Neutrality Tests Using DNA Polymorphism From Multiple Samples

Haipeng Li,*,† Yunwu Zhang,† Ya-Ping Zhang† and Yun-Xin Fu*,‡,1

‡*Laboratory of Bioinformatics, Yunnan University, Kunming 650991, People's Republic of China,* **Human Genetics Center, University of Texas, Houston, Texas 77030 and* †*Laboratory of Molecular Evolution and Genome Diversity, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, People's Republic of China*

Manuscript received April 23, 2002
Accepted for publication November 21, 2002

ABSTRACT

The polymorphism of a gene or a locus is studied with increasing frequency by multiple laboratories or the same group at different times. Such practice results in polymorphism being revealed by different samples at different regions of the locus. Tests of neutrality have been widely conducted for polymorphism data but commonly used statistical tests cannot be applied directly to such data. This article provides a procedure to conduct a neutrality test and details are given for two commonly used tests. Applying the two new tests to the chemokine-receptor gene (*CCR5*) in humans, we found that the hypothesis that all mutations are selectively neutral cannot explain the observed pattern of DNA polymorphism.

THE amount and pattern of polymorphism in DNA sequence samples from a population reflects not only mutations in the ancestors of the sequences but also random genetic drift as well as other evolutionary forces, such as natural selection. How to detect the presence of natural selection in molecular population genetics and evolution is an important issue. It is possible to detect the presence of natural selection because natural selection often causes the pattern of polymorphism to differ from that under the neutral mutation hypothesis, which postulates that the majority of mutations that have contributed significantly to the genetic variation in natural populations is neutral or nearly neutral (Kimura 1983). However, the neutral mutation hypothesis is not sufficiently quantified to be tested rigorously in practice. A narrower definition of neutrality is that all mutations are selectively neutral, which is referred to as the *hypothesis of strict neutrality* (Fu and Li 1993).

A number of statistical tests have been proposed, and almost all of them are designed for a single sample. In reality, polymorphic data can be accumulated over time in the same or different laboratories, which means different sites may be examined using different samples. How to conduct neutrality tests in such situations is the focus of this article. To date, millions of single nucleotide polymorphisms (SNPs) have been identified. It is very likely that SNPs from a single gene or tightly linked regions will be typed for different samples by different research groups over time. The newly developed method will be valuable for analyzing such data. We present an example of such an analysis using data from the *CCR5* gene.

[1]*Corresponding author:* Human Genetics Center, UT School of Public Health, P.O. Box 20186, 1200 Herman Pressler, Houston, TX 77030. E-mail: yunxin.fu@uth.tmc.edu

STATISTICAL TESTS

A number of statistical tests can be extended for multiple samples. However, for the sake of discussion, we focus on two particular tests partly because they are used widely. First, Tajima (1989) proposed using the difference between two estimates of θ ($=4N\mu$), where N is the effective population size, and μ is the mutation rate per sequence per generation, to detect the presence of selection. His test statistic is

$$D_{\mathrm{T}} = \frac{\Pi - K/a_n}{\sqrt{\mathrm{Var}(\Pi - K/a_n)}}, \tag{1}$$

where Π is the mean number of nucleotide differences between two sequences, K is the number of segregating sites, which is equal to the total number of mutations under the infinite-sites model, n is sample size, and

$$a_n = 1 + \frac{1}{2} + \ldots + \frac{1}{n-1}. \tag{2}$$

Second, Fu and Li (1993) suggested several tests of neutrality, one of which is

$$D_{\mathrm{F}} = \frac{K - a_n\xi_1}{\sqrt{\mathrm{Var}(K - a_n\xi_1)}}, \tag{3}$$

where ξ_1 is the number of mutations in the external branches, that is, mutations that are inherited by only one sequence in the sample.

The above two tests can be extended to multiple samples in the following way. Assume that a locus without recombination is divided into m regions that have been surveyed using different or partially overlapping samples (Figure 1). It should be emphasized that the assumption of no recombination is made here to make the null model as simple as possible, similar to the original Tajima and Fu and Li tests. Just as the presence of

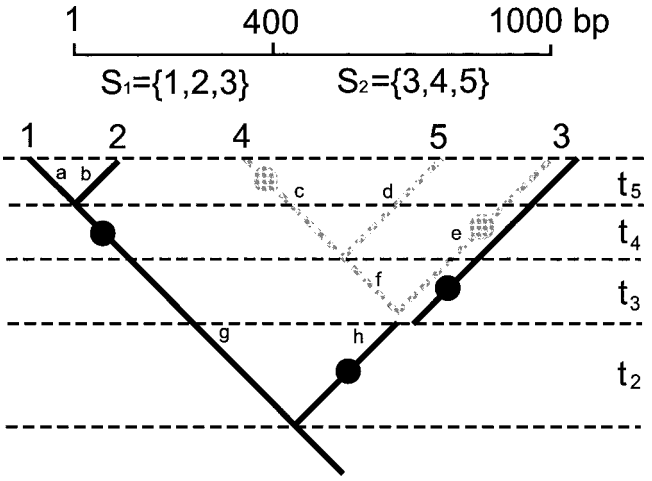

FIGURE 1.—An illustration of a genealogy of multiple samples. The locus is divided into two regions. The sequence length in the first region is 400 and in the second one is 600, so $r_1 = 0.4$ and $r_2 = 0.6$ under the hypothesis of a constant mutation rate. $|S_1| = 3$, $|S_2| = 3$, and $S_1 \cap S_2 = \{3\}$. That means {1, 2} are sequenced from 1 to 400, {4, 5} are sequenced from 401 to 1000, and {3} is sequenced from 1 to 1000. The subtree for the first region is marked by solid lines, and the subtree for the second region is labeled by dashed lines. The lengths of branches can be calculated from coalescent times, and the length of a branch here means a time duration. For example, $g = t_2 + t_3 + t_4$, and $h = t_2$. Let L_i be the branch length of the subtree for the ith region, and then $L_1 = a + b + g + h + e$, and $L_2 = c + d + f + e$. The branch of e is shared among two subtrees. l_{ki} is the length of the k-size branch in the subtree for the ith region, so $l_{11} = a + b + h + e$, $l_{21} = g$, $l_{12} = c + d + e$, and $l_{22} = f$. Five mutations are on the genealogy and are shown as circles, three of which (solid circles) are found in the first region, and two of which (shaded circles) are found in the second region. Therefore, $\xi_{11} = 2$, $\xi_{21} = 1$, $\xi_{12} = 2$, and $\xi_{22} = 0$.

recombination does not invalidate the Tajima and Fu and Li tests, but makes them more conservative, the new tests will likely behave similarly and this will also be applicable to data with recombination.

Define $\theta = 4N\mu$, where μ is the mutation rate per sequence per entire locus, and N is the effective population size. Also define $\theta_i = 4N\mu_i$, where μ_i is the mutation rate for the ith region ($i = 1, 2, \ldots, m$), and $r_i = \mu_i/\mu$, which is the proportion of the mutation rate of the ith region. If the mutation rate per site is constant, the r_i is equal to the proportion of the length of the ith region,

$$\sum_{i=1}^{m} r_i = 1.$$

Furthermore, we define the mean number of nucleotide differences between two sequences in the ith region as Π_i, the number of segregating sites in the ith region as K_i, the number of mutations in the external branches of the genealogy in the ith region as ξ_{1i}, and the sample size in the ith region as n_i (Figure 1). Define

$$\Pi = \sum_{i=1}^{m} \Pi_i, \tag{4}$$

$$K = \sum_{i=1}^{m} K_i, \tag{5}$$

and

$$\xi_1 = \sum_{i=1}^{m} \xi_{1i}. \tag{6}$$

It is obvious that when $m = 1$, these equations reduce to their conventional definitions for a single sample. Note that in these equations the same weight is given to every region. An alternative approach is to give weight to a region according to certain criterion. However, so far we have not found any other weighting scheme to perform better than the equal-weighting scheme.

Furthermore, we define

$$\bar{a} = \sum_{i=1}^{m} r_i a_{n_i}, \tag{7}$$

where a_{n_i} is given by (2). That is, $\bar{a}$ is a weighted average of a_{n_i}. Then the tests (1) and (3) become

$$D_T = \frac{\Pi - K/\bar{a}}{\sqrt{\mathrm{Var}(\Pi - K/\bar{a})}} \tag{8}$$

and

$$D_F = \frac{K - \bar{a}\xi_1}{\sqrt{\mathrm{Var}(K - \bar{a}\xi_1)}}. \tag{9}$$

PERFORMING THE TESTS

Since $\mathrm{Var}(\Pi - K/\bar{a}) = \mathrm{Var}(\Pi) + \mathrm{Var}(K)/\bar{a}^2 - 2\,\mathrm{Cov}(\Pi, K)/\bar{a}$ and $\mathrm{Var}(K - \bar{a}\xi_1) = \mathrm{Var}(K) + \bar{a}^2\,\mathrm{Var}(\xi_1) - 2\bar{a}\,\mathrm{Cov}(K, \xi_1)$, we can compute $\mathrm{Var}(\Pi - K/\bar{a})$ and $\mathrm{Var}(K - \bar{a}\xi_1)$ if we are able to compute $\mathrm{Var}(K)$, $\mathrm{Var}(\Pi)$, $\mathrm{Var}(\xi_1)$, $\mathrm{Cov}(\Pi, K)$, and $\mathrm{Cov}(K, \xi_1)$. Some of these terms can be computed analytically; others have to be estimated.

Analytical result: Assume the total sample consists of n sequences, and let those sequences be numbered from 1 to n. $S = \{1, 2, \ldots, n\}$ will represent the whole sample. The sample, S_i, for the ith region will be a subset of S. That is $S_i \subseteq S$. We do not make any assumption here on the relationship among S_i. We note that in one extreme, we can have $S_i \cap S_j = \phi$ for every pair of i and j, and on the other hand, we may have $S_1 = S_2 = \ldots = S_m = S$. In many situations, it is likely that $S_i \cap S_j \neq \phi$ (*e.g.*, Figure 1). We use $|S_i|$ to represent the number of elements in the set S_i, which is the sample size. Then we have $n_i = |S_i|$, $n = |S|$.

The computation of $\mathrm{Var}(\Pi)$, $\mathrm{Var}(K)$, and other components that we mentioned above requires understanding of the sample genealogy. Let L_i be the total branch length of subtree for the ith region scaled so that 1 unit corresponds to $4N$ generations, and the length of a

branch here means a time duration. ξ_{ki} is the number of k-size mutations in the ith region, and l_{ki} is the length of k-size branches in the subtree for the ith region scaled similarly as L_i (Figure 1). The size of a branch is the number of sequences in the sample that are descendants of that branch, and a mutation is said to be *size k* if it occurs in a branch of size k (Fu 1995). Considering the subtree that is part of the tree shown as solid lines in Figure 1, the branch of g has 2 descendent sequences, 1 and 2, so the size of the branch is 2. A mutation is on the branch of g, so the size of the mutation is 2. Following the definition of θ, θ_i, and r_i, we have

$$\theta_i = 4N\mu_i = 4N\mu(\mu_i/\mu) = r_i\theta. \tag{10}$$

Moreover, since 1 unit in time corresponds to $4N$ generations, we have from Li and Fu (1998)

$$E(L_i) = a_{n_i}. \tag{11}$$

Under the infinite-sites model, we have

$$E(K) = \sum_{i=1}^{m} E(K_i) = \sum_{i=1}^{m} \theta_i E(L_i) = \theta\bar{a}. \tag{12}$$

Conditioning on the coalescent times, the number of mutations in each branch follows a Poisson distribution with parameter θl, where l is the branch length. We thus have

$$E(K_iK_j|t_2, t_3, \ldots) = \theta_i L_i \theta_j L_j = r_i r_j \theta^2 L_i L_j,$$

where t_k $(2 \le k \le n)$ is the time duration required for k sequences to coalesce to $k-1$ sequence, *i.e.*, the so-called k-coalescent time (Figure 1), and $i \ne j$. Then we have

$$E(K_iK_j) = E_{t_2,t_3,\ldots,t_n}[E(K_iK_j|t_2, t_3, \ldots, t_n)] = r_i r_j \theta^2 E(L_iL_j), \tag{13}$$

which leads to

$$\mathrm{Cov}(K_i, K_j) = E(K_iK_j) - E(K_i)E(K_j) = r_i r_j \theta^2 [E(L_iL_j) - a_{n_i}a_{n_j}]. \tag{14}$$

Watterson (1975) showed that K in the case of one sample is

$$\mathrm{Var}(K_i) = a_{n_i}\theta_i + b_{n_i}\theta_i^2 = a_{n_i} r_i \theta + b_{n_i} r_i^2 \theta^2, \tag{15}$$

where

$$b_n = 1 + \frac{1}{4} + \ldots + \frac{1}{(n-1)^2}.$$

Also we can partition $E(L_iL_j)$ further as

$$E(L_iL_j) = E\left(\sum_{k=1}^{n_i=1} l_{ki} \sum_{p=1}^{n_j-1} l_{pj}\right) \sum_{k=1}^{n_i-1} \sum_{p=1}^{n_j-1} E(l_{ki}l_{pj}). \tag{16}$$

Therefore, from (5), (10), (14), (15), and (16), we have

$$\begin{aligned}\mathrm{Var}(K) &= \sum_{i=1}^{m} V(K_i) + 2\sum_{i<j} \mathrm{Cov}(K_i, K_j) \\ &= \theta\bar{a} + \theta^2\left[\sum_{i=1}^{m} b_{n_i} r_i^2 + 2\sum_{i<j} r_i r_j \left(\sum_{k=1}^{n_i-1} \sum_{p=1}^{n_j-1} E(l_{ki}l_{pj}) - a_{n_i}a_{n_j}\right)\right]. \end{aligned} \tag{17}$$

Similar to (13), we have for $i \ne j$ that

$$E(\xi_{ki}\xi_{pj}) = r_i r_j \theta^2 E(l_{ki}l_{pj}). \tag{18}$$

Tajima (1983) showed $E(\Pi_i) = \theta_i$, so we have

$$E(\Pi) = \sum_{i=1}^{m} E(\Pi_i) = \sum_{i=1}^{m} r_i\theta = \theta. \tag{19}$$

Since Π_i is the mean number of nucleotide differences between two sequences in the ith region, it can be calculated from ξ_{ki} as

$$\Pi_i = \frac{2}{n_i(n_i-1)} \sum_{k=1}^{n_i-1} (n_i - k)k\xi_{ki} \tag{20}$$

(Fu 1995). Tajima (1983) derived the variance of Π_i as

$$\mathrm{Var}(\Pi_i) = \frac{n_i+1}{3(n_i-1)} r_i\theta + \frac{2(n_i^2 + n_i + 3)}{9n_i(n_i-1)} r_i^2\theta^2. \tag{21}$$

So we have

$$\begin{aligned} E(\Pi_i\Pi_j) &= E\left[\frac{2}{n_i(n_i-1)} \sum_{k=1}^{n_i-1} (n_i-k)k\xi_{ki} \frac{2}{n_j(n_j-1)} \sum_{p=1}^{n_j-1} (n_j-p)p\xi_{pj}\right] \\ &= \frac{4\theta^2}{n_in_j(n_i-1)(n_j-1)} \left[\sum_{k=1}^{n_i-1} \sum_{p=1}^{n_j-1} kp(n_i-k)(n_j-p) r_i r_j E(l_{ki}l_{pj})\right], \end{aligned} \tag{22}$$

and thus

$$\mathrm{Var}(\Pi) = \sum_{i=1}^{m} \mathrm{Var}(\Pi_i) + 2\sum_{i<j} E(\Pi_i\Pi_j) - 2\sum_{i<j} [E(\Pi_i)E(\Pi_j)], \tag{23}$$

where $\mathrm{Var}(\Pi_i)$ and $E(\Pi_i\Pi_j)$ are given by (21) and (22), respectively. Moreover, from (12) and (19), we have

$$\mathrm{Cov}(K, \Pi) = E(K\Pi) - E(K)E(\Pi) = E(K\Pi) - \theta^2\bar{a}, \tag{24}$$

where $E(K\Pi)$ is given later (Equation 25). Fu (1995) showed the formula to calculate $E(\xi_{ki}\xi_{pi})$. After putting these terms together, we have

$$\begin{aligned} E(K\Pi) &= E\left(\sum_{i=1}^{m} K_i \sum_{i=1}^{m} \Pi_i\right) = \sum_{i,j}^{m} E\left[\left(\sum_{k=1}^{n_i-1} \xi_{ki}\right)\left(\frac{2}{n_j(n_j-1)} \sum_{p=1}^{n_j-1} (n_j-p)p\xi_{pj}\right)\right] \\ &= \sum_{i=1}^{m} \sum_{k=1}^{n_i-1} \sum_{p=1}^{n_i-1} \frac{2(n_i-p)p}{n_i(n_i-1)} E(\xi_{ki}\xi_{pi}) \\ &\quad + \theta^2 \sum_{i\ne j} \sum_{k=1}^{n_i-1} \sum_{p=1}^{n_j-1} \frac{2(n_j-p)p}{n_j(n_j-1)} r_i r_j E(l_{ki}l_{pj}). \end{aligned} \tag{25}$$

From Fu (1995), we have

$$E(\xi_{ki}) = \frac{1}{k}\theta_i, \tag{26}$$

which leads to

$$E(\xi_1) = \sum_{i=1}^{m} E(\xi_{1i}) = \sum_{i=1}^{m} \theta_i = \theta \tag{27}$$

and

$$\mathrm{Var}(\xi_1) = \mathrm{Var}\left(\sum_{i=1}^{m} \xi_{1i}\right) = \sum_{i=1}^{m} \mathrm{Var}(\xi_{1i}) + 2 \sum_{i<j} \mathrm{Cov}(\xi_{1i}, \xi_{1j}). \tag{28}$$

Fu and Li (1993) showed that the variance of the total number of mutations in the external branches is given by

$$\mathrm{Var}(\xi_{1i}) = \theta_i + c_{n_i}\theta_i^2, \tag{29}$$

where $c_n = 1$ when $n = 2$, and when $n > 2$

$$c_n = 2\frac{na_n - 2(n-1)}{(n-1)(n-2)}.$$

From (18) and (26), we have

$$\mathrm{Cov}(\xi_{1i}, \xi_{1j}) = E(\xi_{1i}\xi_{1j}) - E(\xi_{1i})E(\xi_{1j}) = r_i r_j \theta^2 E(l_{1i}l_{1j}) - r_i r_j \theta^2. \tag{30}$$

Substituting (29) and (30) into (28), we have

$$\mathrm{Var}(\xi_1) = \sum_{i=1}^{m} (r_i\theta + c_{n_i} r_i^2 \theta^2) + 2\sum_{i<j} (r_i r_j \theta^2 E(l_{1i}l_{1j}) - r_i r_j \theta^2), \tag{31}$$

and

$$\mathrm{Cov}(K, \xi_1) = E\left(\sum_{i=1}^{m} K_i \sum_{j=1}^{m} \xi_{1j}\right) - E(K)E(\xi_1) = \sum_{i=1}^{m}\sum_{k=1}^{n_i-1} E(\xi_{ki}\xi_{1i}) + \theta^2 \sum_{i\neq j}\sum_{k=1}^{n_i-1} r_i r_j E(l_{ki}l_{1j}) - \theta^2\bar{a}, \tag{32}$$

where $E(\xi_{ki}\xi_{1i})$ is given by Fu (1995).

Numerical estimation: The above derivation shows that $E(l_{ki}l_{pj})$ is critical for computing $\mathrm{Var}(K)$, $\mathrm{Var}(\Pi)$, $\mathrm{Cov}(\Pi, K)$, $\mathrm{Var}(\xi_1)$, and $\mathrm{Cov}(K, \xi_1)$. The value of $E(l_{ki}l_{pj})$ is dependent on the relationship of branches among subtrees of the regions. A simple example of the relationship is given in Figure 1. Although Fu (1995) was able to derive $E(l_{ki}l_{pj})$ when $i = j$, the general formula for $i \neq j$ appears to be analytically intractable. However, an estimate can be obtained relatively easily from the following procedure:

1. Simulate a genealogy g (topology without mutation) of a sample of n sequences numbered from 1 to n.
2. Compute the value of $E_g(l_{ki}l_{pj})$ for the simulated genealogy.
3. Repeat steps 1 and 2 M times. Then $E(l_{ki}l_{pj})$ is finally estimated as

$$\hat{E}(l_{ki}l_{pj}) = \frac{1}{M}\sum_g E_g(l_{ki}l_{pj}).$$

The computation of $E_g(l_{ki}l_{pj})$ in step 2 is done in a similar manner as Fu (1994). As it is obvious, the accuracy of the estimation can be improved by using large values of M.

The components mentioned above can be obtained after $E(l_{ki}l_{pj})$ is estimated, and then D_Γ and D_F can be calculated. Similar to Tajima's (1989) and Fu and Li's (1993) tests, D_Γ and D_F do not follow well-known standard distributions. Since $E(l_{ki}l_{pj})$ have to be estimated from simulated samples, it is natural to use computer simulations to determine the critical points of the tests. Overall, this approach gives more accurate critical values than using approximation by a standard distribution.

TABLE 1

The list of r_i and the frequencies of mutations in 12 regions of the *CCR5* gene

Region (bp)	r	Mutation	Frequency
1 (1–42)	0.038	A25C	1/382
2 (43–91)	0.045	T58A	2/698
3 (92–144)	0.049	A124T	1/170
4 (145–191)	0.043	T164A	29/708
5 (192–355)	0.151	C218T	3/462
6 (356–523)	0.154	C492A	1/98
7 (524–611)	0.081	Δ32	520/5210
8 (612–674)	0.058	G668A	1/64
9 (675–790)	0.107	680del3	1/490
10 (791–901)	0.102	C900A	1/242
11 (902–953)	0.048	G902T	1/90
12 (954–1089)	0.124	C1004T	1/174

CCR5 mutations and their frequencies in Caucasians come from Carrington *et al.* (1997), and the frequency is the number of alleles observed/the total number of chromosomes.

AN EXAMPLE

We consider data from the *CCR5* gene from Caucasians (Carrington *et al.* 1997) to illustrate how the extended tests described in this article can be applied. The *CCR5* encodes a cell-surface chemokine-receptor molecule that serves as a co-receptor for the macrophage-tropic strains of HIV-1. Because of its obvious importance, the *CCR5* gene has been subjected to many studies. One hypothesis is that *CCR5* might have been under natural selection (Carrington *et al.* 1997). In the data from Carrington *et al.* (1997), 12 mutations were documented in Caucasians, and 10 of the discovered mutations alter the amino acid sequence of the protein, and each mutation is typed by different or partially overlapping samples (Table 1). Since the precise relationships among samples were not given in the original article, we consider two extreme cases here. In the first case, we assume that different samples are composed of different individuals. That is, $S_i \cap S_j = \phi$, where $i \neq j$. In the second case, we assume that smaller samples are a subset of larger samples. That is, we assume

TABLE 2

Results of neutrality tests for *CCR5*

	Case 1	Case 2
Var(K)	18.454	18.393
Var(Π)	1.453	1.505
Var(ξ_1)	1.954	1.964
Cov(Π, K)	3.786	3.921
Cov(K, ξ_1)	2.058	2.025
D_T	−1.798***	−1.790***
D_F	−4.579***	−4.555***

Statistical significance was calculated from the empirical distribution of D_T and D_F. For D_T, the critical values for the 1% significance tests are −1.663 (case 1) and −1.660 (case 2), and for D_F the values are −2.829 (case 1) and −2.631 (case 2). The values are estimated from 10,000 simulated samples. ***The test result is significant at 1% level.

S_8 (64) ⊂ S_{11} (90) ⊂ S_6 (98) ⊂ S_3 (170) ⊂ S_{12} (174) ⊂ S_{10} (242) ⊂ S_1 (382) ⊂ S_5 (462) ⊂ S_9 (490) ⊂ S_2 (698) ⊂ S_4 (708) ⊂ S_7 (5210), where the numbers in parentheses are the sample sizes. We refer to those two cases as case 1 and case 2, respectively. By assuming samples are independent in the first case, we basically obtain the maximum possible information from such data. On the other hand, by assuming smaller samples are a subset of larger samples, we have the minimal amount of information.

From (12), we can get the estimate of θ as $\hat{\theta} = K/\bar{a}$, since $\bar{a} = 6.274$, so we have $\hat{\theta} = 1.913$. Also we have $\Pi = \Sigma \Pi_i = 0.392$. Table 2 shows that both tests yield significant negative values in both cases 1 and 2. Thus we conclude that the *CCR5* region has not evolved according to the neutral model. One possibility is that it has evolved under natural selection, which remains to be seen by further study.

We are grateful to Ms. Sara Barton for her help. This work is supported by National Institutes of Health grants R01 GM50428 and R01 GM55759 (Yun-Xin Fu), the Chinese Academy of Sciences (KSCX2-1-05), the National Nature Science Foundation of China, and the Nature Science Foundation of Yunnan Province in China.

LITERATURE CITED

Carrington, M., T. Kissner, B. Gerrard, S. Ivanov, S. J. O'Brien *et al.*, 1997 Novel alleles of the chemokine-receptor gene *CCR5*. Am. J. Hum. Genet. **61:** 1261–1267.

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

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

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

Kimura, M., 1983 *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, UK.

Li, W. H., and Y. X. Fu, 1998 Coalescent theory and its applications in population genetics, pp. 45–79 in *Statistics in Genetics*, edited by E. Halloran. Springer-Verlag, New York.

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

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

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

Communicating editor: G. B. Golding