The Coalescent Process in Models With Selection

Norman L. Kaplan, Thomas Darden and Richard R. Hudson[1]

National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina 27709

Manuscript received October 23, 1987
Revised copy accepted March 9, 1988

ABSTRACT

Statistical properties of the process describing the genealogical history of a random sample of genes are obtained for a class of population genetics models with selection. For models with selection, in contrast to models without selection, the distribution of this process, the coalescent process, depends on the distribution of the frequencies of alleles in the ancestral generations. If the ancestral frequency process can be approximated by a diffusion, then the mean and the variance of the number of segregating sites due to selectively neutral mutations in random samples can be numerically calculated. The calculations are greatly simplified if the frequencies of the alleles are tightly regulated. If the mutation rates between alleles maintained by balancing selection are low, then the number of selectively neutral segregating sites in a random sample of genes is expected to substantially exceed the number predicted under a neutral model.

RESTRICTION mapping and DNA sequencing of genes from populations provides information about variation at the nucleotide level. The selectively neutral infinite-sites model (KIMURA 1969) is often the basis for the analysis of this variation (*e.g.*, SHAW and LANGLEY 1979; KREITMAN 1983; CHAKRAVARTI, ELBEIN and PERMUTT 1986; HUDSON 1987). Recent analyses however, cast doubt on the adequacy of the selectively neutral model to account for the patterns of variation between and within species (*e.g.*, GILLESPIE 1986; HUDSON, KREITMAN and AGUADÉ 1987). It is therefore important to investigate competing population genetic models that might explain the observed genetic variation. The analysis of alternative models could also be useful in the development of hypothesis tests as well as more robust estimation methods.

An important summary statistic for nucleotide variation in a sample of genes from a population is S, the number of segregating sites in the sample. For a variety of selectively neutral infinite sites population genetics models with no recombination, the distribution of S is known (WATTERSON 1975; KINGMAN 1982a, b; TAVARÉ 1984; HUDSON and KAPLAN 1986; KAPLAN and HUDSON 1987). Little, however, is known about the distribution of S for infinite sites models in which some of the genetic variation is not selectively neutral. In these cases S can be written as the sum $S_{neu} + S_{sel}$, where S_{neu} is the number of segregating sites which have no selective effects, and S_{sel} is the number of segregating sites which have selective effects. The work presented here shows that for some models with selection and no recombination, the distribution of S_{neu} is tractable. If S_{sel} is negligible compared to S_{neu}, then the statistical properties of S can be inferred from those of S_{neu}. In the extreme, for example, if selection acts at a single nucleotide site, then S_{sel} is at most one. For some selective models, however, S_{sel} may not be negligible compared to S_{neu}. The statistical properties of S_{sel} for these cases will not be considered here.

Two essential features of the selectively neutral infinite sites model proposed by KIMURA (1969) are (1) each segregating site in a random sample genes is the result of a unique mutation and (2) all mutations are selectively neutral in the sense that they do not affect the sampling mechanisms which determine the population structure each generation (*i.e.*, $S = S_{neu}$). Under these general assumptions it can be shown that for $k \geq 0$

$$P(S = k) = \int_0^\infty e^{-\mu t} \frac{(\mu t)^k}{k!} dF(t), \qquad (1)$$

where μ = the rate of neutral mutation per gene per generation,

$$F(t) = P(T \leq t), \qquad t \geq 0,$$

and T is the sum of the lengths (measured in generations) of all the branches of the ancestral tree describing the genealogical history of the sample. It follows from (1) that the moments of S are immediate from those of T. For example

$$E(S) = \mu E(T), \qquad (2)$$

and

$$\mathrm{Var}(S) = \mu E(T) + \mu^2 \mathrm{Var}(T). \qquad (3)$$

For many selectively neutral models the moments

[1] Present address: Department of Ecology and Evolutionary Biology, University of California, Irvine, California 92717.

of T are known for large populations, since the asymptotic behavior of the stochastic process describing the genealogical history of the sample has been characterized (WATTERSON 1975; KINGMAN 1982a, b; TAVARÉ 1984). For example, for a random sample of n genes from a population of size N whose evolution is described by a neutral Wright-Fisher reproductive scheme, the mean and variance of T, when measured in units of $2N$ generations, are approximately

$$E(T) = \sum_{i=1}^{n-1} \frac{2}{i} \quad \text{and} \quad \text{Var}(T) = \sum_{i=1}^{n-1} \frac{4}{i^2}.$$

For infinite sites population genetic models that are not selectively neutral, Equations 1–3 hold, in general, for S_{neu} but not for S. However, for models with selection, the process describing the genealogical history of a sample has not been characterized and so nothing is known about the distribution of T. In this paper this problem is studied and for many models with selection and no recombination, e.g. overdominant selection or mutation-selection balance, the asymptotic behavior of the genealogical process of a random sample of genes is characterized. These results are presented in the Theory section. In order to fully describe the distribution of the genealogical process for models with selection, certain expectations involving the ancestral frequency process must be calculated. These calculations are in general difficult, but in some special cases explicit formulas can be obtained. These cases as well as some numerical results are discussed in the CALCULATIONS section. Finally, some of the implications of these results are presented in the DISCUSSION.

THEORY

The process which describes the genealogical history of a random sample of n genes is called the coalescent (KINGMAN 1982b). If there is no intragenic recombination, then a realization of this process, referred to as the ancestral tree, can be thought of as a binary tree having a node at the top and n tips at the bottom. Each of the n tips is identified with one of the genes in the sample; thus as one moves up the tree tracing the ancestral genes of members of the sample, time is measured from the present generation into the past. There are $n - 1$ nodes in the tree and they are labeled from 1 to $n - 1$ going from the most recent node to the most ancient one. A node is interpreted as an ancestral generation in the history of the sample when the most recent common ancestral gene of two or more genes in the sample occurred. Let $T(j)$ denote the number of generations between the $(n - j)$th and $(n - j + 1)$th nodes, $2 \le j \le n$. For convenience any of the tips is defined to be the 0th node. An ancestral tree for a sample of size 3 is given in Figure 1.

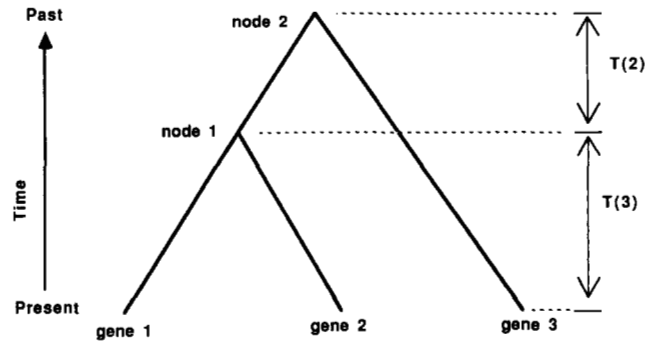

FIGURE 1.—A realization of the coalescent process for a sample of size 3. The first coalescent event occurred at the $T(3)$th ancestral generation and the most recent common ancestor of the sample occurred at the $(T(3) + T(2))$th ancestral generation.

The coalescent process is related to the infinite sites model in the following way (KINGMAN 1982a). If all mutations are unique and neutral in the sense that they do not affect the sampling process, then the distribution of the number of segregating sites when conditioned on the coalescent process is approximately Poisson with mean $\mu(\sum_{j=2}^{n} jT(j))$. Equation 1 is an immediate consequence of this result.

The distribution of the coalescent process is completely characterized for many selectively neutral population genetics models (TAVARÉ 1984). If, for example, a diploid population of size N evolves according to a neutral Wright-Fisher sampling scheme, then the $\{T(j)\}$ are independent random variables and for large N the distribution of each $T(j)$ (when measured in units of $2N$ generations) is approximately negative exponential with mean $2/(j(j-1))$. Furthermore, since the sampling is neutral, any two of the j branches are equally likely to coalesce at the $(n - j + 1)$th node.

The goal of this section is to study the distribution of the coalescent process for population genetic models in which some genetic variation is not selectively neutral. It is instructive for what follows to first present an argument for the neutral case which shows that $T(j)$ has an exponential distribution. For a diploid population of size N that is evolving according to a neutral Wright-Fisher sampling scheme, all parental genes are equally likely to be the parent of a randomly chosen daughter gene. The probability, $1 - Q_j$, that j randomly chosen genes have no common ancestors in the previous generation, can therefore be written as

$$\begin{aligned} 1 - Q_j &= \prod_{i=1}^{j-1} \left(1 - \frac{i}{2N}\right) \\ &= 1 - \frac{\binom{j}{2}}{2N} + O\left(\frac{1}{N^2}\right). \end{aligned} \tag{4}$$

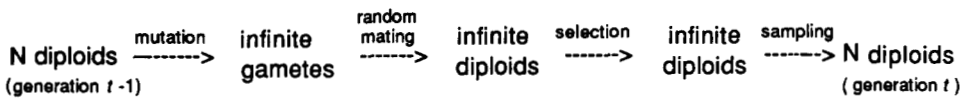

FIGURE 2.—The life cycle.

Hence, for any $t > 0$,

$$P(T(j) > t) = (1 - Q_j)^t \approx e^{-\binom{j}{2}\frac{t}{2N}}. \tag{5}$$

The key to the previous argument is that all parental genes are equally likely to be the parent of a randomly chosen daughter gene. It is this very property which does not hold for populations with genetic variation which is not selectively neutral. For these populations one must keep track of the ancestral allelic frequencies. To simplify matters, it is assumed that at the selected locus A there are two alleles, A_1 and A_2. For generation t, $X(t)$ denotes the fraction of A_1 genes in a diploid population of size N. It is assumed that the population has achieved stationarity and so the current generation from which the random sample is taken is denoted as the 0th generation. The time parameter t thus takes on both positive and negative values, where negative generation times denote ancestral generations and positive generation times future generations.

Each generation the daughter population is obtained by random sampling after mutation and selection have occurred. The life cycle of the process is shown in Figure 2. The fitnesses of the three genotypes A_1A_1, A_1A_2 and A_2A_2 are w_{11}, w_{12} and w_{22}, respectively, and the mean fitness in generation t is denoted by $\bar{w}(t)$. The rates of mutation are $u(A_1$ to $A_2)$ and $v(A_2$ to $A_1)$. Mutations from A_1 to A_2 or A_2 to A_1 will be referred to as selective mutations. It is assumed that

$$w_{11} = w_{12} = w_{22} = 1 + O\left(\frac{1}{N}\right),$$

$$u = \frac{\beta_1}{2N} + O\left(\frac{1}{N^2}\right),$$

and

$$v = \frac{\beta_2}{2N} + O\left(\frac{1}{N^2}\right),$$

where $\beta_1 > 0$, $\beta_2 > 0$.

Let $f_{A_j}(A_k, t)$ denote the probability that a randomly chosen gene from generation t is of allelic type A_k and its parental gene from generation $t - 1$ is of allelic type A_j. For the specified life cycle, it follows from standard population genetic arguments that

$$f_{A_1}(A_1, t) = \frac{1}{\bar{w}(t-1)}\left(X(t-1)^2 w_{11} + X(t-1)(1 - X(t-1))w_{12}\right) + O\left(\frac{1}{N}\right)$$

$$= X(t-1) + O\left(\frac{1}{N}\right),$$

$$f_{A_1}(A_2, t) = \frac{u}{\bar{w}(t-1)}\left(X(t-1)^2 w_{11} + X(t-1)(1 - X(t-1))w_{12}\right) + O\left(\frac{1}{N^2}\right)$$

$$= \frac{\beta_1 X(t-1)}{2N} + O\left(\frac{1}{N^2}\right),$$

$$f_{A_2}(A_2, t) = \frac{1}{\bar{w}(t-1)}\left((1 - X(t-1))^2 w_{22} + X(t-1)(1 - X(t-1))w_{12}\right) + O\left(\frac{1}{N}\right)$$

$$= 1 - X(t-1) + O\left(\frac{1}{N}\right),$$

and

$$f_{A_2}(A_1, t) = \frac{v}{\bar{w}(t-1)}\left((1 - X(t-1))^2 w_{22} + X(t-1)(1 - X(t-1))w_{12}\right) + O\left(\frac{1}{N^2}\right)$$

$$= \frac{\beta_2(1 - X(t-1))}{2N} + O\left(\frac{1}{N^2}\right).$$

Let $f(A_j, t)$ denote the probability of picking a gene of allelic type A_j regardless of the allelic type of the parental gene. It follows that

$$f(A_1, t) = f_{A_1}(A_1, t) + f_{A_2}(A_1, t) = X(t-1) + O\left(\frac{1}{N}\right)$$

and

$$f(A_2, t) = f_{A_1}(A_2, t) + f_{A_2}(A_2, t) = 1 - X(t-1) + O\left(\frac{1}{N}\right).$$

We are now in a position to study the structure of the coalescent process for selective models. Since the ancestral genes can be of allelic type A_1 or A_2, the

coalescent is a two dimensional process. Suppose that n genes are chosen at random from the 0th generation and let $Q(0) = (i, j)$ if the sample consists of i A_1 alleles and j A_2 alleles, $0 \le i, j \le n$, $i + j = n$. For $t < 0$, $Q(t)$ denotes the number of A_1 and A_2 ancestral genes of the sample in generation t. The total number of ancestral genes in generation t is denoted by $|Q(t)|$.

By its very definition Q is a jump process. We define $T_1, T_2, \ldots$ to be the numbers of generations between successive jumps and $Z_1, Z_2, \ldots$ the successive random states to which the process moves. The Q process can therefore be represented as

$$Q(t) = \begin{cases} Q(0) & \text{if} \quad |t| < T_1 \\ Z_k & \text{if} \quad S_k \le |t| < S_k + T_{k+1}, \end{cases}$$

where $S_k = \sum_{i=1}^{k} T_i$, $k \ge 1$. An example of an ancestral tree for a sample of size 4 is given in Figure 3.

It is clear from the definition of the Q process that $|Q(t)|$ never increases. Hence, the process eventually reaches either of the two states (0, 1) or (1, 0). The ancestral generation in which this first occurs is that generation which has the most recent common ancestor of the sample.

We now consider the joint distribution of the $\{T_i\}$ and the $\{Z_i\}$. Toward this end we study the distribution of $Q(t-1)$ conditional on $Q(t)$ and $X(t-1)$. There are two cases to consider:

Case 1: $|Q(t-1)| = |Q(t)|$. The only way that $Q(t-1) \neq Q(t)$ is if the allelic type of at least one of the sampled genes is different (as a result of a selective mutation) than the allelic type of its parental gene. The probability that a sampled A_2 allele from generation t has an A_1 parental gene equals

$$\frac{f_{A_1}(A_2, t)}{f(A_2, t)} + O\left(\frac{1}{N^2}\right).$$

Since j A_2 genes are sampled from generation t,

$$\begin{aligned} P(Q(t-1) = (i+1, j-1) \mid Q(t) = (i, j), X(t-1)) &= \frac{j f_{A_1}(A_2, t)}{f_{A_2}(A_2, t)} + O\left(\frac{1}{N^2}\right) \\ &= j\left(\frac{X(t-1)}{1 - X(t-1)}\right)\frac{\beta_1}{2N} + O\left(\frac{1}{N^2}\right). \end{aligned} \tag{6}$$

Similarly,

$$\begin{aligned} P(Q(t-1) = (i-1, j+1) \mid Q(t) = (i, j), X(t-1)) &= i\left(\frac{1 - X(t-1)}{X(t-1)}\right)\frac{\beta_2}{2N} + O\left(\frac{1}{N^2}\right). \end{aligned} \tag{7}$$

Furthermore, since all the other possible cases where $Q(t-1) \neq Q(t)$ and $|Q(t-1)| = |Q(t)|$ involve at least two selective mutations, these events have probabilities of order $1/N^2$.

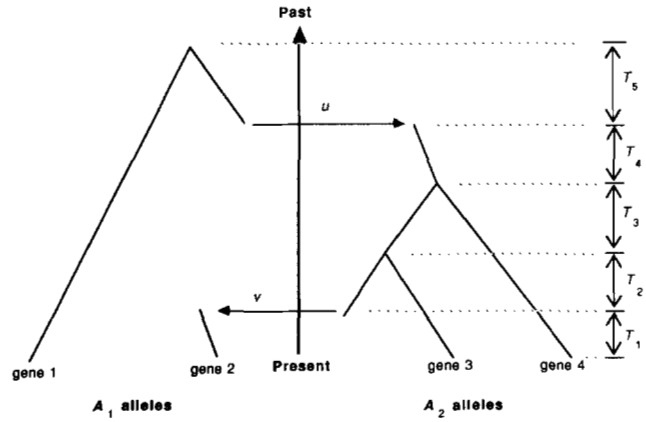

FIGURE 3.—A realization of the coalescent process for a sample of size 4. The Q process changes value at the $S_j = \sum_{i=1}^{j} T_i$ ancestral generations, $1 \le j \le 5$. At the S_1th ancestral generation, an ancestral A_2 allele mutated to an A_1 allele, *i.e.*, the Q process moved from (2, 2) to (1, 3) at the S_2th and S_3th ancestral generations, common ancestors of two ancestral A_2 alleles occurred and so the Q process moved to (1, 2) and then to (1, 1). At the S_4th ancestral generation, an ancestral A_1 allele mutated to an A_2 allele and so the Q process moved to (2, 0). Finally at the S_5th ancestral generation, the most recent common ancestor of the sample occurred and the Q process moved to the state (1, 0).

Case 2: $|Q(t-1)| \neq |Q(t)|$. In this case some of the sampled genes have common parental genes. The fraction of the genes of generation t contributed by a particular A_1 parental gene equals

$$\begin{aligned} \frac{X(t-1)w_{11} + (1 - X(t-1))w_{12}}{2N\bar{w}(t-1)} + O\left(\frac{1}{N^2}\right) \\ = \frac{f_{A_1}(A_1, t)}{2NX(t-1)} + O\left(\frac{1}{N^2}\right). \end{aligned}$$

The probability that two sampled A_1 genes from generation t have the same A_1 parental gene therefore equals

$$\begin{aligned} 2NX(t-1)\left(\frac{f_{A_1}(A_1, t)}{2NX(t-1)f(A_1, t)}\right)^2 + O\left(\frac{1}{N^2}\right) \\ = \frac{1}{2NX(t-1)} + O\left(\frac{1}{N^2}\right). \end{aligned}$$

Since i A_1 genes are sampled from generation t,

$$\begin{aligned} P(Q(t-1) = (i-1, j) \mid Q(t) = (i, j), X(t-1)) \\ = \binom{i}{2}\frac{1}{2NX(t-1)} + O\left(\frac{1}{N^2}\right). \end{aligned} \tag{8}$$

Similarly,

$$\begin{aligned} P(Q(t-1) = (i, j-1) \mid Q(t) = (i, j), X(t-1)) \\ = \binom{j}{2}\frac{1}{2N(1 - X(t-1))} + O\left(\frac{1}{N^2}\right). \end{aligned} \tag{9}$$

Also, since the chance of having more than one coalescent in any generation is of order $1/N^2$,

$$P(|Q(t-1)| < |Q(t)| - 1 \,|\, |Q(t)|, X(t-1)) = O\left(\frac{1}{N^2}\right). \tag{10}$$

It follows from (6)–(10) that the conditional distribution of $Q(t-1)$ up to order $1/N$ is

$$P(Q(t-1) = Q(t) \,|\, Q(t) = (i,j), X(t-1)) = 1 - \frac{h_{ij}(X(t-1))}{2N} + O\left(\frac{1}{N^2}\right), \tag{11}$$

where

$$h_{ij}(x) = \frac{\binom{i}{2}}{x} + \frac{\binom{j}{2}}{1-x} + \frac{j\beta_1 x}{1-x} + \frac{i\beta_2(1-x)}{x}.$$

(If i is less than 2, then $\binom{i}{2}$ is interpreted as 0.)

Furthermore,

$$P(Q(t-1) = (i-1,j) \,|\, Q(t-1) \neq Q(t) = (i,j), X(t-1)) = q_{i-1,j}(X(t-1)) + O\left(\frac{1}{N}\right), \tag{12a}$$

$$P(Q(t-1) = (i,j-1) \,|\, Q(t-1) \neq Q(t) = (i,j), X(t-1)) = q_{i,j-1}(X(t-1)) + O\left(\frac{1}{N}\right), \tag{12b}$$

$$P(Q(t-1) = (i+1,j-1) \,|\, Q(t-1) \neq Q(t) = (i,j), X(t-1)) = q_{i+1,j-1}(X(t-1)) + O\left(\frac{1}{N}\right), \tag{12c}$$

and

$$P(Q(t-1) = (i-1,j+1) \,|\, Q(t-1) \neq Q(t) = (i,j), X(t-1)) = q_{i-1,j+1}(X(t-1)) + O\left(\frac{1}{N}\right), \tag{12d}$$

where

$$q_{i-1,j}(x) = \frac{\binom{i}{2}}{xh_{ij}(x)},$$

$$q_{i,j-1}(x) = \frac{\binom{j}{2}}{(1-x)h_{ij}(x)},$$

$$q_{i+1,j-1}(x) = \frac{j\beta_1 x}{(1-x)h_{ij}(x)}$$

and

$$q_{i-1,j+1}(x) = \frac{i\beta_2(1-x)}{xh_{ij}(x)}.$$

Thus, for large N it can be assumed as a consequence of (12) that when the Q process does jump, there are only four possible states it can jump to: $(i-1,j)$, $(i,j-1)$, $(i+1,j-1)$ and $(i-1,j+1)$. The first two states represent coalescent events and the latter two mutation events.

The formula for the joint distribution of the $\{T_i\}$ and $\{Z_i\}$ follows from (11) and (12). If one conditions on the X process and uses (11) and (12) repeatedly, then

$$P(T_i = t_i, Z_i = z_i, 1 \le i \le k \,|\, Q(0) = z_0) = \frac{1}{(2N)^k} E\left[\prod_{i=1}^{k} (q_{z_i}(X(-s_i)) h_{z_{i-1}}(X(-s_i))) \cdot \prod_{l=s_{i-1}+1}^{s_i-1} \left(1 - \frac{h_{z_{i-1}}(X(-l))}{2N}\right)\right] + O\left(\frac{1}{N^{k+1}}\right), \tag{13}$$

where each $t_i > 0$, $s_i = \sum_{j=1}^{i} t_j$, $s_0 = 0$, each z_i can take on one of four possible values which depend on the value of z_{i-1} and whenever $s_{i-1} + 1$ is greater than $s_i - 1$, the product in (13) is set equal to 1. It should be noted that T_i and Z_i are not independent and that the expectation in (13) is with respect to the distribution of the ancestral frequency process $\{X(t), t \le 0\}$.

We next consider the asymptotic behavior of the expectation in (13). As is customarily done, time is rescaled so that it is measured in units of $2N$ generations. It is a straightforward exercise to show that for N large

$$\prod_{l=2Ns_{i-1}+1}^{2Ns_i-1} \left(1 - \frac{h_{z_{i-1}}(X(-l))}{2N}\right) \approx e^{-\int_{s_{i-1}}^{s_i} h_{z_{i-1}}(X(-2N\tau))\,d\tau}. \tag{14}$$

Hence, it follows from (13) and (14) that

$$P\left(\frac{T_i}{2N} \in [t_i, t_i + \frac{1}{2N}), Z_i = z_i, 1 \le i \le k \,|\, Q(0) = z_0\right) = \frac{1}{(2N)^k} E\left(\prod_{i=1}^{k} (q_{z_i}(X(-2Ns_i)) h_{z_{i-1}}(X(-2Ns_i))) \cdot e^{-\int_{s_{i-1}}^{s_i} h_{z_{i-1}}(X(-2N\tau))\,d\tau}\right) + O\left(\frac{1}{N^{k+1}}\right). \tag{15}$$

If the stationary process $\{X(-2N\tau), \tau > 0\}$ converges weakly to a process $\{Y(\tau), \tau > 0\}$, then it follows from the invariance principle (BILLINGSLEY 1968) that for large N the expectation in (15) can be evaluated with respect to the distribution of the Y process. Thus, in the limit the Q process, when conditioned on the

ancestral frequency process, behaves like a two-dimensional time inhomogeneous Markov jump process.

In many cases of interest the Y process is a diffusion. For example, if the stationary frequency process $\{X(2N\tau),\ \tau > 0\}$ converges to a reversible diffusion, then $\{X(-2N\tau),\ \tau > 0\}$ also converges to a diffusion. Two examples of selective models that lead to limiting diffusions are:

Example 1 (Deleterious selection)

$$w_{11} = 1 + s \qquad w_{12} = 1 + sh \qquad w_{22} = 1,$$

and

Example 2 (Overdominant selection)

$$w_{11} = 1 - s_1 \qquad w_{12} = 1 \qquad w_{22} = 1 - s_2,$$

where s, s_1 and s_2 are of order $1/2N$ and $0 \leq h \leq 1$.

If the limiting process is a diffusion, then it is possible to study properties of the Q process numerically. In the next section a method is described for doing such calculations and some numerical calculations are presented. A special case where this difficult numerical analysis is not necessary is also discussed.

CALCULATIONS

In this section we study the distribution of $T(i, j)$, the sum of the lengths (measured in units of $2N$ generations) of all the branches of the ancestral tree, assuming that the sample consisted of i A_1 alleles and j A_2 alleles ($i + j \geq 2$). We first consider selective models where mutation and selection act in such a way that at equilibrium the frequencies of the two alleles remain essentially constant for long periods of time. In these cases it can be assumed that there is a constant $x_0 (0 < x_0 < 1)$ such that $X(t) = x_0$ for all t. These tightly regulated models are of interest because the moments of $T(i, j)$ are easier to compute and they may be good approximations to the moments of $T(i, j)$ for selective models where the frequencies of the two alleles are not tightly regulated.

If it is assumed that $X(t) = x_0$ for all t, then it follows from (15) that the limiting Q process is a time homogeneous Markov process with the following parameters. For any state (i, j), the holding time in that state, T_{ij}, has a negative exponential distribution with parameter $h_{ij}(x_0)$ and the probability that the process moves from (i, j) to (i', j') equals $q_{i'j'}(x_0)$ where (i', j') can equal $(i - 1, j)$, $(i, j - 1)$, $(i - 1, j + 1)$ *or* $(i + 1, j - 1)$. It is important to note that in the tightly regulated case, the form of selection affects the limiting Q process only to the extent that it affects the value of x_0. Hence, the distribution of the limiting Q process will be the same for any tightly regulated selection model having the same value of x_0 and the same mutation parameters.

For some selection models, *e.g.*, Example 1, x_0 depends on β_1 and β_2. For Example 1, x_0 satisfies the equation

$$\alpha x_0(1 - x_0)(x_0 + h(1 - 2x_0)) - \beta_1 x_0 + \beta_2(1 - x_0) = 0, \tag{16}$$

where $\alpha = 2Ns$ (EWENS 1979). If x_0 is close to one, then it is well known (EWENS 1979) that

$$x_0 \approx 1 - \frac{u}{s(1 - h)}, \tag{16a}$$

while if $s > 0$ and $h = 1$, then

$$x_0 \approx 1 - \sqrt{\frac{u}{s}}. \tag{16b}$$

For other selection models x_0 may not depend on the mutation parameters. This is true for Example 2 if u and v are small when compared to s_1 and s_2, since in this case

$$x_0 = \frac{s_2}{s_1 + s_2}.$$

The conditions under which the approximating diffusions in Examples 1 and 2 behave as if they converge to a deterministic equilibrium point have been discussed by many authors (NORMAN 1975; KURTZ 1981) and the interested reader should consult their papers for details. Loosely speaking, the diffusion will behave in this way if the infinitesimal mean is large compared to the infinitesimal variance.

The following representation for $T(i, j)$ is a direct consequence of the Markov property of the limiting Q process. If we consider what happens when the process first changes value, then

$$T(i, j) = (i + j)T_{ij} + T(Z_{ij}), \tag{17}$$

where T_{ij} is the holding time in state (i, j), Z_{ij} is the random state to which the process moves and if $Z_{ij} = (i', j')$, then $T(Z_{ij})$ is an independent random variable having the same distribution as $T(i', j')$.

Recursions for the mean and variance of $T(i, j)$ are easy to obtain from (17). Let M_{ij} denote the mean of $T(i, j)$ and V_{ij} its variance. Then

$$M_{ij} = \frac{i + j}{h_{ij}(x_0)} + q_{i-1,j}(x_0)M_{i-1j} + q_{i,j-1}(x_0)M_{ij-1} + q_{i+1,j-1}(x_0)M_{i+1j-1} + q_{i-1,j+1}(x_0)M_{i-1j+1} \tag{18}$$

and

$$\begin{aligned} V_{ij} = {} & \frac{(i+j)^2}{h_{ij}^2(x_0)} + q_{i-1,j}(x_0)V_{i-1j} + q_{i,j-1}(x_0)V_{ij-1} \\ & + q_{i+1,j-1}(x_0)V_{i+1j-1} + q_{i-1,j+1}(x_0)V_{i-1j+1} \\ & + q_{i-1,j}(x_0)M_{i-1j}^2 + q_{i,j-1}(x_0)M_{ij-1}^2 \\ & + q_{i+1,j-1}(x_0)M_{i+1j-1}^2 + q_{i-1,j+1}(x_0)M_{i-1j+1}^2 \\ & - (q_{i-1,j}(x_0)M_{i-1j} + q_{i,j-1}(x_0)M_{ij-1} \\ & + q_{i+1,j-1}(x_0)M_{i+1j-1} + q_{i-1,j+1}(x_0)M_{i-1j+1})^2, \end{aligned} \tag{19}$$

where $M_{01} = M_{10} = V_{01} = V_{10} = 0$.

Equations 18 and 19 are not useful for studying the behavior of M_{ij} and V_{ij} if n is large. It is easier to study the evolution of the Q process directly. It is not hard to show that the behavior of the mean and variance of $T(i, j)$ for large samples is the same as in the neutral case. That is, the mean of $T(i, j)$ grows like $\log(i + j)$ and its variance remains bounded.

For small samples, it is necessary to solve the recursions in (18) and (19) for particular parameter values. Since it may commonly be the case that one cannot distinguish between the two alleles, we assume that the number of A_1 alleles in the sample is random with a binomial distribution. The quantities tabulated in Table 1 are therefore

$$M_n(x_0) = \sum_{i=0}^{n} \binom{n}{i} x_0^i (1 - x_0)^{n-i} M_{in-i},$$

and

$$V_n(x_0) = \sum_{i=0}^{n} \binom{n}{i} x_0^i (1 - x_0)^{n-i} (V_{in-i} + (M_{in-i} - M_n(x_0))^2).$$

The results in Table 1 show that the values of $M_n(x_0)$ and $V_n(x_0)$ only differ significantly from their values in the neutral case when both β_1 and β_2 are small. Furthermore, the closer x_0 is to 0 or 1, the smaller β_1 and β_2 have to be in order for $M_n(x_0)$ and $V_n(x_0)$ to differ from their neutral values.

It is easy to explain why small values of β_1 and β_2 lead to ancestral trees which do not look neutral. If selective mutations are rare, then most of the state changes in the Q process result from common ancestor events. Hence, with high probability all the A_1 alleles and all the A_2 alleles will coalesce, resulting in just two ancestral genes; one of each allelic type. In order for these two ancestral genes to have a common ancestor, it is necessary for a selective mutation to occur first. If β_1 and β_2 are small, then this event takes a long time to occur and so the ancestral tree will not look neutral.

In those cases where x_0 depends on β_1 and β_2 the above behavior may not hold. To see what can happen we consider Example 1 of the previous section. For a sample of size 2 the equations for M_{20}, M_{11} and M_{02} are easy to solve. Indeed,

$$M_{20} = \frac{2x_0}{1 + 2\beta_2(1 - x_0)} + \frac{2\beta_2(1 - x_0)M_{11}}{1 + 2\beta_2(1 - x_0)},$$

$$M_{02} = \frac{2(1 - x_0)}{1 + 2\beta_1 x_0} + \frac{2\beta_1 x_0 M_{11}}{1 + 2\beta_1 x_0}$$

and

$$M_{11} = \frac{2\left(x_0(1 - x_0) + \dfrac{\beta_2(1 - x_0)^3}{1 + 2\beta_1 x_0} + \dfrac{\beta_1 x_0^3}{1 + 2\beta_2(1 - x_0)}\right)}{\dfrac{\beta_2(1 - x_0)^2}{1 + 2\beta_1 x_0} + \dfrac{\beta_1 x_0^2}{1 + 2\beta_2(1 - x_0)}}.$$

TABLE 1

Mean ($M_n(x_0)$) and variance ($V_n(x_0)$) of the total time in the genealogy of a random sample of n genes, assuming tight regulation

			$n = 2$		$n = 20$	
β_1	β_2	x_0	$M_2(x_0)$	$V_2(x_0)$	$M_{20}(x_0)$	$V_{20}(x_0)$
100.0	100.0	0.5	2.0	4.0	7.1	6.4
		0.1	1.8	3.4	6.5	5.4
		0.05	1.9	3.6	6.5	5.4
10.0	10.0	0.5	2.1	4.2	7.3	6.7
		0.1	1.8	3.4	6.6	5.4
		0.05	1.9	3.6	6.8	5.8
1.0	1.0	0.5	2.5	6.8	8.4	9.9
		0.1	1.9	3.4	6.8	5.5
		0.05	1.9	3.7	6.8	5.8
0.1	0.1	0.5	7.0	99.0	17.5	127.0
		0.1	2.2	5.2	8.5	10.5
		0.05	2.0	3.9	7.5	6.8
0.01	0.01	0.5	52.0	7704.0	108.0	10208.0
		0.1	5.8	163.0	26.0	483.0
		0.05	2.9	24.0	13.5	102.0
1.0	100.0	0.5	1.0	1.0	3.8	1.7
		0.1	1.8	3.2	6.4	5.2
		0.05	1.9	3.6	6.7	5.8
100.0	1.0	0.1	0.85	0.60	3.6	1.2
		0.05	1.6	2.4	5.8	4.0
80.0	5.0	0.5	1.1	1.3	4.2	2.0
		0.1	2.0	3.8	6.9	6.1

For comparison the mean and variance of the time in the genealogical history of sample from a selectively neutral model is 2.0 and 4.0, for a sample of size 2, and 7.1 and 6.4, for a sample of size 20.

Suppose that β_1 and β_2 are both small and that x_0 satisfies (16a). In this case, α must be large in order for the frequency of the A_1 to be tightly regulated. It is not difficult to show under these conditions that

$$x_0^2 M_{20} \approx 2, \quad x_0(1 - x_0)M_{11} \approx 0 \quad \text{and} \quad (1 - x_0)^2 M_{02} \approx 0.$$

Thus, $M_2(x_0) \approx 2$ regardless of how small β_1 and β_2 are.

Knowing the allelic composition of the sample also affects the conclusions about the sample's genealogy. For example, suppose that both members in a sample of size 2 are the same allelic type. Suppose also that $\beta_1 = \beta_2 = \beta \approx 0$, and $x_0 \approx 1$. Then it is not difficult to show that

$$M_{20} \approx 2$$

and

$$M_{02} \approx 2(1 - x_0) + 4(1 - x_0 + \beta).$$

Thus, if one picked two genes of the common allelic type, then the ancestral tree looks neutral. If, on the other hand, one picked two genes of the rare allelic type, then the ancestral tree looks much smaller than a neutral tree.

Before we consider selection models which are not tightly regulated, we briefly examine a relationship between the coalescent process for tightly regulated selective models and the infinite alleles model. Sup-

pose at locus A there are two *classes* of alleles which are denoted A_1 and A_2. Class A_1 consists of alleles A_{11}, A_{12}, ... and class A_2 consists of alleles A_{21}, A_{22}, ... Suppose that every selective mutation of an A_1 allele gives rise to a new allele of the A_2 class and vice versa. In addition each neutral mutation gives rise to a new allele within the same class. HARTL and CAMPBELL (1982) showed under these assumptions that the distribution of the number and frequencies of A_1 alleles in a random sample from the A_1 class follows an infinite alleles distribution (EWENS 1972) where

$$\theta_{A_1} = 4Nx_0\left(\mu + \frac{\beta_2(1 - x_0)}{2Nx_0}\right).$$

A similar result holds for a sample from the A_2 class with

$$\theta_{A_2} = 4N(1 - x_0)\left(\mu + \frac{\beta_1 x_0}{2N(1 - x_0)}\right).$$

HARTL and CAMPBELL proved this result without using the coalescent process. Their arguments however, provide no information about the distribution of the number of segregating neutral sites in the sample. For the model considered by HARTL and CAMPBELL, our analysis shows that the expected number of neutral segregating sites in a sample of n genes from the A_1 class equals $2N\mu E(T(n, 0))$.

We now examine selective models where the limiting Y process is a diffusion which is not tightly regulated. In these cases one can show that for any $n \geq 2$, the functions, $\{F_{ij}(t, x) = P(T(i, j) > t \mid Y(0) = x),\ 2 \leq i + j \leq n\}$ are the solution of the following system of partial differential equations

$$\begin{aligned}\frac{\partial}{\partial t}F_{ij}(t,x) = \frac{1}{(i+j)}\Bigg[&a(x)\frac{\partial}{\partial x}F_{ij}(t,x) + \frac{b(x)}{2}\frac{\partial^2}{\partial x^2}F_{ij}(t,x)\\ &+\frac{\binom{i}{2}}{x}(F_{i-1j}(t,x) - F_{ij}(t,x))\\ &+\frac{\binom{j}{2}}{(1-x)}(F_{ij-1}(t,x) - F_{ij}(t,x))\\ &+\frac{i\beta_2(1-x)}{x}(F_{i-1j+1}(t,x) - F_{ij}(t,x))\\ &+\frac{j\beta_1 x}{(1-x)}(F_{i+1j-1}(t,x) - F_{ij}(t,x))\Bigg],\end{aligned} \tag{20}$$

where $t > 0$, $0 < x < 1$ and $2 \leq i + j \leq n$.

The derivation of (20) relies on the same kinds of Taylor series arguments that are used to study the distribution of a hitting time for a diffusion process (KARLIN and TAYLOR 1981).

There are several difficulties in solving (20) numerically. First, the coefficient of the second order term equals zero at the boundaries, *i.e.*, $b(0) = b(1) = 0$. Second, the boundary conditions, $\{F_{ij}(t, 0)$ and $F_{ij}(t, 1)$, $t > 0\}$ are unknown. Finally, the right-hand side of (20) explodes as x approaches 0 and 1. For these reasons standard software packages for solving partial differential equations cannot be used and so alternative methods need to be developed.

Since none of the coefficients in (20) are time dependent, it is not difficult to obtain from (20) a system of equations for each moment of $T(i, j)$. For example, if $M_{ij}(x) = E(T(i, j) \mid Y(0) = x)$, then integrating both sides of (20) with respect to t leads to

$$\begin{aligned}-1 = \frac{1}{(i+j)}\Bigg[&a(x)\frac{d}{dx}M_{ij}(x) + \frac{b(x)}{2}\frac{d^2}{dx^2}M_{ij}(x)\\ &+\frac{\binom{i}{2}}{x}(M_{i-1j}(x) - M_{ij}(x)) + \frac{\binom{j}{2}}{(1-x)}(M_{ij-1}(x)\\ &- M_{ij}(x)) + \frac{i\beta_2(1-x)}{x}(M_{i-1j+1}(x) - M_{ij}(x))\\ &+\frac{j\beta_1 x}{(1-x)}(M_{i+1j-1}(x) - M_{ij}(x))\Bigg],\end{aligned} \tag{21}$$

where $0 < x < 1$ and $2 \leq i + j \leq n$. To obtain equations for the higher moments of $T(i, j)$ one multiplies both sides of (20) by an appropriate power of t before integrating.

For the same reasons given above, standard software packages for solving systems of equations such as those in (21) cannot be used. In a forthcoming paper (DARDEN, KAPLAN and HUDSON 1988), a numerical method for solving (21) is described. This method is patterned after the LU decomposition for tridiagonal matrices (PRESS *et al.* 1988).

To illustrate the numerical results the following examples are considered. In Table 2 the analogs of $M_2(x_0)$ and $V_2(x_0)$ are calculated for a sample of size 2, assuming the selection model of example 2. These quantities are

$$M_2 = \sum_{i=0}^{2}\binom{2}{i}\int_0^1 M_{i2-i}(x)x^i(1-x)^{2-i}p(x)dx$$

and

$$V_2 = \sum_{i=0}^{2}\binom{2}{i}\int_0^1 L_{i2-i}(x)x^i(1-x)^{2-i}p(x)dx - (M_2)^2,$$

where $L_{ij}(x) = E(T(i, j)^2 \mid Y(0) = x)$ and $p(x)$ is the stationary density of the diffusion.

Two cases were considered. In the first case $\beta_1 = \beta_2 = 0.1$ and $s_1 = s_2$, while in the second case $\beta_1 = \beta_2 = 0.01$ and $s_1 = 9s_2$. The equilibrium allelic frequency, for the tightly regulated process in the first case equals

TABLE 2

Mean (M_2) and variance (V_2) of the total time in the genealogy of a random sample of two genes, assuming that the selected locus is not tightly regulated

$2Ns_2$	M_2	V_2
Case 1 ($\beta_1 = \beta_2 = 0.1$, $s_1 = s_2$):		
0.0	2.00	4.0
2.5	2.22	5.8
10.0	4.34	36.0
50.0	6.65	90.0
250.0	6.93	97.0
∞	7.00	99.0
Case 2 ($\beta_1 = \beta_2 = 0.1$, $s_1 = 9s_2$):		
0	2.00	4.0
5	2.00	4.2
50	2.45	8.8
100	4.92	107.0
200	5.47	143.0
∞	5.8	163.0

For both of these cases the limiting diffusion is associated with the overdominant selection model of Example 2.

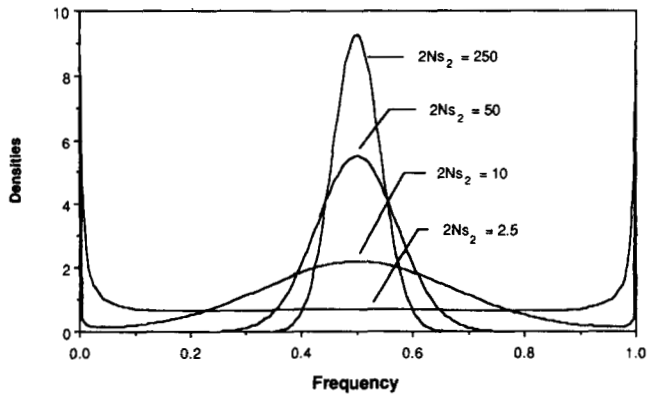

FIGURE 4.—The stationary density of the limiting diffusion associated with the overdominant selection model of Example 2.

0.5, whereas for the second case it equals 0.1. The results in Table 2 show that the values of $M_2(x_0)$ and $V_2(x_0)$ are good approximations to M_2 and V_2 so long as $2Ns_2 > 50$ in case 1 and $2Ns_2 > 200$ in case 2. To provide a visual impression of the tightness of the regulation of the frequency process, the density of the stationary distribution of the limiting diffusion for case 1 is plotted in Figure 4 for four values of $2Ns_2$. It is evident that there is some variability when $2Ns_2$ equals 50, and yet the tight-regulation approximation is quite accurate.

Up until now we have considered selection models where the infinitesimal variance of the approximating diffusion is not large. For some models however, this may not be true. For example, the random environment models studied by GILLESPIE (1978) lead to diffusions whose infinitesimal means and variances are of the form

$$a(x) = x(1 - x)C\left[A + B\left(\frac{1}{2} - x\right)\right]$$

and

$$b(x) = Cx^2(1 - x)^2,$$

where A, B and C are constants.

Diffusions of this type can be thought of running on a different time scale and so it is appropriate to rescale time. Thus, suppose that the ancestral frequency process $\{X(-2N\tau), \tau > 0\}$ converges weakly to a process $\{Y(C\tau), \tau > 0\}$, where Y is a diffusion and C is a constant. If C is sufficiently large, *i.e.*, the diffusion is running on a faster time scale, then it follows from (14) and (15) that the coalescent process can again be approximated by a time homogeneous Markov process. The joint density of the holding time in any state and the random state to which the process jumps is

$$P(T_1 \epsilon [t, t + dt), Z_1 = z_1 \mid Q(t) = z) = e^{-t\int_0^1 h_z(u)p(u)du}\left(\int_0^1 h_z(u)q_{z_1}(u)p(u)du\right)dt, \quad t > 0,$$

where $p(u)$ is the stationary density of the approximating diffusion, Y.

DISCUSSION

Restriction mapping and DNA sequencing of samples of genes from populations give information that is more detailed and less ambiguous than the information from allozyme studies. These new molecular techniques also provide information about the age and genealogical relationships of alleles (*e.g.*, SHAW and LANGLEY 1979; STEPHENS and NEI 1985; AQUADRO *et al.* 1986). Effective use of this new type of information requires an understanding of the genealogical relationships expected under competing population genetic hypotheses that might explain the observed molecular genetic variation. Under some simple genetic models without selection, many statistical properties of the process describing the genealogical history of samples are known (WATTERSON 1975; KINGMAN 1982a, b; TAVARÉ 1984). The purpose of this investigation is to study properties of this process for population genetic models which are not selectively neutral.

The distribution of the coalescent process for models with selection depends on the distribution of the frequencies of alleles in the ancestral generations. For many two-allele selective models, *e.g.*, examples 1 and 2, the ancestral frequency process can be approximated by a diffusion process. In these cases the mean and the variance of T can be computed, but it requires solving a non-standard system of second order differential equations. A computer program was written to solve these equations numerically and some typical results are presented in Table 2.

Some simplification is possible if it can be assumed that the allelic frequencies do not vary from generation to generation, *i.e.*, they are tightly regulated. For selection models of this type, the coalescent process is a time homogeneous Markov jump process whose distribution only depends on the mutation rates and the equilibrium frequency, regardless of the form of selection. In this case the mean and variance of T each satisfy a system of linear equations which is much easier to solve. Furthermore, the results in Tables 1 and 2 suggest that the values obtained assuming tight regulation are good approximations even when the allelic frequencies are not that tightly regulated.

The arguments for the tightly regulated case can be easily generalized to k-allele models, $k > 2$. If the allelic frequencies are not tightly regulated then the results do not generalize since the limiting ancestral frequency process, $\{Y(t), t > 0\}$, is not generally known to be a diffusion for $k > 2$.

If, for the tightly regulated case, the allelic frequencies do not depend on the mutation parameters, $\beta_1(=2Nu)$ and $\beta_2(=2Nv)$, for example under models of strong balancing selection, then the mean and variance of T differ substantially from their values in the neutral case only when the mutation parameters, β_1 and β_2, are small. If, on the other hand, the allelic frequencies depend on the mutation parameters, then the mean and the variance of T may not differ significantly from their neutral values regardless of how small β_1 and β_2 are. The mutation-selection balance model illustrates this behavior.

For neutral models, an unbiased estimate of $2N\mu$ is $S/(2\sum_1^{n-1} 1/j)$, where S is the number of segregating sites in a random sample of n genes and $2\sum_1^{n-1} 1/j$ is the expected value of T for a neutral model for a random sample of n genes. If in fact some of the genetic variation is not selectively neutral, then an unbiased estimate of $2N\mu$ is $S_{neu}/E(T)$, where S_{neu} is the number of segregating selectively neutral sites and $E(T)$ is the expected value of T for the selective model. Thus, under the selective model, the neutral estimate is biased in the following two ways. First, the observed value of S is too large since it is the sum of the numbers of segregating neutral sites (S_{neu}) and segregating selective sites (S_{sel}). If the number of segregating selective sites is small compared to the number of segregating neutral sites, e.g. if β_1 and β_2 are small compared to $2N\mu$, then using the observed value of S instead of S_{neu}, will not introduce much error. Secondly, under a selective model, the expected value of T may, in fact, be much larger than $2\sum_1^{n-1} 1/j$, and so the neutral estimate of $2N\mu$ may be too large. If β_1 and β_2 are very small, then this bias could be substantial (Table 1).

The models studied in this paper assume that all sites are completely linked. Clearly, it is important to introduce recombination into the analysis and in particular to determine the behavior of the coalescent process for neutral sites not completely linked to a site at which selection is operating. In a companion study in this journal this problem is addressed (HUDSON and KAPLAN 1988).

LITERATURE CITED

AQUADRO, C. F., S. F. DEESE, M. M. BLAND, C. H. LANGLEY and C. C. LAURIE-AHLBERG, 1986 Molecular population genetics of the alcohol dehydrogenase gene region of *Drosophila melanogaster*. Genetics **114:** 1165–1190.

BILLINGSLEY, P., 1968 *Convergence of Probability Measures*. Wiley, New York.

CHAKRAVARTI, A., S. C. ELBEIN and M. A. PERMUTT, 1986 Evidence for increased recombination near the human insulin gene: Implication for disease association studies. Proc. Natl. Acad. Sci. USA **83:** 1045–1049.

DARDEN, T., M. L. KAPLAN and R. R. HUDSON, 1988 A numerical method for calculating moments of coalescent times in finite populations with w selection. J Math Biol (in press).

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

EWENS, W. J., 1979 *Mathematical Population Genetics*. Springer-Verlag, New York.

GILLESPIE, J. H., 1978 A general model to account for enzyme variation in natural populations. V. The SAS-CFF model. Theor. Popul. Biol. **14:** 1–45.

GILLESPIE, J. H., 1986 Variability of evolutionary rates of DNA. Genetics **113:** 1077–1091.

HARTL, H. L., and R. B. CAMPBELL, 1982 Allele multiplicity in simple Mendelian disorders. Am. J. Hum. Genet. **34:** 866–873.

HUDSON, R. R., 1987 Estimating the recombination parameter of a finite population model without selection. Genet. Res. **50:** 245–250.

HUDSON, R. R., and N. L. KAPLAN, 1986 On the divergence of alleles in nested subsamples from finite populations. Genetics **113:** 1057–1076.

HUDSON, R. R., and N. L. KAPLAN, 1988 The coalescent process in models with selection and recombination. Genetics **120:** 819–829.

HUDSON, R. R., M. KREITMAN and M. AGUADÉ, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116:** 153–159.

KAPLAN, N. L., and R. R. HUDSON, 1987 On the divergence of genes in multigene families. Theor. Popul. Biol. **31:** 178–194.

KARLIN, S., and H. M. TAYLOR, 1981 *A Second Course in Stochastic Processes*. Academic Press, New York.

KIMURA, M., 1969 The number of heterozygous nucleotide sites maintained in a finite population due to a steady flux of mutations. Genetics **61:** 893.

KINGMAN, J. F. C., 1982a On the genealogy of large populations. J. Appl. Prob. **A19:** 27–43.

KINGMAN, J. F. C., 1982b The coalescent. Stochastic Process. Appl. **13:** 235–248.

KREITMAN, M., 1983 Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. Nature **304:** 412–417.

KURTZ, T. G., 1981 *Approximation of Population Processes*. Society for Industrial and Applied Mathematics, Philadelphia.

NORMAN, F., 1975 Approximation of stochastic processes by Gaussian diffusions and applications to Wright-Fisher genetic models. SIAM J. Appl. Math. **29:** 225–242.

PRESS, W. H., B. P. FLANNERY, S. A. TEUKOLSKY and W. T.

VETTERLING, 1988 *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press.

SHAW, D. M., and C. H. LANGLEY, 1979 Inter- and intraspecific variation in restriction maps of Drosophila mitochondrial DNAs. Nature **281:** 695–699.

STEPHENS, J. C., and M. NEI. 1985 Phylogenetic analysis of polymorphic DNA sequences at the Adh locus in *Drosophila melanogaster* and its sibling species. J. Mol. Evol. **22:** 289–300.

TAVARÉ, S., 1984 Line-of-descent and genealogical processes, and their applications in population genetic models. Theor. Popul. Biol. **26:** 119–164.

WATTERSON, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **10:** 256–276.

Communicating editor: B. S. WEIR