Estimating Effective Population Size or Mutation Rate Using the Frequencies of Mutations of Various Classes in a Sample of DNA Sequences

Yun-Xin Fu

Center for Demographic and Population Genetics, University of Texas, Houston, Texas 77225

Manuscript received February 18, 1994
Accepted for publication August 27, 1994

ABSTRACT

Mutations resulting in segregating sites of a sample of DNA sequences can be classified by size and type and the frequencies of mutations of different sizes and types can be inferred from the sample. A framework for estimating the essential parameter $\theta = 4Nu$ utilizing the frequencies of mutations of various sizes and types is developed in this paper, where N is the effective size of a population and μ is mutation rate per sequence per generation. The framework is a combination of coalescent theory, general linear model and Monte-Carlo integration, which leads to two new estimators $\hat{\theta}_\xi$ and $\hat{\theta}_\eta$ as well as a general Watterson's estimator $\hat{\theta}_K$ and a general Tajima's estimator $\hat{\theta}_\pi$. The greatest strength of the framework is that it can be used under a variety of population models. The properties of the framework and the four estimators $\hat{\theta}_K$, $\hat{\theta}_\pi$, $\hat{\theta}_\xi$ and $\hat{\theta}_\eta$ are investigated under three important population models: the neutral Wright-Fisher model, the neutral model with recombination and the neutral Wright's finite-islands model. Under all these models, it is shown that $\hat{\theta}_\xi$ is the best estimator among the four even when recombination rate or migration rate has to be estimated. Under the neutral Wright-Fisher model, it is shown that the new estimator $\hat{\theta}_\xi$ has a variance close to a lower bound of variances of all unbiased estimators of θ which suggests that $\hat{\theta}_\xi$ is a very efficient estimator.

An important parameter in studying the evolution of a DNA region (locus) of a population is $\theta = 4N\mu$ where N is the effective size of the population and μ is the mutation rate per sequence per generation. From the value of θ, the effective population size N can be obtained if the mutation rate μ is known or vice versa. Until recently inferences about θ have been based largely on two quantities. One is the sequence diversity π, which is the average number of nucleotide differences per pair of sequences; the other is the number K of segregating sites (polymorphic sites). These two quantities lead to respectively TAJIMA's estimate $\hat{\pi}$ of θ and WATTERSON's estimate $\hat{K}$ of θ as follows

$$\hat{\pi} = \pi \qquad \hat{K} = K/a_n$$

where

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

and n is the sample size. Although the computations of these two estimators are easy, they are not very efficient estimators, in particular, the variance of $\hat{\pi}$ does not diminish with increasing sample size. The inefficiency of these estimators and the surging of DNA polymorphism data have been stimulating the development of new methods of estimating θ that make better use of the information in a sample. Phylogenetic information are very useful in developing more accurate estimators (FELSENSTEIN 1992; FU 1994a). For example, FU (1994a) showed that a phylogenetic estimator has a variance close to the minimum variances of all possible unbiased estimators of θ under the neutral Wright-Fisher model. That is, the population under study evolves according to the Wright-Fisher model, all mutations are selectively neutral and there is not recombination and no population subdivision.

Natural populations are usually more complex than described by the neutral Wright-Fisher model. For example, they are often subdivided into small local populations with migrations among them; an autosomal DNA region studied may be very large or consist of several separate regions so that recombinations cannot be neglected; some mutations may not be neutral or mutation rates for different region of a locus may differ. Estimating θ under models other than the neutral Wright-Fisher model is often necessary for detailed analysis of polymorphism data. Although Watterson's estimator $\hat{K}$ and TAJIMA's estimator $\hat{\pi}$ can be modified to provide estimate of θ under various models, they are likely inefficient as they are under the neutral Wright-Fisher model. On the other hand, phylogenetic methods such as FU (1994a) and FELSENSTEIN (1992) are difficult to be extended because they require detailed properties of the branchs of a genealogy which are poorly understood under other models. In this paper, I present a framework for estimating θ which is not only very efficient but can be used under a variety of models provided that additional parameters (if any) can be estimated roughly. The efficiency of the framework is demonstrated using three models: the neutral Wright-Fisher model, the neutral model with recombination and the Wright's finite-islands model.

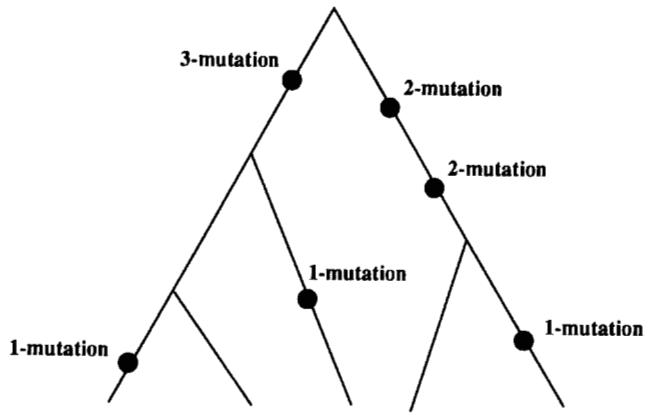

FIGURE 1.—The variants of mutations in a genealogy where each ● represents a mutation.

FREQUENCIES OF MUTATIONS OF VARIOUS CLASSES

Consider a random sample of n sequences from a population. Corresponding to each homologous site of the sequences, there is a genealogy connecting the n sequences (nucleotides) to their most recent common ancestor and the genealogy consists of $2(n-1)$ branches. A branch is said to be *size i* if exactly i sequences in the sample are descendents of the branch. A mutation occurred in a branch of size i is said to be of *size i* or an *i*-mutation. Therefore, mutations leading to segregating sites in a sample can be classified into $n-1$ sizes. An illustration of the definition is given in Figure 1.

Let ξ_i be the summation of i-mutations over all the sites and $\boldsymbol{\xi}$ be a vector given by

$$\boldsymbol{\xi} = (\xi_1, \ldots, \xi_{n-1})^T \tag{2}$$

where T stands for transpose. The vector $\boldsymbol{\xi}$ is a primary source of information that will be utilized in this study to estimate θ. When the infinite-sites model is assumed and an outgroup sequence is available, the value of $\boldsymbol{\xi}$ can be inferred directly from the sequences of a sample. This is because that a mutation results in a segregating site of two segregating nucleotides and the nucleotide that is not the same as that of the outgroup sequence must be the size of the mutation, otherwise it will contradicts with the infinite-sites model. However, when the infinite-sites model does not hold and there is no outgroup sequence available, the value of $\boldsymbol{\xi}$ has to be inferred by reconstructing a genealogy of the sample; when the length of sequences are not sufficiently long or there are recombinations the inferred value of $\boldsymbol{\xi}$ will often contain errors. Although we wish to address all important issues related to the estimation of θ, we shall assume as an initial investigation that the value of $\boldsymbol{\xi}$ is known for a given sample.

Another vector of information that will be utilized to estimate θ is

$$\boldsymbol{\eta} = (\eta_1, \ldots, \eta_{[n/2]})^T \tag{3}$$

where $[n/2]$ denotes the largest integer contained in $n/2$ and the i-th element η_i of $\boldsymbol{\eta}$ is given by

$$\eta_i = \frac{\xi_i + \xi_{n-1}}{1 + \delta_{i,n-i}} \tag{4}$$

where $\delta_{i,n-i}$ is the Kronecker delta, *i.e.*, it is equal to 1 if $i = n - i$ and 0 otherwise. Under the infinite-sites model, η_i is simply the number of such segregating sites at which the frequencies of the two segregating nucleotides are i and $n - i (i < n - i)$ respectively. Such a segregating site is said to be of *type i* or *i*-segregating site. We thus call η_i the frequency of *i*-segregating sites. It is easy to see that when the infinite-sites model holds the value of $\boldsymbol{\eta}$ can be obtained directly from a sample without the help of an outgroup sequence. Because of this reason, it is easier to use an estimator based on $\boldsymbol{\eta}$ than an estimator based on $\boldsymbol{\xi}$, although the former is inferior to the latter, as shall be demonstrated later.

BEST LINEAR ESTIMATORS OF θ FROM $\boldsymbol{\xi}$ AND $\boldsymbol{\eta}$

Because the number of segregating site K is equal to $\xi_1 + \ldots + \xi_{n-1}$ and $\eta_1 + \ldots + \eta_{[n/2]}$, WATTERSON's estimator $\hat{K}$ can be written as

$$\hat{K} = \sum_{i-1}^{n-1} \left(\frac{1}{a_n}\right)\xi_i = \sum_{i=1}^{[n/2]} \left(\frac{1}{a_n}\right)\eta_i \tag{5}$$

where a_n is given by (1). Therefore, $\hat{K}$ is a linear function of $\xi_1, \ldots, \xi_{n-1}$ or a linear function of $\eta_1, \ldots, \eta_{[n/2]}$. Since a mutation of size i is counted in $i(n-i)$ pairwise comparisons, TAJIMA's estimator $\hat{\pi}$ can be written as

$$\hat{\pi} = \sum_{i=1}^{n-1} \left[\frac{2i(n-i)}{n(n-1)}\right]\xi_i = \sum_{i=1}^{[n/2]} \left[\frac{2i(n-i)}{n(n-1)}\right]\eta_i \tag{6}$$

which is also linear function of $\xi_1, \ldots, \xi_{n-1}$, or $\eta_1, \ldots, \eta_{[n/2]}$. One common feature of these linear estimators is that their coefficients, that is $1/a_n, \ldots, 1/a_n$ for $\hat{K}$ and $2i(n-i)/n(n-1)$, $(i = 1, \ldots, n-1)$ for $\hat{\pi}$, are predetermined constants. A linear estimator with predetermined coefficients is not in general a best linear estimator. To demonstrate this, we consider a sample of only three sequences. There is only one topology for three sequences (Figure 2) and $\boldsymbol{\xi} = (\xi_1, \xi_2)^T$ where ξ_1 is the sum of the numbers of mutations in all the three external branches and ξ_2 is the number of mutations in the internal branch. From KINGMAN's (1982a,b) coalescent theory, it is easy to show [for example, FU and LI (1993b) that

$$E(\xi_1) = \theta \qquad E(\xi_2) = \tfrac{1}{2}\theta$$

$$\mathrm{Var}(\xi_1) = \theta + \tfrac{1}{2}\theta^2 \qquad \mathrm{Var}(\xi_2) = \tfrac{1}{2}\theta + \tfrac{1}{4}\theta^2$$

$$\mathrm{Cov}(\xi_1, \xi_2) = \tfrac{1}{4}\theta^2$$

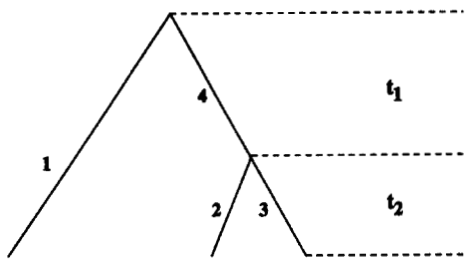

FIGURE 2.—A genealogy of three sequences. ξ_1 is the sum of the numbers of mutations in branch 1, 2 and 3 while ξ_2 is the number of mutations in branch 4. See texts for the explanation of t_1 and t_2.

Consider the following linear function of ξ_1 and ξ_2

$$f = a\xi_1 + b\xi_2 \tag{7}$$

In order for f to be an unbiased estimator of θ, it must satisfy

$$E(f) = a\theta + (b/2)\theta = \theta$$

Therefore, $b = 2(1 - a)$. Substituting $2(1 - a)$ for b in (7), we obtain the variance of f as

$$\begin{aligned}\mathrm{Var}(f) &= a^2\mathrm{Var}(f_1) + 4(1-a)^2\mathrm{Var}(f_2) \\ &\quad + 4a(1-a)\mathrm{Cov}(f_1, f_2) \\ &= [a^2 + 2(1-a)^2]\theta + \tfrac{1}{2}[1 + (1-a)^2]\theta^2\end{aligned}$$

It is easy to show that the value of a for the most efficient estimator, that is, the one with smallest variance, must be $(4 + \theta)/(6 + \theta)$. Therefore,

$$f = \frac{4+\theta}{6+\theta}\xi_1 + \frac{4}{6+\theta}\xi_2 \tag{8}$$

is the best linear unbiased estimator of θ and the coefficients of f are not predetermined. Equation 8 suggests that to have an optimal estimating scheme, we should give more weight to ξ_1 than to ξ_2 when there are many segregating sites in a sample (indicating that θ is large) and we should give about equal weights to both ξ_1 and ξ_2 when there are few segregating sites (indicating that θ is small). Nevertheless, we should not fix the coefficients in advance.

Similar analyses can be made for samples of more than three sequences and a general framework by FU (1994a) can be applied. The framework was developed for a vector of variables whose means are linear functions of θ and variances and covariances are quadratic functions of θ. In next section, we shall show that

$$E(\boldsymbol{\xi}) = \theta\boldsymbol{\alpha} \tag{9}$$

$$\mathrm{Var}(\boldsymbol{\xi}) = \theta\mathbf{D}_\alpha + \theta^2\boldsymbol{\Sigma} \tag{10}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{n-1})$; $\mathbf{D}_\alpha = \mathrm{diag}(\alpha_1, \ldots, \alpha_{n-1})$, *i.e.*, a matrix whose diagonal elements are $\alpha_1, \ldots, \alpha_{n-1}$ and all non-diagonal elements are zero, and $\boldsymbol{\Sigma} = \{\sigma_{ij}\}$, $i, j = 1, \ldots, n-1$; α_i and σ_{ij} are all constants independent of θ once the model is given. With some modifications of the notation in FU (1994a), we have that the best linear unbiased estimator of θ from $\boldsymbol{\xi}$ is given by

$$\frac{\boldsymbol{\alpha}^T(\mathbf{D}_\alpha + \theta\boldsymbol{\Sigma})^{-1}}{\boldsymbol{\alpha}^T(\mathbf{D}_\alpha + \theta\boldsymbol{\Sigma})^{-1}\boldsymbol{\alpha}}\boldsymbol{\xi} \tag{11}$$

However, Equation 11 does not provide a direct estimate of θ because its computation requires the value of θ which is unknown. The problem can be circumvented by the following iteration procedure. Define a series

$$\theta_k = \frac{\boldsymbol{\alpha}^T(\mathbf{D}_\alpha + \theta_{k-1}\boldsymbol{\Sigma})^{-1}}{\boldsymbol{\alpha}^T(\mathbf{D}_\alpha + \theta_{k-1}\boldsymbol{\Sigma})^{-1}\boldsymbol{\alpha}}\xi, \quad k = 1, \ldots \tag{12}$$

and assign an arbitrary non-negative value to θ_0. Then the limit $\hat{\theta}_\xi$ of the series is taken as our estimate of θ. $\hat{\theta}_\xi$ will be referred to as the BLUE (best linear unbiased estimator) of θ from $\boldsymbol{\xi}$ and the estimation procedure as the BLUE procedure, though strictly speaking the estimator is not a linear function in $\boldsymbol{\xi}$ because of the iteration.

From the definition of $\boldsymbol{\eta}$ and Equations 9 and 10, it is simple to see that

$$E(\boldsymbol{\eta}) = \theta\boldsymbol{\beta} \tag{13}$$

$$\mathrm{Var}(\boldsymbol{\eta}) = \theta\mathbf{D}_\beta + \theta^2\boldsymbol{\Gamma} \tag{14}$$

where $\mathbf{D}_\beta = \mathrm{diag}(\beta_1, \ldots, \beta_{[n/2]})$, $\boldsymbol{\Gamma} = \{\gamma_{ij}\}$ and

$$\beta_i = \frac{\alpha_i + \alpha_{i,n-i}}{1 + \delta_{i,n-i}}$$

$$\gamma_{ij} = \frac{\sigma_{ij} + \sigma_{i,n-j} + \sigma_{n-i,j} + \sigma_{n-i,n-j}}{(1 + \delta_{i,n-i})(1 + \delta_{j,n-j})}$$

The BLUE $\hat{\theta}_\eta$ of θ from $\boldsymbol{\eta}$ is the limit of the series

$$\theta_k = \frac{\boldsymbol{\beta}^T(\mathbf{D}_\beta + \theta_{k-1}\boldsymbol{\Gamma})^{-1}}{\boldsymbol{\beta}^T(\mathbf{D}_\beta + \theta_{k-1}\boldsymbol{\Gamma})^{-1}\boldsymbol{\beta}}\boldsymbol{\eta}, \quad k = 1, \ldots \tag{15}$$

and θ_0 can be any non-negative value.

From (9), we have that

$$E(\hat{K}) = \theta\frac{\sum_i \alpha_i}{a_n} \qquad E(\hat{\pi}) = \theta\sum_i \frac{2i(n-i)}{n(n-1)}\alpha_i$$

Therefore we define a general WATTERSON estimator $\hat{\theta}_K$ and a general TAJIMA estimator $\hat{\theta}_\pi$ as

$$\hat{\theta}_K = \frac{a_n}{\sum \alpha_i}\hat{K} \tag{16}$$

$$\hat{\theta}_\pi = \left(\sum_i \frac{2i(n-i)}{n(n-1)}\alpha_i\right)^{-1}\hat{\pi} \tag{17}$$

Under the neutral Wright-Fisher model we have that $\hat{\theta}_K = \hat{K}$ and $\hat{\theta}_\pi = \hat{\pi}$, which are also true under the neutral model with recombination which will be shown later.

When θ is close to zero, it is easy to see from (11) or (12) that

$$\hat{\theta}_\xi \approx \frac{\boldsymbol{\alpha}^T\mathbf{D}_\alpha^{-1}}{\boldsymbol{\alpha}^T\mathbf{D}_\alpha^{-1}\boldsymbol{\alpha}}\boldsymbol{\xi} = \frac{K}{\sum_k \alpha_k} = \hat{\theta}_K.$$

It thus follows that when θ is small, the general WATTERSON estimator is approximately the best linear estimator of θ.

Treating the vector of the coefficients of $\hat{\theta}_\xi$

$$\mathbf{u} = \frac{\boldsymbol{\alpha}^T(\mathbf{D}_\alpha + \hat{\theta}_\xi\boldsymbol{\Sigma})^{-1}}{\boldsymbol{\alpha}^T(\mathbf{D}_\alpha + \hat{\theta}_\xi\boldsymbol{\Sigma})^{-1}\boldsymbol{\alpha}} \tag{18}$$

as a vector of constants as did in FU (1994a), we have that

$$\mathrm{Var}(\hat{\theta}_\xi) = a_\xi\theta + b_\xi\theta^2$$

where $a_\xi = \mathbf{u}^T\mathbf{D}_\alpha\mathbf{u}$ and $b_\xi = \mathbf{u}^T\boldsymbol{\Sigma}\mathbf{u}$. It is easy to see that a nearly unbiased estimate of the variance of $\hat{\theta}_\xi$ is

$$V_\xi = a_\xi\hat{\theta}_\xi + \frac{\hat{\theta}_\xi(\hat{\theta}_\xi - a_\xi)b_\xi}{1 + b_\xi}. \tag{19}$$

Similarly, a nearly unbiased estimate of the variance of $\hat{\theta}_\eta$ is given by

$$V_\eta = a_\eta\hat{\theta}_\eta + \frac{\hat{\theta}_\eta(\hat{\theta}_\eta - a_\eta)b_\eta}{1 + b_\eta} \tag{20}$$

where $a_\eta = \mathbf{v}^T\mathbf{D}_\beta\mathbf{v}$, $b_\eta = \mathbf{v}^T\boldsymbol{\Gamma}\mathbf{v}$ and $\mathbf{v}$ is the vector of the coefficients of $\hat{\theta}_\eta$ given by

$$\mathbf{v} = \frac{\boldsymbol{\beta}^T(\mathbf{D}_\beta + \hat{\theta}_\eta\boldsymbol{\Gamma})^{-1}}{\boldsymbol{\beta}^T(\mathbf{D}_\beta + \hat{\theta}_\eta\boldsymbol{\Gamma})^{-1}\boldsymbol{\beta}).} \tag{21}$$

ESTIMATION OF THE MEAN AND VARIANCE OF $\boldsymbol{\xi}$ AND $\boldsymbol{\eta}$

Consider a sample of n sequences. The time interval between the moment at which the sample is taken and the time representing the most recent common ancestor of the n sequences can be divided to a number of periods by the events occurred. In this paper, we mean, by an event, a coalescence, a recombination or a migration, but not a mutation. For convenience of notations, we treat the moment at which the sample is taken as an event. Let the events be numbered according to their orders of occurrence and t_k be the time length (in terms of the number of generations) between the kth event and $(k+1)$th event. Under the neutral Wright-Fisher model, there are only coalescent events, so t_k represents the time length for the period during which the sample has $k+1$ ancestral sequences; therefore t_k is a $(k+1)$-coalescent time.

Suppose there are in total L sites in each sequence. Each site can be regarded as a locus and there is a genealogy for each site which connects the n nucleotides at the site to their most recent common ancestor. Consider the genealogy of the lth site. We can see that the time length of a branch of the genealogy must be of the form

$$t_i + \ldots + t_j$$

where i is the number of the event after which the branch starts and j is the number of the event at which the branch ends. For the genealogy of lth site, the length $l_k^{(l)}$ of the branchs of size k is given by

$$l_k^{(l)} = s_{k1}^{(l)}t_1 + \ldots + s_{kp}^{(l)}t_p$$

where $s_{ki}^{(l)}$ is an index variable representing the number of times t_i appears in the time lengths of branchs of size k; p is the total number of events. The average time length of branches of size k over all sites is therefore

$$l_k = \frac{1}{L}\sum_l l_k^{(l)} = \sum s_{ki}t_i \qquad \text{where} \qquad s_{ki} = \frac{1}{L}\sum_{l=1}^{L} s_{ki}^{(l)}.$$

When there is no recombination, all the L genealogies are the same and consequently $s_{ki} = s_{ki}^{(l)} = \ldots = s_{ki}^{(L)}$. For example we have for the genealogy in Figure 2 that $s_{12} = 3$, $s_{11} = 1$, $s_{22} = 0$ and $s_{21} = 1$.

Because μ is defined as the mutation rate per sequence per generation, the mutation rate per site per generation is therefore μ/L. Assume that the number of mutations in a branch of the genealogy of a site is a Poisson variable with parameter $l\mu/L$ where l is the length of the branch. Then the number of k-mutations in all the L genealogies is a Poisson variable with parameter $(\mu/L)\sum_l l_k^{(l)} = l_k\mu$. In other words, we assume that

$$\Pr(\xi_i = k \mid l_i) = \frac{e^{l_i\mu}(l_i\mu)^k}{k!}$$

For each sample, we define the following two quantities

$$\omega_i = \sum_k s_{ik}(t_k), \tag{22}$$

$$\phi_{ij} = \sum_k s_{ik}s_{jk}E^2(t_k). \tag{23}$$

Then we have for each sample

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

$$\begin{aligned}\mathrm{Var}(\xi_i) &= E(\xi_i^2) - E^2(\xi_i)\\ &= \omega_i\theta + \left\{E\left[\left(\sum_{k=2}^{n} s_{ik}t_k\right)^2\right] - \left[\sum_{k=2}^{n} s_{ik}E(t_k)\right]^2\right\}\theta^2\\ &= \omega_i\theta + \phi_{ii}\theta^2\end{aligned}$$

and

$$\begin{aligned}\mathrm{Cov}(\xi_i, \xi_j) &= E(\xi_i\xi_j) - E(\xi_i)E(\xi_j)\\ &= E_{l_i,l_j}(\mu^2 l_i l_j) - E(\xi_i)E(\xi_j)\\ &= \mu^2 E(\textstyle\sum s_{ik}t_k \sum s_{jk}t_k) - \omega_i\omega_j\theta^2\\ &= \mu^2[\textstyle\sum s_{ik}E(t_k) \sum s_{jk}E(t_k) + \sum s_{ik}s_{jk}E^2(t_k)] - \omega_i\omega_j\theta^2\\ &= \phi_{ij}\theta^2\end{aligned}$$

TABLE 1

Variances of four estimators under the neutral Wright-Fisher model

θ	n	$V_{\min}$	$sv(\hat{\theta}_\pi)$	$sv(\hat{\theta}_K)$	$sv(\hat{\theta}_\xi)$	$sv(\hat{\theta}_\eta)$	$\bar{V}_\xi$	$\bar{V}_\eta$
2	5	2.11	2.48	2.28	2.21	2.28	1.87	1.76
	10	1.32	1.94	1.48	1.39	1.46	1.42	1.50
	20	0.93	1.72	1.07	0.98	1.04	1.02	1.12
	50	0.66	1.63	0.78	0.70	0.74	0.74	0.80
	300	0.42	1.54	0.48	0.43	0.45	0.45	0.47
5	5	9.16	11.75	10.65	9.73	10.59	9.54	9.60
	10	5.16	9.10	6.59	5.60	6.41	5.66	6.53
	20	3.35	7.98	4.56	3.63	4.33	3.67	4.33
	50	2.18	7.65	3.17	2.38	2.84	2.38	2.86
	300	1.25	7.30	1.84	1.33	1.51	1.32	1.49
10	5	31.00	41.83	37.71	32.50	37.44	32.43	36.04
	10	16.16	32.21	23.02	17.70	22.33	17.67	22.17
	20	9.68	28.67	15.48	10.79	14.27	10.91	14.26
	50	5.77	26.96	10.31	6.48	5.85	6.53	8.79
	300	2.96	25.39	5.73	3.25	4.122	3.23	4.04
20	5	112.22	157.28	141.46	117.00	140.44	116.88	139.15
	10	54.96	121.83	84.94	59.79	81.26	59.64	80.65
	20	30.50	107.08	56.58	34.51	51.35	34.68	51.01
	50	16.38	100.52	37.24	19.23	29.58	19.23	29.41
	300	7.28	95.83	19.79	8.44	11.85	8.43	11.87
50	5	655.95	943.55	841.26	672.60	832.82	672.01	834.17
	10	304.88	719.55	500.64	329.28	479.03	328.65	477.61
	20	156.57	638.37	330.25	179.90	294.36	179.45	294.23
	50	73.73	603.11	214.87	92.84	166.87	92.74	164.16
	300	25.85	571.70	112.51	34.25	57.44	33.99	056.29

Note: Results are based on 50,000 simulated samples for each combination of θ and n. sv, sampling variance. $\bar{V}_\xi$ and $\bar{V}_\eta$ are, respectively, the mean of V_ξ given by Equation 19 and the mean of V_η given by Equation 20.

Suppose there are in total T genealogies for a sample of n sequences and let $\omega_i^{(k)}$, $\phi_{ij}^{(k)}$ be the ω_i, ϕ_{ij} defined above for genealogy k and p_k be the probability of observing genealogy k. Define

$$\alpha_i = \sum_k^T \omega_i^{(k)} p_k \tag{24}$$

$$\sigma_{ij} = \sum_k^T (\phi_{ij}^{(k)} + \omega_i^{(k)}\omega_k^{(k)})p_k - \alpha_i\alpha_j \tag{25}$$

Then, it is easy to see that the mean and variance of $\boldsymbol{\xi}$ are given respectively by (9) and (10). Since T is a very large number for even a modest sample size and p_k is not always easy to compute, it is impractical to obtain $\boldsymbol{\alpha}$ and $\boldsymbol{\Sigma}$ by examining all possible genealogies. To date, analytical solutions for $\boldsymbol{\alpha}$ and $\boldsymbol{\Sigma}$ are available only for the neutral Wright-Fisher model (Fu 1994b). Analytical solutions for $\boldsymbol{\alpha}$ and $\boldsymbol{\Sigma}$ simplify computations tremendously, but when they are not available, the values of $\boldsymbol{\alpha}$ and $\boldsymbol{\Sigma}$ have to be estimated, which can be done as follows.

Suppose an algorithm is available to generate genealogies of samples under a given population model and let G be the total number of genealogies randomly generated. Then according to the standard theory of Monte-Carlo integration [for example, Hammersley and Handscomb (1965)], one can estimate α_i and σ_{ij} by

$$\hat{\alpha}_i = \frac{1}{G}\sum_k^G \omega_i^{(k)} \qquad \hat{\sigma}_{ij} = \frac{1}{G}\sum_k^G (\phi_{ij}^{(k)} + \omega_i^{(k)}\omega_k^{(k)}) - \hat{\alpha}_i\hat{\alpha}_j,$$

respectively. Once the estimate $\hat{\boldsymbol{\alpha}}$ of $\boldsymbol{\alpha}$ and the estimate $\hat{\boldsymbol{\Sigma}}$ of $\boldsymbol{\Sigma}$ are obtained, we can obtain the estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\Gamma}$. The accuracies of these estimations obviously should increase with the number of genealogies examined. My experience suggests that the number G of genealogies in the proximity of 10,000 is usually sufficient for the purpose of estimating θ. Because algorithms based on coalescent theory are usually very efficient, the need to estimate $\boldsymbol{\alpha}$ and $\boldsymbol{\Sigma}$ does not pose a serious burden of computation.

THE NEUTRAL WRIGHT-FISHER MODEL

The neutral Wright-Fisher model is the simplest model in coalescent theory and is often selected to be the null model in studying DNA polymorphisms. Because the lower bound of the variance of all unbiased estimators of θ under this model is known to be

$$V_{\min} = \theta\left(\sum_{k=1}^{n-1}\frac{1}{\theta + k}\right)^{-1} \tag{26}$$

(Fu and Li 1993a) and because $\boldsymbol{\alpha}$ and $\boldsymbol{\Sigma}$ are known analytically (Fu 1994b), we can obtain the efficiencies of

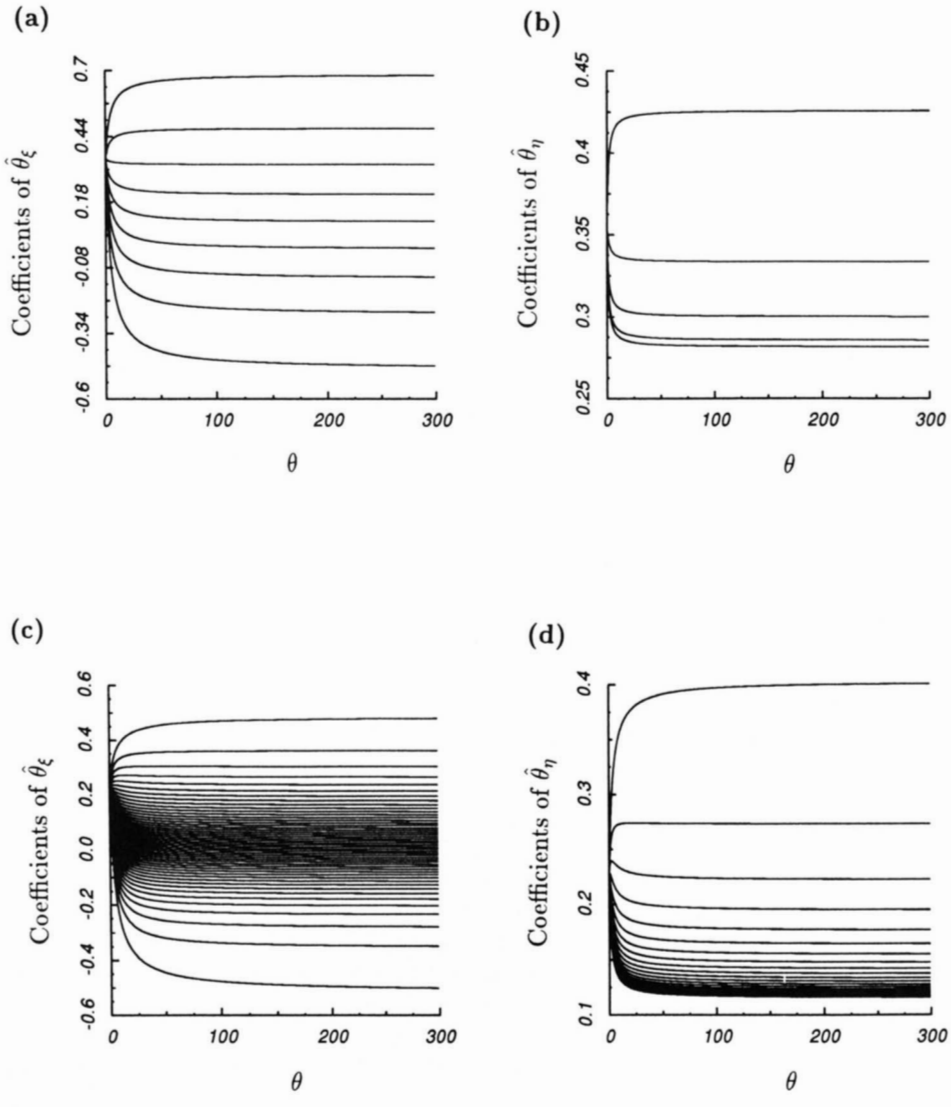

FIGURE 3.—Coefficients of $\hat{\theta}_\xi$ (panels a and c) and $\hat{\theta}_\eta$ (panel b and d) as functions of θ. The sample size n is 10 for panels a and b, and 50 for panels c and d. In panel a and c the curves from top down are u_1, u_2, ..., respectively, and in panels b and d the curves from top down are v_1, v_2, ... respectively.

estimators, $\hat{\theta}_\pi$, $\hat{\theta}_K$, $\hat{\theta}_\xi$ and $\hat{\theta}_\eta$ by comparing their variances to the lower bound $V_{\min}$.

Simulated samples were used to measure the performances of the four estimators. For a given values of θ and sample size n, we generated a large number of samples using straightforward coalescent algorithm (KINGMAN 1982a,b; HUDSON 1983; TAJIMA 1983); the values of the vector $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ from each sample were then used to obtain $\hat{\theta}_\pi$, $\hat{\theta}_K$, $\hat{\theta}_\xi$ and $\hat{\theta}_\eta$ with both analytical values or estimated values of $\boldsymbol{\alpha}$ and $\boldsymbol{\Sigma}$. The results using analytical $\boldsymbol{\alpha}$ and $\boldsymbol{\Sigma}$ are summarized in Table 1.

Table 1 shows that the variance of $\hat{\theta}_\xi$ is only marginally larger than $V_{\min}$, suggesting that $\hat{\theta}_\xi$ is a very efficient estimator of θ. Comparing the variance of $\hat{\theta}_\xi$ to that of the estimator UPBLUE [Table 3 of FU (1994a)], we find that $\hat{\theta}_\xi$ is slightly less efficient than UPBLUE of θ. This is expected because UPBLUE uses more information than $\hat{\theta}_\xi$. Nevertheless, $\hat{\theta}_\xi$ is substantially better than WATTERSON's estimator $\hat{K}$ and TAJIMA's estimate $\hat{\pi}$. The latter always has the largest variance among the four estimators considered. The percentage of variance reduction by $\hat{\theta}_\xi$ over $\hat{K}$ increases with the value of θ and sample size n. In comparison $\hat{\theta}_\eta$ is only marginally better than $\hat{\theta}_K$ when sample size is small but become considerably better than $\hat{\theta}_K$ when sample size and θ are both large. Table 1 also shows that V_ξ given by (19) and V_η given by (20) are nearly unbiased estimators of $\mathrm{Var}(\hat{\theta}_\xi)$ and $\mathrm{Var}(\hat{\theta}_\eta)$ respectively.

The performances of $\hat{\theta}_\xi$ and $\hat{\theta}_\eta$ using estimated values of $\boldsymbol{\alpha}$ and $\boldsymbol{\Sigma}$ (data not shown) are found to be almost indistinguishable from those using analytical results of $\boldsymbol{\alpha}$ and $\boldsymbol{\Sigma}$, as long as reasonably large number of genealogies (for example 10,000) are used to estimate $\boldsymbol{\alpha}$ and $\boldsymbol{\Sigma}$. This indicates that the BLUE procedure proposed in this paper is a powerful Monte-Carlo method for estimation θ.

We also examined the coefficient vectors $\mathbf{u}$ and $\mathbf{v}$ to see the relative contributions of ξ_i ($i = 1, \ldots, n-1$) and $\eta_i (i = 1, \ldots, [n/2])$ to $\hat{\theta}_\xi$ and $\hat{\theta}_\eta$. We have shown earlier that when θ is small, the $\hat{\theta}_\xi$ is close to WATTERSON's estimate, which gives the same weight to all the frequencies, $\xi_1, \ldots, \xi_{n-1}$. On the other hand, when θ is

TABLE 2

Variances of four estimators under the neutral model with recombinations

n	P	$sv(\hat{\theta}_\pi)$	$sv(\hat{\theta}_K)$	$sv(\hat{\theta}_\xi)$	$sv(\hat{\theta}_\eta)$	$\tilde{V}_\xi$	$\tilde{V}_\eta$
$\theta = 10$							
10	0.0	31.97	22.82	17.56	22.11	17.41	21.69
	1.0	25.72[a]	18.84	15.89	18.41	16.35	19.11
		26.74[b]	19.27	16.05	18.69	16.22	18.61
	10.0	19.15	14.18	12.47	13.99	12.61	14.24
		13.78	10.67	10.45	10.60	10.96	11.15
50	0.0	26.81	10.50	6.87	9.06	6.16	8.24
	1.0	21.58	8.71	6.50	7.83	5.60	6.90
		22.58	8.96	6.57	7.96	5.85	7.70
	10.0	15.85	6.69	5.48	6.15	5.01	5.89
		11.77	5.36	5.01	5.12	4.84	5.17
$\theta = 50$							
10	0.0	716.10	495.55	328.34	475.53	327.81	479.45
	1.0	555.77	395.04	292.66	381.75	305.07	402.32
		593.73	413.02	299.67	394.47	306.55	393.28
	10.0	398.12	283.30	216.62	276.72	219.41	284.75
		269.01	199.08	184.07	194.71	194.53	206.04
50	0.0	598.35	214.02	99.41	167.92	82.86	150.20
	1.0	462.39	172.48	97.45	142.86	72.73	118.84
		492.79	179.08	98.08	146.37	74.66	139.10
	10.0	327.22	123.96	80.11	105.95	60.23	95.60
		222.13	90.20	73.09	80.30	60.50	76.52

[a] Two-loci model.

[b] The infinite-sites model. Covariance matrix for each combination of n and R is estimated from 10,000 simulated samples. See the footnote to Table 1 for sv, $\tilde{V}_\xi$ and $\tilde{V}_\eta$.

extremely large, $\hat{\theta}_\xi$ is approximately

$$\frac{\boldsymbol{\alpha}^T\Sigma^{-1}}{\boldsymbol{\alpha}^T\Sigma^{-1}\boldsymbol{\alpha}}\boldsymbol{\xi}$$

while the coefficients for intermediate values of θ should lie between the two extremes. This is confirmed by Figure 3 where the coefficients of estimates for several sample sizes are plotted. The coefficients are very sensitive to θ when θ is small but become stable when θ is large.

Suppose u_i and v_i are the ith elements of the coefficient vector $\mathbf{u}$ and $\mathbf{v}$, respectively. Then Figure 3 shows for $i < j$, we have

$$u_i \geq u_j \quad \text{and} \quad v_i \geq v_j. \tag{27}$$

This suggests that the information from ξ_i (or η_i) is more reliable than the information from ξ_j (or η_j) for the purpose of estimating θ. This is indeed the case because although ξ_i/α_i, $(i = 1, \ldots, n-1)$ and η_i/β_i, $(i = 1, \ldots, [n/2])$ all have the same expectation θ, their variances have the relationships

$$\mathrm{Var}\left(\frac{\xi_i}{\alpha_i}\right) < \mathrm{Var}\left(\frac{\xi_j}{\alpha_j}\right) \qquad \mathrm{Var}\left(\frac{\eta_i}{\beta_i}\right) < \mathrm{Var}\left(\frac{\eta_j}{\beta_j}\right).$$

for $i < j$. However it is found, for example from Figure 3, that some of the coefficients of $\hat{\theta}_\xi$ are negative when θ is not too small. This is surprising because in comparison the coefficients of $\hat{\theta}_\eta$ are all positive and so are all the coefficients of the UPBLUE of θ (Fu 1994a). It is possible to construct a linear estimator similar to $\hat{\theta}_\xi$ by restricting all coefficients to be non-negative, but such an estimator was found to have larger variance than that of $\hat{\theta}_\xi$.

THE NEUTRAL MODEL WITH RECOMBINATIONS

When an autosomal locus studied is either very large or consisting of several separate regions, recombinations cannot be neglected. The coalescent theory of the neutral model with recombination and algorithms to generate samples under this model have been developed by Hudson (1983), Hudson and Kaplan (1985) and Kaplan and Hudson (1987). An event under the this model is either a coalescence or a recombination. Suppose between two consecutive events, there are m ancestral sequences. Hudson (1983) showed that the expected time length between the two events is

$$\frac{4N}{Rmc + m(m-1)}$$

where the recombination parameter $R = 4Nr$ and r is the recombination rate per sequence per generation; $c(0 \leq c \leq 1)$ is the average proportion of recombinable sites at which recombinations are relevant to the sample [see Hudson (1983) for detail].

Since the coalescence process of a single site is identical to that under the neutral Wright-Fisher model, expectation of the frequency $\xi_i^{(k)}$ of i-mutations in the genealogy of kth site is

$$E(\xi_i^{(k)}) = \frac{1}{i}\frac{\theta}{L} \tag{28}$$

The expectation of the total number ξ_i of i-mutations is

$$E(\xi_i) = \sum E(\xi_i^{(k)}) = \frac{1}{i}\theta$$

It follows that recombinations do not change the expectation of the frequency of i-mutations. Consequently, expectation of the total number of mutations remains unchanged, which has long been known. However, recombinations change the variances of $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ as they change the variance of the total number K of segregating sites.

We again used simulated samples to investigate the performances of $\hat{\theta}_\pi$, $\hat{\theta}_K$, $\hat{\theta}_\xi$ and $\hat{\theta}_\eta$ under the neutral model with recombination. We focused on two cases. The first is a two loci model so that recombination occurs only between the loci and each locus follows the infinite-sites model. The second is an infinite-loci model such that recombinations can occur between any two sites. In the first case, we assume that the values of θ for the two loci are the same. In both cases, it reduces to the neutral WRIGHT-FISHER model when recombination rate r is zero. Table 2 presents the results of simulations for several sets of parameters.

Because recombinations reduce the correlation between two sites, we expect for any estimator of θ that the larger the recombination parameter R is, the smaller the variance of the estimator is. This is confirmed by simulation results as it is clear in Table 2 that the variance of each of the four estimators decreases with increasing value of R. Table 2 also shows that $\hat{\theta}_\xi$ is the best estimator among the four; the second best is the $\hat{\theta}_\eta$ which is traced closely by $\hat{\theta}_K$; the worsest estimator is $\hat{\theta}_\pi$. Comparing the results of the two-loci and the infinite-loci models shows that when R is small, the variance of an estimator under the two-loci model is smaller than under the infinite-loci model; while when R is large, the reverse is true. This seems to be logical because when the number of recombinations is very small, for example only one recombination, the best site for recombination is close to the middle of the sequence as far as estimation of θ is concerned. The estimators perform better under the two loci model when R is small because the recombination occurs precisely at the middle of the sequences. On the other hand, when the number of recombinations are large, it is better to have them occurred at as many sites as possible. Therefore, these estimators perform better when R is large under the infinite-loci model than under the two-loci model. Table 2 also shows that V_ξ and V_η are in general adequate estimators of $\mathrm{Var}(\hat{\theta}_\xi)$ and $\mathrm{Var}(\hat{\theta}_\eta)$ respectively, but they become biased when both θ and sample size n are large.

Under the infinite-loci model, WATTERSON's estimator is getting closer to $\hat{\theta}_\xi$ with the increase of R. This suggests that when recombinations are frequent, WATTERSON's estimator $\hat{\theta}_K$ is quite efficient even for modest sample sizes. Examining the coefficients of $\hat{\theta}_\xi$ (Figure 4) shows that, for a given sample size n, the coefficients of $\hat{\theta}_\xi$ are moving towards those of $\hat{\theta}_K$ with increasing R. This implies that $\hat{\theta}_K$ is becoming not only efficient but the same estimator as $\hat{\theta}_\xi$ and $\hat{\theta}_\eta$. This can be explained as follows.

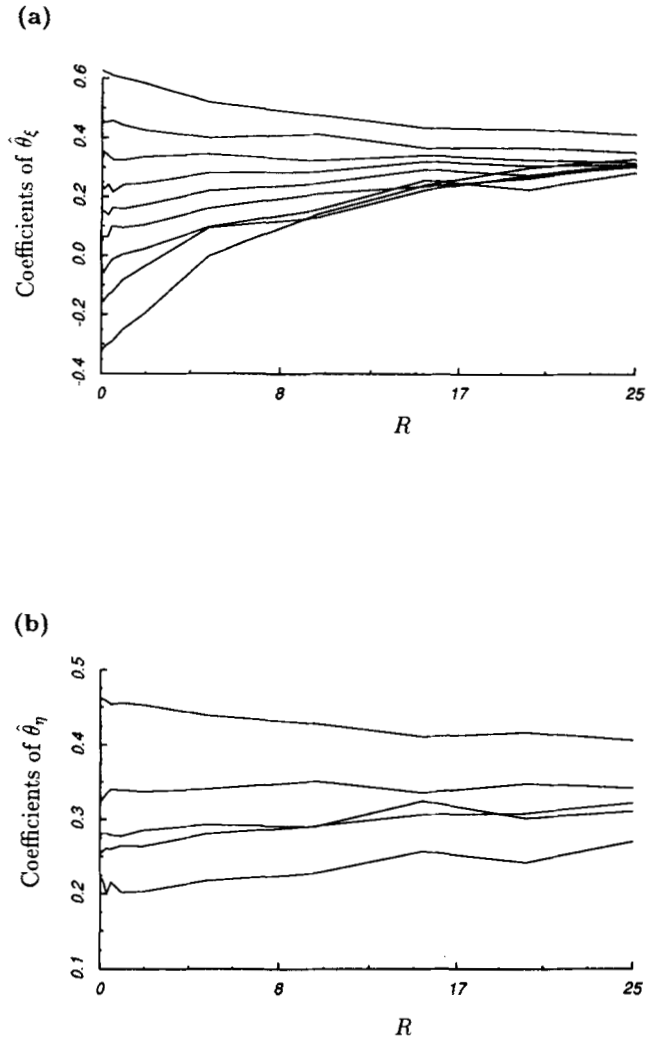

FIGURE 4.—Coefficients of $\hat{\theta}_\xi$ (panel a) and $\hat{\theta}_\eta$ (panel b) as functions of recombination parameter R under the inifite-loci model with $\theta = 20$ and $n = 10$. The curves from top down (when $R = 0$) represent $u_1, \ldots, u_9$ in panel a and $v_1, \ldots, v_5$ in panel b. The coefficients for each value of R are averaged over 10,000 samples.

When R is large, the coalescence processes of different sites are nearly independent which means that the best estimator of θ should be close to the sum of best estimator of θ of each site; because the θ per site is very small, the best estimator for each site is close to WATTERSON's estimator which has been shown before and thus the best estimator of θ is close to WATTERSON's estimator when R is large. Under the neutral model with recombination, the coefficients of $\hat{\theta}_\xi$ and $\hat{\theta}_\eta$ also satisfy the relationship given by (27).

THE NEUTRAL WRIGHT'S FINITE-ISLANDS MODEL

Let d be the number of islands and m be the overall migration rate. The neutral Wright's finite-islands model assumes that at each generation, each individual has probability $m/(d-1)$ to migrate from his current island to each of the other $d-1$ islands. We shall assume

TABLE 3

Variances of four estimators under the neutral Wright' finite-islands model

θ	M	n	Scheme	$sv(\hat{\theta}_\pi)$	$sv(\hat{\theta}_K)$	$sv(\hat{\theta}_\xi)$	$sv(\hat{\theta}_\eta)$	$\tilde{V}_\xi$	$\tilde{V}_\eta$
Two islands									
5	0.1	20	*A*	49.05	29.03	6.50	28.58	7.21	29.88
5	1.0	20	*A*	10.53	6.25	4.54	6.00	5.60	8.17
5	5.0	20	*A*	7.28	4.28	3.37	4.14	4.43	6.30
5	0.1	20	*B*	15.35	10.52	2.66	8.39	2.68	8.22
5	1.0	20	*B*	7.32	4.09	2.62	3.89	3.24	5.04
5	5.0	20	*B*	7.09	3.88	2.52	3.67	3.71	5.32
20	0.1	50	*A*	851.70	439.10	36.69	432.95	37.52	442.44
20	1.0	50	*A*	144.27	62.67	31.14	59.14	38.18	80.55
20	5.0	50	*A*	101.82	41.91	22.45	37.35	29.81	56.50
20	0.1	50	*B*	236.94	133.96	15.80	69.53	16.11	81.32
20	1.0	50	*B*	100.03	39.68	15.73	30.92	18.37	42.98
20	5.0	50	*B*	95.71	34.80	15.63	26.87	20.53	42.25
Five islands									
5	0.1	20	*A*	114.19	68.84	32.99	68.86	36.95	73.34
5	1.0	20	*A*	16.49	11.03	9.46	11.00	12.23	17.25
5	5.0	20	*A*	8.09	4.98	3.66	4.92	5.14	8.16
5	0.1	20	*B*	8.50	6.54	3.29	5.80	4.72	10.36
5	1.0	20	*B*	6.79	3.75	1.95	3.39	2.87	6.06
5	5.0	20	*B*	6.18	3.27	1.94	3.00	2.84	5.20
20	0.1	50	*A*	1709.09	1052.06	360.56	1085.31	373.79	1227.03
20	1.0	50	*A*	253.37	126.98	86.01	124.31	111.24	205.75
20	5.0	50	*A*	117.48	55.45	31.35	53.59	44.27	90.85
20	0.1	50	*B*	133.40	95.90	30.76	86.94	40.33	127.38
20	1.0	50	*B*	99.82	41.75	14.91	31.65	18.91	55.11
20	5.0	50	*B*	92.79	33.54	13.79	25.45	17.87	42.23

Note; α and Σ are estimated from 10,000 samples and the results in each row of the table is from 10,000 samples. *A*, extreme sampling scheme; *B*, balanced sampling scheme. Also see the footnote to Table 1 for sv, $\tilde{V}_\xi$ and $\tilde{V}_\eta$.

that all the islands have the same effective population size N and that the overall migration rate m is sufficiently small so that the probability that two or more individuals in a sample migrate at the same generation can be neglected. Also we assume that there is no recombination.

An event under the WRIGHT's finite-islands model is either a coalescence or a migration. A coalescence event reduces by one the number of ancestral genes in the island where the coalescence event occurs while a migration event reduces by one the number of ancestral genes in one island but increases by one the number of ancestral genes in another island, thus the total number of ancestral genes remains the same. Let $M = 4Nm$ and b_k $(k = 1, \ldots, d)$ be the number of ancestral genes in island k between two consecutive events. STROBECK (1987) showed that the expected time length between the two events is

$$\frac{4N}{M\sum b_k + \sum_k b_k(b_k - 1)}.$$

Algorithms for generating samples under the neutral Wright's finite-islands model were developed by STROBECK (1987) and SLATKIN and MADDISON (1989).

Since there is no recombination, all the sites in the sequences have the same genealogy. However, unlike the model with recombination, the expectation of the number of i-mutations is no longer the same as that under the neutral Wright-Fisher model. Therefore, $\hat{\theta}_K$ is no longer the same as $\hat{K}$, neither is $\hat{\theta}_\pi$ as $\hat{\pi}$. Because migrations tends to increase the time between two coalescent events, $\hat{K}$ and $\hat{\pi}$ both overestimate θ.

We focus on the cases of two islands and five islands and consider two sampling schemes. The first is that all the sequences are taken from one island and the second is that a sample is taken from each island. These two sampling schemes will be referred to as the extreme-sampling scheme and the balanced-sampling scheme respectively and consequently a sample from the extreme-sampling scheme and a sample from the balanced sampling scheme will be will be referred to as an extreme sample and a balanced sample, respectively.

We again use simulated samples to evaluate the performances of the four estimators. Table 3 summarizes the results of simulations. It is obviously that $\hat{\theta}_\xi$ is again the best estimator among the four estimators, but all of them have smaller variances when migration rate is large than when migration rate is small. Similar to the neutral WRIGHT-FISHER model and the neutral model with recombination, superiority of $\hat{\theta}_\xi$ is mostly evident when θ and sample size n are both large and in such cases, $\hat{\theta}_\eta$ also shows considerable improvement over $\hat{\theta}_K$. $\hat{\theta}_\pi$ is again the worst estimator among the four.

Comparing the results for the cases of two islands and five islands, we find that under the extreme-sampling

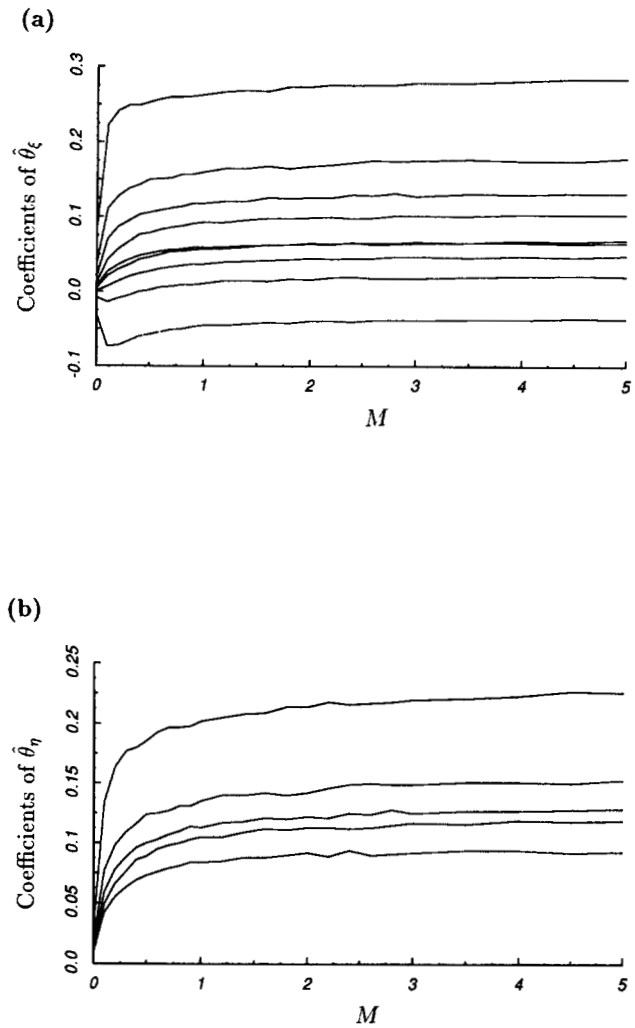

FIGURE 5.—Coefficients of $\hat{\theta}_\xi$ (panel a) and $\hat{\theta}_\eta$ (panel b) as functions of migration parameter M for balanced samples of size 10 from two islands. The curves from top down represent $u_1, \ldots, u_9$ in panel a and $v_1, \ldots, v_5$ in panel b. The coefficients for each value of M are averaged over 10,000 samples and $\theta = 20$.

scheme, all the estimators performs better in the case of two islands than in the case of five islands, suggesting that in general the variance of an estimator of θ increases with the number of islands under the extreme-sampling scheme. For the balanced-sampling scheme, the patterns of the four estimators are not all the same. When M is small, $\hat{\theta}_\xi$ and $\hat{\theta}_\eta$ perform better with smaller number of islands than with larger number of islands but the reverse are true when M is relatively large. In comparison, $\hat{\theta}_K$ and $\hat{\theta}_\pi$ seem to perform better with larger number of islands for all the migration rates we have considered.

Table 3 also shows that the balanced-sampling scheme is always better than the extreme-sampling scheme for the purpose of estimating θ because all the estimators perform better under the former scheme than under the latter one. The extreme sampling scheme should be avoided particularly when migration rate is small. It should be pointed out that we have assumed that all islands have the same effective population size in our

TABLE 4

Variances of four estimators under the infinite-loci neutral model with recombination when R has to be estimated

θ	R	$sv(\hat{\theta}_\pi)$	$sv(\hat{\theta}_K)$	$sv(\hat{\theta}_\xi)$	$sv(\hat{\theta}_\eta)$
		$n = 20$ and $\hat{R} = 5$			
10	0.0	27.84	15.33	11.50	14.65
	3.0	19.15	10.93	9.35	10.30
	8.0	13.25	8.11	7.52	7.81
	10.0	12.60	7.73	7.35	7.47
	15.0	10.40	6.72	6.73	6.64
50	0.0	625.78	329.67	194.21	313.01
	3.0	398.84	214.24	154.07	196.82
	8.0	266.53	150.94	124.43	141.21
	10.0	237.02	136.15	119.21	128.96
	15.0	184.32	109.32	102.25	105.49
		$n = 50$ and $\hat{R} = 10$			
10	0.0	26.68	10.45	7.81	9.61
	3.0	17.77	7.47	6.31	6.93
	8.0	12.45	5.64	5.12	5.32
	10.0	11.43	5.23	4.92	5.01
	15.0	9.75	4.74	4.65	4.63
50	0.0	609.04	220.29	144.36	200.84
	3.0	367.31	139.85	98.50	127.21
	8.0	248.52	97.72	78.56	86.43
	10.0	215.08	90.74	75.06	82.60
	15.0	177.47	75.84	67.16	69.44

Estimations of α and Σ are based on estimated $\hat{R}$ and 10,000 samples. Each row is obtained from 10,000 simulated samples with recombination rate R. Also see the footnote to Table 1 for sv, $\hat{V}_\xi$ and $\hat{V}_\eta$.

simulations. When this is not true, the best sampling scheme is likely the generalized balanced-sampling scheme in which the relative size of a sample from an island with respect to the overall sample size is the same as the relative value of the effective population size of the island with respect to the sum of all effective population sizes.

We also present in Table 3 the means of estimated variances of $\hat{\theta}_\xi$ and $\hat{\theta}_\eta$. It is found that variances of $\hat{\theta}_\xi$ and $\hat{\theta}_\eta$ computed from Equations 19 and 20, respectively are on average overestimates of the true variances under the neutral Wright's finite-islands model.

The coefficients of $\hat{\theta}_\xi$ and $\hat{\theta}_\eta$ for a balanced sample of size 10 from two islands are given in Figure 5. It is interesting to see that the coefficients quickly become stable with the value of M. This example and many others we examined indicate that as far as θ is concerned, samples from islands may be treated as from a panmictic population even when migration parameter M is as small as 1, particularly when the estimator $\hat{\theta}_\xi$ is used. This observation is in accordance with a large body of literature showing that migration is extremely powerful in eliminating differences among local populations. Under the neutral Wright's finite-islands model, the coefficients of $\hat{\theta}_\xi$ and $\hat{\theta}_\eta$ also satisfy the relationship given by (27) which seems to be true in general.

DISCUSSION

We have demonstrated under the neutral WRIGHT-FISHER model, the neutral model with recombination

TABLE 5

Means and MSE of the four estimators under the neutral Wright's finite-islands model for balanced samples

	θ	$\hat{M}$	M	$\tilde{\hat{\theta}}_\pi$	MSE	$\tilde{\hat{\theta}}_K$	MSE	$\tilde{\hat{\theta}}_\xi$	MSE	$\tilde{\hat{\theta}}_\eta$	MSE
A											
	5	0.5	0.1	11.8	131.4	9.2	52.7	6.1	5.3	7.7	26.5
			0.3	6.2	16.6	5.7	8.1	5.2	2.8	5.5	6.2
			0.5	5.1	8.8	5.0	4.8	5.0	2.7	5.0	4.1
			0.7	4.5	6.5	4.7	3.9	4.9	2.5	4.8	3.5
			1.0	4.1	5.9	4.4	3.6	4.8	2.6	4.6	3.4
			5.0	3.5	5.8	3.9	3.6	4.5	2.6	4.2	3.2
	20	5.0	0.1	69.5	5382.4	47.7	1708.0	24.7	69.3	35.7	656.4
			1.0	23.9	162.8	22.6	79.7	21.2	35.3	21.7	64.6
			2.0	21.3	114.9	21.0	59.6	20.6	32.4	20.8	53.5
			5.0	20.1	106.0	20.1	55.6	20.1	30.5	20.1	50.2
			7.0	19.6	102.2	19.7	52.9	19.8	30.4	19.8	46.8
			10.0	19.7	97.7	19.8	51.9	19.9	29.9	19.9	49.1
B											
	5	0.5	0.1	16.1	207.1	13.8	124.6	9.1	26.8	11.1	65.7
			0.3	6.8	17.1	6.5	10.3	5.7	3.1	6.1	7.5
			0.5	4.9	6.9	5.0	4.0	5.0	1.9	5.0	3.4
			0.7	4.1	5.4	4.3	3.3	4.7	1.6	4.5	3.1
			1.0	3.5	5.4	3.8	3.4	4.4	1.7	4.1	2.9
			5.0	2.4	8.1	2.8	5.6	3.7	2.7	3.3	4.0
	20	5.0	0.1	133.9	18949.9	97.7	8458.5	50.3	1228.9	67.4	3291.1
			1.0	29.0	284.0	26.8	138.7	23.8	47.6	24.8	93.2
			2.0	23.5	146.3	22.8	71.6	21.8	29.7	22.2	57.6
			5.0	19.6	88.9	19.8	41.9	20.0	22.0	19.9	36.7
			7.0	19.4	92.1	19.4	42.8	19.5	21.3	19.5	37.1
			10.0	18.5	84.5	18.8	42.2	19.1	21.9	19.0	38.6

$\tilde{\hat{\theta}}_\pi$, $\tilde{\hat{\theta}}_K$, $\tilde{\hat{\theta}}_\xi$ and $\tilde{\hat{\theta}}_\eta$ are, respectively, the means of $\hat{\theta}_\pi$, $\hat{\theta}_K$, $\hat{\theta}_\xi$ and $\hat{\theta}_\eta$. Estimations of $\boldsymbol{\alpha}$ and $\boldsymbol{\Sigma}$ are based on estimated value $\hat{M}$ of M and 10,000 samples. Each row is obtained from 10,000 samples. *A*, Two islands and 10 sequences are taken from each island; *B*, five islands and five sequences are taken from each island.

and the neutral Wright's finite-islands model that the $\hat{\theta}_\xi$ is an efficient estimator of θ. However, comparisons of the performances of various estimators were conducted assuming that the values of parameters other than θ are known. Since in practice the value of migration parameter M and the value of recombination parameter R are likely unknown as well, one has to estimate their values in order to obtain estimates of θ by $\hat{\theta}_\xi$ or $\hat{\theta}_\eta$. It then raises a question: will $\hat{\theta}_\xi$ be still superior when M and R are to be estimated? Although the recombination parameter R can be estimated by, for example, HUDSON and KAPLAN's (1987) method or HUDSON's (1993) method and the migration parameter M by, for example, SLATKIN and MADDISON's (1989) method, but their inferences are still in their infancies and better methods for estimating M and R are likely to appear in the near future. Therefore I decide to examine the performances of the four estimators by using directly erroneous values of M or R to estimate $\boldsymbol{\alpha}$ and $\boldsymbol{\Sigma}$.

Consider the neutral model with recombination first. Since recombinations do not change the expectation of $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$, it is easy to see that $\hat{\theta}_\xi$ and $\hat{\theta}_\eta$ is still unbiased estimators of θ, when the estimated value of R is inaccurate. Table 4 gives a few examples on how the variances of $\hat{\theta}_\xi$ and $\hat{\theta}_\eta$ are affected when the estimate of R is inaccurate. It is found from Table 4 that $\hat{\theta}_\xi$ remains to be the best estimator even when the estimate $\hat{R}$ of R is considerably different from its true value. However, when $\hat{R}$ is different from R, the variances of $\hat{\theta}_\xi$ and $\hat{\theta}_\eta$ increase. For example, consider the case of $\theta = 50$ and sample size $n = 50$, when R is equal to 0 while it is estimated to be 10, the variances of $\hat{\theta}_\xi$ and $\hat{\theta}_\eta$ are, respectively, 144.4 and 200.8 (Table 4), but if the estimation of R is accurate, *i.e.*, $\hat{R} = 0$, the variances become 92.3 and 166.9 respectively (Table 1). These simulations show that the effort to compute $\hat{\theta}_\xi$ is undoubtedly worthwhile and one must make effort to obtain good estimate of R to make most of $\hat{\theta}_\xi$.

Next we consider the neutral Wright's finite-islands model. When the estimate $\hat{M}$ of M is not accurate, all the four estimators become biased. To measure the performance of a biased estimator we must consider its bias as well as its variance. A proper measure of the accuracy of a biased estimator $\hat{\theta}$ of θ is the mean square error (MSE), defined by

$$\text{MSE} = E(\hat{\theta} - \theta)^2 = \text{Var}(\hat{\theta}) + (\hat{\theta} - \theta)^2.$$

Note that when $\hat{\theta}$ is unbiased, MSE is simply the variance of $\hat{\theta}$. Since the balanced-sampling scheme is much better than the extreme-sampling scheme and its use is strongly recommended, I shall examine the performances of the four estimators for balanced samples only. Table 5 gives a few examples of the effect on estimations of θ when $\hat{M}$ is not accurate.

It is clear from Table 5 that all the four estimators, $\hat{\theta}_\pi$, $\hat{\theta}_K$, $\hat{\theta}_\xi$ and $\hat{\theta}_\eta$, are biased when $\hat{R}$ is inaccurate but $\hat{\theta}_\xi$ remains to be the best estimator among the four. $\hat{\theta}_\xi$ is

a superior estimator not only because its MSE is the smallest in all the cases we examined but also because it is the least biased estimator. Table 5 also shows that the MSE of each estimator becomes relatively large when $\hat{M}$ is much larger than M; while on the other hand, the MSE of an estimator does not change substantially when $\hat{M}$ is considerably smaller than M. This suggests that for the purpose of estimating θ it is better to underestimate M than to overestimate M. We pointed out earlier by examining the coefficients of $\hat{\theta}_\xi$ and $\hat{\theta}_\eta$ that samples from islands may well be treated as from a panmictic population when M is not extremely small and Table 5 reinforces this conclusion. For example, when $\hat{M}$ is 5 which may be regarded to represent a nearly panmictic population, the bias of $\hat{\theta}_\xi$ is not substantial when M is as small as 1, through the bias seems to increase with the number of islands.

We assumed in this paper that the frequencies of mutations of various sizes, *i.e.* $\boldsymbol{\xi}$, can be inferred accurately, which implies that either the infinite-sites model is appropriate and an outgroup sequence is available, or the sequences are sufficiently long so that the phylogeny reconstruction is accurate. Samples of sequences of modest length without an outgroup are common and thus procedure for estimating θ for such samples will be of practical importance. Because $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ are also likely to be important in constructing more powerful tests of the hypothesis of neutral mutations than those by TAJIMA (1989) and FU and LI (1993b), and may even lead to better estimators of R and M, we are currently searching for the best methods to infer the values of $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ under various population models. But without going further, we expect that when the values of $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ can not be inferred accurately, some simple equations for bias correction may be sufficient in practice, as found for a BLUE of θ by FU (1994a). For the neutral Wright's finite-islands model, it may turn out that the bias due to estimating M is larger than that due to estimating $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ therefore finding a good estimate of M may be more critical.

The programs written in C language for estimating θ under models considered in this paper are available from the author whose E-mail address is fu@gsbs18.gs.uth.tmc.edu.

I thank an anonymous referee for suggestions. This work is supported in part by a FIRST AWARD from the National Institutes of Health.

LITERATURE CITED

FELSENSTEIN, J., 1992 Estimating effective population size from samples of sequences: a bootstrap monte carlo integration method. Gen. Res. **60:** 209–220.

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

Fu, Y. X., 1994b Statistical properties of segregating sites. Theor. Popul. Biol. (in press).

FU, Y. X., and W. H. LI, 1993a Maximum likelihood estimation of population parameters. Genetics **134:** 1261–1270.

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

HAMMERSLEY, J. M., and D. C. HANDSCOMB, 1965 *Monte-Carlo* Methods. John Wiley & Sons, New York.

HUDSON, R. R., 1983 Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. **23:** 183–201.

HUDSON, R. R., 1993 The how and why of generating gene genealogies, pp 23–38 in *Mechanisms of Molecular Evolution*, edited by N. TAKAHATA and A. G. CLARK. Sinaur Associates, Sunderland, Mass.

HUDSON, R. R., and N. L. KAPLAN, 1985 Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics **111:** 147–164.

KINGMAN, J. F. C., 1982a The coalescent. Stochast. Proc. Their Appl. **13:** 235–248.

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

SLATKIN, M., and W. P. MADDISON, 1989 A cladistic measure of gene flow inferred from the phylogenies of alleles. Genetics **123:** 603–613.

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

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

TAJIMA, F., 1989 DNA polymorphism in a subdivided population: the expected number of segregating sites in the two-subpopulations model. Genetics **123:** 229–240.

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

Communicating editor: G. B. GOLDING