THEORETICAL POPULATION BIOLOGY 48, 172–197 (1995)

Statistical Properties of Segregating Sites

YUN-XIN FU

Center for Demographic and Population Genetics, University of Texas, P.O. Box 20334, Houston, Texas 77225

Received February 17, 1994

A mutation leading to a segregating site of a sample can be classified by the number of sequences in the sample that inherits the mutant nucleotide; it can also be classified by the frequencies of the two segregating nucleotides at the resulting segregating side. We define the size of a mutation to be the number of sequences in the sample that inherits the mutant nucleotide and the type of mutation (segregating site) to be the smallest value of the frequencies of segregating nucleotides. Each of these two classifications of mutations is analogous to allelic types in a sample of genes. Assuming the neutral Wright–Fisher model, we derived in this paper the mean and variance of the frequency of mutations of each size and type, and the covariance between the numbers of mutations of two different sizes and two different types. Potential applications of these results are discussed.

INTRODUCTION

Segregating sites in a set of homologous DNA sequences are sites at which there are two or more different nucleotides. The number of segregating sites in a random sample of DNA sequences from a population is an important statistic for studying DNA polymorphisms because it leads to a simple estimator of the essential parameter $\theta = 4N\mu$, where N is the effective population size and μ is the mutation rate per sequence (locus) per generation (Watterson, 1975). The number of segregating sites in a sample of DNA sequences is analogous to the number of alleles in a sample of genes. The latter is a sufficient statistic for θ under the infinite alleles model (Ewens, 1972) but the former is only an asymptotic sufficient statistic for θ under the infinite sites model (Fu and Li, 1993a) and the efficiency of estimating θ based on only the number of segregating sites can be astonishingly low for finite samples (Felsenstein, 1992; Fu and Li, 1993a). Just as alleles in a sample can be classified into a number of allelic types, segregating sites can be classified by size and type whose meanings will become clear shortly. However, despite the popularity of the infinite-sites model for DNA sequences and that the different sizes or types of segregating sites should play more important roles in the infinite-sites model than

0040-5809/95 $12.00

different alleles do in the infinite-alleles model because of the insufficiency of the number segregating sites, statistical properties of segregating sites of various sizes and types are poorly understood. The purpose of this paper is to derive the means, variances, and covariances of the numbers of segregating sites of various sizes and types.

We assume that samples are taken from a population that evolves according to the Wright–Fisher model, that all mutations at the locus under study are selectively neutral and that there is no recombination. Each sample of n sequences under these assumptions has a genealogy of $2(n-1)$ branches which connect the sequences to their most recent common ancestor. We define the size of a branch to be the number of sequences in the sample that are descendents of the branch. Therefore, the size of a branch of a genealogy is a positive integer less than the number of sequences in the genealogy. A mutation is said to be *size k* if it occurs in a branch of size k. The mutations in the genealogy of a sample can thus be grouped into $n-1$ different sizes. If the root of a genealogy, i.e., the most recent common ancestor, is removed, then one cannot uniquely determine the size of a branch because it can be either k or $n-k$. Consequently, one cannot distinguish a mutation of size k from a mutation of size $n-k$. Therefore, another useful classification of mutations is to divide them into $n/2$ groups so that group k consists all mutations that are either mutations of size k or mutations of size $n-k$. A mutation is said to be *type i* if it is either size i or size $n-i$.

It should be emphasized that we do not assume the infinite-sites model when defining size and type of a mutation. If the infinite-sites model is further assumed, each segregating site then results from exactly one mutation and therefore the sizes and types of mutations are synonymous to those of segregating sites. Although the definitions of size and type of a mutation take the sample genealogy into consideration, the numbers of mutations of different sizes and particularly the numbers of mutations of different types can often be inferred without knowing the genealogy of a sample. To illustrate, consider the following hypothetical sample of five sequences:

Sequence				
1	$---A---G---C---T--$			
2	$---G---G---C---T--$			
3	$---G---C---C---A--$			
4	$---G---C---C---A--$			
5	$---G---C---T---A--$			
Segrating Site	1	2	3	4

and assume the infinite-sites model. Because a mutation in a branch of size k causes all its k descendent sequences to inherit the mutant nucleotide and all other $n-k$ sequences to inherit the ancestral nucleotide, it is clear that, for example, the first segregating site must result from a mutation in a branch of size 1 or 4. That is, a mutation of type 1. Similarly the second, the third, and the fourth segregating sites are respectively mutations of types 2, 1, and 2. Suppose that an outgroup sequence is available. Then the nucleotide at a segregating site that is identical to that of the outgroup sequence must be the ancestral nucleotide; therefore, the frequencies of the other nucleotide is the size of the mutation. For example suppose that an outgroup sequences has nucleotides G, C, T, and T, respectively, at the four segregating sites. Then the four segregating sites correspond to mutations of size 1, 2, 4, and 3, respectively.

The rest of the paper is organized as follows. The means, variances, and covariances of the numbers of mutations of various sizes are presented first, together with the some results derived from them; the preparation of the derivatives of these results is then given and followed by the derivations. Several numerical examples are presented for illustration and potential applications of the main results are discussed, and finally an Appendix gives some identities that are used in the derivations.

The Results

Let θ be twice the expected number of mutations at a locus in a population per generation and let ξ_i be the number of mutations of size i in a random sample of n sequences. Then

$$\begin{aligned} E(\xi_i) &= \frac{1}{i}\theta \\ \mathrm{Var}(\xi_i) &= \frac{1}{i}\theta + \sigma_{ii}\theta^2 \\ \mathrm{Cov}(\xi_i, \xi_j) &= \sigma_{ij}\theta^2, \end{aligned} \tag{1}$$

where E, Var, and Cov represent respectively the mathematical expectation, variance, and covariance; σ_{ii} is given by

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

and σ_{ij} $(i>j)$ is given by

$$\sigma_{ij}=\begin{cases}\dfrac{\beta_n(i+1)-\beta_n(i)}{2}, & \text{if } i+j<n,\\ \dfrac{a_n-a_i}{n-i}+\dfrac{a_n-a_j}{n-j}-\dfrac{\beta_n(i)+\beta_n(j+1)}{2}-\dfrac{1}{ij}, & \text{if } i+j=n,\\ \dfrac{\beta_n(j)-\beta_n(j+1)}{2}-\dfrac{1}{ij}, & \text{if } i+j>n,\end{cases} \tag{3}$$

where

$$a_n=1+\frac{1}{2}+\cdots+\frac{1}{n-1} \tag{4}$$

and

$$\beta_n(i)=\frac{2n}{(n-i+1)(n-i)}(a_{n+1}-a_i)-\frac{2}{n-i}. \tag{5}$$

Let μ be the mutation rate per locus per generation. Then $\theta=4N\mu$ if the locus is autosomal, where N is the effective population size, and $\theta=2N\mu$ if the locus is a mitochondrial locus where N is the effective size of the female population.

The above results allow one to obtain the mean and variance of a linear function of $\xi_1, \ldots, \xi_{n-1}$ and the covariance between two linear functions of $\xi_1, \ldots, \xi_{n-1}$. A particularly interesting class of linear functions are those leading the numbers of mutations of various types. Let η_i be the number of mutations of type i, i.e., mutations of size i or size $n-i$. It is obvious that

$$\eta_i=\frac{\xi_i+\xi_{n-i}}{1+\delta_{i,n-i}}, \tag{6}$$

where $\delta_{i,j}$ is the Kroneker delta. That is, it is equal to 1 if $i=j$ and 0 otherwise. From (1), it is easy to see that

$$\begin{aligned}E(\eta_i)&=\phi_i\theta\\ \operatorname{Var}(\eta_i)&=\phi_i\theta+\rho_{ii}\theta^2\\ \operatorname{Cov}(\eta_i,\eta_j)&=\rho_{ij}\theta^2,\end{aligned} \tag{7}$$

where

$$\phi_i = \frac{1}{1+\delta_{i,n-i}}\left(\frac{1}{i}+\frac{1}{n-i}\right), \tag{8}$$

$$\rho_{ij} = \frac{\sigma_{ij}+\sigma_{in-j}+\sigma_{n-ij}+\sigma_{n-in-j}}{(1+\delta_{i,n-i})(1+\delta_{j,n-j})}. \tag{9}$$

Let $\eta = \xi_1 + \cdots + \xi_{n-1}$ be total number of mutations in the genealogy of a sample. Then the covariance between ξ_i and η is given by

$$\mathrm{Cov}(\xi_i, \eta) = \frac{1}{i}\theta + \frac{a_n - a_i}{n-i}\theta^2. \tag{10}$$

Consequently the covariance between η_i and η is

$$\mathrm{Cov}(\eta_i, \eta) = \phi_i\theta + \frac{1}{1+\delta_{i,n-i}}\left(\frac{a_n - a_i}{n-i}+\frac{a_n - a_{n-i}}{i}\right)\theta^2. \tag{11}$$

Let Π_n be the average number of mutations separating two sequences in a sample, which is the same as the number of nucleotide differences between two sequences under the infinite-sites model. Then

$$\Pi_n = \frac{2}{n(n-1)}\sum_{i=1}^{n-1} i(n-i)\,\xi_i = \frac{2}{n(n-1)}\sum_{i=1}^{[n/2]} i(n-i)\,\eta_i.$$

This is because the number of mutations on a branch of size k is counted in $k(n-k)$ pairwise comparisons between the k descendent sequences of the branch and $n-k$ sequences that are not descendents of the branch. Thus

$$\mathrm{Cov}(\xi_k, \Pi_n) = \frac{2}{n(n-1)}\left[(n-k)\,\theta + \theta^2\sum_{i=1}^{n-1} i(n-i)\,\sigma_{ki}\right] \tag{12}$$

and

$$\mathrm{Cov}(\eta_k, \Pi_n) = \frac{2}{n(n-1)(1+\delta_{k,n-k})}\left[n\theta + \theta^2\sum_{i=1}^{n-1} i(n-i)(\sigma_{ki}+\sigma_{n-ki})\right]. \tag{13}$$

It should be noted that not all the results above are new. Ewens (1979, Eqs. (9.68)–(9.70)) has essentially derived $E(\xi_i)$ while $E(\eta_i)$ which can be derived from Ewens's equations was explicitly given by Tajima (1989). Also $\mathrm{Var}(\xi_1)$ and $\mathrm{Var}(\eta_1)$ were obtained by Fu and Li (1993b).

THE LINES OF DESCENTS AND PÓLYA URN MODEL

The period from the time at which a sample of n sequences is taken to the most recent common ancestor of the sequences in the sameple can be divided into $n-1$ states such that state i ($i=2, ..., n$) represents the time period during which the sample has exactly i ancestral sequences. For the genealogy of a sample, state i corresponds to the section in which there are exactly i lines (Fig. 1). A line of a state is simply a segment of a branch of the genealogy and is said to be of size i at state k' if there are exactly i lines of state k' that are descendents of the line.

The lines of a state of a genealogy can be viewed as balls of colors. Therefore, as far as the sizes of lines are concerned, the genealogy of a sample can be described as a sampling path of the Pólya urn scheme in which the ball drawn randomly from the urn is returned, together with one ball of the same color (see Feller, 1969, p. 119); the two balls returned represent a branching in the genealogy. The results of the Pólya urn scheme (Feller, 1969; Johnson and Kotz, 1977) can thus be used to obtain the probabilities concerning the sizes of lines, which are necessary for the derivation of the results presented in the previous section. Several Pólya urn schemes have been used to study various aspects of the infinite-alleles model and the finite-sites model (Hoppe, 1984; Donnely, 1986; Ethier and Griffiths, 1987; Griffiths, 1989).

Treating the k lines of state k as balls of k different colors and let $\lambda_1, ..., \lambda_k$ be the size of these lines at state k'; i.e., λ_i is the number of balls of color i after $(k'-k)$th drawing from a urn of k balls of different color, we have from Johnson and Kotz (1977) that all possible values of $\lambda_1, ..., \lambda_k$ are equally probable. The number of possible choices of $\lambda_1, ..., \lambda_k$ is the same as the number of ways to put k' balls into k boxes so that none of the boxes is empty. Therefore, the number is given by the binomial coefficient

$$\binom{k'-1}{k-1} = \frac{(k'-1)!}{(k-1)!\,(k'-k)!}.$$

Each choice of $\lambda_1, ..., \lambda_k$ thus has the probability

$$\frac{1}{\binom{k'-1}{k-1}}.$$

Since the number of ways that a randomly chosen line of state k is of size i at state n is

$$\binom{n-i-1}{k-2}.$$

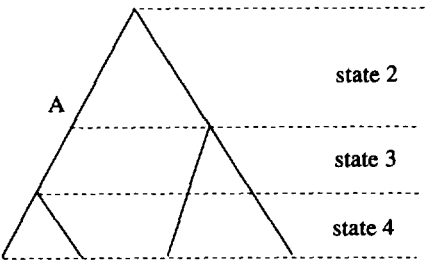

FIG. 1. The states of a genealogy of four sequences. State i represents the section in which there are i lines. Branch A is partitioned into two lines of states 2 and 3, respectively.

The probability $p(k, i)$ that a randomly chosen line of state k is of size i at state n is

$$p(k, i) = \frac{\binom{n-i-1}{k-2}}{\binom{n-1}{k-1}} = \frac{\binom{n-k}{i-1}}{\binom{n-1}{i}} \frac{k-1}{i}. \tag{14}$$

Similarly, the number of possible ways that two randomly chosen lines of state k are of size i and j, respectively, at state n is

$$\binom{n-i-j-1}{k-3},$$

so the probability $p(k, i; k, j)$ that two randomly chosen lines of state k are size i and j, respectively, at state n is

$$p(k, i; k, j) = \frac{\binom{n-i-j-1}{k-3}}{\binom{n-1}{k-1}}. \tag{15}$$

The probability $p(k, i; k', j)$ that a randomly chosen line of state k and a randomly chosen line of state k' are of size i and j, respectively, at state n is more complicated and we will break it down to two cases.

Without loss of generality, we assume that $k < k'$. It is clear from the above analysis that the joint probability that the line of state k is of size

t at state k' and that the line of state k' is a descendent of the line of state k (case 1) is

$$\frac{\binom{k'-t-1}{k-2}}{\binom{k'-1}{k-1}}\frac{t}{k'}=\frac{\binom{k'-k}{t-1}}{\binom{k'-1}{t}}\frac{(k-1)}{k'}, \tag{16}$$

and that the joint probability that the line of state k is of size t at state k' and that the line of state k' is not a descendent of the line of state k (case 2) is

$$\frac{\binom{k'-t-1}{k-2}}{\binom{k'-1}{k-1}}\frac{k'-t}{k'}=\frac{\binom{k'-k}{t-1}}{\binom{k'-1}{t}}\frac{(k-1)(k'-t)}{tk'}. \tag{17}$$

Suppose that case 1 is true. Then the probability that the line of state k' is of size j at state n and the other $t-1$ descendent lines of the line of state k have in total $i-j$ descendent lines at state n is

$$\frac{\binom{i-j-1}{(t-2)_+}\binom{n-i-1}{k'-t-1}}{\binom{n-1}{k'-1}}$$

if $i \geqslant j$ and zero otherwise, where $a_+ = \max(0, a)$.

Suppose that case 2 is true. Then the probability that the line of state k' is of size j at state n and the t descendent lines of the line of state k have in total i descendent lines at state n is

$$\frac{\binom{i-1}{t-1}\binom{n-i-j-1}{(k'-t-2)_+}}{\binom{n-1}{k'-1}}.$$

The probability $p_a(k, i; k', j)$ that the line of state k' is a descendent of the line of state k and that the lines of state k and k' are size i and j, respectively, is given by

$$p_a(k, i; k', j)=\sum_t \frac{\binom{k'-k}{t-1}}{\binom{k'-1}{t}}\frac{(k-1)}{k'}\frac{\binom{i-j-1}{(t-2)_+}\binom{n-i-1}{k'-t-1}}{\binom{n-1}{k'-1}} \tag{18}$$

where $i \geqslant j$ and $p_a(k, i; k', j) = 0$ otherwise. The summation in (18) is taken over all meaningful values of t, i.e., $1 \leqslant t \leqslant \min\{k' - k + 1, i - j + 1\}$.

The probability $p_b(k, i; k', j)$ that the line of state k and the line of k' are of size i and j, respectively, and the latter is not a descendent of the former is obviously equal to zero when $i + j > n$. When $i + j < n$ we have

$$p_b(k, i; k', j) = \sum_t \frac{\binom{k'-k}{t-1}}{\binom{k'-1}{t}} \frac{(k-1)(k'-t)}{tk'} \frac{\binom{i-1}{t-1}\binom{n-i-j-1}{(k'-t-2)_+}}{\binom{n-1}{k'-1}}, \tag{19}$$

where the summation is taken over $1 \leqslant t \leqslant \min\{k' - 2, k' - k + 1, i\}$. In order for $i + j = n$, the line of state k must be of size $k' - 1$ at state k'; it is obvious that k must be equal to 2. Therefore when $i + j = n$

$$p_b(2, i; k', j) = \frac{\binom{k'-2-1}{k'-3}}{\binom{k'-1}{k'-2}} \frac{k' - k' + 1}{k'} \frac{\binom{n-j-1}{k'-2}}{\binom{n-1}{k-1}} = \frac{1}{jk'} \frac{\binom{n-k'}{j-1}}{\binom{n-1}{j}} \tag{20}$$

and $p_b(k, i; k', n - i) = 0$ when $2 < k$.

Assembling these probabilities, we have that

$$p(k, i; k', j) = \delta_{(i \leqslant j)} p_a(k, i; k', j) + \delta_{(i+j \leqslant n)} p_b(k, i; k', j), \tag{21}$$

where $\delta_{(r)}$ is a delta function that takes the value 1 if the relationship r is true and 0 otherwise.

The Pólya urn scheme has another interesting consequence. Suppose the n sequences in a sample are labeled. Then there are

$$\frac{n!}{\lambda_1! \cdots \lambda_k!}$$

ways to divide the sample into k groups so that group i has λ_i sequences. On the other hand, there are $k!$ ways to color the k lines of state k. Therefore, each path from state k to the n labeled sequences so that after renumbering the lines of state k the size of line i is λ_i has the probability

$$k! \Big/ \left[\binom{n-1}{k-1} \frac{n!}{\lambda_1! \cdots \lambda_k!} \right] = \frac{(n-k)!\, k!(k-1)!}{n!(n-1)!} \lambda_1! \cdots \lambda_k!.$$

which is Kingsman's (1982a) distribution.

Derivation of the Main Results

Let the k lines of state k be numbered from 1 to k and ξ_{kl} be the number of mutations occurred in the lth line of state k. We define $\varepsilon_{kl}(i)$ be an index variable so that it takes the value 1 if the lth line of state k is of size i and takes the value 0 otherwise. Then the number ξ_i of mutations of size i can be written as

$$\xi_i = \sum_{k=2}^{n} \sum_{l=1}^{k} \varepsilon_{kl}(i)\, \xi_{kl}.$$

To obtain the moments of ξ_i, we need to know the moments of ξ_{kl}'s which can be obtained using the coalescent theory (Kingman, 1982b, see Tavaré, 1984, or Hudson, 1991, for reviews). It is simple to show that

$$E(\xi_{kl}) = \frac{1}{k(k-1)}\,\theta$$

$$E(\xi_{kl}^2) = \frac{1}{k(k-1)}\,\theta + \frac{2}{k^2(k-1)^2}\,\theta^2$$

$$E(\xi_{kl}\xi_{kl'}) = \frac{2}{k^2(k-1)^2}\,\theta^2$$

$$E(\xi_{kl}\xi_{k'l'}) = \frac{1}{k(k-1)\,k'(k'-1)}\,\theta^2.$$

Note that the probability that $\varepsilon_{kl}(i) = 1$ is simply $p(k, i)$ because the number l assigned to a line is arbitrary. Similarly, we have

$$p(k, i; k, j) = \Pr[\varepsilon_{kl}(i)\,\varepsilon_{kl'}(j) = 1], \qquad l \neq l',$$

$$p(k, i; k', j) = \Pr[\varepsilon_{kl}(i)\,\varepsilon_{k'l'}(j) = 1], \qquad k < k'$$

We are now ready to derive the main results (1). The mean of ξ_i is simply

$$E(\xi_i) = \sum_{k=2}^{n} \sum_{l=1}^{k} \Pr(\varepsilon_{kl}(i) = 1)\, E(\xi_{kl}) = \sum_{k=2}^{n} kp(k, i)\, E(\xi_{k1})$$

$$= \frac{\theta}{i} \binom{n-1}{i}^{-1} \sum_{k=2}^{n} \binom{n-k}{i-1}$$

$$= \frac{1}{i}\,\theta. \tag{22}$$

To obtain the variances and covariances of ξ_i's, we need to consider the product of ξ_i and ξ_j, which is

$$\xi_i\xi_j=\Big(\sum_{k=2}^{n}\sum_{l=1}^{k}\varepsilon_{kl}(i)\,\xi_{kl}\Big)\Big(\sum_{k=2}^{n}\sum_{l=1}^{k}\varepsilon_{kl}(j)\,\xi_{kl}\Big)$$
$$=\sum_{k,k'=2}^{n}\sum_{l,l'}\varepsilon_{kl}(i)\,\varepsilon_{k'l'}(j)\,\xi_{kl}\xi_{k'l'}.$$

It follows that

$$E(\xi_i\xi_j)=\sum_{k,k'=2}^{n}\sum_{l,l'}\Pr(\varepsilon_{kl}(i)\,\varepsilon_{k'l'}(j)=1)\,E(\xi_{kl}\xi_{k'l'})$$
$$=\delta_{(i=j)}\sum_{k=2}^{n}kp(k,i)\,E(\xi_{k1}^2)+\sum_{k=2}^{n}k(k-1)\,p(k,i;k,j)\,E(\xi_{k1}\xi_{k2})$$
$$+\sum_{k<k'}^{n}kk'[\,p(k,i;k',j)+p(k,j;k',i)]\,E(\xi_{k1}\xi_{k'1}). \tag{23}$$

The subsequent derivations will make use of some identities involving binomial coefficients which are given in the Appendix. We now consider the terms of (23) in turn. First, we have that

$$\sum_{k=2}^{n}kp(k,i)\,E(\xi_{k1}^2)$$
$$=\sum_{k=2}^{n}\frac{\binom{n-k}{i-1}(k-1)}{\binom{n-1}{i}i}\left[\frac{1}{k-1}\theta+\frac{2}{k(k-1)^2}\theta^2\right]$$
$$=\frac{1}{i}\theta+\frac{2\theta^2}{\binom{n-1}{i}i}\left[\sum_{k=2}^{n}\frac{\binom{n-1-(k-1)}{i-1}}{k-1}-\sum_{k=2}^{n}\frac{\binom{n-k}{i-1}}{k}\right].$$

Further simplification of above equation, making use of identity (34), leads to

$$\delta_{(i=j)}\sum_{k=2}^{n}kp(k,i)\,E(\xi_{k1}^2)$$
$$=\frac{\delta_{(i=j)}}{i}\theta+\delta_{(i=j)}\left[\frac{2}{n-1}+\frac{n-i-1}{n-i}\beta_{n-1}(i)-\beta_n(i)\right]\theta^2. \tag{24}$$

Consider the second term in (23). Because of (15) and (36), we have

$$\sum_{k=2}^{n} k(k-1)\, p(k,i;k,j)\, E(\xi_{k1}\xi_{k2})$$

$$= \frac{\delta_{(i+j=n)}}{n-1}\theta^2 + \delta_{(i+j<n)}\, 2 \sum_{k=3}^{n} \frac{\binom{n-(i+j-1)-2}{k-3}}{\binom{n-1}{k-1}} \frac{\theta^2}{k(k-1)}$$

$$= \frac{\delta_{(n=i+j)}}{n-1}\theta^2 + \delta_{(i+j<n)}[\beta(i+j-1) - \beta_n(i+j)]\,\theta^2. \tag{25}$$

Because of (21), the third term in (23) can further be broken into three terms:

$$\sum_{k<k'}^{n} kk'(\, p(k,i;k',j) + p(k,j;k',i))\, E(\eta_{k1}\eta_{k'1}) \tag{26}$$

$$= \sum_{k'=3}^{n} \sum_{k=2}^{k'-1} [\delta_{(j\geq i)} p_a(k,i;k',j) + \delta_{(i\geq j)} p_a(k,j;k',i)] \frac{\theta^2}{(k-1)(k'-1)}$$

$$+ \delta_{(i+j<n)} \sum_{k'=3}^{n} \sum_{k=2}^{k'-1} [\, p_b(k,i;k',j) + p_b(k,j;k',i)] \frac{\theta^2}{(k-1)(k'-1)}$$

$$+ \delta_{(n=i+j)} \sum_{k'=3}^{n} [\, p_b(2,i;k',j) + p_b(2,j;k',i)] \frac{\theta^2}{(k-1)(k'-1)}. \tag{27}$$

Because

$$\sum_{k'=3}^{n} \sum_{k'-k=1}^{k'-2} \frac{p_x(k,i;k',j)}{(k-1)(k'-1)}$$

$$= \sum_{k'=3}^{n} \sum_{k'-k=2}^{k'-2} \sum_{t} \frac{\binom{k'-k}{t-1}}{\binom{k'-1}{t}} \frac{\binom{i-j-1}{(t-2)_+}\binom{n-i-1}{k'-t-1}}{k'(k'-1)\binom{n-1}{k'-1}}$$

$$= \sum_{k'=3}^{n} \sum_{t\geq 2} \sum_{k'-k=t-1}^{k'-2} \frac{\binom{k'-k}{t-1}}{\binom{k'-1}{t}} \frac{\binom{i-j-1}{t-2}\binom{n-i-1}{k'-t-1}}{k'(k'-1)\binom{n-1}{k'-1}}$$

$$+\delta_{(i=j)}\sum_{k'=3}^{n}\sum_{k=1}^{k'-2}\frac{1}{k'(k'-1)^2}\frac{\binom{n-i-1}{k'-2}}{\binom{n-1}{k'-1}}$$

$$=\sum_{k'=3}^{n}\sum_{t\geq 2}\frac{\binom{i-j-1}{t-2}\binom{n-i-1}{k'-3-(t-2)}}{k'(k'-1)\binom{n-1}{k'-1}}$$

$$+\delta_{(i=j)}\frac{1}{\binom{n-1}{i}i}\sum_{k'=3}^{n}\frac{k'-2}{k'(k-1)}\binom{n-k}{i-1}$$

$$=\sum_{k'=3}^{n}\frac{\binom{n-j-2}{k'-3}}{k'(k'-1)\binom{n-1}{k'-1}}+\frac{\delta_{(i=j)}}{\binom{n-1}{i}i}$$

$$\times\left[2\sum_{k'=3}^{n}\frac{\binom{n-k}{i-1}}{k}-\sum_{k'=3}^{n}\frac{\binom{n-1-(k-1)}{i-1}}{k-1}\right]$$

$$=\frac{\beta_n(j)-\beta_n(j+1)}{2}+\delta_{(i=j)}\left[\beta_n(i)-\frac{1}{n-1}-\frac{n-i-1}{2(n-1)}\beta_{n-1}(i)\right]$$

and the fact that we can obtain the summation for $p_a(k,j;k',i)$ by exchanging i and j in the equation, it follows that

$$\sum_{k'=3}^{n}\sum_{k=2}^{k'-1}[\delta_{(j\leq 1)}p_a(k,i;k',j)+\delta_{(i\leq j)}p_a(k,j;k',i)]\frac{\theta^2}{(k-1)(k'-1)}$$

$$=\delta_{(i>j)}\frac{\beta_n(j)-\beta_n(j+1)}{2}\theta^2+\delta_{(i<j)}\frac{\beta_n(i)-\beta_n(i+1)}{2}\theta^2$$

$$+\delta_{(i=j)}\left[2\beta_n(i)-\frac{2}{n-1}-\frac{n-i-1}{n-1}\beta_{n-1}(i)\right]\theta^2. \tag{28}$$

Because

$$\sum_{k'=3}^{n}\sum_{k=1}^{k'-2}\frac{p_b(k,i;k',j)}{(k-1)(k'-1)}$$

$$=\sum_{k'=3}^{n}\sum_{k=1}^{k'-2}\sum_{t\geq 2}\frac{\binom{k'-k}{t-1}}{\binom{k'-1}{t}}\frac{(k'-t)\binom{i-1}{t-1}\binom{n-i-j-1}{k'-t-2}}{tk'(k'-1)\binom{n-1}{k'-1}}$$

$$+\sum_{k'=3}^{n}\sum_{k'-k=1}^{k'-2}\frac{1}{k'(k'-1)}\frac{\binom{n-i-j-1}{k'-3}}{\binom{n-1}{k'-1}}$$

$$=\sum_{k'=3}^{n}\sum_{t\geq 2}\sum_{k'-k=t-1}^{k'-2}\frac{\binom{k'-k}{t-1}}{\binom{k'-1}{t}}\frac{k'-t}{k'}\frac{\binom{i-1}{t-1}\binom{n-i-j-1}{k'-t-2}}{(k'-1)\binom{n-1}{k'-1}}$$

$$+\sum_{k'=3}^{n}\frac{k'-2}{k'(k'-1)}\frac{\binom{n-i-j-1}{k'-3}}{\binom{n-1}{k'-1}}$$

$$=\sum_{k'=3}^{n}\sum_{t\geq 2}\frac{\binom{i}{t}\binom{n-i-j-1}{k'-t-2}}{(k'-1)\binom{n-1}{k'-1}}-\sum_{k'=3}^{n}\sum_{t\geq 2}\frac{\binom{i-1}{t-1}\binom{n-i-j--1}{k'-t-2}}{k'(k'-1\binom{n-1}{k'-1}}$$

$$+\sum_{k'=3}^{n}\frac{k'-2}{k'(k'-1)}\frac{\binom{n-i-j-1}{k'-3}}{\binom{n-1}{k'-1}}$$

$$=\sum_{k'=3}^{n}\sum_{t\geq 1}\frac{\binom{i}{t}\binom{n-i-j-1}{k'-t-2}}{k'(k'-1)\binom{n-1}{k'-1}}$$

$$-\sum_{k'=3}^{n}\left[\frac{\binom{n-k'-2}{k'-3}}{k'(k'-1)\binom{n-1}{k'-1}}+\frac{\binom{n-i-j-1}{k'-3}}{k'(k'-1)\binom{n-1}{k'-1}}\right]$$

$$=\frac{1}{i}\sum_{k'=3}^{n}\left[\frac{\binom{n-k'-1}{k'-2}}{(k'-1)\binom{n-1}{k'-1}}-\frac{\binom{n-i-j-1}{k'-2}}{(k'-1)\binom{n-1}{k'-1}}\right]$$

$$-\frac{1}{2}[\beta_n(j)+\beta_n(i+j-1)-\beta_n(j+1)-\beta_n(i+j)]$$

$$=\left[\frac{1}{ij}-\frac{1}{i(i+j)}\right]-\frac{1}{2}[\beta_n(j)+\beta_n(i+j-1)-\beta_n(j+1)-\beta_n(i+j)]$$

and the fact that again we can obtain the summation for $p_b(k,j;k',i)$ by exchanging i and j in the above equation, we have

$$\delta_{(i+j<n)}\sum_{k'=3}^{n}\sum_{k=2}^{k'-1}[p_b(k,i;k',j)+p_b(k,j;k',i)]\frac{\theta^2}{(k-1)(k'-1)}$$

$$=\frac{\delta_{(i+j<n)}}{ij}\theta^2-\delta_{(i+j<n)}[\beta_n(i+j-1)-\beta_n(i+j)]\theta^2$$

$$-\frac{\delta_{(i+j<n)}}{2}[\beta_n(i)-\beta_n(i+1)+\beta_n(j)-\beta_n(j+1)]\theta^2. \tag{29}$$

Finally, because

$$\sum_{k'=3}\frac{p_b(2,i;k',j)}{k'-1}=\sum_{k'=3}\frac{\binom{n-k'}{j-1}}{k_j'\binom{n-1}{j}}\frac{1}{k'-1}$$

$$=\frac{1}{\binom{n-1}{j}j}\left[\sum_{k'=3}\frac{\binom{n-k'}{j-1}}{k'-1}-\sum_{k'=3}\frac{\binom{n-k'}{j-1}}{k'}\right]$$

$$=\frac{\binom{n-2}{j}j}{2\binom{n-1}{j}j}\beta_{n-1}(j)-\frac{1}{2}\beta_n(j)+\frac{\binom{n-2}{j-1}}{2\binom{n-1}{j}j}$$

$$=\frac{1}{n-j}(a_n-a_j)-\frac{1}{2}\beta_n(j)-\frac{1}{2(n-1)}$$

it follows that

$$\delta_{(n=i+j)}\sum_{k'=3}^{n}[p_b(2,i;k',j)+p_b(2,j;k',i)]\frac{\theta^2}{(k-1)(k'-1)}$$

$$=\delta_{(n=i+j)}\left[\frac{a_n-a_i}{n-i}+\frac{a_n-a_j}{n-j}-\frac{\beta_n(i)+\beta_n(j)}{2}-\frac{1}{n-1}\right]\theta^2. \tag{30}$$

We can now assemble these results. Combining (24) to (30), we arrive at

$$E(\xi_i\xi_j) = \frac{\delta_{(i=j)}}{i}\theta + \delta_{(i=j)}\beta_n(i)\,\theta^2$$
$$+\delta_{(i>j)}\frac{\beta_n(j)-\beta_n(j+1)}{2}\theta^2 + \delta_{(i<j)}\frac{\beta_n(i)-\beta_n(i+1)}{2}\theta^2$$
$$+\frac{\delta_{(i+j<n)}}{ij}\theta^2 - \frac{\delta_{(i+j<n)}}{2}[\beta_n(i)-\beta_n(i+1)+\beta_n(j)-\beta_n(j+1)]\,\theta^2$$
$$+\delta_{(n=i+j)}\left[\frac{a_n-a_i}{n-i}+\frac{a_n-a_j}{n-j}-\frac{\beta_n(i)+\beta_n(j)}{2}\right]\theta^2. \tag{31}$$

A separation of above equation into different cases, noting that $E(\xi_i)\,E(\xi_j) = (1/ij)\,\theta^2$, completes the proof of Eq. (1).

The covariant between ξ_i and η can be obtained easily from the σ_{ij}'s and by making use of the identity (37). We shall demonstrate the case $i < n/2$. Note that

$$\operatorname{cov}(\xi_i,\eta) = \sum_{k=1}^{n-1}\sigma_{ki} = \sum_{k=1}^{i-1}\sigma_{ki} + \sigma_{ii} + \sum_{k=i+1}^{n-i-1}\sigma_{ki} + \sigma_{n-i\,i} + \sum_{k=n-i+1}^{n-1}\sigma_{ki}.$$

From (3), it is easy to see that

$$\sum_{k=1}^{n-1}\sigma_{ki} + \sum_{k=n-i+1}^{n-1}\sigma_{ki} = -\frac{a_n-a_{n-i+1}}{i}$$

and because of identity (37), we have

$$\sum_{k=i+1}^{n-i-1}\sigma_{ki} = \frac{\beta_n(n-i)+\beta_n(i+1)}{2}.$$

Therefore,

$$\operatorname{cov}(\xi_i,\eta) = -\frac{a_n-a_{n-i+1}}{i} + \sigma_{ii} + \frac{\beta_n(n-i)+\beta_n(i+1)}{2} + \sigma_{n-i\,i}$$
$$= \frac{a_n-a_i}{n-i}.$$

Numerical Examples

The quantity $\beta_n(i)$ plays a role in almost all the variances and covariances we have considered. It follows from Eq. (36) that $\beta_n(i) > \beta_n(i+1)$ which implies that σ_{ij} is negative for $i+j<n$, $i \neq j$. However, it is not easy

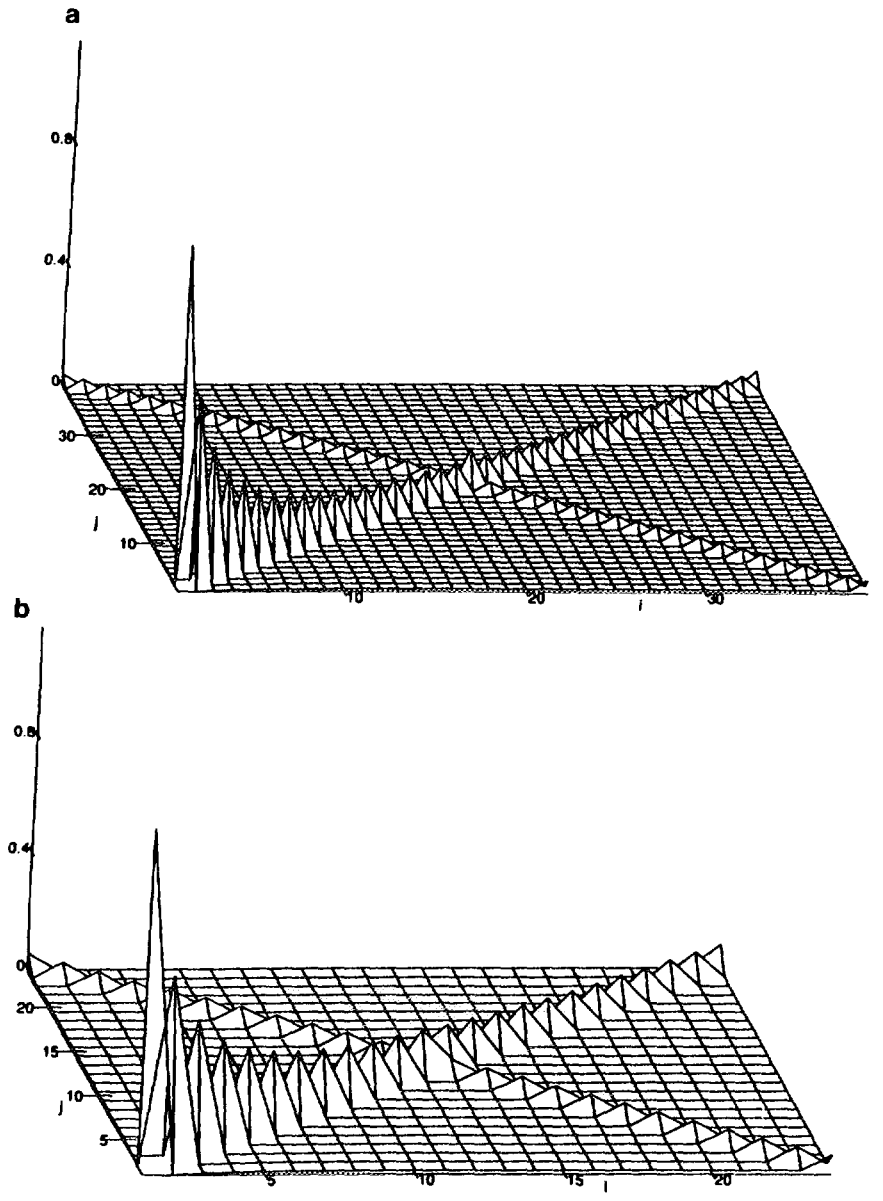

FIG. 2. Variances and covariances of ξ_i's with $\theta = 1$, $n = 40$ (a) and $n = 25$ (b).

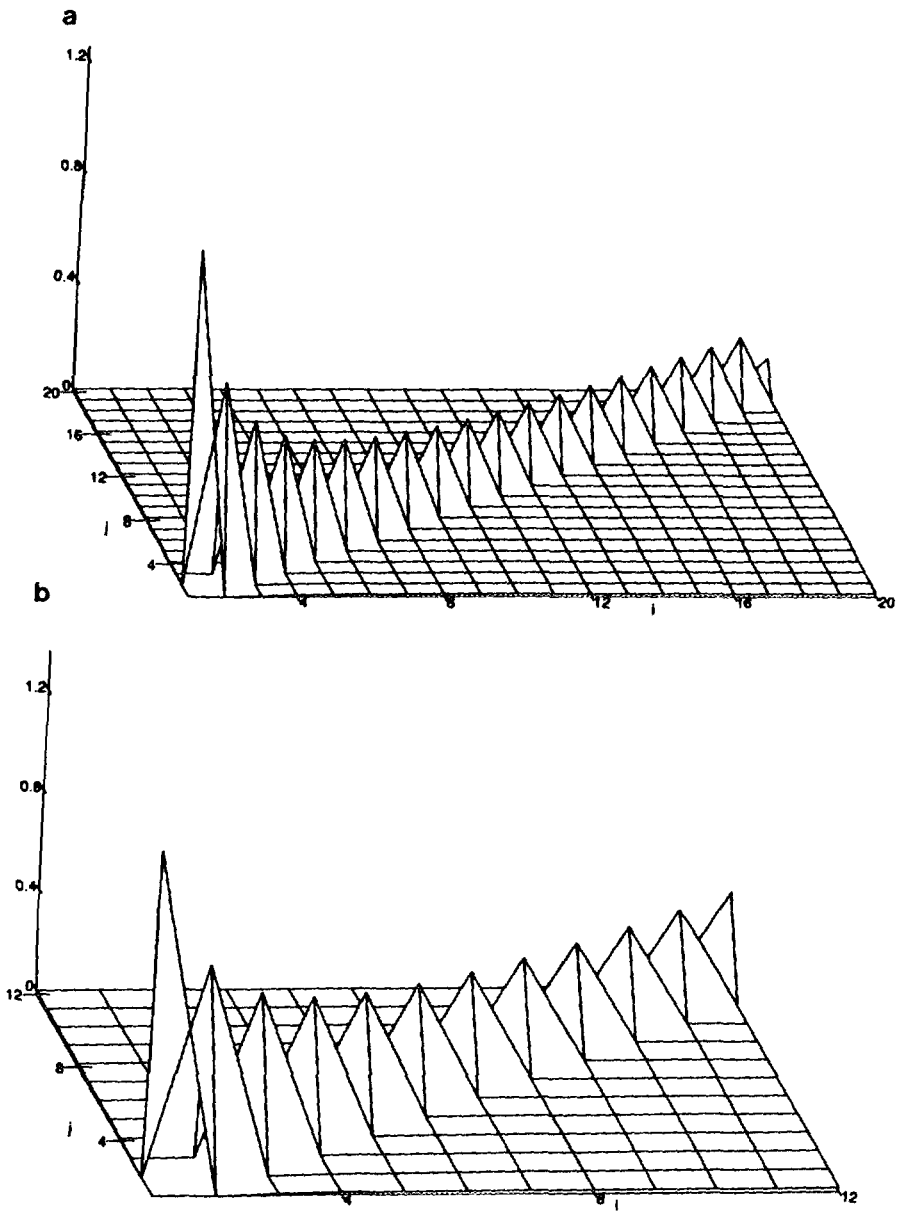

FIG. 3. Variances and covariances of η_i's with $\theta = 1$, $n = 40$ (a) and $n = 25$ (b).

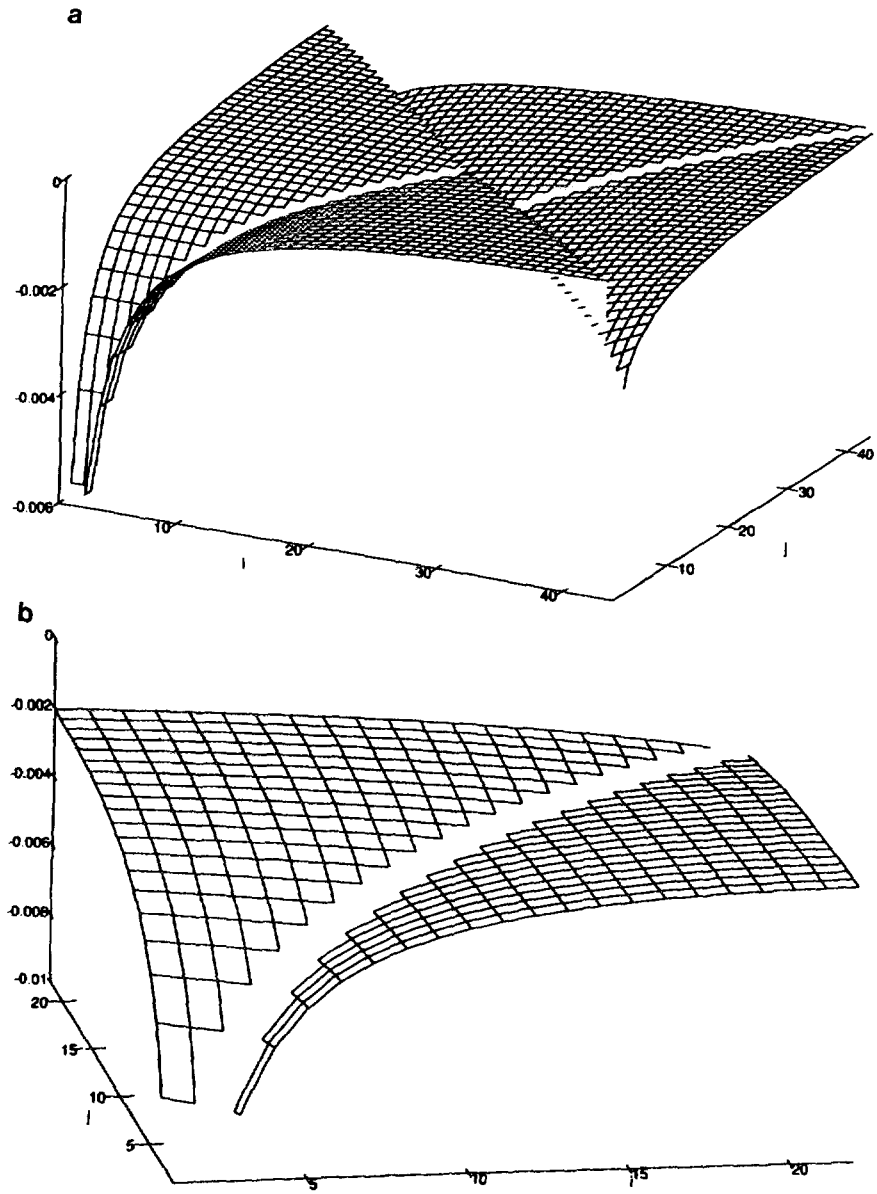

FIG. 4. (a) Covariances of ξ's without showing the covariance between ξ_k and ξ_{n-k} $(k = 1, ..., n-1)$ with $\theta = 1$ and $n = 45$. (b) Covariances of η's with $\theta = 1$ and $n = 25$.

to see the relative values and patterns of the variences and covariances from (2), (3), and particularly from (9). We thus present a few examples. Figure 2 shows the variances and covariances of ξ_i's with $\theta = 1$; panels *a* and *b* correspond to samples of size 40 and 25, respectively.

Figure 2 shows that the variance of ξ_i decreases with i, except that $\xi_{n/2}$ is larger than $\xi_{n/2-1}$ when n is even. It is clear that all the covariances are relatively small compared to the variances except for the covariance between ξ_i and ξ_{n-i}. This is proabably due to the fact that there is a considerable chance that the two branches leading to the most recent common ancestor are of size i and $n-i$, respectively; because these two branches share the common coalescent time t_2 which is the longest coalescent time, it leads to a relatively large positive covarance between ξ_i and ξ_{n-i}.

Figure 3 shows the variances and covariances of η_i's with $\theta = 1$ and panels a and b correspond to samples of sizes 40 and 25, respectively. It can be seen that the variance of η_i decreases with i but the difference between $\eta_{n/2-1}$ and $\eta_{n/2}$ when n is even is considerably larger than the difference between $\eta_{n/2-2}$ and $\eta_{n/2-1}$. This is apparently due to the fact that $\eta_{n/2} = \xi_{n/2}$ while $\eta_i = \xi_i + \xi_{n-i}$ when $i \neq n/2$.

Because of the dominance of variances, the relative values of the covariances can not be seen in Figs. 2 and 3. We plot in Fig. 4 only the covariances. Panel a of Fig. 4 shows all the covariances in the case $\theta = 1$ and $n = 45$, except for $\mathrm{cov}(\xi_k, \xi_{n-k})$ $(k = 1, ..., n-1)$. It can be seen that the covariance between ξ_i and ξ_j is negative and, among all the covariances with $i + j \neq n$, $\mathrm{cov}(\xi_1, \xi_2)$ is the smallest and $\mathrm{cov}(\xi_1, \xi_{n-2})$ is the largest. Similarly, panel b of Fig. 4 shows all the covariances in the case $\theta = 1$ and $n = 45$. It can be seen from this panel that all the covariances are negative and among them $\mathrm{cov}(\eta_1, \eta_2)$ is the smallest and $\mathrm{cov}(\eta_1, \eta_{n/2})$ is the largest.

Concluding Remarks

There are at least two important applications of our results. The first is to estimate the parameter θ. The general least square theory used in Fu (1994a) can be applied to the numbers of mutations of different sizes and types to obtain estimates of θ. This approach is computationally simple and can be extended to deal with more complicated population models, such as recombination and population subdivision (Fu, 1994b). The second application of our result is to construct tests of the neutral Wright–Fisher model, which are often referred to as tests of neutrality of mutations. Several tests of neutrality of mutations have been proposed by Tajima (1989) and Fu and Li (1993b) by utilizing available statistical properties of linear functions of the numbers of mutations of various sizes and types. The

availability of the variances and covariances of these quentities will certainly help to develop more powerful tests in the future.

The variance of ξ_i can written as

$$\begin{aligned}\mathrm{Var}(\xi_i) &= E(\xi_i^2) - E_{\text{genealogy}}[E^2(\xi_i \mid \text{genealogy})] \\ &\quad + E_{\text{genealogy}}[E^2(\xi_i \mid \text{genealogy})] - E^2(\xi_i) \\ &= V_w(\xi_i) + V_b(\xi_i), \end{aligned} \tag{32}$$

where $E_{\text{genealogy}}$ means the expectation over genealogies and $E(\xi_i \mid \text{genealogy})$ is the expectation conditioning on a genealogy. The first term of (32) is the mean variance of ξ_i over all possible genealogies and the second term is the variance of the mean of ξ_i among the genealogies. In other words, the first term represents the variation within a genealogy and the second term is the variation between the genealogies. Similar partition can be found for the covariance between ξ_i and ξ_j. It is not clear what uses such a partition may have, but it is worthy of pointing out that the analytical results for each term can be derived by using recurrent relationships similar to those in Fu and Li (1993b). In fact, Eqs. (2) and (3) were obtained first by this approach, but the derivations are considerably more lengthy than the present one.

Ultimately we would like to have the joint distribution of ξ_i's and the joint distribution of η_i's. In the light of the complexity of the distribution of the numer of segregating sites (Taveré, 1984), these joint distributions are likely to be very complicated and even the mariginal distribution of ξ_i does not seem to be simple. Therefore, it may be helpful to consider first marginal distributions for the large sample size. It is easy to show using Chebyshev's inequality that $\lim_{n\to\infty} \beta_n(i) = 0$ for any constant i, which indicates that the mean and variance of ξ_i are about the same when the sample size is large. Therefore, the large sample distribution of ξ_i can be approximated by a Poisson distribution with mean θ/i which is independent of the sample size n. It is also easy to show that the large sample distribution of η_i can be approximated by a Possoin distribution with mean $\theta[1/i + 1/(n-i)]$. Apparently, the distributions of η_i and ξ_i converge to the same distribution when the sample size n approaches infinity. Interestingly the results by Watterson (1974a, 1974b) and Chakraborty and Griffiths (1982) on the moments associated with alleles in the infinite-alleles model suggest that the number of the allele that has exactly i copies in a large sample under the infinite-alleles model also follows approximately a Poisson distribution with a mean that is dependent on both i and the sample size n. However, the above analysis does not mean that the joint distribution of ξ_i $(i = 1, \ldots, n-1)$ approaches the product of $n-1$ independent Poisson distributions when the sample size is large. This is because not all the

marginal distributions approach the Poisson limit. Take ξ_{n-1}, for example; the ratio of its variance to its mean is $(n-1)\beta_n(n-1)-1/(n-1)$ which can easily be shown to approach infinity instead of one.

APPENDIX: Some Identities Involving Binomial Coefficients

1. For $i \leqslant n$,

$$\sum_{k=1}^{n} \frac{1}{k}\binom{n-k}{i-1} = \binom{n}{i-1}(a_{n+1}-a_i). \tag{33}$$

Proof. We shall prove this identity by induction. The identity obviously holds for $i=1$. Suppose it is true for $i=m<n$. Then because

$$\frac{1}{k}\binom{n-k}{m} = \frac{n-k-m+1}{km}\binom{n-k}{m-1}$$

we have

$$\begin{aligned}\sum_{k=1}^{n} \frac{1}{k}\binom{n-k}{m} &= \frac{n-m+1}{m}\sum_{k=1}^{n}\frac{1}{k}\binom{n-k}{m-1} - \frac{1}{m}\sum_{k=1}^{n}\binom{n-k}{m-1} \\ &= \frac{n-m+1}{m}\binom{n}{m-1}(a_{n+1}-a_m) - \frac{1}{m}\binom{n}{m} \\ &= \binom{n}{m}(a_{n+1}-a_m).\end{aligned}$$

Therefore, the identity holds for all $m \leqslant n$ by induction.

2. For $i<n$,

$$\beta_n(i) = \frac{2}{\binom{n-1}{i} i}\sum_{k=2}^{n}\frac{\binom{n-k}{i-1}}{k}. \tag{34}$$

Proof.

$$\begin{aligned}\frac{2}{\binom{n-1}{i} i}\sum_{k=2}^{n}\frac{\binom{n-k}{i-1}}{k} &= \frac{2}{\binom{n-1}{i} i}\left[\binom{n}{i-1}(a_{n+1}-a_i) - \binom{n-1}{i-1}\right] \\ &= \frac{2n}{(n-i)(n-i+1)}(a_{n+1}-a_i) - \frac{2}{n-i}.\end{aligned}$$

3. For $i<n-1$,

$$\frac{\beta_n(i)-\beta_n(i+1)}{2}=\frac{1}{(n-i-1)}\left[\frac{1}{i}-\beta_n(i)\right]. \tag{35}$$

Proof.

$$\begin{aligned}
&\frac{1}{(n-i-1)}\left[\frac{1}{i}-\beta_n(i)\right]-\frac{1}{2}\beta_n(i)\\
&\quad=\frac{1}{i(n-i-1)}-\frac{n-i+1}{2(n-i-1)}\beta_n(i)\\
&\quad=\frac{1}{(n-i-1)}-\frac{n}{(n-i)(n-i-1)}(a_{n+1}-a_i)+\frac{n-i+1}{(n-i-1)(n-i)}\\
&\quad=\frac{1}{i(n-i-1)}-\frac{n}{(n-i)(n-i-1)}(a_{n+1}-a_{i+1})\\
&\qquad-\frac{n}{(n-i-1)(n-i)\,i}+\frac{n-i+1}{(n-i-1)(n-i)}\\
&\quad=-\frac{1}{2}\beta_n(i+1).
\end{aligned}$$

4. For $i<n-1$,

$$\sum_{k=3}^{n}\frac{\binom{n-i-2}{k-3}}{k(k-1)\binom{n-1}{k-1}}=\frac{\beta_n(i)-\beta_n(i+1)}{2}. \tag{36}$$

Proof.

$$\begin{aligned}
\sum_{k=3}^{n}\frac{\binom{n-i-2}{k-3}}{k(k-1)\binom{n-1}{k-1}}&=\sum_{k=3}^{n}\frac{(n-k-2)!\,(k-1)!\,(i-1)!\,(n-i)!}{k(k-1)(k-3)!\,(n-i-k+1)!\,(n-1)!\,(k-1)!}\\
&=\frac{1}{\binom{n-1}{i-1}(n-i)(n-i-1)}\sum_{k=3}^{n}\binom{n-k}{k-1}\frac{k-2}{k}
\end{aligned}$$

$$= \frac{1}{\binom{n-1}{i-1}(n-i)(n-i-1)} \times \left[\binom{n-2}{i} - \binom{n-1}{i} i\beta_n(i) + \binom{n-2}{i-1} \right]$$

$$= \frac{1}{(n-i-1)} \left[\frac{1}{i} - \beta_n(i) \right].$$

5. For $u \leqslant v < n-1$,

$$\sum_{k=u}^{v} [\beta_n(k+1) - \beta_n(k)] = \beta_n(v+1) + \beta_n(u). \tag{37}$$

Proof. Since

$$\sum_{k=u}^{v} \frac{a_k}{(n-k+1)(n-k)}$$

$$= \sum_{i=1}^{v} \frac{1}{i} \sum_{k=u}^{v} \frac{1}{(n-k+1)(n-k)} - \sum_{i=u}^{v} \frac{1}{i} \sum_{k=u}^{i} \frac{1}{(n-k+1)(n-k)}$$

$$= \left(\frac{1}{n-v} - \frac{1}{n-u+1} \right) a_{v+1} - \frac{1}{n} \sum_{k=u}^{v} \left(\frac{1}{k} + \frac{1}{n-k} \right) + \frac{1}{n-u+1} \sum_{k=u}^{v} \frac{1}{k}$$

$$= \left(\frac{1}{n-v} - \frac{1}{n-u+1} \right) a_{v+1} - \frac{1}{n} (a_{v+1} - a_u + a_{n-u+1} - a_{n-v}) + \frac{1}{n-a+1} (a_{v+1} - a_u)$$

$$= \frac{1}{n-b} a_{v+1} - \frac{1}{n} (a_{v+1} - a_u + a_{n-u+1} - a_{n-v}) - \frac{1}{n-u+1} a_u,$$

we have

$$\sum_{k=u}^{v} \beta_n(k) = 2n \left(\frac{1}{n-v} - \frac{1}{n-u+1} \right) a_{n+1} - 2n \left[\frac{1}{n-v} a_{v+1} - \frac{1}{n} (a_{v+1} - a_u + a_{n-u+1} - a_{n-v}) \right.$$

$$-\frac{1}{n-u+1}a_u\Bigg] - 2(a_{n-u+1} - a_{n-v})$$

$$= \frac{2n}{n-v}(a_{n+1} - a_{v+1}) - \frac{2n}{n-u+1}(a_{n+1} - a_u) + 2(a_{v+1} - a_u).$$

From this equation it is simple to show that (37) is true.

ACKNOWLEDGMENT

I thank the anonymous referees for their suggestions that improved the manuscript considerably. This work is partly supported by a FIRST AWARD from NIH.

REFERENCES

CHAKRABORTY, R., AND GRIFFITHS, R. C. (1982). Correlation of heterozygosity and the number of alleles in different frequency classes, *Theor. Pop. Biol.* **21**, 205–218.

DONNELY, P. (1986). Partition structures, polya urns, the Ewens' sampling formula and the age of alleles, *Theor. Pop. Biol.* **30**, 271–288.

ETHIER, S. N., AND GRIFFITHS, R. C. (1987). The infinitely-many-sites model as a measure-valued diffusion, *Ann. Probab.* **15**, 515–545.

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

EWENS, W. J. (1979). "Mathematical Population Genetics," Springer-Verlag, Berlin.

FELLER, W. (1968). "An Introduction to Probability: Theory and Applications," Vol. 1, 3rd ed., Wiley, New York.

FELSENSTEIN, J. (1992). Estimating effective population size from samples of sequences: Inefficiency of pairwise and segregation sites as compared to phylogenetic estimates, *Genet. Res.* **56**, 139–147.

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

FU, Y. X. (1994b). 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., AND LI, W. H. (1993a). Maximum likelihood estimation of population parameters, *Genetics* **134**, 1261–1270.

FU, Y. X. AND LI, W. H. (1993b). Statistical test of neutrality of mutations, *Genetics* **133**, 693–709.

GRIFFITHS, R. C. (1989). Genealogical tree probabilities in the infinitely-many-site model, *J. Math. Biol.* **27**, 667–680.

HOPPE, F. M. (1984). Polya-like urns and the Ewens' sampling formula, *J. Math. Biol.* **20**, 91–94.

HUDSON, R. R. (1991). Gene genealogies and the coalescent process, *in* "Oxford Surveys in Evolutionary Biology" (D. Futuyama and J. Antonovics, Eds), Vol. 7, pp. 1–44, Oxford Univ. Press, Oxford.

JOHNSON, N. L., AND KOTZ, S. (1977). "Urn Models and Their Application," Wiley, New York.

KINGMAN, J. F. C. (1982a). The coalescent, *Stochastic Process. Appl.* **13**, 235–248.

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

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

TAVARÉ, S. (1984). Line of descent and genealogical process and their applications in population genetics models, *Theor. Pop. Biol.*, **26**, 119–164.

WATERSON, G. A. (1974a). The sampling theory of selectively neutral alleles, *Adv. Appl. Probab.* **6**, 463–488.

WATTERSON, G. A. (1974b). Models for the logarithmic species abundance distributions, *Theor. Pop. Biol.* **6**, 217–250.

WATTERSON, G. A. (1975). On the number of segregation sites, *Theor. Pop. Biol.* **7**, 256–276.