GENE GENEALOGY AND VARIANCE OF INTERPOPULATIONAL NUCLEOTIDE DIFFERENCES

NAOYUKI TAKAHATA[1] AND MASATOSHI NEI

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

Manuscript received November 2, 1984
Accepted February 8, 1985

ABSTRACT

A mathematical theory is developed for computing the probability that m genes sampled from one population (species) and n genes sampled from another are derived from l genes that existed at the time of population splitting. The expected time of divergence between the two most closely related genes sampled from two different populations and the time of divergence (coalescence) of all genes sampled are studied by using this theory. It is shown that the time of divergence between the two most closely related genes can be used as an approximate estimate of the time of population splitting (T) only when $T \equiv t/(2N)$ is small, where t and N are the number of generations and the effective population size, respectively. The variance of Nei and Li's estimate (d) of the number of net nucleotide differences between two populations is also studied. It is shown that the standard error (s_d) of d is larger than the mean when T is small ($T \ll 1$). In this case, s_d is reduced considerably by increasing sample size. When T is large ($T > 1$), however, a large proportion of the variance of d is caused by stochastic factors, and increase in the sample size does not help to reduce s_d. To reduce the stochastic variance of d, one must use data from many independent unlinked gene loci.

AFTER the discovery of the molecular clock (ZUCKERKANDL and PAULING 1965), many authors have attempted to estimate the time of divergence between species or populations from amino acid or nucleotide sequence data. Although the molecular clock does not run as regularly as the ordinary clock, it gives a rough idea about the divergence time (FITCH 1976). When this method of estimating divergence time is applied to closely related species or populations, however, some caution is necessary because the divergence time between a pair of genes (or proteins) sampled from different populations may be substantially greater than the time of population splitting (Figure 1). This situation occurs when the ancestral population is polymorphic. When more than two genes are sampled from each population, a correction for the effect of ancestral polymorphism can be made, and the number of nucleotide substitutions (net nucleotide substitutions), d, that occurred after population splitting is estimated by subtracting within-population differences (NEI and LI 1979).

[1] Present address: National Institute of Genetics, Mishima, Shizouka-ken, 411 Japan.

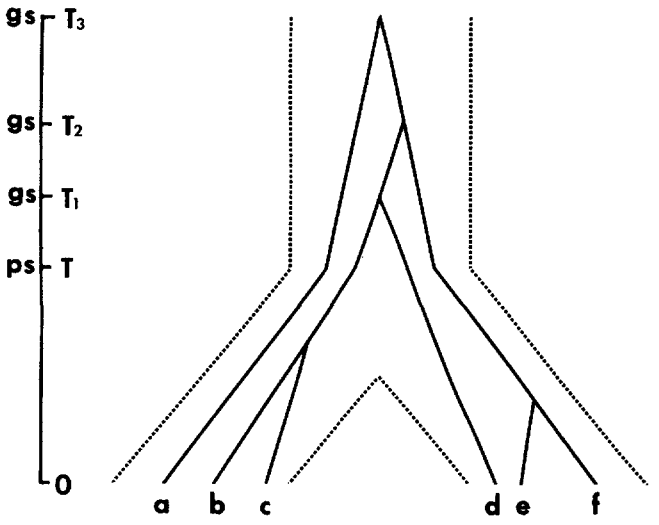

FIGURE 1.—Genealogy of six sampled genes, three (a, b and c) from one population and three (d, e and f) from the other population. The two populations are assumed to have diverged at time T. T_1, T_2 and T_3 represent the times of gene splitting.

When nucleotide substitution occurs by mutation and genetic drift, the expectation of d is given by $E(d) = 2vt$, where v is the mutation rate per gene and t is the time since divergence between populations X and Y (LI 1977; NEI and LI 1979). Therefore, if we know v, we can estimate t from d. However, to evaluate the accuracy of this method, we must determine the variance of d generated by both sampling and stochastic errors. The sampling variance was studied by NEI and TAJIMA (1981), but the variance due to stochastic errors has not been worked out.

To evaluate the variance of d generated by both sampling and stochastic errors, we must first know the expected genealogy of sampled genes. Knowledge of the expected genealogy of sampled genes is also important for estimating the difference between the time of population splitting and the time of gene splitting. If this difference is small, the time of population splitting may be estimated approximately by the time of gene splitting.

The main purpose of this paper is to study the above two problems. We first examine the relationship between population and gene splitting times and then investigate the variance of d. Throughout the paper, we assume that one population splits into two (populations X and Y) t generations ago and, thereafter, no migration occurs between them. We also assume that the effective population size (N) remains constant throughout the evolutionary process, and, thus, the populations are in steady state with respect to the effects of mutation and genetic drift. We use the infinite-site model of neutral mutations with no recombination (KIMURA 1971; WATTERSON 1975).

GENEALOGY OF GENES SAMPLED FROM TWO POPULATIONS

We first consider the expected genealogy of m genes sampled from population X (or Y) and derive a formula for the probability that the m genes are

descended from m_0 at the time of population splitting. Obviously, $1 \leq m_0 \leq m$. We start with the formula for the probability density $[f_{p-1}(s)]$ of waiting time s at which p genes are descended from $p-1$ genes. We use the continuous time approximation to the Wright-Fisher model and apply KINGMAN'S (1982) equation

$$f_{p-1}(s) = \frac{\alpha_p}{2N} e^{-\frac{\alpha_p s}{2N}}, \tag{1}$$

where $\alpha_p = p(p-1)/2$ (see also HUDSON 1983; TAJIMA 1983; TAVARÉ 1984). Equation 1 may be written as

$$f_{p-1}(\tau) = \alpha_p e^{-\alpha_p \tau}, \tag{2}$$

if we measure time in terms of $\tau = s/(2N)$.

To obtain the probability distribution of the number of ancestral genes (m_0) at the time of population splitting $[T = t/(2N)$ units of time ago], we introduce a set of random variables (τ_p). τ_{p-1} is the waiting time at which p genes are descended from $p-1$ ancestral genes and follows the exponential density function given by (2). For $1 \leq m_0 \leq m-1$, we define the sum

$$S_{mm_0} = \sum_{p=m_0}^{m-1} \tau_p \tag{3}$$

and denote by $p(S_{mm_0} = \tau)$ the probability density of $S_{mm_0} = \tau$. Equation (3) represents the waiting time at which m genes are descended from m_0 ancestral genes. $p(S_{mm_0} = \tau)$ can be obtained by the convolution of $f_{m_0}, \ldots, f_{m-2}, f_{m-1}$. Using the Laplace transform and the partial fraction expansion, we obtain

$$p(S_{mm_0} = \tau) = \left(\prod_{p=m_0+1}^{m} \alpha_p \right) \sum_{p=m_0+1}^{m} \rho_p(m, m_0) e^{-\alpha_p \tau}, \tag{4}$$

where

$$\rho_p(m, m_0) = \prod_{\substack{r=m_0+1 \\ r \neq p}}^{m} (\alpha_r - \alpha_p)^{-1}. \tag{5}$$

The probability, $P_{mm_0}(T)$, that m genes are descended from m_0 ancestral genes T units of time ago can be obtained by using (4). That is,

$$\begin{aligned} P_{mm_0}(T) &= p\{S_{mm_0} < T \leq S_{mm_0} + \tau_{m_0-1}\} \\ &= \int_0^T d\tau p\{S_{mm_0} = \tau\} \int_{T-\tau}^{\infty} d\zeta \alpha_{m_0} \exp(-\alpha_{m_0} \zeta), \end{aligned} \tag{6}$$

which becomes

$$P_{mm_0}(T) = \left(\prod_{p=m_0+1}^{m} \alpha_p \right) \sum_{p=m_0}^{m} \rho_p(m, m_0 - 1) \exp(-\alpha_p T) \tag{7a}$$

for $2 \leq m_0 \leq m-1$. $P_{m1}(T)$ and $P_{mm}(T)$ are computed separately.

TABLE 1

Probabilities, $P_{mm_0}(T)$, *that* m *genes sampled from a population are descended from* m_0 *ancestral genes* T *units of time* (2NT *generations*) *ago*

		m_0				
T	m	1	2	3	4	$5 \leq m_0 \leq m$
0.5	2	0.3935	0.6065			
	5	0.0813	0.4013	0.4030	0.1076	0.0068
	10	0.0248	0.2047	0.4051	0.2788	0.0866
1.0	2	0.6321	0.3679			
	5	0.3341	0.5298	0.1300	0.0061	0.0000
	10	0.2278	0.5257	0.2221	0.0237	0.0007
1.5	2	0.7769	0.2231			
	5	0.5695	0.3990	0.0312	0.0003	0.0000
	10	0.4824	0.4583	0.0581	0.0012	0.0000
2.0	2	0.8647	0.1353			
	5	0.7329	0.2601	0.0070	0.0000	0.0000
	10	0.6746	0.3120	0.0134	0.0000	0.0000

$$P_{m1}(T) = P(S_{m1} \leq T) = \left(\prod_{p=2}^{m} \alpha_p\right) \sum_{p=2}^{m} \frac{1}{\alpha_p} \rho_p(m, 1)\{1 - e^{\alpha_p T}\}, \tag{7b}$$

$$P_{mm}(T) = P(T \leq \tau_{m-1}) = e^{-\alpha_m T}. \tag{7c}$$

Equations (7a)–(7c) are equivalent to those of GRIFFITHS (1980), TAVARÉ (1984) and WATTERSON'S (1984) when mutation is absent. In particular, we have

$$P_{21}(T) = 1 - e^{-T}, \qquad P_{22}(T) = e^{-T} \tag{8}$$

for $m = 2$, and

$$P_{31}(T) = 1 - \frac{3}{2} e^{-T} + \frac{1}{2} e^{-3T},$$

$$P_{32}(T) = \frac{3}{2} (e^{-T} - e^{-3T}), \qquad P_{33}(T) = e^{-3T} \tag{9}$$

for $m = 3$. Some numerical values of (7a)–(9) are given in Table 1. It is noted that m_0 decreases quite rapidly as T increases.

We are now in a position to compute the probability, $P_l(T)$, that m genes sampled from population X and n genes sampled from population Y are descended from l genes at the time of population splitting (T units of time ago). Obviously, the probability, $P_{nn_0}(T)$, that the genes sampled from population Y are descended from n_0 genes T units of time ago is given by (7a)–(7c), if we replace m and m_0 by n and n_0, respectively. Therefore, we have

TABLE 2

Probabilities, $P_l(T)$, *that* m *genes sampled from population* X *and* n *genes sampled from population* Y *are descended from* l *ancestral genes at the time of population splitting* (2NT *generations ago*)

			l				
T	m	n	2	3	4	5	$6 \leq l \leq m + n$
0.5	2	2	0.1548	0.4773	0.3679		
	5	5	0.0066	0.0652	0.2266	0.3410	0.3606
	2	10	0.0098	0.0956	0.2836	0.3554	0.2556
	10	10	0.0006	0.0102	0.0620	0.1797	0.7475
1.0	2	2	0.3996	0.4651	0.1353		
	5	5	0.1116	0.3540	0.3675	0.1418	0.0251
	2	10	0.1440	0.4161	0.3338	0.0967	0.0094
	10	10	0.0519	0.2395	0.3775	0.2443	0.0868
1.5	2	2	0.6035	0.3467	0.0498		
	5	5	0.3244	0.4545	0.1947	0.0252	0.0012
	2	10	0.3747	0.4637	0.1474	0.0139	0.0003
	10	10	0.2327	0.4421	0.2661	0.0544	0.0047
2.0	2	2	0.7476	0.2340	0.0184		
	5	5	0.5371	0.3812	0.0780	0.0037	0.0000
	2	10	0.5833	0.3611	0.0538	0.0018	0.0000
	10	10	0.4550	0.4209	0.1154	0.0084	0.0003

$$P_l(T) = \sum_{s=1}^{l-1} P_{ns}(T) P_{m,l-s}(T) \tag{10}$$

where $l = m_0 + n_0 \geq 2$. Note that, when only one gene is sampled from one of the two populations, say X, $P_l(T)$ becomes identical with $P_{nn_0}(T)$. In (10), the probability that the genes in the two populations are derived from the same genes at exactly T units of time ago is neglected, because this probability can be shown to be very small. Table 2 gives some numerical values of (10). It is seen that, when T is small and m and n are relatively large, l is expected to be quite large, but as T increases, the expected value of l declines.

Let us now consider the time of divergence (or coalescence) of all genes sampled from populations X and Y. KINGMAN (1982) and TAJIMA (1983) studied the time of coalescence of genes sampled from a single population, and their theory applies to the l genes at the time of population splitting. That is, the expected time of coalescence of l genes *prior to population splitting* is given by

$$E(\tau_{\max}) = 2(1 - 1/l) \tag{11a}$$

(KINGMAN 1982; TAJIMA 1983), whereas the variance is

$$V(\tau_{\max}) = \sum_{i=1}^{l-1} \left\{ \frac{2}{i(i+1)} \right\}^2 \tag{11b}$$

TABLE 3

Expected minimum and maximum divergence times of genes prior to population splitting when m *genes are sampled from population* X *and* n *genes are sampled from population* Y

		n			
T	*m*	2	3	5	10
0.1		0.20	0.13	0.07	0.032
	2	1.47	0.09	0.05	0.026
	3	1.56	1.63	0.04	0.019
	5	1.66	1.70	1.76	0.012
	10	1.77	1.79	1.82	1.850
1.0		0.58	0.49	0.42	0.35
	2	1.22	0.43	0.36	0.30
	3	1.27	1.32	0.31	0.26
	5	1.32	1.36	1.40	0.23
	10	1.37	1.41	1.44	1.47
2.0		0.83	0.79	0.75	0.71
	2	1.09	0.75	0.71	0.68
	3	1.11	1.13	0.68	0.65
	5	1.13	1.15	1.17	0.62
	10	1.15	1.17	1.19	1.20

The figures above the diagonal refer to the minimum divergence times, whereas those below the diagonal refer to the maximum divergence times (times of coalescence). Time is measured in $2N$ generations.

(TAJIMA 1983). Here, the time is still measured in terms of $2N$ generations, *i.e.*, $\tau = t/(2N)$. If time is measured in generations, (11a) and (11b) must be multiplied by $2N$ and $(2N)^2$, respectively. $E(\tau_{max})$ is identical with the expected time of divergence (prior to population splitting) of the two genes which are most distantly related (*e.g.*, a and e or f in Figure 1). On the other hand, the mean and variance of the time of divergence of the two most closely related genes (*e.g.*, d and b or c in Figure 1) prior to population splitting are given by

$$E(\tau_{min}) = \frac{2}{l(l-1)}, \qquad V(\tau_{min}) = \left\{\frac{2}{l(l-1)}\right\}^2 \tag{12}$$

(LITTLER 1975; GRIFFITHS 1980; TAJIMA 1983).

In the above, we treated the number of ancestral genes as a constant. Actually, this is a random variable so that we have to take the expectation with respect to the distribution $P_l(T)$. For example, when $m = n = 2$, the distribution is given by $P_2(T) = (1 - e^{-T})^2$, $P_3(T) = 2e^{-T}(1 - e^{-T})$ and $P_4(T) = e^{-2T}$ from (8) and (10). Table 3 gives numerical values of the unconditional means of τ_{min} and τ_{max} when multiple samples are involved. When T is small, $E(\tau_{min})$ decreases substantially with increasing m and n. Therefore, the divergence time between the two closest genes can be used as an approximate estimate of the

time of population splitting. However, this estimate gradually becomes inaccurate as T increases, and when T is as large as 2.0, the divergence time between the two closest genes is substantially larger than the time of population splitting. It is also noted that $E(\tau_{max})$ is much larger than $E(\tau_{min})$ when T is small, but the difference is gradually diminished as T increases. Both $E(\tau_{min})$ and $E(\tau_{max})$ approach 1 with increasing T regardless of m and n, since m_0 and n_0 become 1 and, therefore, $l = 2$ because of coalescence.

VARIANCE OF THE NUMBER OF NUCLEOTIDE DIFFERENCES

NEI and LI's (1979) estimate of the number of net nucleotide substitutions between two populations is given by

$$d = d_{XY} - \frac{1}{2}(d_X + d_Y), \tag{13}$$

where d_{XY} is the mean number of nucleotide substitutions between genes from populations X and Y, and d_X and d_Y are the number of nucleotide differences (substitutions) between two randomly chosen genes within populations X and Y, respectively. In actual data analysis, d_{XY}, d_X and d_Y are estimated from the proportion of nucleotide differences or restriction site data (J. C. STEPHENS and M. NEI, unpublished results). In the present case, however, we are considering the infinite-site model so that the number of nucleotide substitutions between a pair of genes is identical with the number of nucleotide differences.

Suppose that m and n genes are sampled from populations X and Y, respectively, and let k_{ij} be the number of nucleotide differences between the ith gene from population X and the jth gene from population Y. d_{XY} in (13) is then given by

$$d_{XY} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} k_{ij}. \tag{14}$$

On the other hand, d_X is given by

$$d_X = \frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{i'\neq i}^{m} k_{ii'}, \tag{15}$$

where $k_{ii'}$ is the number of nucleotide differences between the ith and i'th genes sampled from population X. d_Y can be obtained in the same way. The estimates of d_{XY}, d_X and d_Y are obtained from (14) and (15) by using the observed values of k_{ij} and $k_{ii'}$.

The expectation of d_{XY} is given by $M + MT$, where $M = 4Nv$ (LI 1977; GILLESPIE and LANGLEY 1979) and that of d_X or d_Y by M regardless of sample size (KIMURA 1969; WATTERSON 1975; NEI and TAJIMA 1981). Therefore, the expectation of d in (13) is $E(d) = MT = 2vt$. To derive the variance of d in (13), we first note that $V(d)$ can be written as

$$V(d) = \frac{1}{4}[V(d_X) + V(d_Y)] + V(d_{XY}) + \frac{1}{2}\mathrm{Cov}(d_X, d_Y) - \mathrm{Cov}(d_{XY}, d_X) - \mathrm{Cov}(d_{XY}, d_Y), \tag{16}$$

where $V(\cdot)$ and $\mathrm{Cov}(\cdot, \cdot)$ refer to the variance and covariance of the quantities concerned. Thus, $V(d)$ can be obtained by evaluating all terms on the right-hand side of (16).

TAJIMA (1983) derived $V(d_X)$, which is given by

$$V(d_X) = \frac{m+1}{3(m-1)} M + \frac{2(m^2+m+3)}{9m(m-1)} M^2. \tag{17}$$

The variance of d_Y is also given by (17) if we replace m by n.

To calculate the remaining quantities in (16), we need to evaluate $E(d_{XY}^2)$, $E(d_X d_Y)$, $E(d_{XY} d_X)$ and $E(d_{XY} d_Y)$, which can be written as

$$E(d_{XY}^2) = \frac{1}{mn}[E(k_{ij}^2) + (m+n-2)E(k_{ij}k_{ij'}) + (m-1)(n-1)E(k_{ij}k_{i'j'})], \tag{18a}$$

$$E(d_X d_Y) = E(k_{ii'}k_{jj'}), \tag{18b}$$

$$E(d_{XY} d_X) = \frac{2}{m} E(k_{ii'}k_{ij}) + \frac{m-2}{m} E(k_{ii'}k_{i''j}), \tag{18c}$$

$$E(d_{XY} d_Y) = \frac{2}{n} E(k_{jj'}k_{ji}) + \frac{n-2}{n} E(k_{jj'}k_{j''i}), \tag{18d}$$

where subscripts i and j stand for the ith gene from population X and the jth gene from population Y, respectively. Obviously, $E(k_{ij}^2) = E(k_{i'j'}^2)$, $E(k_{ij}k_{ij'}) = E(k_{ij}k_{i'j})$, etc. In the evaluation of (18a)–(18d), it is necessary to consider the genealogical relationship of the genes sampled. For example, in the case of one gene sampled from population X and two genes sampled from population Y, there are four possible genealogical relationships as shown in Figure 3. Therefore, we must consider all of these possibilities in evaluating $E(k_{ij}^2)$, $E(k_{ij}k_{ij'})$, etc.

We note that in the infinite-site model, mutations accumulate in a nucleotide sequence following the Poisson distribution. Therefore, the mean and variance of the number of accumulated mutations (x) for a given time period τ are both $M\tau/2 = vt$. To facilitate the computation of $E(k_{ij}^2)$, $E(k_{ij}k_{ij'})$, etc., we introduce the following random variable

$$\xi = x - M\tau/2. \tag{19}$$

Obviously, $E(\xi) = 0$, and $E(\xi^2) = M\tau/2$. We use ξ for each evolutionary time which should be considered separately.

The first term $[E(k_{ij}^2)]$ on the right-hand side of (18a) refers to the case in which two randomly chosen genes, one from each of X and Y, are compared. In this case, we consider four ξ's, *i.e.*, ξ_1, ξ_2, ξ_3 and ξ_4 as shown in Figure 2, where T is the time of population splitting and $T + \tau_1'$ is the time when gene C from population X and gene D from population Y diverged. The number of nucleotide differences (k_{ij}) between C and D is then given by

$$k_{ij} = k_{OC} + k_{OD} = M(T + \tau_1') + \sum_{i=1}^{4} \xi_i, \tag{20}$$

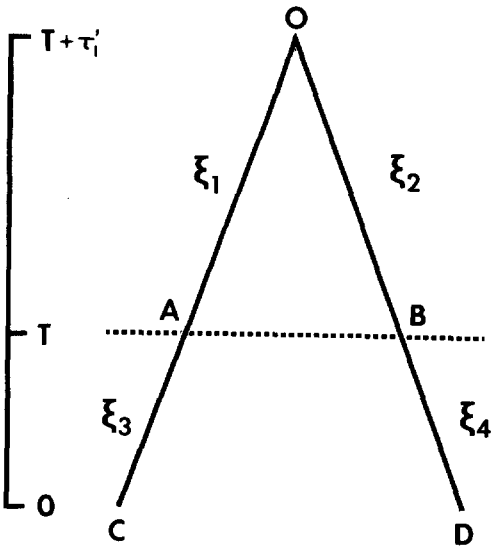

FIGURE 2.—Two genes (C and D) derived from the ancestral gene O that existed at time $T + \tau_1'$. T and the dotted line represent the time of population splitting. ξ_1, ξ_2, ξ_3 and ξ_4 are random variables representing the deviation of the number of nucleotide substitutions from the expected number.

where k_{OC} and k_{OD} are the numbers of nucleotide differences between O and C and O and D in Figure 2, respectively. Obviously, the expectations of k_{ij} and k_{ij}^2 conditional on τ_1' are $E_c(k_{ij}) = M(T + \tau_1')$, and $E_c(k_{ij}^2) = (T + \tau_1')M + (T + \tau_1')^2M^2$, respectively, because $\sum_{i=1}^{4} E(\xi_i^2) = (T + \tau_1')M$. If we note that τ_1' follows the distribution of $f_1(\tau_1') = \exp(-\tau_1')$ in (2), the unconditional expectations become

$$E(k_{ij}) = (T + 1)M, \tag{21a}$$

$$E(k_{ij}^2) = (T + 1)M + (T^2 + 2T + 2)M^2, \tag{21b}$$

and, thus, the variance is

$$V(k_{ij}) = (T + 1)M + M^2. \tag{22}$$

This agrees with the result of LI (1977), who used the generating function method.

We have already obtained the expectation of k_{ij}^2 in the first term of (18a). The second term can be computed in a similar way. In this case, however, we have to consider four different types of gene genealogies as given in Figure 3. Type (a) occurs when the divergence time (T_1) between genes j and j' from population Y is shorter than the time (T) of population splitting. This probability is given by $1 - e^{-T}$ [equation (8)]. Therefore, with the condition of $\tau_1 \leq T$, we have

$$\begin{aligned} E(k_{ij}) &= E(k_{ij'}) = (T + 1)M, \\ E(k_{ij}k_{ij'}) &= (T + 1 - F)M + (T^2 + 2T + 2)M^2, \\ \mathrm{Cov}(k_{ij}, k_{ij'}) &= (T + 1 - F)M + M^2, \end{aligned} \tag{23}$$

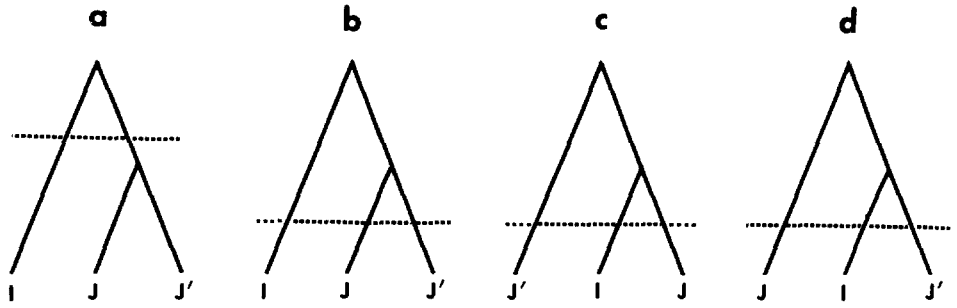

FIGURE 3.—Four different types of gene genealogies possible for three sampled genes. Gene I is from population X, and genes J and J' from population Y. The dotted line represents the time of population splitting.

where $2F = \int_0^T \tau_1 e^{-\tau_1} d\tau_1 = 1 - (T + 1)e^{-T}$. The remaining three types occur when $\tau_1 > T$, and each has a probability of $e^{-T}/3$. Therefore, we obtain

$$E(k_{ij}) = E(k_{ij'}) = \left(T + \frac{4}{3}\right)M$$
$$E(k_{ij}k_{ij'}) = \left(\frac{1}{2}T + \frac{7}{6}\right)M + \left(T^2 + \frac{8}{3}T + \frac{26}{9}\right)M^2 \tag{24}$$

for type (b). We also have

$$E(k_{ij}) = \left(T + \frac{1}{3}\right)M,\ E(k_{ij'}) = \left(T + \frac{4}{3}\right)M$$
$$E(k_{ij}k_{ij'}) = \left(\frac{1}{2}T + \frac{1}{6}\right)M + \left(T^2 + \frac{5}{3}T + \frac{5}{9}\right)M^2 \tag{25}$$

for type (c) and

$$E(k_{ij}) = \left(T + \frac{4}{3}\right)M,\ E(k_{ij'}) = \left(T + \frac{1}{3}\right)M$$
$$E(k_{ij}k_{ij'}) = \left(\frac{1}{2}T + \frac{1}{6}\right)M + \left(T^2 + \frac{5}{3} + \frac{5}{9}\right)M^2 \tag{26}$$

for type (d). $\mathrm{Cov}(k_{ij}, k_{ij'})$ can be obtained by evaluating the mean of $E(k_{ij})$, $E(k_{ij'})$ and $E(k_{ij}k_{ij'})$ over all four types. It becomes

$$\mathrm{Cov}(k_{ij}, k_{ij'}) = (1 - e^{-T})\{(T + 1 - F)M + M^2\} + e^{-T}\left\{\frac{1}{2}(T + 1)M + \frac{1}{3}M^2\right\}.$$

In the case of four genes, we have to consider 14 different types of gene genealogies if all the genes are to be distinguished. However, if we do not distinguish between two genes sampled from the same population, there are seven different types (Figure 4). Table 4 summarizes the probabilities of occurrence of the seven types. The formulas for $E(k_{ij})$, $E(k_{i'j'})$ and $E(k_{ij}k_{i'j'})$ for each type of genealogy are given in the APPENDIX. The final formula for $\mathrm{Cov}(k_{ij}k_{i'j'})$ becomes

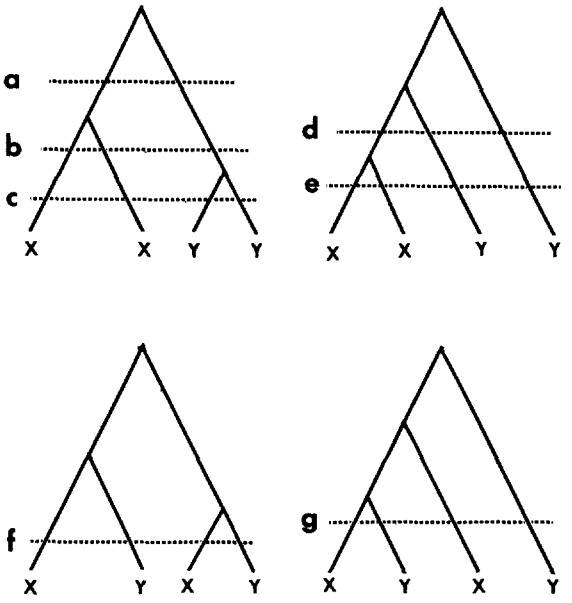

FIGURE 4.—Seven different types of gene genealogies possible for four sampled genes, two from each of populations *X* and *Y*. The dotted line represents the time of population splitting.

TABLE 4

Probabilities of seven different types of gene genealogies for two genes from each of populations X *and* Y *in relation to the time of population splitting*

Genealogy class	Genealogy type	Probability of class	Probability within class
I	(a)	$(1 - e^{-T})^2$	1
II	(b)	$2e^{-T}(1 - e^{-T})$	1/3
II	(d)	$2e^{-T}(1 - e^{-T})$	2/3
III	(c)	e^{-2T}	1/9
III	(e)	e^{-2T}	2/9
III	(f)	e^{-2T}	2/9
III	(g)	e^{-2T}	4/9

See Figure 4. Genealogy class I, two gene divergences occur after population splitting; genealogy class II, one gene divergence occurs before population splitting and the other occurs after population splitting; genealogy class III, two gene divergences occur before population splitting. These probabilities are identical with those given by TAJIMA (1983).

$$\mathrm{Cov}(k_{ij}, k_{i'j'}) = (1 - e^{-T})^2\{(T + 1 - 2F)M + M^2\} + 2e^{-T}(1 - e^{-T})\left\{\left(\frac{1}{2}T + \frac{1}{2} - F\right)M + \frac{1}{3}M^2\right\} + e^{-2T}\left\{\frac{1}{3}M + \frac{2}{9}M^2\right\}. \quad (28)$$

We are now in a position to compute $V(d_{XY}) = E(d_{XY}^2) - E^2(d_{XY})$. It becomes

$$V(d_{XY}) = \frac{1}{mn}\Bigg[(T+1)M+M^2+(m+n-2)\Bigg\{(1-e^{-T})((T+1-F)M+M^2) + e^{-T}\left(\frac{1}{2}(T+1)M+\frac{1}{3}M^2\right)\Bigg\}+(m-1)(n-1)\Bigg\{(1-e^{-T})^2 \times((T+1-2F)M+M^2)+2e^{-T}(1-e^{-T}) \cdot\left(\left(\frac{1}{2}T+\frac{1}{2}-F\right)M+\frac{1}{3}M^2\right)+e^{-2T}\left(\frac{1}{3}M+\frac{2}{9}M^2\right)\Bigg\}\Bigg]. \tag{29}$$

Before going further, let us examine a few properties of this variance. We note that, when $m = n = 1$, (29) reduces to (22), whereas, when $m, n \gg 1$, (29) becomes

$$V(d_{XY}) = (1 - e^{-T})^2\{(T + 1 - 2F)M + M^2\} + 2e^{-T}(1 - e^{-T}) \times \left\{\left(\frac{1}{2}T + \tfrac{1}{2} - F\right)M + \frac{1}{3}M^2\right\} + e^{-2T}\left\{\frac{1}{3}M + \frac{2}{9}M^2\right\}. \tag{30}$$

Furthermore, when $T \ll 1$, equation (30) becomes

$$V(d_{XY}) = \frac{1}{3}M + \frac{2}{9}M^2, \tag{31a}$$

and, when $T \gg 1$,

$$V(d_{XY}) = MT + M^2. \tag{31b}$$

Equation (31a) is identical with (17) when m is large, whereas (31b) is smaller than (22) by M when $m = n = 1$. Thus, multiple sampling of genes does reduce the variance for any T, but the extent of the reduction is small when T is large.

Using formulas (25), (26) and those in the APPENDIX, we can also derive the formula of $V(d_X)$. As noted by TAJIMA (1983), $V(d_X)$ can be written as

$$V(d_X) = \frac{1}{m(m-1)}[2V(k_{ij}) + 4(m-2)\mathrm{Cov}(k_{ij}, k_{ij'}) + (m-2)(m-3)\mathrm{Cov}(k_{ij}, k_{i'j'})].$$

The first and second terms of the right-hand side are given by (22) and (27) with $T = 0$, respectively. The third term is

$$\mathrm{Cov}(k_{ij}, k_{i'j'}) = \frac{1}{3}M + \frac{2}{9}M^2$$

from the required moments for genealogy types (c), (e), (f) and (g) given in the APPENDIX and letting $T = 0$. Substitution of these formulas into $V(d_X)$ gives (17).

The same procedure as that for computing $V(d_{XY})$ can be used to obtain the

covariances in (16). We first consider $\mathrm{Cov}(d_X, d_Y)$ and then $\mathrm{Cov}(d_{XY}, d_X)$, which is the same as $\mathrm{Cov}(d_{XY}, d_Y)$. The expectation of the product of d_X and d_Y is independent of m and n values. This is because $\mathrm{Cov}(d_X, d_Y)$ is generated only when four different genes are involved and any permutation of either i and i' or j and j' or both does not affect the value of $E(k_{ii'}k_{jj'})$. Putting this another way, there is no sampling covariance for this quantity (NEI and TAJIMA 1981). Therefore, it suffices to consider only the genealogies given in Figure 4. We present only the final result:

$$\mathrm{Cov}(d_X, d_Y) = e^{-2T}\left(\frac{1}{3}M + \frac{2}{9}M^2\right). \tag{32}$$

Equation (32) indicates that the covariance equals (31a) for $T = 0$, and it disappears when T is large.

The computation of $\mathrm{Cov}(d_{XY}, d_X)$ is as cumbersome as that of $V(d_{XY})$. We use Figure 3 with $i' = j'$ to obtain the first term of the right-hand side of (18c). We then have

$$\mathrm{Cov}(k_{ii'}, k_{ij}) = (1 - e^{-T})MF + e^{-T}\left\{\frac{1}{2}(T + 1)M + \frac{1}{3}M^2\right\}. \tag{33}$$

Next, we consider $\mathrm{Cov}(k_{ii'}, k_{i''j})$ corresponding to the case in which one gene is sampled from population Y and three genes are sampled from population X. The possible genealogical relationships of the four genes sampled in relation to the population divergence time are presented in Figure 5, and the probability of each type of genealogy is given in Table 5. This probability was obtained by using (9).

The values of $E(k_{ii'}k_{i''j})$, $E(k_{ii'})$ and $E(k_{i''j})$ for each type of genealogy of Figure 5 are given in the APPENDIX. Using these values and the probabilities of different types of genealogies in Table 5, we finally obtain

$$\begin{aligned}\mathrm{Cov}(d_{XY}, d_X) = {} & \frac{2}{m}\left[(1 - e^{-T})MF + e^{-T}\left\{\frac{1}{2}(T + 1)M + \frac{1}{3}M^2\right\}\right] \\ & + \frac{m - 2}{m}\left[\frac{1}{3}\left(1 - \frac{3}{2}e^{-T} + \frac{1}{2}e^{-3T}\right)MS_1 + \frac{3}{2}(e^{-T} - e^{-3T})\right. \\ & \times \left\{\frac{1}{3}(T + 1 - S_2)M + \frac{2}{9}M^2\right\} \\ & \left. + e^{-3T}\left(\frac{1}{3}M + \frac{2}{9}M^2\right)\right],\end{aligned} \tag{34}$$

where $S_1 = 1 - (\frac{3}{2})(1 + T)e^{-T} + (\frac{1}{2})(1 + 3T)e^{-3T}$ and $S_2 = (\frac{1}{5})\{1 - (1 + 3T)e^{-3T}\}$.

$\mathrm{Cov}(d_{XY}, d_Y)$ can be obtained by replacing m in (34) by n. Substituting (17),

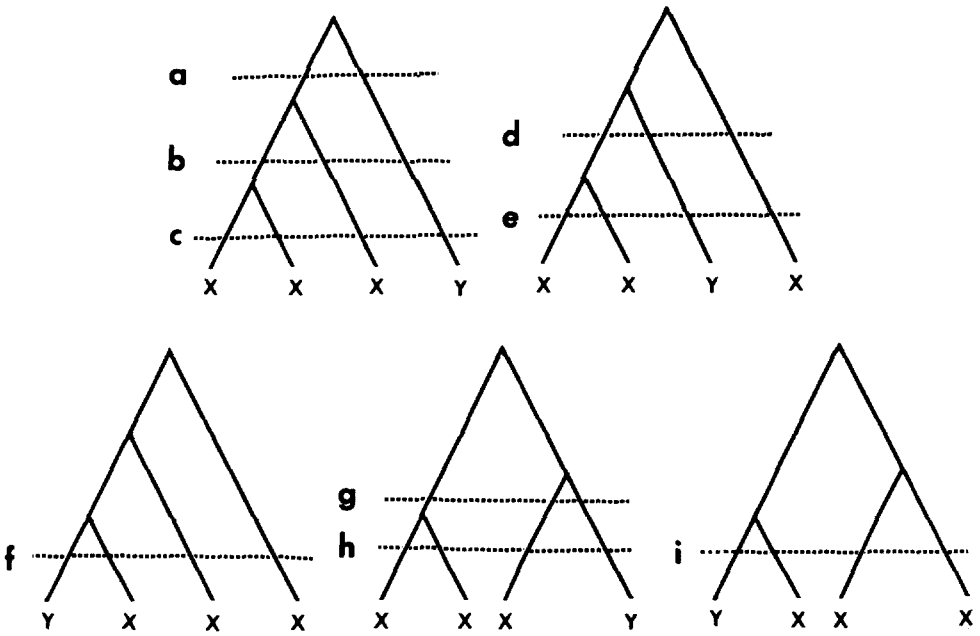

FIGURE 5.—Nine different types of gene genealogies possible for four sampled genes, three from population X and one from population Y. The dotted line represents the time of population splitting.

(29), (32), (34) and $\mathrm{Cov}(d_{XY}, d_Y)$ into (16), we have the formula of the variance of d. The general formula is quite complicated, but, when $T \ll 1$, it becomes

$$V(d) = \frac{(m + n - 1)(m + n - 2)}{6mn(m - 1)(n - 1)} \left(M + \frac{5}{3} M^2\right). \tag{35}$$

On the other hand, when $T \gg 1$,

$$V(d) = MT + M^2.$$

Furthermore, when $m, n \gg 1$, it becomes

$$\begin{aligned} V(d) = {} & \frac{1}{2}\left(\frac{1}{3}M+\frac{2}{9}M^2\right)+(1-e^{-T})^2\{(T+1-2F)M+M^2\}+2e^{-T}(1-e^{-T}) \\ & \times\left\{\frac{1}{2}(T+1-2F)M+\frac{1}{3}M^2\right\}+\frac{3}{2}e^{-2T}\left(\frac{1}{3}M+\frac{2}{9}M^2\right) \\ & -2\left[\frac{1}{3}\left(1-\frac{3}{2}e^{-T}+\frac{1}{2}e^{-3T}\right)MS_1+\frac{3}{2}(e^{-T}-e^{-3T})\left\{\frac{1}{3}(T+1-S_2)M\right.\right. \\ & \left.\left.+\frac{2}{9}M^2\right\}+e^{-3T}\left(\frac{1}{3}M+\frac{2}{9}M^2\right)\right] \end{aligned} \tag{36}$$

for any value of T. When T is small, this becomes 0 as expected.

TABLE 5

Probabilities of nine different types of gene genealogies for three genes from population X *and one gene from population* Y *in relation to the time of population splitting*

Genealogy class	Genealogy type	Probability of class	Probability within class
I	(a)	$1 - \frac{3}{2}e^{-T} + \frac{1}{2}e^{-3T}$	1
II	(b)	$\frac{3}{2}(e^{-T} - e^{-3T})$	1/3
II	(d)	$\frac{3}{2}(e^{-T} - e^{-3T})$	1/3
II	(g)	$\frac{3}{2}(e^{-T} - e^{-3T})$	1/3
III	(c)	e^{-3T}	1/6
III	(e)	e^{-3T}	1/6
III	(f)	e^{-3T}	1/3
III	(h)	e^{-3T}	1/6
III	(i)	e^{-3T}	1/6

See Figure 5. Genealogy class I, three X genes split after population splitting; genealogy class II, one X gene splitting occurs before population splitting and the other X gene splitting occurs after population splitting; genealogy class III, both X gene splittings occur before population splitting regardless of Y gene splitting.

Table 6 gives the standard errors of d and d_{XY} for various values of m, n, M and T. It is seen that when T [or $E(d)$] is small, the standard error (s_d) of d is very large relative to its mean [$E(d)$] unless the sample size is large. Namely, in this case, a large sample size is required for obtaining a reliable estimate of $E(d)$. This large standard error is caused by the variation of $(d_X + d_Y)/2$ in (13). As T increases, however, the ratio of s_d to $E(d)$ declines rapidly. When $M = 10$, $E(d) = 10$ and $m = n = 2$, s_d is of the same order of magnitude as $E(d)$. When T is large, s_d does not decrease very much with increasing sample size.

Properties of the standard error (s_{dXY}) of d_{XY} are somewhat different from those of s_d. When sample size is small, s_{dXY} is slightly larger than s_d for $T \leq 1$, but since s_{dXY} does not decrease appreciably with increasing sample size, the difference between s_d and s_{dXY} for a large sample size is substantial. The reason for the larger standard error of d_{XY} is that d_{XY} includes the nucleotide differences that existed in the ancestral population (effects of polymorphism), and the variance of the number of nucleotide differences between two randomly chosen genes from a population is not reduced very much by increasing sample size, as shown by TAJIMA (1983). When M is large and $T = 10$, s_{dXY} is slightly smaller than s_d. This is because, when T is large, the covariance between d_{XY} and $(d_X + d_Y)/2$ in (34) is reduced subtantially.

It should be noted that $E(d)$, s_d and s_{dXY} in Table 6 refer to the number of nucleotide (or amino acid) differences per gene. If one is interested in the number of nucleotide differences per nucleotide site, their values should be divided by the total number of nucleotides involved. For example, if the gene studied consists of 1000 nucleotides, M, $E(d)$, s_d and s_{dXY} should all be divided by 1000.

TABLE 6

Standard errors of d *and* d_{XY}

M	T	$E(d)$	m = 2	m = 10	m = 10^2	m = 10^3
100	0.01	1.0	65.4 (69.4)	11.8 (50.7)	2.15 (48.0)	1.25 (47.8)
	0.1	10.0	70.8 (71.9)	20.5 (53.6)	11.8 (50.5)	11.0 (50.2)
	1.0	100.0	103 (88.8)	73.8 (79.5)	70.0 (77.4)	69.6 (77.2)
	10.0	1000.0	126 (105)	111 (105)	110 (105)	110 (105)
10	0.01	0.1	6.71 (7.32)	1.22 (5.39)	0.22 (5.11)	0.13 (5.08)
	0.1	1.0	7.28 (7.60)	2.11 (5.70)	1.22 (5.38)	1.14 (5.35)
	1.0	10.0	10.8 (9.64)	7.86 (8.61)	7.47 (8.39)	7.43 (8.36)
	10.0	100.0	15.8 (14.3)	14.5 (14.2)	14.4 (14.1)	14.4 (14.1)
1	0.01	0.01	0.83 (1.03)	0.15 (0.79)	0.03 (0.75)	0.02 (0.75)
	0.1	0.1	0.90 (1.09)	0.26 (0.84)	0.15 (0.80)	0.14 (0.79)
	1.0	1.0	1.49 (1.53)	1.16 (1.36)	1.11 (1.32)	1.11 (1.31)
	10.0	10.0	3.39 (3.39)	3.27 (3.33)	3.26 (3.32)	3.26 (3.32)
0.1	0.01	0.001	0.17 (0.25)	0.03 (0.20)	0.01 (0.19)	0.00[a] (0.19)
	0.1	0.01	0.19 (0.27)	0.06 (0.21)	0.03 (0.20)	0.03 (0.20)
	1.0	0.1	0.36 (0.41)	0.29 (0.36)	0.28 (0.35)	0.28 (0.35)
	10.0	1.0	1.01 (1.03)	0.98 (1.01)	0.98 (1.01)	0.98 (1.01)

The standard errors of d_{XY} are given in parentheses. $M = 4Nv$, $T = t/(2N)$ and $m = n$ have been used. $E(d_{XY}) = E(d) + M$.

[a] The actual value is less than 0.001.

DISCUSSION

We have seen that, when the time (T) since divergence between two populations is relatively short and the numbers of genes (m and n) sampled from the two populations are relatively large, the expected number of ancestral genes (l) at the time of population splitting is quite large (Table 1). This number gradually declines as T increases, but even for $T = 2$, *i.e.*, $4N$ generations, the probability of l larger than 2 is 0.545. NEI and ROYCHOUDHURY (1974) have estimated that the Negroid and Mongoloid populations of man diverged about 120,000 yr ago. If the average effective size for the populations and the generation time in the past have been $N = 10{,}000$ and 25 yr, respectively, a period of 120,000 yr corresponds to 4800 generations or $T = 0.24$. Therefore, if we sample five genes from each of the two populations, the probability of $l \geq 5$ is 0.857 from equation (10). Thus, the genes from the Negroid and Mongoloid populations are expected to share many common ancestral genes. That this is indeed the case has been subtantiated by the data of CANN, BROWN and WILSON (1982) on the genealogical relationship of mitochondrial DNA sequences.

J. C. STEPHENS and M. NEI (unpublished results) analyzed the nucleotide sequences of the alcohol dehydrogenase genes from *Drosophila melanogaster* and *D. simulans* (data from KREITMAN 1983; COHN, THOMPSON and MOORE

1984; BODMER and ASHBURNER 1984). Using a rate of nucleotide substitution of 4×10^{-9} per site per year (LI, LUO and WU 1984), they estimated that the time since divergence between the two species is about 2 million yr and the average effective population size (N) is about 2×10^6. A period of two million yr corresponds to 12 million generations or $T = 3$ if there are six generations per year in nature. In this case, the probability of $l = 2$ is as high as 0.77 even for $m = n = 10$ (see also Table 2). This prediction is supported by the genealogical relationship of 11 genes from *D. melanogaster* and two genes from *D. simulans*. In this case, the number of ancestral genes (l) at the time of population splitting has been estimated to be two (J. C. STEPHENS and M. NEI, unpublished data).

When the number of ancestral genes at the time of population splitting is large, one can compute the maximum and minimum times of divergence of genes from information on nucleotide sequences. The minimum time can be used as an estimate of the time of population splitting when the latter is not known. However, this minimum time is expected to be a serious overestimate when T is large. In practice, of course, we usually do not know whether T is large or not, but if the number of common ancestral genes shared by the two populations is small, T is expected to be large. Therefore, in this case, this method should not be used. It should also be noted that the estimate has a large stochastic variance, the standard error being the same as the mean [equation (12)].

In general, a more reliable estimate of the time of population splitting is obtained from (13), provided the rate of nucleotide substitution (or mutation rate) is known. As long as the rate of nucleotide substitution is constant and the populations are in equilibrium with respect to mutation and genetic drift, (13) gives an unbiased estimate. However, d is also subject to a large variance when d is small.

J. C. STEPHENS and M. NEI (unpublished data) compared the nucleotide sequences (807 nucleotides long) of the alcohol dehydrogenase genes from *D. melanogaster*, *D. simulans* and *D. mauritiana* and estimated the d values. In the comparison of *D. melanogaster* (X) and *D. simulans* (Y), $m = 11$ and $n = 2$, and they obtained $\hat{d}_{XY} = 20.32$, $\hat{d}_X = 5.75$, $\hat{d}_Y = 7$, $\hat{d} = 13.94$ and $\hat{T} = 2.18$, where $\hat{}$ refers to an estimate. We can, therefore, estimate M by $(\hat{d}_X + \hat{d}_Y)/2$, which is 6.38. Using this estimate together with $m = 11$, $n = 2$ and $\hat{T} = 2.18$, we can compute the expected standard error of d by using (16). It becomes 7.74. Therefore, s_d is about half of $\hat{d}$. Similarly, we can estimate s_d of $\hat{d}$ for the comparison of *D. melanogaster* and *D. mauritiana* ($n = 2$; $\hat{d}_Y = 6$). In this case, $\hat{d}$ and $\hat{s}_d$ become 17.94 and 7.75, respectively. In the case of comparison of *D. simulans* and *D. mauritiana*, however, we obtained $\hat{d} = 6.0 \pm 7.1$. That is, $\hat{s}_d$ was as large as $\hat{d}$.

The large value of s_d relative to d for the *D. simulans* to *D. mauritiana* comparison is partly due to the small sample sizes ($m = n = 2$) used for these two species. However, equation (16) indicates that, even if $m = n = 1000$, s_d is reduced only slightly and becomes 4.8. This indicates that a large part of the variance of d is due to stochastic factors. It should be noted that the

stochastic variance can be reduced only by using sequence data from many independent (unlinked) gene loci.

We thank CLAY STEPHENS for his help in computing the numerical values in Table 6 and for his comments on the manuscript. We also thank NARUYA SAITOU, SIMON TAVARÉ and GEOFF WATTERSON for their comments. This work was supported by research grants from the National Institutes of Health and the National Science Foundation.

LITERATURE CITED

BODMER, M. and M. ASHBURNER, 1984 Conservation and change in the DNA sequences coding for alcohol dehydrogenase in sibling species of *Drosophila*. Nature **309:** 425–430.

CANN, R. L., W. M. BROWN and A. C. WILSON, 1982 Evolution of human mitochondrial DNA: a preliminary report. pp. 157–164. In: *Human Genetics, part A: The Unfolding Genome*, Edited by B. BONNÉ-TAMIR, P. COHEN and R. N. GOODMAN. Alan R. Liss, New York.

COHN, V. H., M. A. THOMPSON and G. P. MOORE, 1984 Nucleotide sequence comparison of the Adh gene in three drosophilids. J. Mol. Evol. **20:** 31–37.

FITCH, W. M., 1976 Molecular evolutionary clocks. pp. 160–178. In: *Molecular Evolution*, Edited by F. J. AYALA. Sinauer Associates, Sunderland, Massachusetts.

GILLEPSIE, J. H. and C. H. LANGLEY, 1979 Are evolutionary rates really variable? J. Mol. Evol. **13:** 27–34.

GRIFFITHS, R. C., 1980 Lines of descent in the diffusion approximation of neutral Wright-Fisher models. Theor. Pop. Biol. **17:** 37–50.

HUDSON, R. R., 1983 Testing the constant rate neutral allele model with protein sequence data. Evolution **37:** 203–217.

KIMURA, M., 1969 The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics **61:** 893–903.

KIMURA, M., 1971 Theoretical foundations of population genetics at the molecular level. Theor. Pop. Biol. **2:** 174–208.

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

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

LI, W.-H., 1977 Distribution of nucleotide differences between two randomly chosen cistrons in a finite population. Genetics **85:** 331–337.

LI, W.-H., C.-C. LUO and C.-I. WU, 1984 Evolution of DNA sequences. In: *Molecular Evolutionary Genetics*, Edited by R. I. MACINTYRE. Plenum Press, New York. In press.

LITTLER, R. A., 1975 Loss of variability in a finite population. Math. Biosci. **25:** 151–163.

NEI, M. and W.-H. LI, 1979 Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. USA **76:** 5269–5273.

NEI, M. and A. K. ROYCHOUDHURY, 1974 Genic variation within and between the three major races of man, Caucasoids, Negroids, and Mongoloids. Am. J. Hum. Genet. **26:** 421–443.

NEI, M. and F. TAJIMA, 1981 DNA polymorphism detectable by restriction endonucleases. Genetics **97:** 145–163.

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

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

WATTERSON, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Pop. Biol. **7:** 256–276.

WATTERSON, G. A., 1984 Lines of descent and the coalescent. Theor. Pop. Biol. **26:** 77–92.

ZUCKERKANDL, E. and L. PAULING, 1965 Evolutionary divergence and convergence in proteins. pp. 97–166. In: *Evolving Genes and Proteins*, Edited by V. BRYSON and H. J. VOGEL. Academic Press, New York.

Communicating editor: B. S. WEIR

APPENDIX

According to the classification of gene genealogies in Figure 4, we present the formulas of $E(k_{ij})$, $E(k_{i'j'})$ and $E(k_{ij}k_{i'j'})$ required for computing $V(d_{XY})$. Here, the arrow sign between $E(k_{ij})$ and $E(k_{i'j'})$ indicates that the two values occur with an equal probability for each of $E(k_{ij})$ and $E(k_{i'j'})$. The probability of each of the seven different types of genealogies in Figure 4 is given in Table 4. The formulas are:

Genealogy type	Probability	$E(k_{ij})$		$E(k_{i'j'})$	$E(k_{ij}k_{i'j'})$
(a)		$M(T+1)$		$M(T+1)$	$M(T+1-2F)+M^2(T^2+2T+2)$
(b)		$M(T+\frac{4}{3})$		$M(T+\frac{4}{3})$	$M(\frac{1}{2}T+\frac{7}{6}-F)+M^2(T^2+\frac{8}{3}T+\frac{26}{9})$
(c)		$M(T+\frac{3}{2})$		$M(T+\frac{3}{2})$	$\frac{7}{6}M+M^2(T^2+3T+\frac{1}{9})$
(d)		$M(T+\frac{1}{3})$	$\leftrightarrow$	$M(T+\frac{4}{3})$	$M(\frac{1}{2}T+\frac{1}{6}-F)+M^2(T^2+\frac{5}{3}T+\frac{5}{9})$
(e)		$M(T+\frac{1}{2})$	$\leftrightarrow$	$M(T+\frac{3}{2})$	$\frac{1}{6}M+M^2(T^2+2T+\frac{2}{9})$
(f)	$\frac{1}{2}$	$M(T+\frac{1}{2})$	$\leftrightarrow$	$M(T+\frac{1}{6})$	$M^2(T^2+\frac{2}{3}T+\frac{1}{9})$
	$\frac{1}{2}$	$M(T+\frac{3}{2})$		$M(T+\frac{3}{2})$	$\frac{7}{6}M+M^2(T^2+3T+\frac{2}{9})$
(g)	$\frac{1}{2}$	$M(T+\frac{1}{6})$	$\leftrightarrow$	$M(T+\frac{3}{2})$	$M^2(T^2+\frac{5}{3}T+\frac{5}{18})$
	$\frac{1}{2}$	$M(T+\frac{1}{2})$	$\leftrightarrow$	$M(T+\frac{3}{2})$	$\frac{1}{6}M+M^2(T^2+2T+\frac{4}{9})$

Formulas of $E(k_{ii'})$, $E(k_{jj'})$ and $E(k_{ii'}k_{jj'})$ required for computing $\mathrm{Cov}(d_X, d_Y)$ are:

Genealogy type	$E(k_{ii'})$		$E(k_{jj'})$	$E(k_{ii'}k_{jj'})$
(a)	$2MF$		$2MF$	$4M^2F^2$
(b)	$M(T+\frac{1}{3})$		$2MF$	$2M^2(T+\frac{1}{3})F$
(c)	$M(T+\frac{1}{2})$	$\leftrightarrow$	$M(T+\frac{1}{6})$	$M^2(T^2+\frac{2}{3}T+\frac{1}{9})$
(d)	$2MF$		$M(T+\frac{4}{3})$	$2M^2(T+\frac{4}{3})F$
(e)	$M(T+\frac{1}{6})$	$\leftrightarrow$	$M(T+\frac{3}{2})$	$M^2(T^2+\frac{5}{3}T+\frac{5}{18})$
(f)	$M(T+\frac{3}{2})$		$M(T+\frac{3}{2})$	$\frac{7}{6}M+M^2(T^2+3T+\frac{61}{18})$
(g)	$M(T+\frac{1}{2})$	$\leftrightarrow$	$M(T+\frac{3}{2})$	$\frac{1}{6}M+M^2(T^2+2T+\frac{8}{9})$

Formulas of $E(k_{ii'})$, $E(k_{i'j})$ and $E(k_{ii'}k_{i'j})$ required for computing $\mathrm{Cov}(d_{XY}, d_X)$ are:

Genealogy type	Probability	$E(k_{ii'})$		$E(k_{i'j})$	$E(k_{ii'}k_{ij})$
(a)	$\frac{1}{3}$	MS_2		$M(T+1)$	$M^2(T+1)S_2$
	$\frac{2}{3}$	$M(S_1+S_2)$		$M(T+1)$	$\frac{1}{2}MS_1 + M^2(T+1)(S_1+S_2)$
(b)	$\frac{1}{3}$	MS_2		$M(T+\frac{4}{3})$	$M^2(T+\frac{4}{3})S_2$
	$\frac{2}{3}$	$M(T+\frac{1}{3})$		$M(T+\frac{4}{3})$	$\frac{1}{2}M(T+\frac{1}{3}-S_2) + M^2(T^2+\frac{5}{3}T+\frac{5}{9})$
(c)	$\frac{1}{3}$	$M(T+\frac{1}{6})$		$M(T+\frac{3}{2})$	$M^2(T^2+\frac{5}{3}T+\frac{5}{18})$
	$\frac{2}{3}$	$M(T+\frac{1}{2})$	$\leftrightarrow$	$M(T+\frac{3}{2})$	$\frac{1}{6}M + M^2(T^2+2T+\frac{8}{9})$
(d)	$\frac{1}{3}$	MS_2		$M(T+\frac{6}{3})$	$M^2(T+\frac{4}{3})S_2$
	$\frac{2}{3}$	$M(T+\frac{4}{3})$		$M(T+\frac{1}{3})$	$\frac{M}{2}(T+\frac{1}{3}-S_2) + M^2(T^2+\frac{5}{3}T+\frac{5}{9})$
(e)	$\frac{1}{3}$	$M(T+\frac{1}{6})$	$\leftrightarrow$	$M(T+\frac{3}{2})$	$M^2(T^2+\frac{5}{3}T+\frac{5}{18})$
	$\frac{2}{3}$	$M(T+\frac{3}{2})$	$\leftrightarrow$	$M(T+\frac{1}{2})$	$\frac{M}{6} + M^2(T^2+2T+\frac{8}{9})$
(f)	$\frac{1}{3}$	$M(T+\frac{3}{2})$		$M(T+\frac{1}{6})$	$M^2(T^2+\frac{5}{3}T+\frac{5}{18})$
	$\frac{2}{3}$	$M(T+\frac{3}{2})$	$\leftrightarrow$	$M(T+\frac{1}{2})$	$\frac{1}{6}M + M^2(T^2+2T+\frac{8}{9})$
(g)	$\frac{1}{3}$	MS_2		$M(T+\frac{1}{3})$	$M^2(T+\frac{1}{3})S_2$
	$\frac{2}{3}$	$M(T+\frac{4}{3})$		$M(T+\frac{4}{3})$	$\frac{1}{2}M(T+\frac{7}{3}-S_2) + M^2(T^2+\frac{8}{3}+\frac{26}{9})$
(h)	$\frac{1}{3}$	$M(T+\frac{1}{6})$	$\leftrightarrow$	$M(T+\frac{1}{2})$	$M^2(T^2+\frac{2}{3}T+\frac{1}{9})$
	$\frac{2}{3}$	$M(T+\frac{3}{2})$		$M(T+\frac{3}{2})$	$\frac{7}{6}M + M^2(T^2+3T+\frac{61}{18})$
(i)	$\frac{1}{3}$	$M(T+\frac{1}{2})$	$\leftrightarrow$	$M(T+\frac{1}{6})$	$M^2(T+\frac{2}{3}T+\frac{1}{9})$
	$\frac{2}{3}$	$M(T+\frac{3}{2})$		$M(T+\frac{3}{2})$	$\frac{7}{6}M + M^2(T^2+3T+\frac{61}{18})$

In the above tabulation, genealogy type refers to those in Figure 5, and $S_1 = 1 - (\frac{3}{2})(1+T)e^{-T} + (\frac{1}{2})(1+3T)e^{-3T}$ and $S_2 = (\frac{1}{3})\{1-(1+3T)e^{-3T}\}$.