta}_\xi$ is considerably smaller than that of $\hat{\theta}_W$, so is its variance.

A preliminary study of these statistical tests confirmed above analysis of the five estimators. $\hat{\theta}_h$ is indeed the least desirable choice, $\hat{\theta}_\eta$ and $\hat{\theta}_\xi$ are both similar to $\hat{\theta}_W$, although $\hat{\theta}_W$ is slightly superior. Therefore, we shall focus on only STROBECK's test S and the new test W from this class of statistical tests. It should be noted that the well-known homozygosity test by WATTERSON (1978) is also inspired by EWENS sampling formula. Our preliminary study showed that WATTERSON's test has little power in detecting the kind of models considered in this paper, thus we shall not discuss this test further.

Tests using the frequencies of segregating sites of various classes: There are $2(n-1)$ branches in the genealogy of a sample of n sequences. We define the *size of a branch* as the number of sequences in the sample, represented by external nodes in the genealogy, that are descendents of that branch. Mutations resulting in segregating sites of a sample must occur in the branches of the sample genealogy. The *size of a mutation* is defined as the size of the branch in which the mutation occurs. Therefore, mutations in a genealogy of n sequences are classified into $n-1$ different size groups. We define ξ_i as the number of mutations of size i and let $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_{n-1})$.

A mutation is said to be *type i* if its size is either i or $n-i$ $(i < n-i)$. Therefore, a mutation belongs to one of $[n/2]$ types, where $[n/2]$ is equal to $n/2$ if n is even and is equal to $(n-1)/2$ when n is odd. We define η_i as the number of segregating sites of type i and let $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_{[n/2]})$. By definition, we have

$$\eta_i = \begin{cases} \xi_i + \xi_{n-i}, & \text{when } i \neq n-i \\ \xi_i, & \text{when } i = n-i. \end{cases}$$

Under the infinite-sites model, one segregating site corresponds to exactly one mutation in the sample genealogy and vice versa. Therefore, under the infinite-sites model, a mutation of type i is simply a segregating site at which the two segregating nucleotides are present in i and $n-i$ sequences, respectively. This means that the value of vector $\boldsymbol{\eta}$ can be found directly from the sample under the infinite-sites model. The value of $\boldsymbol{\xi}$ can also be obtained directly from the sample if the ancestral nucleotide of each segregating site can be determined, for example, by using an outgroup sequence or by phylogenetic reconstruction. FU (1995) derived the means, variances and covariances of ξ_i's and η_i's for a panminctic population. The two estimators

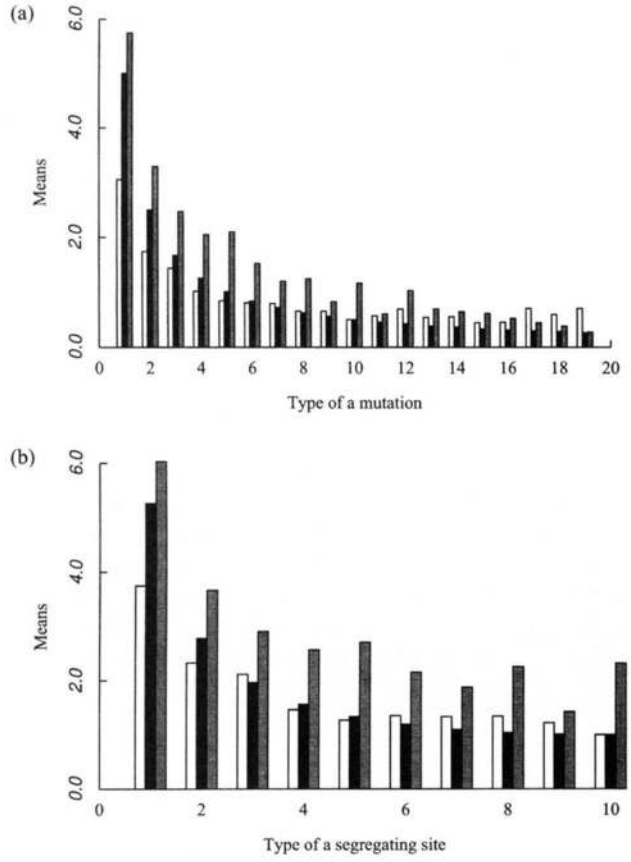

FIGURE 1.—Means of ξ_i (a) and η_i (b) for a sample of 20 sequences. Black bars, panminctic population; white bar; four-islands population with $Nm = 0.125$ and all sequences from a single island; grey bars, four-islands population with $Nm = 0.125$ and five sequences from each island. Five thousand independent samples were simulated to compute the means of ξ_i's and η_i's for the finite-islands model.

$\hat{\theta}_\eta$ and $\hat{\theta}_\xi$ developed by FU (1994a) make use of $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$, respectively.

Because the two arrays $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ contain rich information about θ, it is tempting to use them to construct statistical tests of the neutral model. An informal approach is to compare the expected values of $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ to their observed ones graphically as did in TAJIMA (1989) for detecting natural selection. Figure 1 shows the means of ξ_i and η_i for a sample of 20 sequences from a panminctic population and a subdivided population. There are clear differences between the means of ξ_i as well as the means of η_i for samples from panminctic and subdivided populations. It is also obvious that samples from subdivided populations by different sampling schemes show different patterns of deviation from the panminctic means. This suggests that visual inspection of the frequencies of ξ_i and η_i will be an useful supplement to more formal statistical tests of the neutral model.

FU and LI (1993) proposed several tests that make use of the frequencies of mutations of different classes, one of which is defined as

$$D = \frac{\xi_1 - (K - \xi_1)/(a_n - 1)}{\sqrt{\mathrm{Var}[\xi_1 - (K - \xi_1)/(a_n - 1)]}}. \qquad (7)$$

Since the means, variances and covariances of η_i's and η_i's are now known, one can explore new tests utilizing these frequencies. From statistical point of view, an appealing test statistic using $\boldsymbol{\xi}$ is given by

$$G = \sum_{ij} g^{ij}(\xi_i - \theta/i)(\xi_j - \theta/j),$$

where g^{ij} is the ij element of the inverse of the matrix $\mathrm{Var}(\boldsymbol{\xi})$. Note that θ/i is the expected value of ξ_i. Test of this form is often called *Hotelling test* and is a natural choice for frequencies that do not follow a multinomial distribution. Asymptotically, this test statistics (multiplied by a factor) follows a χ^2 distribution. However, this large sample distribution is not helpful for a finite DNA sample. The main disadvantage of test G is that it requires inversion of a large matrix, which is inconvenient and time consuming. FU (1995) showed that the nondiagonal elements of $\mathrm{Var}(\boldsymbol{\xi})$ are very small in comparison with the diagonal elements, suggesting that $\mathrm{Diag}(\mathrm{Var}^{-1}(\xi_1), \ldots, \mathrm{Var}^{-1}(\xi_{n-1}))$ is close to the inverse of $\mathrm{Var}(\boldsymbol{\xi})$. For this reason, I propose to use the following test statistic

$$G_\xi = \frac{1}{n-1}\sum_{i}^{n-1}\frac{(\xi_i - \theta/i)^2}{\mathrm{Var}(\xi_i)}, \qquad (8)$$

which is the mean of squares of standardized frequencies, and FU (1995) showed that the variance of ξ_i is given by

$$\mathrm{Var}(\xi_i) = \frac{1}{i}\theta + \sigma_i\theta^2,$$

where

$$\sigma_i = \begin{cases} \beta_n(i+1) & \text{if } i < \dfrac{n}{2} \\ 2\dfrac{a_n - a_i}{n-i} - \dfrac{1}{i^2} & \text{if } i = \dfrac{n}{2} \\ \beta_n(i) - \dfrac{1}{i^2} & \text{if } i > \dfrac{n}{2} \end{cases}, \qquad (9)$$

and

$$\beta_n(i) = \frac{2n}{(n-i+1)(n-i)}(a_{n+1} - a_i) - \frac{2}{n-i}. \qquad (10)$$

The analogous test statistic for $\boldsymbol{\eta}$ is given by

$$G_\eta = \frac{1}{\left[\frac{n}{2}\right]}\sum_{i=1}^{[n/2]}\frac{(\eta_i - \theta/\alpha_i)^2}{\mathrm{Var}(\eta_i)} \qquad (11)$$

where θ/α_i is the expected value of η_i and therefore is equal to $\theta[1/i + 1/(n-i)]$ when $i \neq n-i$ and θ/i when $i = n-i$. The computation of $\mathrm{Var}(\eta_i)$ requires

the covariance between ξ_i and ξ_{n-i} ($i \neq n - i$). FU (1995) showed that

$$\mathrm{Cov}(\xi_i, \xi_{n-i}) = \left[\frac{a_n - a_i}{n - i} + \frac{a_n - a_{n-i}}{i} - \frac{\beta_n(i) + \beta_n(n - i + 1)}{2} - \frac{1}{i(n - i)} \right] \theta^2$$

for $i < n - i$.

Because a sample from a population that has an excess of old mutations or a reduction of young mutations or both will result in larger values of G_η and G_ξ than expected under the neutral model, therefore, the neutral model should only be rejected when the value of G_η or the value of G_ξ is large, which suggests that an one-sided test should be used. Note that to use either of the tests, one also has to estimate the value of θ. An estimator whose expectation is insensitive to different models and has relatively small variance would be a good choice because by substituting this estimator for θ, it will result in smaller variances of ξ_i and η_i, which in turn will yield larger values of the test statistics. It is not difficult to see from Table 1 that the EWENS' estimator $\hat{\theta}_E$ is likely to be the best choice and in the rest of the paper $\hat{\theta}_E$ is used as the substitute of θ in these two tests.

DETERMINING THE CRITICAL VALUES OF A TEST

We proposed three new test statistics in the previous section, but in order to use them to test the neutral model, we have to determine their critical values. If the distribution of a test statistic is known, for example for a normal distribution, the critical values can be obtained easily, otherwise perhaps the best one can do is to obtain the critical values from a distribution that is close to that of the test statistic. An approximate to the distribution of a test statistic does not have to be analytical. For example, the empirical distribution of a test statistic is a useful approximation, which can be obtained if samples under the null hypothesis can be simulated by a computer. However, when the distribution of a test statistic has a parameter (θ) whose value is unknown, one must choose some values of the parameter to simulate the samples from which the empirical distribution is compiled. One approach is to obtain the critical values for a number of values of the parameter and to take either the minimum or maximum of these critical values as the critical value of the test, as did by FU and LI (1993) and SIMONSON *et al.* (1995). This approach is appropriate when the critical values do not differ very much for plausible values of θ; otherwise, it may result in a test that is unnecessarily too conservative and consequently become less powerful.

An alternative approach is to obtain an unbiased estimate $\hat{\theta}$ of θ and to calculate the critical values from the empirical distribution derived from samples generated with $\theta = \hat{\theta}$. This approach seems to be logically better than the first one because the inference is made from the most likely distribution of the test statistic. However, this approach is rather burdensome because many samples have to be simulated to obtain the empirical distribution every time the test is used, and presenting the critical values by tables for a large number of combinations of the values of θ and sample size n is clearly out of question. A solution to this dilemma is to summarize the critical values of different values of θ and n by regression analyses so that reasonably accurate critical values can be computed from the regression equation once available. This is the approach used in this paper.

The first step is to obtain the critical values of the four tests S, W, G_η and G_ξ for a large combinations of the values of θ and n. We considered 49 different sample sizes [$n = 10(5)100(10)300$, *i.e.*, 10, 15, 20, ..., 100, 110, ..., 300] and 39 different values of θ [$\theta = 0.2(0.1)1(0.2)3\ (0.5)4(1)20(5)50(10)80$]. For each of the $49 \times 39 = 1911$ combinations, 20,000 independent samples under the neutral model were generated, from which the empirical distribution of each of the four tests was obtained. This is a large scale computer experiment because in total $1911 \times 20{,}000 \approx 3.8$ millions samples have to be generated. What makes such a large-scale simulation possible is the coalescent algorithms (*e.g.*, HUDSON 1982). After the empirical distributions are obtained, the critical values of each test can be determined easily. For example, the critical value for test W at 5% significance level is the maximum value s such that the proportion of samples, in the 20,000 independent samples, with $W \leq s$ is not more than 5%.

The second step of the approach is to summarize the critical values from step one by regression analyses. Let $c_{n,\theta}(\alpha)$ be the critical value of a test at α significance level for given values of θ and n. Consider first the tests S and W. These two statistics are both the probabilities of having no more than the observed number of alleles, so their values are between 0 and 1. Therefore the critical values $c_{n,\theta}(\alpha)$ are also between by 0 and 1, which suggests that the $c_{n,\theta}(\alpha)$ of both tests may be fitted well by the following logistic regression:

$$\log\left[\frac{c_{n,\theta}(\alpha)}{1 - c_{n,\theta}(\alpha)}\right] = \sum_{i+j=0,\ldots,t} u_{ij}(\log n)^i(\log \theta)^j, \quad (12)$$

where t is the degree of the polynomial. Regression analysis of this kind can be performed by many available statistical packages and our analyses were carried out using the *S* package (BECKER *et al.* 1988). We first examined polynomial of degree one, then degree two and so on till we found a polynomial that fit the critical values sufficiently well. We found that polynomials of degree 5 ($t = 5$) were sufficiently accurate to summarize the critical values of both tests ($R^2 \approx 0.998$). The resulting coefficients u_{ij} of both tests for $\alpha = 0.01$, 0.05 and 0.10 are given in Table 2.

TABLE 2

The values of u_{ij} for statistical test S and W

		Test S			Test W		
i	j	1%	5%	10%	1%	5%	10%
0	0	−185.111298	109.082016	15.044644	−3.489784	0.000511	−1.280793
0	1	33.753254	−32.225147	−0.618113	−1.537124	0.764512	0.266559
0	2	−3.015477	3.739171	−0.433905	−0.054939	−0.391339	−0.239910
0	3	0.890137	−0.386910	−0.088669	0.000804	0.133860	0.002830
0	4	−0.034225	0.022612	0.026116	0.026263	−0.030685	0.027011
0	5	−0.001008	0.005592	0.001569	0.002010	0.005321	−0.002389
1	0	261.882416	−146.268860	−21.304512	6.584112	0.337418	1.559456
1	1	−36.910984	35.535301	1.560557	2.832137	−0.150552	0.352623
1	2	−36.910984	35.535301	0.411562	−0.104963	0.183202	0.046478
1	3	−0.539223	0.168822	0.002215	−0.056296	−0.044648	−0.037627
1	4	0.020277	−0.018626	−0.011392	−0.012927	−0.004339	−0.002933
2	0	−143.417999	76.568031	11.873906	−3.509265	0.158736	−0.490870
2	1	14.868989	−14.013808	−0.787185	−1.133289	0.032517	−0.124980
2	2	−0.214840	0.697717	−0.127098	0.069633	−0.023248	0.016870
2	3	0.057726	−0.013337	0.006860	0.015236	0.008780	0.006474
3	0	38.511288	−19.401344	−3.137545	1.016457	−0.076385	0.075944
3	1	−2.527603	2.399835	0.172957	0.199899	0.003182	0.019126
3	2	−0.003391	−0.052053	0.013197	−0.007544	0.000839	−0.002819
4	0	−5.058973	2.398004	0.404350	−0.150550	0.010343	−0.005717
4	1	0.156266	−0.150378	−0.013435	−0.013245	−0.001002	−0.001181
5	0	0.260767	−0.115992	−0.020435	0.008827	−0.000438	0.000170

It should be emphasized that the purpose of our regression analyses is different from that of a typical statistical analysis of data, in which finding a simple relationship between a dependent variable and independent variables is often the focus. Our purpose is to find a regression equation that can be used to regenerate accurate critical values. Therefore, once the degree of polynomial was determined, we did not go further to see if some of the terms in the polynomial can be dropped without significantly reducing the goodness of fit. Also since users of these tests are not expected to calculate the critical values by hand, it does not matter for a regression equation to have a few more terms.

The high accuracy of the regression equation in recovering the critical values of test W can be seen from the two examples in Figure 2 where the observed values of $c_{n,\theta}(0.05)$ and the values computed from the regression equation are plotted for sample sizes $n = 30$ and 100.

From the logistic regression Equation 12, the critical values of test S or test W at significant levels $\alpha = 0.01$, 0.05 and 0.10 can be computed as

$$c_{n,\theta}(\alpha) = \frac{\exp[\sum u_{ij}(\log n)^i(\log \theta)^j]}{1 + \exp[\sum u_{ij}(\log n)^i(\log \theta)^j]}, \quad (13)$$

where u_{ij} are from an appropriate column of Table 2.

If $c_{n,\theta}(\alpha)$ of test W or test S is larger than α, test W or test S will be conservative when the critical value is set to be α; on the other hand, if $c_{n,\theta}(\alpha)$ is smaller than α, the test W or test S will be too liberal when the critical value is set to be α. Figure 3 gives values of $c_{n,\theta}(0.05)$ of test W. It is clear that when θ and sample size n are small, test W will be conservative, and when n are both large, the test W will be too liberal, if one set $c_{n,\theta}(0.05) = 0.05$. Similar pattern was found for test S.

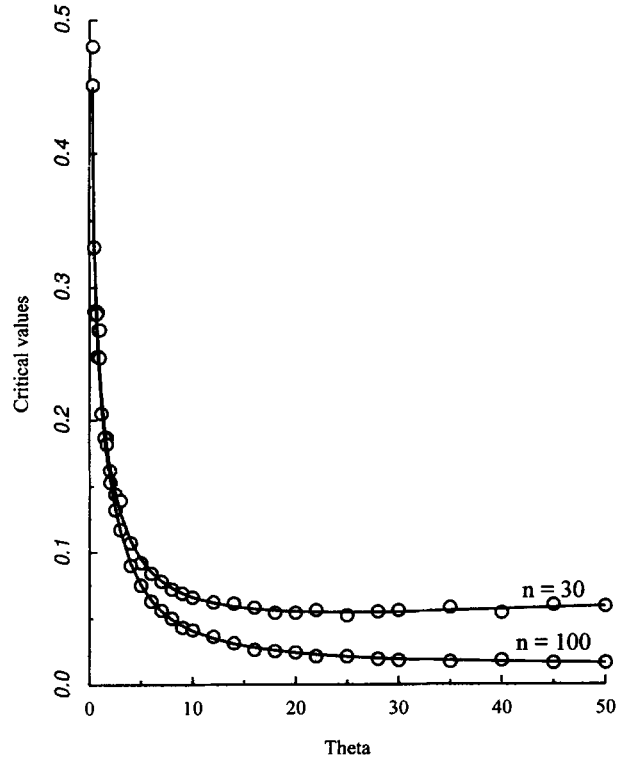

FIGURE 2.—Critical values $c_{n,\theta}(0.05)$ of test W for sample sizes 30 and 100. The circles are observed values and curves are from the regression equation.

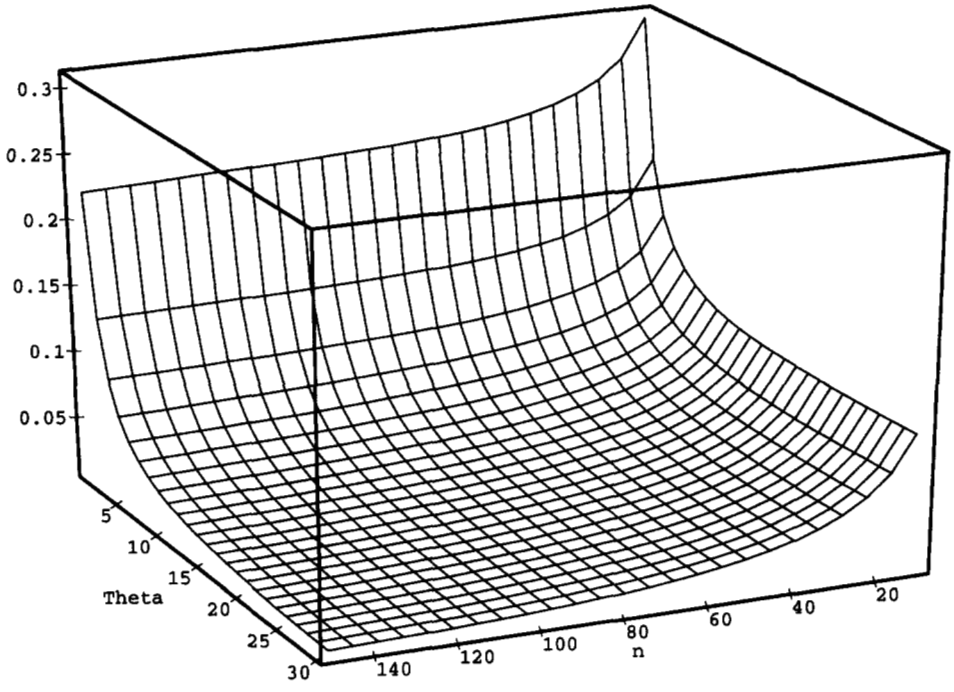

FIGURE 3.—The values $c_{n,\theta}(0.05)$ of test W computed from the regression equation (13).

Next we consider the regression analyses for the critical values of tests G_η and G_ξ. Since both of them are nonnegative and theoretically have no upper limits, we choose the following regression equation

$$\log[c_{n,\theta}(\alpha)] = \sum_{i+j=0,\ldots,t} v_{ij}(\log n)^i(\log \theta)^j. \quad (14)$$

The model fitting was carried out again using the S package and polynomials of 5 degree were found to be sufficiently accurate for both tests ($R^2 \approx 0.99$). The resulting coefficients v_{ij} of both tests are given in Table 3, from which the critical values of these two tests can be computed as

$$c_{n,\theta}(\alpha) = \exp\left[\sum_{i+j=0\cdots5} v_{ij}(\log n)^i(\log \theta)^j\right]. \quad (15)$$

THE ACHIEVED LEVELS OF SIGNIFICANCE

We obtained in the previous section regression equations for computing the critical values of tests S, W, G_η and G_ξ for a given sample size and θ. When performing these tests, we have to substitute estimates for θ. Therefore, we cannot claim yet that the achieved levels of significance with critical values computed from the regression equations will be close enough to the nominal levels of significance and thus the tests are properly constructed. Before we go on to verify the validity of these tests, an examination of the achieved levels of significance of STROBECK's original S (Figure 4) should help to appreciate the difficulty in constructing a proper test of the neutral model. It is clear from Figure 4 that the achieved levels of significance of STROBECK's original test vary considerably for different values of θ, and they can be considerably larger than the nominal level of significance, rendering the test invalid frequently.

To see the effect of estimating θ on the achieved levels of significance of a test, we applied the four tests to samples from panminctic populations simulated independently from those used to compile the critical values. For each simulated sample, we computed the value of statistics S and compared it to the critical value obtained from Equation 13 by substituting $\hat{\theta}_\pi$ for θ. Similar computations were performed for the other three tests.

Figure 5a shows the achieved levels of significance of STROBECK's test S using the critical values computed from (13). We can see that at 5 and 10% nominal levels of significance, the critical values appear to be properly constructed because the achieved levels of significance are close to 5 and 10%, respectively, although the test appears slightly conservative for some values of θ. However, at 1% nominal significant level, the achieved levels of significance are still too high for some values of θ and sample size. This seems due to the fact that the true critical values of the test at 1% significance level are so close to zero such that their estimates from the empirical distributions are not accurate. Because of this drawback, I do not recommend to use test S at 1% significance level.

The relatively large variation in the achieved levels of significance of test S for different values of θ is disappointing, but regression equations are not the causes because they fit the critical values derived from the empirical distributions extremely well. To show this is indeed true, I examined the achieved levels of significance of test S using the critical values from the empirical distribution directly plus interpolations, the results were almost identical to those in Figure 5a. Therefore,

TABLE 3

The values of v_{ij} for tests G_η and G_ξ

		G_η			G_ξ		
i	j	1%	5%	10%	1%	5%	10%
0	0	0.967084	1.752913	−1.050296	−1.479156	−0.219901	0.408483
0	1	1.309209	0.538587	1.662626	1.368190	0.831832	0.543068
0	2	−0.101838	−0.228294	0.135360	−0.114900	0.160098	−0.089139
0	3	−0.010869	0.029290	0.105359	0.031198	0.089510	0.078124
0	4	0.000703	0.003140	0.007196	0.009747	0.001371	0.010644
0	5	0.000475	0.000991	0.000758	0.000426	0.000415	0.000433
1	0	0.955308	−0.758499	1.645483	4.194960	1.730981	0.474256
1	1	−0.669673	0.022883	−1.549584	−0.818370	−0.665446	−0.125440
1	2	0.011283	0.083208	−0.232490	−0.031969	−0.246778	−0.046099
1	3	0.016080	−0.010227	−0.047392	−0.018335	−0.029942	−0.042342
1	4	−0.001604	−0.002664	−0.002974	−0.003218	−0.001296	−0.002971
2	0	−0.699278	0.232020	−0.524068	−2.376007	−0.959304	−0.405576
2	1	0.114329	−0.123740	0.598341	0.214444	0.259561	−0.034746
2	2	−0.005919	−0.010299	0.074490	0.024084	0.071735	0.030626
2	3	−0.001200	0.002367	0.006368	0.003670	0.003329	0.005900
3	0	0.225428	−0.026298	0.039726	0.645058	0.248595	0.149916
3	1	−0.005174	0.029694	−0.105090	−0.031794	−0.049336	0.011864
3	2	0.000885	0.000118	−0.007013	−0.003023	−0.006089	−0.003692
4	0	−0.034200	0.000076	0.007092	−0.085000	−0.031077	−0.025276
4	1	−0.000268	−0.002035	0.006836	0.002063	0.003469	−0.000878
5	0	0.001982	0.000102	−0.000940	0.004371	0.001514	0.001590

it seems that the large variation in the achieved levels of significance of test S for different θ is due to the large variance of $\hat{\theta}_\pi$.

Comparing a and b of Figure 5, we can see that the new test W has achieved levels of significance much closer to the nominal levels of significance than test S does, indicating that this test together with the regression equations for computing its critical values is well constructed. The variation in the achieved levels of significance for different values of θ and n is much smaller than that of test S.

Figure 6 shows the achieved levels of significance of tests G_η and G_ξ. It is clear that the variations in the achieved levels of significance of tests G_η and G_ξ are very similar. These achieved levels of significance are all close to their nominal levels, except for small values of θ ($\theta \leq 1$). Also when sample size is too small relative to the value of θ, the achieved levels of significance tend to be larger than the nominal levels of significance. This is because when θ is large and sample size is small, EWENS' estimator $\hat{\theta}_E$ becomes biased downward, which leads to the use of biased critical values. But how can we judge whether θ is too large for these two tests for a given sample size? A rule of thumb is that it is likely so when $\hat{\theta}_E$ is close to $^2/_3$ of the sample size. Nevertheless, we can conclude that the two tests G_η and G_ξ are in general properly constructed.

We have thus demonstrated that the simulation-regression method for computing the critical values of a test statistic is an effective approach to construct a test and that tests constructed by this approach tend to have achieved levels of significance close to the nominal lev-

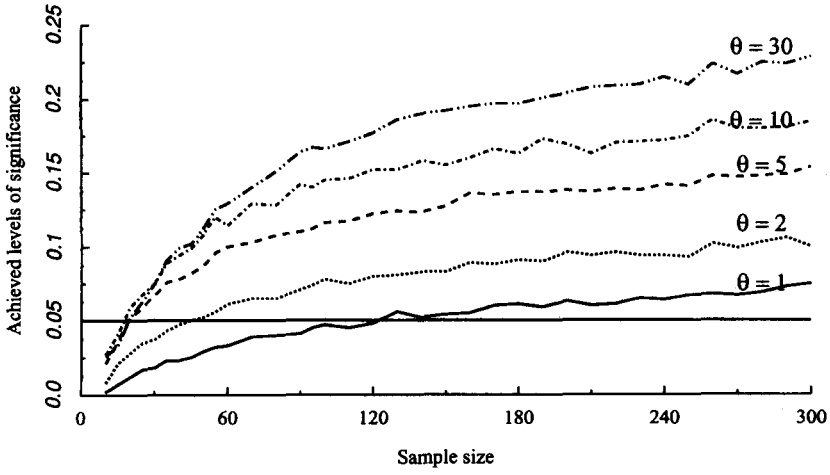

FIGURE 4.—The achieved levels of significance of Strobeck's original test at 5% significance level. Ten thousand independent samples from a panminctic population were simulated for each combination of θ and sample size n.

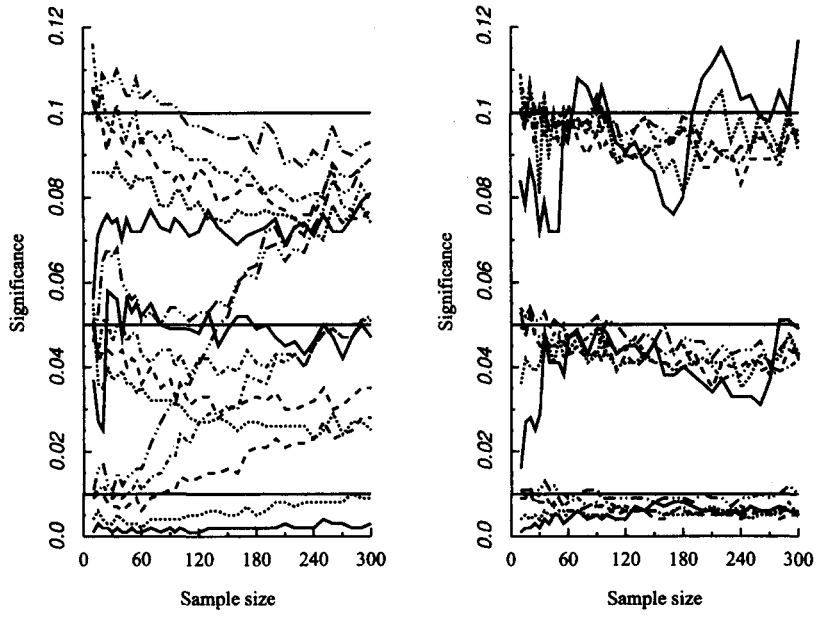

FIGURE 5.—The achieved levels of significance of tests S (a) and W (b) at 1, 5 and 10% nominal significance levels. The five curves at each nominal significance level in each panel correspond to samples with θ = 1, 2, 5, 10 and 30 and their legends are the same as those of Figure 4. Twenty thousand independent samples were simulated for each combination of θ and n.

els of significance for a wide range of values of θ and sample size.

POWERS OF THE TESTS

In this section, we will investigate the powers of the tests S, W, G_η and G_ξ, as well as TAJIMA's test T and the tests by FU and LI (1993), for rejecting the neutral model when it is not true. We shall consider three alternative population genetic models. The first is WRIGHT's finite-islands model with migration, the second is a model with decreasing effective population size and the third is a model of selection and mutation balance. In all these studies, the regression equations [(13) and (15)] are used to compute the critical values of tests S, W, G_η and G_ξ, and similar regression equations (not presented) are used to compute the critical values of TAJIMA's T and FU and LI's tests. Although all the four tests by FU and LI (1993) were examined in our simulations, it was found that their powers are very similar and I will therefore present only the results for test D.

Structured population: Among various models of structured populations, WRIGHT's finite-islands model (WRIGHT 1931) is the one mostly studied. Consider a population that is subdivided into k local subpopulations (islands). WRIGHT's finite-islands model assumes that each individual in an island has probability $m/(k$

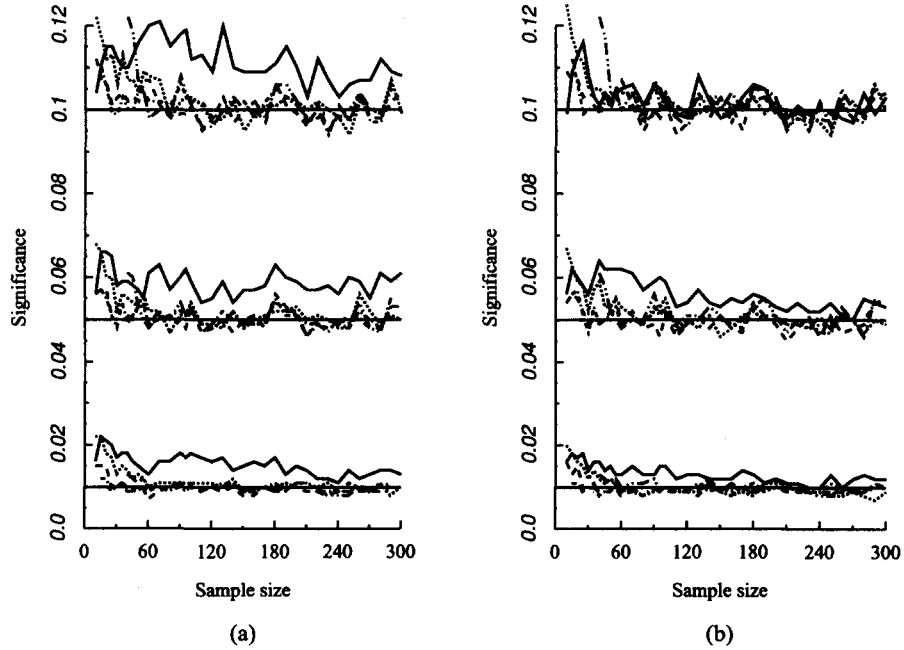

FIGURE 6.—The achieved levels of significance of tests G_η (a) and G_ξ (b) at 1, 5 and 10% nominal significance levels. See footnote to Figure 5 for explanations of legends.

TABLE 4

Powers of detecting population structure when samples are from island populations with $\theta = 5$

Sampling	Islands	Nm	S	W	G_η	G_ξ	D	T
a	2	0.10	0.16	0.13	0.18	0.16	0.15	0.21
	2	0.50	0.05	0.05	0.06	0.06	0.08	0.06
	2	2.00	0.04	0.04	0.05	0.05	0.06	0.06
	5	0.10	0.36	0.31	0.32	0.32	0.39	0.29
	5	0.50	0.08	0.07	0.08	0.08	0.11	0.08
	5	2.00	0.05	0.05	0.06	0.06	0.07	0.06
b	2	0.10	0.19	0.25	0.25	0.26	0.18	0.14
	2	0.50	0.12	0.14	0.14	0.14	0.14	0.11
	2	2.00	0.07	0.08	0.08	0.08	0.11	0.08
	5	0.10	0.38	0.48	0.48	0.49	0.30	0.20
	5	0.50	0.32	0.41	0.37	0.39	0.28	0.22
	5	2.00	0.16	0.19	0.17	0.17	0.20	0.13
c	4	0.10	0.36	0.34	0.35	0.34	0.33	0.36
	4	0.50	0.07	0.08	0.08	0.08	0.11	0.10
	4	2.00	0.04	0.05	0.05	0.05	0.07	0.06
	10	0.10	0.69	0.68	0.61	0.61	0.72	0.54
	10	0.50	0.13	0.14	0.15	0.14	0.19	0.15
	10	2.00	0.05	0.06	0.06	0.06	0.08	0.07
d	4	0.10	0.36	0.56	0.51	0.55	0.33	0.22
	4	0.50	0.29	0.42	0.35	0.37	0.26	0.23
	4	2.00	0.14	0.19	0.16	0.16	0.21	0.16
	10	0.10	0.42	0.54	0.52	0.56	0.34	0.22
	10	0.50	0.49	0.66	0.60	0.64	0.40	0.33
	10	2.00	0.30	0.42	0.33	0.35	0.33	0.25

(a) Sample size is 30 with equal number of sequences drawn from each island. (b) Sample size is 30 with all 30 sequences drawn from a single island. (c) Sample size is 80 with equal number of sequences drawn from each island. (d) Sample size is 80 with all 80 sequences drawn from a single island. Five thousand independent samples were simulated for each combination of parameters.

-1) of migrating to one of the other $k-1$ islands at each generation, where m is the overall migration rate.

We shall consider the case where all the islands have the same effective population size N/k. Samples under WRIGHT's finite-islands model can be simulated using the coalescent algorithm developed by STROBECK (1987). With some modifications, STROBECK's algorithm can also be used to generate samples from circular stepping-stone model (MARUYAMA 1970) and linear-islands model. Since the powers of the six tests were found to be very similar for these models from our preliminary study, we thus focused on WRIGHT's finite-islands model.

Several factors can influence the power of a test to detect population structure, among which are sample size, values of Nm and $\theta = 4N\mu$, number of islands and sampling strategy. Although our main interest here is to identify test(s) that are powerful for detecting hidden population structure from a sample drawn at one geographic location (island), we shall also examine the case of multiple samples for the purpose of comparison.

Table 4 gives the powers of the six tests for detecting population structure for several combinations of parameters. Comparing the results of different settings (a–d), one can see that the powers of these tests are indeed affected by the factors mentioned above. We summarize the results as follows:

- The structure of a sample and the level of migration have strong effects on the powers of these tests. When a subsample of equal size is taken from each island (cases a and c), the power of each test decreases rapidly with the value of Nm and all these tests have little power when Nm is larger than 0.5. On the other hand, when only one sample from a single island is taken (cases b and d), these tests have considerable powers even when Mm is as large as 2.0, and the power of each test does not increase or decrease monotonically with Nm. The powers of these tests are generally larger for the case of a single sample than those for the case of multiple subsamples unless Mm is very small.
- The number of islands also has considerable effect on the powers of these tests. When other things are equal, the powers of these tests appear to increase with the number of islands, but at different rates.
- In the case of a single sample, tests W, G_η and G_ξ are the most powerful tests, and among them W appears to be most powerful when the number of islands is large. Test S is less powerful than these three tests, but overall is more powerful than FU and LI's D test. Among the six tests, TAJIMA's test T is the least powerful one. The difference in the powers of test T and each of the three tests W, G_η and G_ξ is substantial in several combinations of parameters.

- In the case of multiple subsamples of equal size, TAJIMA's test is slightly more powerful than the rest five tests when there are only two islands with low migration rate, but the difference in the powers of these tests is not substantial. TAJIMA's test T gradually becomes the least powerful test with increasing number of islands.

Recently SIMONSEN *et al.* (1995) examined TAJIMA's test T and FU and LI's test D^* for detecting population structure in the case of two populations evolving independently since their separation some time ago, *i.e.*, no migration between them. They studied the situation in which a sample of 25 sequences was drawn from each subpopulation and found that test T is more powerful than test D^*. Their model should be similar to WRIGHT's finite-islands model with two islands and small migration rate. Our study shows that TAJIMA's test T is indeed slightly more powerful than FU and LI's test D (and test D^*, result not shown) in this situation, but as we can see from Table 4, test T becomes less powerful than test D (and test D^*, result not shown) when the number of island is 5. Further simulations (results not shown) showed that even with two islands, TAJIMA's test can be considerably less powerful than the other tests when the two subsamples are of different sizes or when the effective sizes of the two island populations are not the same.

The relatively small powers of these tests in detecting population structure for multiple subsamples do not necessarily mean that it is a better strategy to take only one sample from a single island. HUDSON *et al.* (1992) developed a test of population structure for the case of multiple subsamples and our preliminary study showed that their test is in general more powerful than the six tests considered in this paper. Therefore, when multiple subsamples from different geographic regions (islands) are available and the interest is to test whether there is significant genetic difference among these populations, HUDSON *et al.*'s (1992) test should be preferred.

Population with decreasing effective size: We shall consider a model in which the size of a population changes in a deterministic manner. The coalescent theory for such populations was developed by GRIFFITHS and TAVARÉ (1994), in which exponentially growing populations were examined in detail. An obvious candidate model for us to consider is one in which the effective population size decreases exponentially over generations or grows exponentially looking backward in time. Assume that a sample is taken from the current population and let N_t be the effective population size at the tth generation prior to the current generation. Then this model can be specified by

$$N_t = e^{\beta t} N,$$

where N is the effective size of the present population and $\beta > 0$. However, this model implies that the effective population size at some generations ago was effectively infinite so that there is certain probability that coalescence to the common ancestor does not occur (R. C. GRIFFITHS, personal communication). Because only finite populations are of interest here, we shall examine a model in which the effective population size decreases linearly over generations, *i.e.*,

$$N_t = N(\beta t + 1), \tag{16}$$

where $\beta > 0$. Following GRIFFITHS and TAVARÉ (1994), the kth coalescent time t_k under this model (one unit corresponds to $2N$ generations) has the following density function

$$g(t) = \frac{\binom{k}{2}}{\beta(s_{k+1} + t) + 1} \exp\left[-\binom{k}{2} \times \int_{s_{k+1}}^{s_{k+1}+t} \frac{1}{\beta x + 1}\, dx \right], \tag{17}$$

where $s_{k+1} = t_n + \cdots + t_{k+1}$ with $s_{n+1} = 0$. It is easy to show from $g(t)$ that

$$E(t_k | s_{k+1}) = \int_0^\infty \left(\frac{\beta s_{k+1} + 1}{\beta t + \beta s_{k+1} + 1} \right)^{\binom{k}{2}/\beta} dt, \tag{18}$$

which is finite only when $\beta < \binom{k}{2}$. Therefore, as long as $\beta < 1$, the age of the most recent common ancestor (MRCA) of a sample will be finite.

It is well known that the mean age of the MRCA of a sample is about $4N$ generations in a population of constant effective size (*e.g.*, TAJIMA 1983). When N_t increases with t, the mean age of the MRCA will be larger than $4N$. Table 5 lists, for different values of β, the mean age (MA) of the MRCA and the effective size (RS) of the population at the generation in which the MRCA lived relative to the size of present population. Table 5 shows that both MA and RS increase with β as

TABLE 5

The mean age of the MRCA and the relative size of the common ancestral population

	$n = 20$		$n = 50$	
β	MA	*RS*	MA	*RS*
0.10	1.1	1.2	1.1	1.2
0.20	1.3	1.5	1.3	1.5
0.30	1.5	1.9	1.5	1.9
0.40	1.8	2.4	1.8	2.5
0.50	2.2	3.2	2.3	3.3
0.60	2.8	4.4	3.1	4.7
0.70	3.7	6.2	4.5	7.3
0.80	6.2	10.9	5.7	10.1
0.90	8.4	16.1	8.9	17.1
0.95	11.1	22.1	12.6	25.0

Four thousand samples were simulated for each values of β and n. One unit in MA corresponds to 4N generations.

TABLE 6

Powers of detecting linear change in population sizes at 5% significance level when sample size is 50

β	S	W	G_η	G_ξ	D	T
			$\theta = 3$			
0.10	0.05	0.06	0.07	0.07	0.07	0.07
0.20	0.09	0.09	0.11	0.11	0.10	0.10
0.30	0.11	0.13	0.14	0.14	0.12	0.11
0.40	0.14	0.16	0.18	0.18	0.14	0.14
0.50	0.18	0.21	0.23	0.23	0.18	0.17
0.60	0.21	0.24	0.25	0.15	0.21	0.18
0.70	0.25	0.28	0.30	0.30	0.24	0.21
0.80	0.30	0.33	0.35	0.35	0.28	0.24
0.90	0.32	0.36	0.37	0.37	0.31	0.25
0.95	0.35	0.39	0.39	0.40	0.33	0.26
			$\theta = 10$			
0.10	0.07	0.08	0.08	0.08	0.07	0.08
0.20	0.11	0.11	0.12	0.12	0.10	0.11
0.30	0.14	0.14	0.16	0.17	0.15	0.11
0.40	0.18	0.18	0.20	0.20	0.18	0.15
0.50	0.23	0.25	0.26	0.26	0.23	0.18
0.60	0.26	0.28	0.29	0.30	0.26	0.19
0.70	0.31	0.33	0.34	0.34	0.30	0.21
0.80	0.36	0.38	0.39	0.39	0.35	0.25
0.90	0.38	0.41	0.41	0.42	0.38	0.26
0.95	0.41	0.45	0.44	0.45	0.40	0.27

Four thousand samples are generated for each values of β and θ.

expected, but none of them are sensitive to the sample size. Also note that even with a large value of β, say 0.9, the size of the common ancestral population is only ~17 times as large as the size of present population. A 17-fold reduction in population size over about $9 \times 2N$ generations is certainly possible for populations on their ways to extinction.

Table 6 gives the powers of the six tests for detecting linear change in population size for a sample of 50 sequences with $\theta = 3$ and $\theta = 10$. It is clear that the powers of all these tests increase with the value of β but at different rates. Tests W, G_η and G_ξ are the most powerful among the six tests, trailing closely behind are the tests S and test D. Test T is the least powerful test among the six tests and when β is close to 1, the difference in the powers of test T and the other five tests is considerable. Table 6 also shows that larger value of θ does not improve the powers of these test substantially and in the case of TAJIMA's test, increasing θ from 3 to 10 has nearly no effect on the power of the test.

Mutation and selection balance: Consider a sample of DNA sequences from a neutral locus that is completely linked to a locus of two alleles subject to natural selection, and assume that the frequencies of the two alleles are at equilibrium due to the balance of natural selection and mutation. The coalescent theory and a simulation algorithm under this model were developed by KAPLAN *et al.* (1987). Balance of selection and mutation can be reached in a number of population genetics models, including the deleterious mutation model and the balancing (over-dominant) selection model. Let w_{AA}, w_{Aa} and w_{aa} be the fitness of genotypes AA, Aa and aa, respectively. Then the deleterious mutation model corresponds to the fitness scheme

$$w_{AA} = 1 + s, \quad w_{Aa} = 1 + sh, \quad w_{aa} = 1,$$

and the balancing selection model corresponds to the fitness scheme

$$w_{AA} = 1 - s_1, \quad w_{Aa} = 1, \quad w_{aa} = 1 - s_2,$$

where s, s_1 and s_2 are of order $1/(2N)$ and h is between 0 and 1.

Let $b1 = 2Nu_A$ and $b2 = 2Nu_a$, where u_A and u_a are the mutation rate from A to a and from a to A, respectively. Then the frequency ($x0$) of allele A at equilibrium may depend on both the selection coefficients and the mutation rates for some models but can also be independent of the mutation rates for other models. In general, when the mutation parameters $b1$ and $b2$ are given, $x0$ can still take different values depending on the values of the selection coefficients. Therefore, it is of interest to examine the powers of the six tests for various values of $x0$ for given values of $b1$ and $b2$. Figure 7 plots the powers of these tests with respect to different values of $x0$ for three settings of parameters. The samples used in these comparisons were generated using the coalescent algorithm by KAPLAN *et al.* (1987).

Figure 7 shows that no single test is most powerful

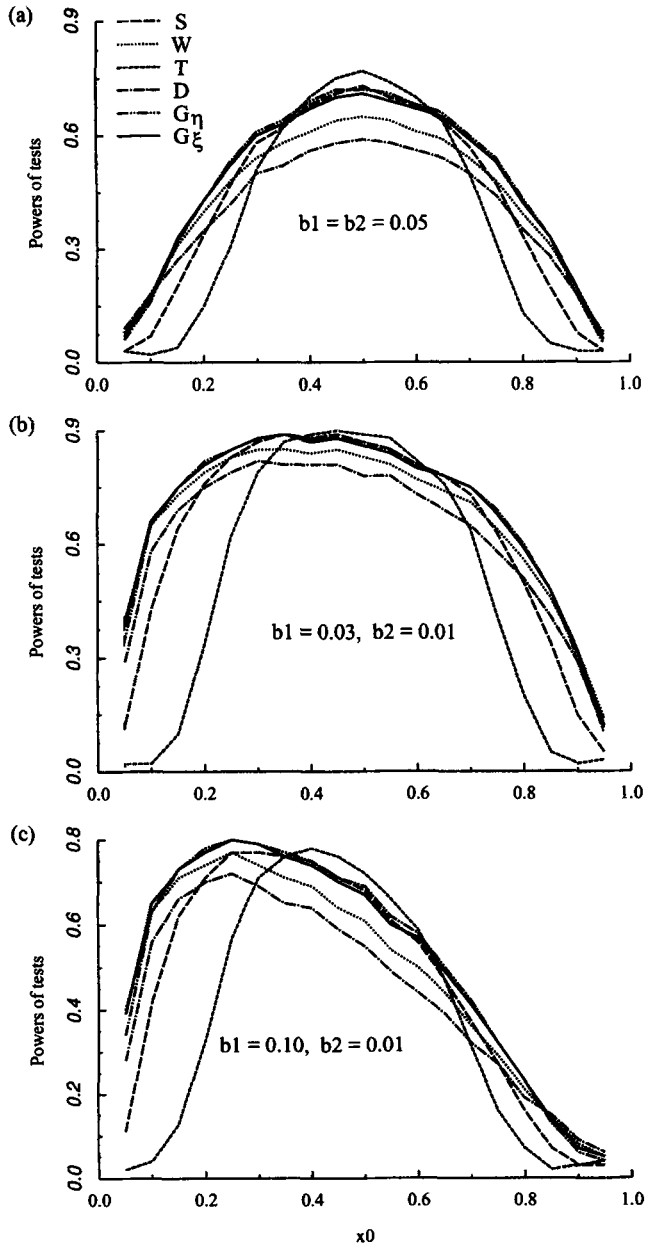

FIGURE 7.—Powers of tests at 5% significance level under mutation-selection balance for a sample of 50 sequences and $\theta = 5$. Four thousand independent samples were generated for each value of $x0$ in each panel.

for all values of $x0$ in each of the three settings. TAJIMA's test T performs very well when $x0$ is ~0.5 but becomes the least powerful test when $x0$ is close to 0 or to 1. In comparison, FU and LI's test D performs well when $x0$ is close to 0 or to 1, but becomes the least powerful tests among the six tests when $x0$ is around 0.5. Figure 7 also shows that tests G_η and G_ξ are very similar in their powers and they are slightly less powerful than TAJIMA's test T when $x0$ is ~0.5 but are substantially more powerful than TAJIMA's test T when $x0$ is close to 0 or to 1. For example, the difference in the powers of test G_η and test T is only 0.05 when $x0 = 0.5$, $b1 = b2 = 0.05$ but is 0.64 when $x0 = 0.1$, $b1 = 0.03$ and $b2 = 0.01$. Test S is as powerful as tests G_η and G_ξ when x_0 is ~0.5 but is less powerful when x_0 is close to 0 or to 1. On the other hand, test W is as powerful as tests G_η and G_ξ when x_0 is close either to 0 or to 1 but is less powerful than G_η, G_ξ and S when $x0$ is around 0.5.

To summarize these results, we conclude that tests G_η and G_ξ are overall the most powerful tests when selection and mutations are balanced. Tests W and S are slightly less powerful than tests G_η and G_ξ. TAJIMA's test T and FU and LI's test D (and D^*, F and F^*, results not shown) have strength and weakness but are overall less powerful than the other four tests.

The models of selection and mutation balance are in many aspects similar to WRIGHT's finite-islands model with migration. Therefore, the relative powers of the six tests (Table 4) in the cases of many islands suggest that under a selection model with more than two alleles, the tests G_η and G_ξ should continue to be the most powerful tests among the six tests, and the difference in the powers of these two tests and TAJIMA's test is expected to be even more substantial.

DISCUSSION AND CONCLUSIONS

We proposed three new statistical tests W, G_η and G_ξ in this paper and reformulated STROBECK's (1987) test S, which was originally designed to detect population structure from a single sample. Despite the fact that the two tests S and W, which are based on EWENS' sampling formula, are very different from the two tests G_η and G_ξ, which are based on the frequencies of mutations of various classes, our simulation and regression approach for determining the critical values of these tests are quite successful. The advantage of this approach is that it can bring the achieved levels of significance close to the nominal levels for a wide range of values of the unknown parameter in the distribution of a test statistic. Therefore, this approach should be useful for determining the critical values of future statistical tests of the neutral model.

The new tests are designed to detect such departures from the neutral model that there is an excess of old mutations or a reduction of young mutations or both. We demonstrated that the four new tests are overall more powerful than TAJIMA's test T and FU and LI's tests using simulated samples from structured populations, populations with linearly decreasing sizes and a model of selection and mutation balance. We found that when a test is powerful for detecting one nonneutral model, it is generally also powerful for detecting other nonneutral models of similar characteristics. The three new tests W, G_η and G_ξ are the most powerful among the tests examined. Among them, test W is slightly more powerful than the other two when a sample is from a structured population, while tests G_η and G_ξ are more powerful than test W when the sample is from a locus that is linked to another locus whose allelic frequencies are at equilibrium due to selection and mutation balance. The modified STROBECK test S is also quite power-

ful, although it is overall less powerful than the three new tests. This may be partly due to the large variation in the achieved levels of significance of this test.

That the powers of tests G_η and G_ξ differ little is quite a surprise because the latter utilizes more information than the former and because FU (1994a) showed that the estimator $\hat{\theta}_\xi$ of θ, which is based on ξ is considerably better than the estimator $\hat{\theta}_\eta$, which is based on η. Since the values of η_i can be found directly from a sample, G_η is easier to compute than G_ξ and therefore should be preferred over G_ξ. However, when using either of the two tests, one should be aware that the sample size needs to be reasonably larger than the value of θ so that EWENS' estimator of θ will be unbiased. All things considered, I recommend W and G_η as general tests of the neutral model against the alternative models that are likely to give rise an excess of old mutations or a reduction of young mutations or both.

It should be pointed out that the infinite-sites model was implicitly assumed when we simulated samples to determine the critical values and to compare the powers of the tests. When multiple hits at some sites of DNA sequences of a sample are evident, some corrections should be done before applying these new tests, as well as TAJIMA's test T and FU and LI's (1993) tests. One effective way to deal with multiple hits is to calculate the values of the variables used by these tests from the sample genealogy estimated by maximum parsimony method. This approach has been used by FU (1994b) to estimate the value of θ from human mitochondria sequences.

We have also assumed in this paper that there is no recombination within the locus from which sequences are obtained. This assumption is likely incorrect for an autosomal locus that is large or consisting of multiple regions. It is therefore important to understand the effects of recombinations on these new tests. We note that recombinations do not change the expectations of $\hat{\theta}_\pi$ and $\hat{\theta}_W$ but reduce their variances and increase the number of alleles in a sample. The larger the number of alleles in a sample is, the less likely the neutral model will be rejected by tests S and W. Therefore, both tests S and W are likely conservative in the presence of recombinations. The effects of recombinations on tests G_η and G_ξ are less clear because recombinations inflate both the value of the numerator and the value of denominator of each term in the summations of G_η and G_ξ. Since both the numerator and the denominator are quadratic functions of θ, it appears that recombinations would affect them by about the same order of magnitude; we thus expect that recombinations do not affect these two tests much unless they are frequent. With more sequences available from large locus or multiple loci, the construction of statistical tests of neutrality of mutations taking recombinations into consideration is an area of considerable importance and deserves further investigations.

I thank Dr. B. GRIFFITHS for discussions on the coalescent algorithm for a sample from a population of changing effective population size and two referees for their comments. This study is supported in part by a First Award from National Institutes of Health.

LITERATURE CITED

BECKER, R. A., J. M. CHAMBERS and A. R. WILKS, 1988 *The New S Language.* Wadsworth, Pacific Grove, CA.

CHAKRABORTY, R., and K. M. WEISS, 1991 Genetic variation of the mitochondrial DNA genome in american indians is at mutation-drift equilibrium. Am. J. Phys. Anth. **86:** 497–506.

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

FU, Y. X., 1994a 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., 1994b A phylogenetic estimator of effective population size or mutation rate. Genetics **136:** 685–692.

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

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

GOLDING, B., and C. STROBECK, 1983 Variance and covariance of homozygosity in a structured population. Genetics **104:** 533–545.

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., 1982 Testing the constant-rate neutral allele model with protein sequence data. Evolution **37:** 203–217.

HUDSON, R. R., D. D. BOOS and N. L. KAPLAN, 1992 A statistical test for detecting geographic subdivision. Mol. Biol. Evol. **9:** 138–151.

KAPLAN, N. L., T. DARDEN and R. R. HUDSON, 1987 The coalescent process in models with selection. Genetics **120:** 819–829.

KARLIN, S., and J. L. MCGREGOR, 1972 Addendum to a paper of W. EWENS. Theor. Popul. Biol. **5:** 95–105.

KINGMAN, J. F. C., 1982a The coalescent. Stochastic Processes Applications **13:** 235–248.

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

KUHNER, M. K., J. YAMATO and J. FELSENSTEIN, 1995 Estimating effective population size and mutation rate from sequence data using METROPOLIS-HASTINGS sampling. Genetics **140:** 1421–1430.

LI, W. H., 1976 Distribution of nucleotide differences between two randomly chosen cistrons in a subdivided population: the finite island model. Theor. Popul. Biol. **10:** 303–308.

MARUYAMA, T., 1970 Analysis of population structure. i. one dimensional stepping-stone models of finite length. Ann. Hum. Genet. **34:** 201–219.

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

SLATKIN, M., 1982 Testing neutrality in a subdivided population. Genetics **100:** 533–545.

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 sequnences 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. Theor. Popul. Biol. **7:** 256–276.

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

WORKMAN, P. L., and J. D. NISWANDER, 1970 Population studies on southwestern indian tribes. ii. local genetic differentiation in the papago. Am. J. Hum. Genet. **22:** 24–49.

WRIGHT, S., 1931 Evolution in mendelian populations. Genetics **16:** 97–159.

ZOUROS, E., 1979 Mutation rates, population sizes, and amounts of electrophoretic variation of enzyme loci in natural populations. Genetics **92:** 623–649.

Communicating editor: G. B. GOLDING