DOI: 10.1534/genetics.104.032797

Gene Genealogy and Properties of Test Statistics of Neutrality Under Population Growth

Akinori Sano* and Hidenori Tachida†,[1]

*Department of Biology, Graduate School of Sciences, Kyushu University, Fukuoka 810-8560, Japan and †Department of Biology, Faculty of Sciences, Kyushu University, Fukuoka 810-8560, Japan

Manuscript received June 24, 2004
Accepted for publication November 13, 2004

ABSTRACT

We consider the Wright-Fisher model with exponential population growth and investigate effects of population growth on the shape of genealogy and the distributions of several test statistics of neutrality. In the limiting case as the population grows rapidly, the rapid-growth-limit genealogy is characterized. We obtained approximate expressions for expectations and variances of test statistics in the rapid-growth-limit genealogy and star genealogy. The distributions in the star genealogy are narrower than those in the cases of the simulated and rapid-growth-limit genealogies. The expectations and variances of the test statistics are monotone decreasing functions of the time length of the expansion, and the higher power of R_2 against population growth is suggested to be due to their smaller variances rather than to change of the expectations. We also investigated by simulation how quickly the distributions of test statistics approach those of the rapid-growth-limit genealogy.

THE coalescent theory is one of the most powerful tools for analyzing DNA sequence data (Kingman 1982; Hudson 1990; Donnelly and Tavaré 1995). One of the most important issues of this development is to relate the genetic diversity with the population genetic processes. For this purpose, some test statistics of neutrality that detect deviations of DNA polymorphisms from neutral expectations have been developed under the assumptions of constant population size, infinite sites, nonoverlapping generations, random mating, and no recombination.

However, either one or more assumptions do not hold in most organismic populations. In recent years, properties of those test statistics with some alternative assumptions have been investigated by a number of authors. For example, Simonsen *et al.* (1995) studied power properties of three test statistics, Tajima's D (1989) and D^* and F^* proposed by Fu and Li (1993), and proposed a new method of constructing critical values of these test statistics. They conducted simulations under several alternatives, that is, a selective sweep event, population bottleneck, and population subdivision, and then they concluded that the most powerful test against the specific alternative hypotheses is Tajima's D-statistic. Also, Ramos-Onsins and Rozas (2002) classified some of those tests into three categories (those based on the distribution of the mutation frequencies, on the haplotype, and on the mismatch distribution) and studied the statistical powers of those tests for two population-growth models (sudden and logistic population-growth models). In addition, they newly defined a statistic R_2 on the basis of the difference between the number of singleton mutations, the average number of nucleotide differences, and the number of segregating sites and concluded that R_2 was the most powerful statistic. The effect of a recent spatial expansion on the pattern of molecular diversity within a deme was studied by Ray *et al.* (2003). They found that both the age of the expansion and the rate of migration between neighboring demes affect the shape of gene genealogies. However, all those studies on the test statistics are based only on simulations.

Here, we investigate effects of the past growth of population size on statistical properties of test statistics of neutrality. One of the reasons why we concentrate on the change of population size is that some data show this violation of the assumptions. For example, Stephens *et al.* (2001) analyzed 3899 single-nucleotide polymorphisms within 313 genes from 82 unrelated human individuals of diverse ancestry. Of the 313 genes analyzed in their study, 281 showed a negative Tajima's D-value. They interpreted this as strong evidence for a recent expansion of the human population. Also, Purugganan and Suddith (1999) analyzed intraspecific sequence variation at the *Apetala3* and *Pistillata* genes of *Arabidopsis thaliana* and a recent rapid expansion was suggested by the result of their analysis. Expansion of the *A. thaliana* population was also suggested by Innan and Stephan (2000).

So far, properties of the test statistics have been investigated in sudden bottleneck, sudden growth, and logistic growth models by computer simulation (Simonsen *et al.* 1995; Ramos-Onsins and Rozas 2002). In this ar-

[1]*Corresponding author:* Department of Biology, Faculty of Sciences, Kyushu University, 4-2-1 Ropponmatsu, Chuo-ku, Fukuoka, 810-8560, Japan. E-mail: htachscb@mbox.nc.kyushu-u.ac.jp

ticle, we mainly analyze an exponential growth model but give some analytical treatments of the problem. We obtain the limiting density of the coalescence time and shape of the genealogy when the population grows rapidly. In this limiting case, approximate expectations and variances of some test statistics, D_T (Tajima 1989), D_{FL}, D^*, F, F^* (Fu and Li 1993), and R_2 (Ramos-Onsins and Rozas 2002), can be calculated for a specified mutation rate μ (the definitions of those statistics are given in the appendix). Those statistics belong to class I statistics that use information from the mutation frequency and two other classes of statistics are listed in Ramos-Onsins and Rozas (2002). Because the class I statistics are more powerful than the other class statistics for detecting population growth, we investigated only the class I statistics here.

To see how quickly those statistics converge to the limiting values, genealogies are generated by computer simulation on the basis of the coalescent (Hudson 1990), assuming finite growth rates of the population. Also the "rapid-growth-limit genealogy" is generated and means and variances of the statistics under these two types of genealogies with given μ are compared to check the approximations. Since Marjoram and Donnelly (1994) stated that exponential population growth makes the genealogy be star-like (see also Rosenberg and Hirsh 2003), we also calculate the expectations and variances of the statistics for the "star genealogy" and compare statistical properties of the test statistics under these genealogies.

THEORY AND METHODS

Coalescent process and growth of population size: We consider a haploid population that evolves according to the Wright-Fisher model with increasing population size. Suppose that there were N_t individuals t generations ago. For any fixed time point r, define the relative size function $\nu_{N_r}(x)$ by

$$\nu_{N_r}(x) = \frac{N_{[N_r x]}}{N_r} = \frac{N_t}{N_r}, \quad \frac{t}{N_r} \le x < \frac{t+1}{N_r}, \quad t = 0, 1, \ldots,$$

where $[x]$ denotes the greatest integer less than or equal to x. We define the coalescence time T_i as the time interval leading from a coalescent with i alleles to that with $i-1$ alleles and T as the time interval from the present up to the most recent common ancestor of the whole sample. Obviously, $T = \sum_{i=2}^{n} T_i$ holds. Griffiths and Tavaré (1994) showed that a diffusion approximation of the Wright-Fisher model exists for the deterministic fluctuation ν_{N_r}, assuming that ν_{N_r} is approximated by the population size function ν with a continuous-time parameter as the limit as N_r tends to infinity. Here, time is scaled in units of N_r generations. They also showed that the joint density of $(T_n, \ldots, T_2)$ is given by

$$g(t_n, \ldots, t_2) = \prod_{i=2}^{n} \frac{c_i}{\nu(x_i)} e^{-c_i(I(x_i)-I(x_{i+1}))}, \tag{1}$$

where

$$c_i = \frac{i(i-1)}{2},$$

$$I(t) = \int_0^t \frac{1}{\nu(u)}\, du,$$

and

$$x_i = t_i + t_{i+1} + \ldots + t_n, \quad x_{n+1} = 0$$

(for more intuitive derivation, see Slatkin and Hudson 1991). Here, we consider a process under the exponential population-growth scenario; that is, we assume that the population size, originally of size N_A, increases as an exponential function. Namely, the relative size function can be written as

$$\nu(t) = \begin{cases} e^{\alpha(t_e - t)} & \text{if } t < t_e, \\ 1 & \text{otherwise,} \end{cases}$$

where t_e denotes the time interval of the population growth and t and t_e are measured in units of N_A generations. This model was considered by Weiss and Von Haeseler (1998) and Prichard *et al.* (1999). By (1), the joint density under the exponential population growth is given as follows:

$$g(t_n, \ldots, t_2) = \begin{cases} \left(\prod_{i=2}^{n} \frac{c_i}{e^{\alpha(t_e - x_i)}}\right) \exp\left(-\frac{1}{\alpha}\sum_{i=2}^{n}(i-1)e^{\alpha(x_i - t_e)}\right), \\ \qquad \text{if } x_2 < t_e \\ \left(\prod_{i=2}^{n} c_i\right)\left(\prod_{i=j+1}^{n} \frac{1}{e^{\alpha(t_e - x_i)}}\right) \\ \quad \times \exp\left(c_j(t_e - x_j) - \sum_{i=2}^{j}(i-1)x_i\right) \\ \quad \times \exp\left(\frac{c_n e^{-\alpha t_e} - c_j}{\alpha} - \sum_{i=j+1}^{n}\left\{\frac{i-1}{\alpha}e^{-\alpha(t_e - x_i)}\right\}\right), \\ \qquad \text{if } x_{j+1} < t_e \le x_j \\ \left(\prod_{i=2}^{n} c_i\right)\exp\left(-\sum_{i=2}^{n}(i-1)(x_i - t_e) - \frac{c_n}{\alpha}(1 - e^{-\alpha t_e})\right), \\ \qquad \text{if } t_e \le x_n. \end{cases} \tag{2}$$

As α tends to infinity, we obtain

$$g(t_n, \ldots, t_2) \to \begin{cases} 0, \\ \qquad \text{if } x_n < t_e \\ \left(\prod_{i=2}^{n} c_i\right)\exp\left(-\sum_{i=2}^{n}(i-1)(x_i - t_e)\right), \\ \qquad \text{if } t_e \le x_n. \end{cases} \tag{3}$$

If we set $y_i = x_i - t_e$, the limiting joint density of y_i as α tends to infinity is the same as the joint density under the constant population size. This limiting joint density indicates that there is no coalescence from the present to time t_e, and the coalescent process before the population expansion is the same as the process under the case of constant population size with size N_A. In other

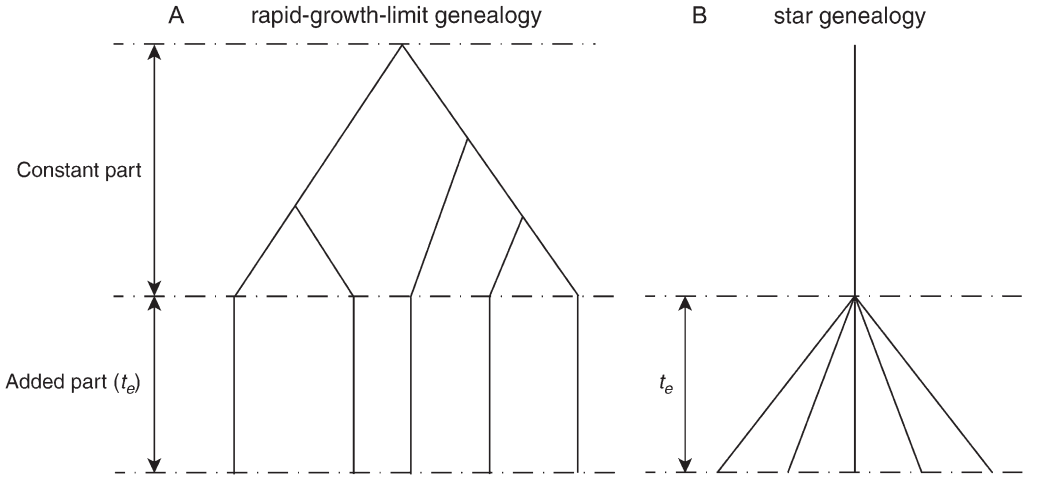

FIGURE 1.—The shape of rapid-growth-limit genealogy and star genealogy.

words, the limiting genealogy is represented by adding a branch of length t_e to each external branch of the genealogy under constant population size (Figure 1A). We call this limiting genealogy the rapid-growth-limit genealogy. We can also calculate the limiting values of the expectation of T_{n-m} $(0 \le m < n-1)$ as α tends to infinity,

$$E_n[T_{n-m}] \to \frac{1}{c_{n-m}} \quad \text{as } \alpha \to \infty \ (m > 0), \tag{4}$$

$$E_n[T_n] \to t_e + \frac{1}{c_n} \quad \text{as } \alpha \to \infty. \tag{5}$$

By (4) and (5), the limiting value of the expectation of T can be computed as

$$E_n[T] \to t_e + 2\left(1 - \frac{1}{n}\right) \quad \text{as } \alpha \to \infty. \tag{6}$$

The second term of the right-hand side of (6) is equal to the expectation of T under constant population size. When the population size is large, it can be thought intuitively that any coalescences are difficult to occur in the expansion period, so the rapid-growth-limit genealogy has no coalescence in that period.

Averages and variances of the statistics: We study properties of six test statistics with a given mutation rate under three different genealogies, the genealogy for the exponential population-growth model (the "simulated genealogy"), genealogy constructed from the limiting joint density of coalescence times as α tends to infinity (the rapid-growth-limit genealogy), and the star genealogy. For the rapid-growth-limit genealogy and star genealogy, we can compute approximate averages and variances of the test statistics as shown below.

Rapid-growth-limit genealogy: We divide the rapid-growth-limit genealogy into two parts, a genealogy with constant population size (constant part) and added branches (added part) as illustrated in Figure 1A. Assume that X_i $(1 \le i \le n)$ is the number of mutations in the added branch of the ith sample and write $Y = \sum_{i=1}^{n} X_i$. The number of nucleotide differences, d_{ij}, between the ith and jth samples is given by

$$d_{ij} = d_{ij}^{(c)} + X_i + X_j,$$

where $d_{ij}^{(c)}$ denotes the number of nucleotide differences between the ith and jth samples in the constant part. Then the average number of pairwise differences between sample sequences $\hat{k}$ can be written as

$$\hat{k} = (1/\tbinom{n}{2}) \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d_{ij} = (1/\tbinom{n}{2}) \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (d_{ij}^{(c)} + X_i + X_j)$$
$$= \hat{k}^{(c)} + \frac{2}{n} Y,$$

where $\hat{k}^{(c)}$ denotes the average number of nucleotide differences due to the constant part. The distribution of each X_i is approximately Poisson with a mean μt_e (KINGMAN 1982). By letting $\theta = 2N_A\mu$ and using d_i $(i = 1, 2)$ given in the APPENDIX, the average and variance of $\hat{k}$ are given as

$$E[\hat{k}] = \theta + 2\mu t_e, \tag{7}$$

$$V[\hat{k}] = V[\hat{k}^{(c)}] + V\left[\frac{2}{n} Y\right] + 2\,\mathrm{Cov}\left[\hat{k}^{(c)}, \frac{2}{n} Y\right]$$
$$= d_1\theta + d_2\theta^2 + \frac{4}{n}\mu t_e, \tag{8}$$

since the covariance term vanishes. Although the mutation parameter θ is usually defined using the current population size, we here define it using the initial size N_A since the current population size tends to infinity in the rapid-growth limit. The number of segregating sites S in the sample is written as

$$S = S^{(c)} + Y,$$

where $S^{(c)}$ denotes the number of segregating sites in the constant part. The expectation and variance of S are given by

$$E[S] = E[S^{(c)}] + E[Y] = a_n\theta + n\mu t_e, \quad (9)$$

$$V[S] = V[S^{(c)}] + V[Y] = a_n\theta + b_n\theta^2 + n\mu t_e. \quad (10)$$

Similarly, the covariance between $\hat{k}$ and S is given by

$$\mathrm{Cov}(\hat{k}, S) = \theta + \left(\frac{1}{2} + \frac{1}{n}\right)\theta^2 + 2\mu t_e. \quad (11)$$

Therefore, by (7)–(11), we can obtain the expectations of the numerator and the square of the denominator of D_T as

$$E\left[\hat{k} - \frac{S}{a_n}\right] = \mu t_e\left(2 - \frac{n}{a_n}\right), \quad (12)$$

$$E[e_1S + e_2S(S-1)] = (e_1 - e_2)(a_n\theta + n\mu t_e) + e_2(a_n\theta + b_n\theta^2 + n\mu t_e + (a_n\theta + n\mu t_e)^2). \quad (13)$$

An approximate expression for the expectation of D_T is obtained by dividing (12) by the square root of (13). The approximation and the δ-method used later give good estimates if the coefficients of variation of respective variables are much smaller than one. We check the accuracy of the approximation by simulation later. Similarly, we obtain the expectations of the numerators and squares of the denominators of the other test statistics except R_2 as

$$E[S - a_n\eta_e] = n\mu t_e(1 - a_n), \quad (14)$$

$$E\left[\frac{n}{n-1}S - a_n\eta_s\right] = n\mu t_e(1 - a_n), \quad (15)$$

$$E[\hat{k} - \eta_e] = \mu t_e(2 - n), \quad (16)$$

$$E\left[\hat{k} - \frac{n-1}{n}\eta_s\right] = \mu t_e(3 - n), \quad (17)$$

$$E[uS + vS^2] = u(a_n\theta + n\mu t_e) + v(a_n\theta + b_n\theta^2 + n\mu t_e + (a_n\theta + n\mu t_e)^2). \quad (18)$$

When we calculate values for D_{FL}, D^*, F, and F^*, we put (u_D, v_D), (u_{D^*}, v_{D^*}), (u_F, v_F), and (u_{F^*}, v_{F^*}), respectively, into (u, v) of (18). The expectation of the denominator of R_2 is given by (9) and the expectation of the square of the numerator can be calculated by using the equation

$$E\left[\sum_{i=1}^{n}\left(\xi_i - \frac{\hat{k}}{2}\right)^2\right] = \left(\frac{2n^2 + (18a_{n+1} - 25)n + 6}{9n(n-1)}\right)\theta^2 + \left(\frac{n^2 + (12 - 6a_{n+1})n - 6a_{n+1} - 1}{3(n-1)^2} - \frac{\mu t_e(n^2 - 2n - 2)}{n}\right)\theta + \mu t_e(1 + (5 - 2n)\mu t_e). \quad (19)$$

In those calculations, we used an approximation of a form,

$$E\left[\frac{x_1}{x_2}\right] \approx \frac{E[x_1]}{E[x_2]}.$$

If x_1 and x_2 are independent, the left-hand side of the equation equals $E[x_1] \times E[1/x_2]$ and this shows the right-hand side of the equation underestimates its left-hand side in absolute values. Although we know x_1 and x_2 are dependent, this argument seems to hold approximately as shown in Table 1. Indeed, absolute values of most estimated values for the rapid-growth-limit genealogy are smaller than the values of simulations.

We can also compute covariances between any pair of random variables $\hat{k}$, S, η_e, η_s, and ξ_i; then we are able to obtain approximate estimates of variances of test statistics by the δ-method,

$$V\left[\frac{x_1}{x_2}\right] \approx \left(\frac{E[x_1]}{E[x_2]}\right)^2\left(\frac{V[x_1]}{(E[x_1])^2} + \frac{V[x_2]}{(E[x_2])^2} - \frac{2\,\mathrm{Cov}[x_1x_2]}{E[x_1]E[x_2]}\right) \quad (20)$$

(Stuart and Ord 1987).

Star genealogy: Since all mutations are singletons in the star genealogy (Figure 1B), we have a relation between $\hat{k}$ and S,

$$\hat{k} = \frac{2}{n}S,$$

and obviously $\eta_e = \eta_s = S$ hold. Then the test statistics can be written as a function of S as follows:

$$D_T = \frac{(2/n - 1/a_n)}{\sqrt{(1/S)(e_1 - e_2) + e_2}}, \quad (21)$$

$$D_{FL} = \frac{1 - a_n}{\sqrt{(1/S)u_D + v_D}}, \quad (22)$$

$$D^* = \frac{n/(n-1) - a_n}{\sqrt{(1/S)u_{D^*} + v_{D^*}}}, \quad (23)$$

$$F = \frac{(2/n) - 1}{\sqrt{(1/S)u_F + v_F}}, \quad (24)$$

$$F^* = \frac{(3/n) - 1}{\sqrt{(1/S)u_{F^*} + v_{F^*}}}. \quad (25)$$

Let Z_i $(i = 1, 2, \ldots, n)$ be the number of mutations on the external branch of the ith sample. We can approximate the distribution of Z_i as Poisson with mean μt_e (Kingman 1982); then the distribution of $S = \sum_{i=1}^{n} Z_i$ is also Poisson with mean $n\mu t_e$. Because (21)–(25) contain only one random variable S and each equation has the same form,

$$Q = \frac{z}{\sqrt{(1/S)x + y}}, \quad (26)$$

it is easy to obtain distributions of the test statistics except for R_2. The distribution is given by

$$P(Q = l) = \frac{1}{\gamma!}e^{-n\mu t_e}(n\mu t_e)^\gamma, \quad (27)$$

where

TABLE 1

The expectations and variances of test statistics

		D_T		D_{FL}		D^*		F		F^*		R_2	
t_e	α	Ave.	Var.	Ave.	Var.	Ave.	Var.	Ave.	Var.	Ave.	Var.	Ave.	Var.
Constant		−0.041	0.901	−0.033	0.973	−0.030	0.950	−0.040	1.167	−0.028	0.721	0.20209	0.00600
0.5	0.5	−0.121	0.873	−0.121	1.011	−0.107	0.969	−0.142	1.200	−0.128	1.126	0.14936	0.00183
	1	−0.201	0.844	−0.212	1.042	−0.187	0.982	−0.245	1.223	−0.220	1.131	0.14458	0.00170
	3	−0.509	0.739	−0.624	1.112	−0.553	0.995	−0.698	1.267	−0.623	1.120	0.12816	0.00133
	7	−0.902	0.595	−1.308	1.058	−1.150	0.898	−1.411	1.186	−1.248	1.000	0.10895	0.00093
	∞ (sim)	−1.294	0.444	−2.170	0.842	−1.884	0.673	−2.276	0.952	−1.985	0.764	0.09071	0.00059
	∞ (cal)	−1.232	0.505	−2.124	0.898	−1.860	0.728	−2.163	0.721	−1.914	0.772	0.08106	0.00127
	Star	−1.464	0.265	−2.176	0.686	−2.018	0.544	−2.232	0.716	−2.146	0.607	0.09747	0.00033
1	0.5	−0.205	0.817	−0.206	1.032	−0.180	0.968	−0.242	1.204	−0.215	1.108	0.14271	0.00161
	1	−0.395	0.732	−0.434	1.068	−0.376	0.959	−0.499	1.219	−0.440	1.080	0.13158	0.00132
	3	−1.059	0.460	−1.463	0.998	−1.257	0.810	−1.601	1.074	−1.389	0.870	0.09782	0.00064
	7	−1.522	0.305	−2.512	0.719	−2.130	0.546	−2.649	0.775	−2.266	0.594	0.07710	0.00036
	∞ (sim)	−1.738	0.237	−3.099	0.532	−2.603	0.390	−3.220	0.587	−2.729	0.438	0.06789	0.00025
	∞ (cal)	−1.685	0.261	−3.052	0.560	−2.597	0.415	−3.098	0.470	−2.663	0.440	0.05742	0.00048
	Star	−1.900	0.083	−3.081	0.339	−2.858	0.214	−3.081	0.347	−2.858	0.230	0.06892	0.00040
3	0.5	−0.520	0.574	−0.597	1.002	−0.495	0.843	−0.685	1.108	−0.581	0.926	0.11795	0.00087
	1	−1.170	0.300	−1.596	0.864	−1.308	0.646	−1.763	0.881	−1.471	0.663	0.08664	0.00039
	3	−2.065	0.085	−3.746	0.314	−3.003	0.210	−3.890	0.311	−3.171	0.213	0.04795	0.00009
	7	−2.222	0.062	−4.261	0.190	−3.397	0.124	−4.374	0.197	−3.548	0.133	0.04205	0.00006
	∞ (sim)	−2.278	0.054	−4.456	0.152	−3.542	0.098	−4.458	0.162	−3.686	0.109	0.03997	0.00006
	∞ (cal)	−2.246	0.057	−4.416	0.155	−3.575	0.101	−4.458	0.132	−3.642	0.106	0.03134	0.00009
	Star	−2.549	.001 >	−5.006	0.014	−3.971	0.004	−4.344	0.073	−3.734	0.027	0.03979	0.00045

The expectations and variances of test statistics in the simulated genealogy, rapid-growth-limit genealogy (α = ∞), and star genealogy are shown. The values in the simulated genealogy and limit (sim) are generated by simulation. The values in the star genealogy and the value limit (cal) are obtained by formulas (19)–(28) in the text. Ave., average; Var, variance.

$$\gamma = x\left(\frac{z^2}{l^2} - y\right)^{-1}.$$

To compute the statistic R_2, however, we need each value of Z_i, so it is not possible to express it as a function of S. The expectation of R_2 is represented by

$$E[R_2] = \frac{((1/n)(n-1)\mu t_e)^{1/2}}{n\mu t_e}, \qquad (28)$$

and the variance can be obtained approximately by the δ-method.

SIMULATION RESULTS

We examined effects of population growth on the expectations, variances, shapes of distributions, and critical values of the six test statistics for various values of n, t_e, α, and $\theta = 2N_A\mu$. We simulated the polymorphism data by using methods based on HUDSON (1990). For the rapid-growth limit, we generated its genealogy by adding branches to the constant genealogy and then assigned mutations. Although we investigated the distribution for several parameter sets, typical examples are shown in the figures with fixed parameters $n = 20$ and $\theta = 1$.

Figure 2 shows the effect of α on the distributions of each test statistic with $t_e = 3$. The parameter α represents the magnitude of expansion. It can be observed that the distributions of all statistical tests are shifted toward the negative side, kurtosises of distributions become higher, and shapes of the distributions converge to those of the rapid-growth-limit genealogies with increasing α. We can see that R_2 has the narrowest distribution, and the variances converge to those of the rapid-growth-limit genealogies more rapidly in D_T and R_2 than in the other test statistics. The distributions under the star genealogy are not shown in this figure since the variances are so small that the distributions are much narrower than those in the case of the simulated and rapid-growth-limit genealogies. In Figure 3, the distribution of D_T when one of the parameters n, θ, or t_e is changed is illustrated. Generally the same tendency as those illustrated in Figure 2 can be seen in those distributions. Although the convergence to the rapid-growth limit is slightly slower for smaller t_e, the distributions converge to the limits (see also Table 1). Small sample size and short time length of the expansion result in higher kur-

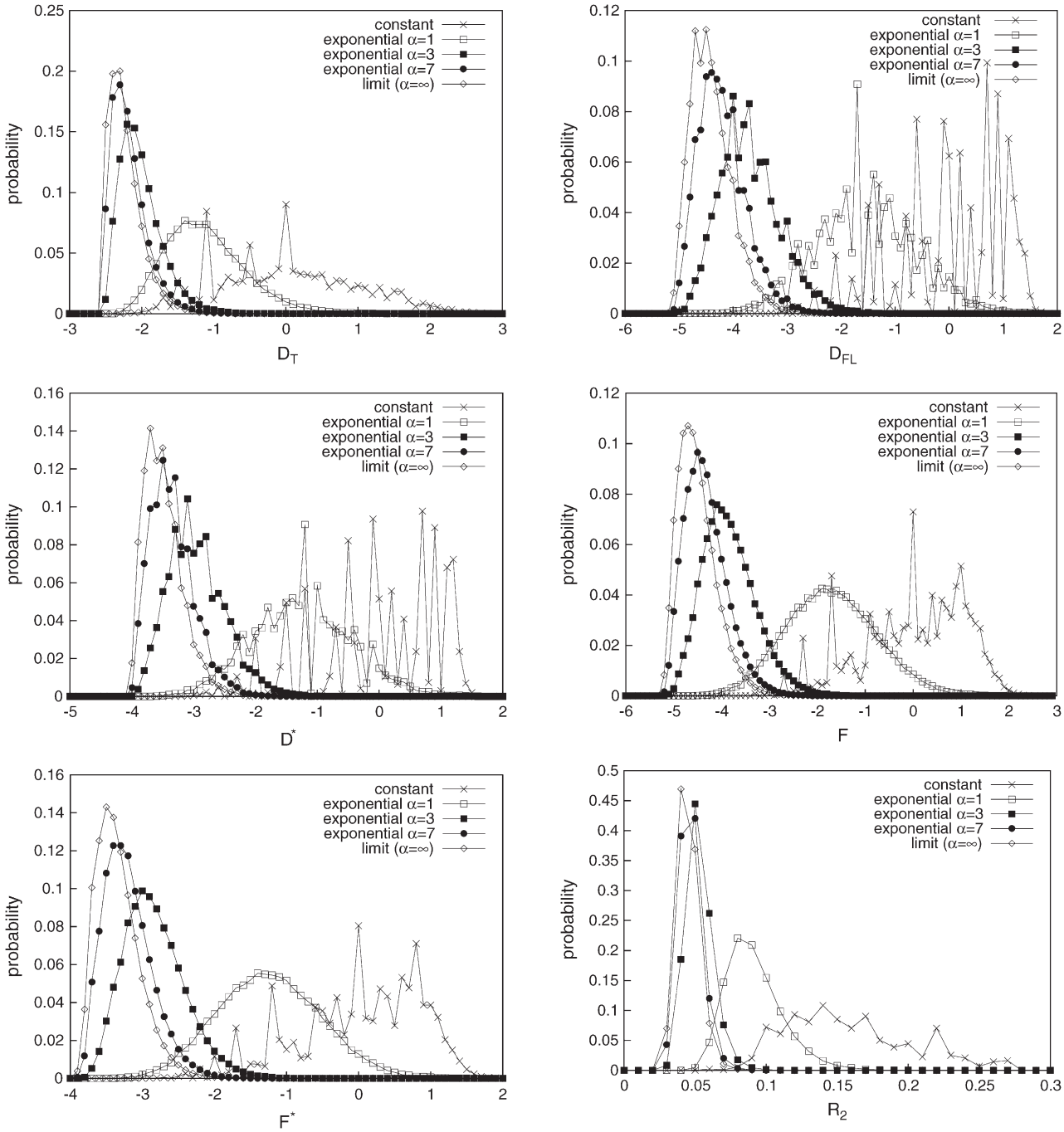

FIGURE 2.—The effects of exponential population growth on the distributions of test statistics. Data are based on 100,000 simulations with parameters $n = 20$; $\theta = 1$; $t_e = 3$; and $\alpha = 1, 3, 7$. The distribution for the rapid-growth-limit genealogy is also illustrated.

tosises of the distribution, and smaller mutation rates make the distribution more rugged.

The averages and variances of test statistics under exponential population growth and star genealogy are presented in Table 1. Unless the values of the parameters t_e and α are small, the values for the rapid-growth-limit genealogy are similar to the values of the simulated genealogy in both average and variance. Moreover, the averages for the star genealogy are similar to those for the simulated genealogy, but their variances are one order less than those of the simulated and rapid-growth-limit genealogies. We also note a difference among test statistics in the amount of the reduction of variances caused by population expansion. The variance of R_2 is 0.006 for constant population size and 0.0004 for the rapid-growth-limit genealogy when $t_e = 1$. Therefore, ~93% reduction is recognized. In the other test statistics (D_T, D_{FL}, D^*, F, and F^*), amounts of reduction are <75% (74, 45, 59, 50, and 40%, respectively).

We also investigated the power of the test statistics

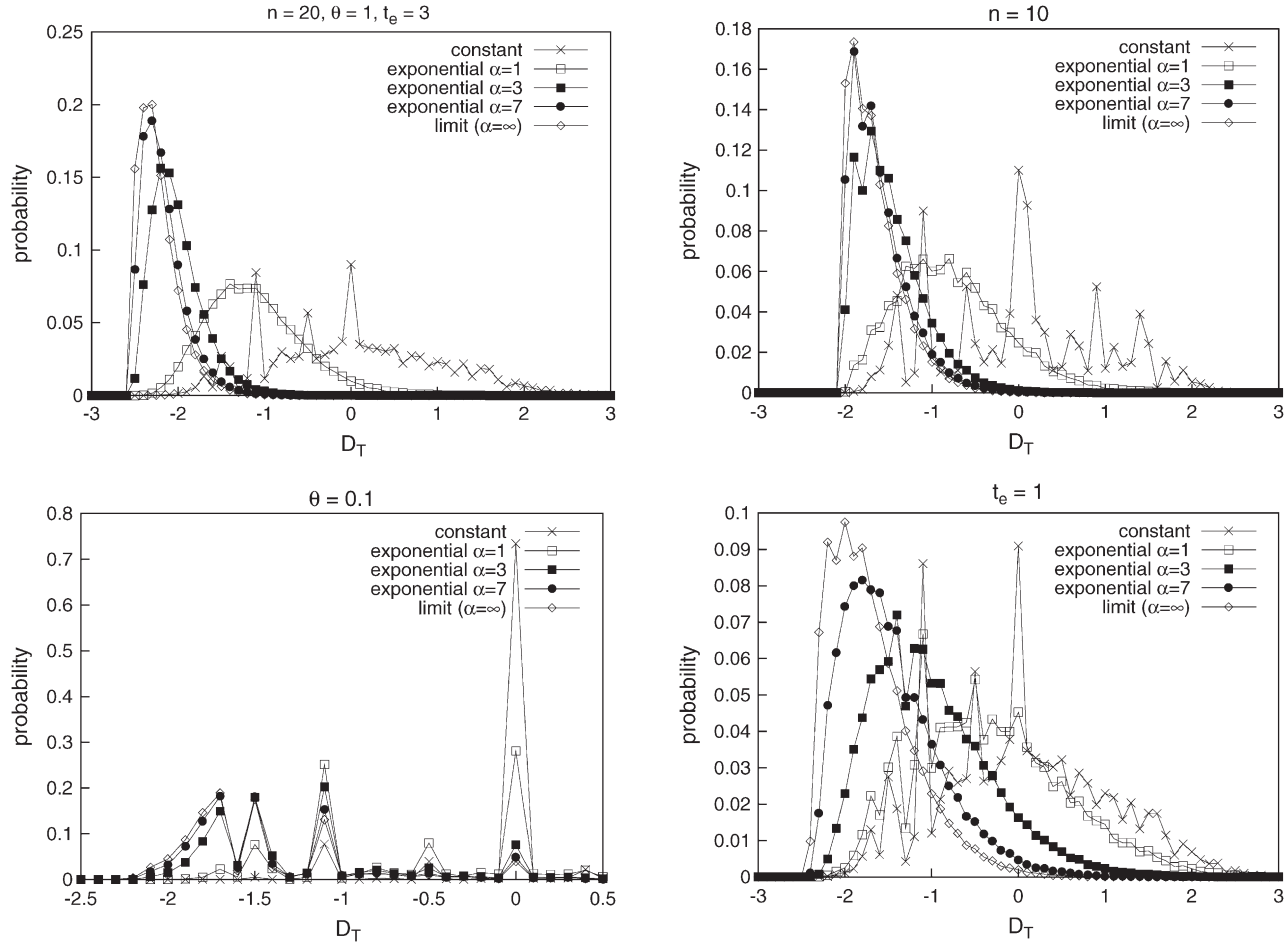

FIGURE 3.—The effects of changing one of the parameters (n, θ, or t_e) on the distributions of Tajima's D-statistics.

and some of the results are shown in Figure 4. First, data were generated assuming the exponential-growth model and then 95% confidence intervals for given S of respective statistics were obtained to compute their powers by computer simulation. We do not recognize much difference among the powers of the test statistics except R_2. In RAMOS-ONSINS and ROZAS (2002), R_2 was characterized as the most powerful test statistics for sudden and logistic population-growth models, and Figure 4 shows that R_2 is also most powerful for the exponential popula-

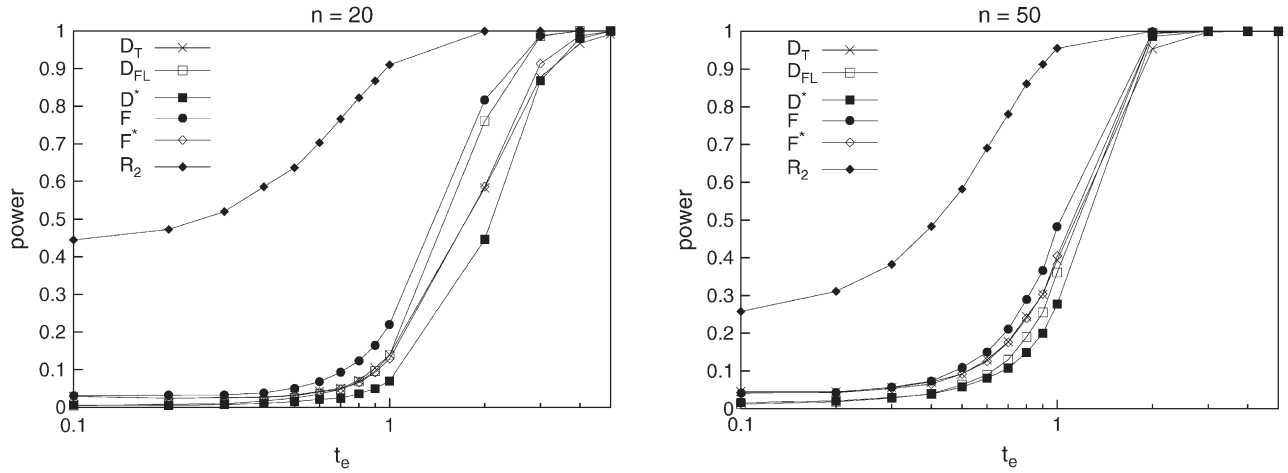

FIGURE 4.—The power of test statistics with parameters θ = 1 and α = 3.

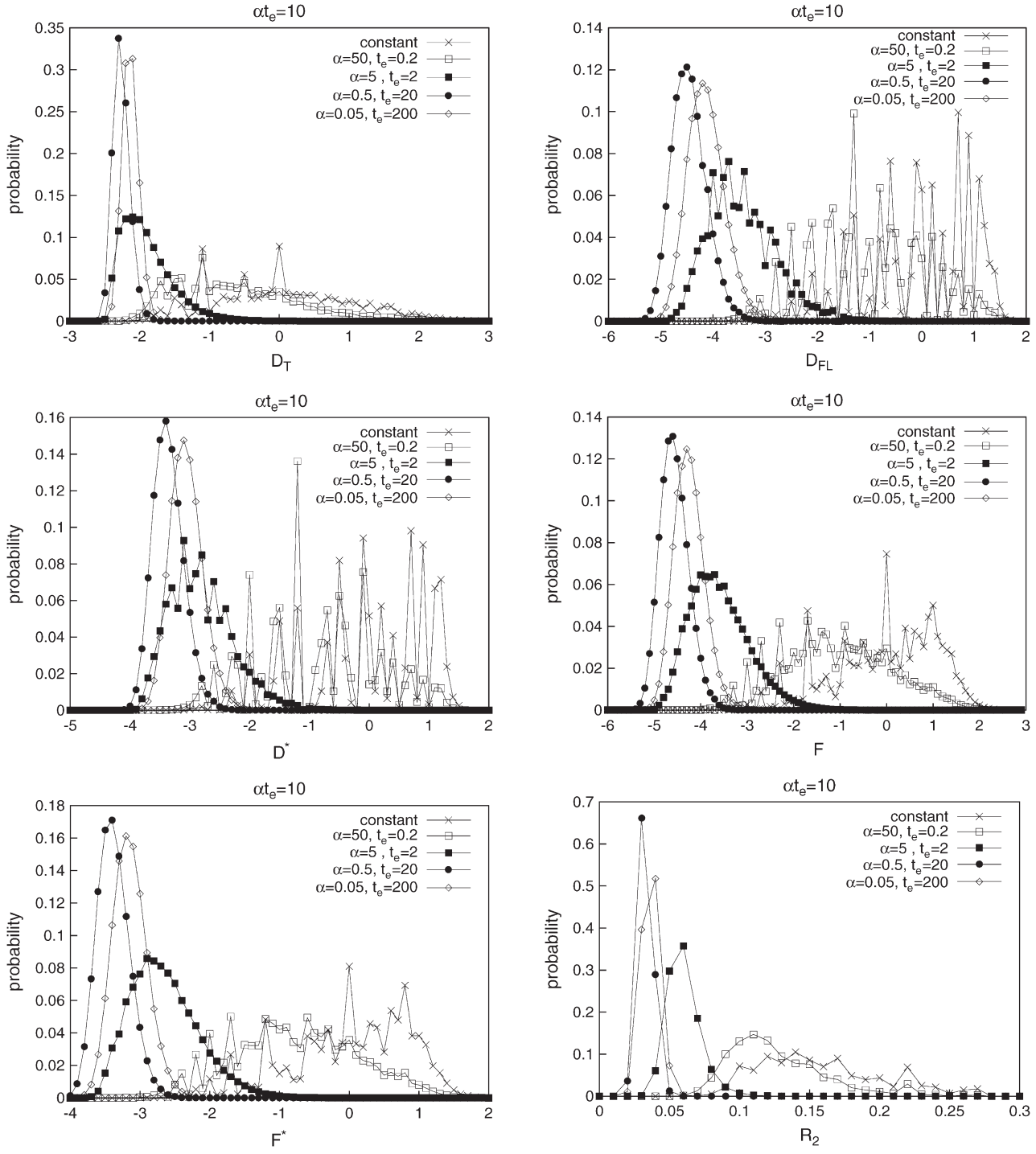

FIGURE 5.—The distributions of test statistics under exponential population growth for fixed αt_e (= 10).

tion-growth model. Our analysis indicates that the more powerful test statistics are characterized by larger reduction of variances in the population-growth models.

In Figures 2 and 3, we examined the convergence of distributions of the statistics to those of the rapid-growth limit increasing α independently. However, sometimes we may have some information about interrelationship among parameters. For example, PLUZHNIKOV *et al.* (2002) state that high growth rates are compatible with the data of a human population for only small growth periods and vice versa. Although such interdependencies among parameters are not usually simple, here we investigate the convergence of the distributions as α grows, keeping αt_e constant as a simple example of interdependency. Recall that t_e is the length of the growth period and α is the rate of increase of population size, so the ratio of the present population size to initial population size is represented by $e^{\alpha t_e}$; namely, the present

population size is represented by $N_A e^{\alpha t_e}$. Figure 5 shows the results with $\alpha t_e = 10$. In all statistics, the changes of modes of distributions are largest in the case of $\alpha = 0.5$, $t_e = 20$. For the distributions to be shifted toward minus, mutations must accumulate after the expansion starts. Since not many mutations accumulate if t_e is short, changes of the modes will be small in these cases.

DISCUSSION

In this article, we investigated the limiting behavior of test statistics of neutrality and obtained approximate formulas for their moments in the limits. In the limiting case, a population, originally of size N_A, instantaneously increases to an infinite size at time t_e ago so that no common ancestry can occur in the growth period. Although such extreme cases may be found rarely in nature, we can gain some insights into the previous simulation results obtained and intuitive arguments given by other authors from these formulas.

Pluzhnikov *et al.* (2002) surveyed 10 unlinked noncoding regions of individuals from three human populations. To analyze these data, averages and variances of a few test statistics were estimated using coalescent simulation. From the simulation, they found that recent population growth shifts the distribution of D_T and D^* toward smaller values and the variance of D_T decreases as the time, t_e, elapsed since the expansion event increases. We can explain these behaviors on the basis of our analytical results. As shown in the previous section, the expectations of test statistics are written as functions of t_e by using Equations 12–18. From these expressions, we can calculate their derivatives with respect to the variable t_e. For example,

$$\frac{dE[D_T]}{dt_e} = \frac{\mu(2a_n - n)(nt_e\mu e_1 + 2e_2\theta^2(a_n^2 + b_n) + 2a_n\theta(e_1 + nt_e\mu e_2))}{2a_nF_n\sqrt{F_n}}, \tag{29}$$

where F_n denotes the right-hand side of (13) and is positive. This expression is negative for any n (≥ 2). In the case of D_{FL},

$$\frac{dE[D_{FL}]}{dt_e} = \frac{n\mu(a_n - 1)(n\mu t_e(u_d + v_d) + 2v_d\theta^2(a_n^2 + b_n)2a_n\theta(u_d + v_d + n\mu t_e v_d))}{2G_n\sqrt{G_n}}, \tag{30}$$

and this also takes a negative value. Then we can see that those expectations are monotone decreasing with increasing t_e. For the other test statistics, we can show the same behavior (equations not shown). Besides, monotone decrease of approximate variances obtained by the δ-method can be shown similarly. Of course, since the monotonicity with regard to t_e shown here is for the rapid-growth limit, caution must be taken to extrapolate the results to cases with finite α but we expect the behavior of the limit reflects at least those when α is large.

Generally, it has been considered that sampled genes will have a star-like genealogy if there was rapid population growth (Kaplan *et al.* 1989). Also Marjoram and Donnelly (1994) pointed out that most coalescences of ancestors occur within a relatively short period under the exponential growth of the population size, and this makes a genealogy star-like. In fact, the star-like genealogy is a special case of the rapid-growth-limit genealogy, so we consider in what situation the star-like genealogy will be obtained. If t_e satisfies

$$\frac{1}{t_e} \ll 1,$$

then, from (12) and (13), the expectation of D_T under the case of the rapid-growth-limit genealogy is approximately given by

$$\frac{(2/n - 1/a_n)}{\sqrt{(e_1 - e_2)(1/n\mu t_e) + e_2}}. \tag{31}$$

Since the expectation of S is equal to $n\mu t_e$ in the rapid-growth-limit genealogy, (31) is approximately equal to the expectation of D_T under the case of the star genealogy given by (21). Recall that time is measured in units of N_A generations. So $1/t_e \to 0$ amounts to $N_A \to 0$.

We can show the same convergence for the variance. Arguments go similarly for the other test statistics introduced in theory and methods. So the star genealogy can be considered as "the secondary limiting genealogy" as N_A tends to infinitely small, so that all common ancestry events occur instantaneously.

In this article, only the case of exponential growth was considered thus far but the method can be extended to the sudden population growth case. Let α_s be the ratio of the population size after the growth to that before the growth and t_e be the time since the growth event. Then, the joint density, $g(t_n, \ldots, t_2)$, is expressed as

$$\begin{cases} \left(\prod_{i=2}^{n}\frac{c_i}{\alpha_s}\right)\exp\left(-\sum_{i=2}^{n}\frac{(i-1)x_i}{\alpha_s}\right), & \text{if } x_2 < t_e \\ \left(\prod_{i=2}^{n}c_i\right)\left(\frac{1}{\alpha_s}\right)^{n-(j+1)} \times \exp\left(\left(1-\frac{1}{\alpha_s}\right)c_j t_e - \sum_{i=2}^{j}(i-1)x_i \right. & \\ \qquad \left. - \sum_{i=j+1}^{n}\frac{(i-1)x_i}{\alpha_s}\right), & \text{if } x_{j+1} \leq t_e \leq x_j \\ \left(\prod_{i=2}^{n}c_i\right)\exp\left(-\sum_{i=2}^{n}(i-1)(x_i - t_e) - \frac{c_n t_e}{\alpha_s}\right), & \text{if } t_e < x_n, \end{cases} \tag{32}$$

where

$$x_i = t_i + t_{i+1} + \ldots + t_n, \quad x_{n+1} = 0.$$

It is obvious that the limiting joint density in this case equals (3) as α_s tends to infinity, so the limiting cases of sudden growth and exponential growth are the same. However, if α and α_s are finite, effects of changing t_e are different in (2) and (32). In both equations, the density takes the last expression ($t_e < x_n$) if $t_e = 0$ and has the same form as that of the constant-size case. As t_e increases, the density starts to differ from that of the constant case and the power to detect deviations

increases. However, note that the density (32) for sudden population growth is not a function of t_e when $x_2 < t_e$ and has the same form as that for constant population size. Therefore, the probability that $x_2 < t_e$ constantly increases as t_e increases and eventually the density converges to that of the constant size. Thus, the power to detect deviations decreases as t_e tends to infinity. This is why Simonsen *et al.* (1995) and Ramos-Onsins and Rozas (2002) found peaks of the power of test statistics under sudden population growth at intermediate values of t_e. Under exponential growth, it is difficult to predict the behavior of the power of test statistics from (2). However, the behavior of the power as t_e changes for large α is expected to be similar to that for the limiting case. Since both the means and variances of the statistics decrease as t_e increases in the rapid-growth-limit genealogy, the power increases as t_e increases in the limiting case. Therefore, the effect of t_e on the power in exponential growth is considered very different from that in sudden growth. Note that a logistic growth is a more realistic population-growth scenario for many organisms and the degree of shift of distribution should be small under a logistic-growth model.

We thank M. Iizuka for discussion and comments on the manuscript. We also thank two anonymous referees for their helpful comments. This work was partially supported by a grant-in-aid for scientific research on priority areas, "Medical Genome Science" no. 15012239.

LITERATURE CITED

Donnelly, P., and S. Tavaré, 1995 Coalescent and genealogical structure under neutrality. Annu. Rev. Genet. **29:** 401–421.

Fu, X.-Y., 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. Proc. R. Soc. Lond. Ser. B **344:** 403–410.

Hudson, R. R., 1990 Gene genealogies and the coalescent process, pp. 1–44 in *Oxford Surveys in Evolutionary Biology*, Vol. 7, edited by D. J. Futuyma and J. Antonovics. Oxford University Press, Oxford.

Innan, H., and W. Stephan, 2000 The coalescent in an exponentially growing metapopulation and its application to *Arabidopsis thaliana*. Genetics **155:** 2015–2019.

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

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

Marjoram, P., and P. Donnelly, 1994 Pairwise comparisons of mitochondrial-DNA sequences in subdivided populations and implications for early human-evolution. Genetics **136:** 673–683.

Pluzhnikov, A., A. Di Rienzo and R. R. Hudson, 2002 Inferences about human demography based on multilocus analyses of noncoding sequences. Genetics **161:** 1209–1218.

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

Purugganan, M. D., and J. I. Suddith, 1999 Molecular population genetics of floral homeotic loci: departures from the equilibrium-neutral model at the *APETALA3* and *PISTILLATA* genes of *Arabidopsis thaliana*. Genetics **151:** 839–848.

Ramos-Onsins, S. E., and J. Rozas, 2002 Statistical properties of new neutrality tests against population growth. Mol. Biol. Evol. **19:** 2092–2100.

Ray, N., M. Currat and L. Excoffier, 2003 Intra-deme molecular diversity in spatially expanding populations. Mol. Biol. Evol. **20:** 76–86.

Rosenberg, N. A., and A. E. Hirsh, 2003 On the use of star-shaped genealogies in inference of coalescence times. Genetics **164:** 1677–1682.

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

Slatkin, M., and R. R. Hudson, 1991 Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics **129:** 555–562.

Stephens, J. C., J. A. Schneider, D. A. Tanguay, J. Choi, T. Acharya *et al.*, 2001 Haplotype variation and linkage disequilibrium in 313 human genes. Science **293:** 489–493.

Stuart, A., and J. K. Ord, 1987 Standard errors, pp. 320–344 in *Kendall's Advanced Theory of Statistics*, Vol. I, Ed. 5. Charles Griffin & Co., London.

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

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

Communicating editor: J. Wakeley

APPENDIX

$$D_{\mathrm{T}} = \frac{\hat{k} - S/a_1}{\sqrt{e_1 S + e_2 S(S-1)}},$$

$$D_{\mathrm{FL}} = \frac{S - a_n \eta_e}{\sqrt{u_D S + v_D S^2}},$$

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

$$F = \frac{\hat{k} - \eta_e}{\sqrt{u_F S + v_F S^2}},$$

$$F^* = \frac{\hat{k} - ((n-1)/n)\eta_s}{\sqrt{u_{F^*} S + v_{F^*} S^2}},$$

$$R_2 = \frac{((1/n)\sum_{i=1}^{n}(\xi_i - (\hat{k}/2))^2)^{1/2}}{S},$$

$$a_n = \sum_{i=1}^{n-1} \frac{1}{i},$$

$$b_n = \sum_{i=1}^{n-1} \frac{1}{i^2},$$

$$c_1 = d_1 - \frac{1}{a_n},$$

$$c_2 = d_2 - \frac{n+2}{a_n n} + \frac{b_n}{a_n^2},$$

$$c_n = 2\,\frac{n a_n - 2(n-1)}{(n-1)(n-2)},$$

$$d_1 = \frac{n+1}{3(n-1)},$$

$$d_2 = \frac{2(n^2+n+3)}{9n(n-1)},$$

$$d_n = c_n + \frac{n-2}{(n-1)^2} + \frac{2}{n-1}\left(\frac{3}{2} - \frac{2a_{n+1}-3}{n-2} - \frac{1}{n}\right),$$

$$e_1 = \frac{c_1}{a_n},$$

$$e_2 = \frac{c_2}{a_n^2 + b_n},$$

$$v_D = 1 + \frac{a_n^2}{b_n + a_n^2}\left(c_n - \frac{n+1}{n-1}\right),$$

$$u_D = a_n - 1 - v_D,$$

$$v_{D^*} = \frac{1}{a_n^2 + b_n}\left(\left(\frac{n}{n-1}\right)^2 b_n + a_n^2 d_n - 2\frac{na_n(a_n+1)}{(n-1)^2}\right),$$

$$u_{D^*} = \frac{n}{n-1}\left(a_n - \frac{n}{n-1}\right) - v_{D^*},$$

$$v_F = \frac{1}{a_n^2 + b_n}\left(c_n + d_2 - \frac{2}{n-1}\right),$$

$$u_F = \frac{1}{a_n}\left(1 + d_1 - 4\frac{n+1}{(n-1)^2}\left(a_{n+1} - \frac{2n}{n+1}\right)\right) - v_F,$$

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

$$u_{F^*} = \frac{4n^2 + 19n + 3 - 12(n+1)a_{n+1}}{3n(n-1)a_n} - v_{F^*}.$$