Genet. Res., Camb. (1975), **25**, *pp.* 229–248
Printed in Great Britain

Drift variances of heterozygosity and genetic distance in transient states

By WEN-HSIUNG LI and MASATOSHI NEI
Centre for Demographic and Population Genetics University of Texas at Houston, Texas 77025

(*Received* 29 *July* 1974)

SUMMARY

Using the moments of gene frequencies, the drift variances of heterozygosity and genetic distance in transient states have been studied under the assumption that all mutations are selectively neutral. Interestingly, this approach provides a simple derivation of Stewart's formula for the variance of heterozygosity at steady state. The results obtained indicate that if all alleles in the initial population are equally frequent, the standard derivation of heterozygosity is very small and increases linearly with time in the early generations. On the other hand, if the initial allele frequencies deviate appreciably from equality, then the standard deviation in the early generations is much larger but increases linearly with the square root of time. Under certain conditions, the standard deviation of genetic distance also increases linearly with time. Numerical computations have shown that the standard deviations of heterozygosity and genetic distance relative to their means are so large that a large number of loci must be used in estimating the average heterozygosity and genetic distance per locus.

1. INTRODUCTION

The genetic variability of a population is usually measured by the average heterozygosity per locus, while the gene differences between two populations may be measured by the genetic distance proposed by Nei (1972). The expected value of heterozygosity of a locus maintained by selectively neutral mutations in a finite population has been studied by Malécot (1948), Kimura & Crow (1964) and Kimura (1968) for both transient and steady states while the drift variance at steady state has recently been obtained by Stewart (1974). However, no one seems to have studied the variance of heterozygosity in transient states. Since the steady state is reached only when population size remains constant for a long time, this variance should be worked out. On the other hand, the expected genetic distance between two populations that have been isolated for an arbitrary number of generations has been studied for some important cases by Nei (1972), Nei & Feldman (1972) and Chakraborty & Nei (1974). However, the drift variance of this quantity has yet to be studied. Empirical data from natural populations suggest that this variance is generally very large (cf. Nei & Roychoudhury, 1974*a*). The purpose of this paper is to study this variance as well as the variance of heterozygosity in transient states.

It should be noted that the variances to be studied here are those due to random genetic drift. The sampling variances of these quantities at the time of survey have already been studied by Nei & Roychoudhury (1974*b*).

2. VARIANCE OF HETEROZYGOSITY

Let us consider a randomly mating diploid population of effective size N. We assume that N is sufficiently large so that $1/N^2$ and higher powers of $1/N$ are negligible compared with $1/N$. Following Kimura (1968), we assume that there are k possible allelic states at a locus and each allele mutates at the rate of v per generation to any one of the $l = k-1$ other allelic types with equal probability. We assume that in each generation the gene frequency changes first by mutation deterministically and then by random sampling of gametes stochastically. Selection will not be considered in this paper.

The homozygosity $j(t)$ and heterozygosity $h(t)$ of a locus in generation t is defined as $\Sigma x_i^2(t)$ and $1-\Sigma x_i^2(t)$, respectively, where $x_i(t)$ is the frequency of the ith allele and Σ stands for the summation over all alleles. Clearly, the variance of $h(t)$ is equal to that of $j(t)$, and we shall determine the variance of $h(t)$ by studying the moments of $j(t)$. The mean $\bar{j}(t)$ and variance $V(j(t))$ of $j(t)$ are given by

$$\bar{j}(t) \equiv E\{j(t)\} = E\left\{\sum_{i=1}^{k} x_i^2(t)\right\} = \sum_{i=1}^{k} E\{x_i^2(t)\}, \tag{1}$$

$$\begin{aligned} V(j(t)) &= E\left\{\sum_{i=1}^{k} x_i^2(t)\right\}^2 - \bar{j}^2(t) \\ &= \sum_{i=1}^{k} E\{x_i^4(t)\} + \sum_{i\neq j}^{k} E\{x_i^2(t)\, x_j^2(t)\} - \bar{j}^2(t). \end{aligned} \tag{2}$$

Thus, if we know the single and joint moments of gene frequencies, then $\bar{j}(t)$ and $V(j(t))$ can be computed.

The single nth moment of gene frequency $x_i(t)$ has been studied by Kimura (Crow & Kimura, 1956). It is approximately given by

$$\begin{aligned} \mu_n'^{(t)} = \sum_{i=0}^{\infty} \binom{n}{i} & \frac{\Gamma(B+n)\,\Gamma(A+2i)\,\Gamma(A-B+i)\,\Gamma(A+i-1)}{\Gamma(A+n+i)\,\Gamma(B+i)\,\Gamma(A-B)\,\Gamma(A+2i-1)} \\ & \times F(A+i-1, -i, A-B, 1-p) \exp\left\{-i\left(c+\frac{i-1}{4N}\right)t\right\}, \end{aligned} \tag{3}$$

where $p = x_j(0)$, $A = 4Nc$ and $B = 4Nd$, in which $d = v/l$ and $c = kd$, while $F(\cdot, \cdot, \cdot, \cdot)$ denotes the hypergeometric function.

The joint moments of two gene frequencies can be obtained by extending Kimura's method. Let $\mu_{mn}'^{(t)} = E\{x_i^m(t)\, x_j^n(t)\}$ be the m, nth moments of $x_i(t)$ and $x_j(t)$. In our mathematical model, $x_i(t)$ and $x_j(t)$ satisfy the following recurrence equations:

$$x_i(t+1) = X_i(t) + \delta X_i(t), \tag{4a}$$

$$x_j(t+1) = X_j(t) + \delta X_j(t), \tag{4b}$$

where

$$X_i(t) = (1-c)\, x_i(t) + d,$$

$$E\{\delta X_i(t)\} = 0,$$

https://doi.org/10.1017/S0016672300015664 Published online by Cambridge University Press

$$E[\{\delta X_i(t)\}^2] = X_i(t)\,(1-X_i(t))/(2N),$$

and
$$E(\delta X_i(t)\,\delta X_j(t)) = -X_i(t)\,X_j(t)/(2N).$$

Approximating $\mu'^{(t+1)}_{mn}-\mu'^{(t)}_{mn}$ by $\mathrm{d}\mu'^{(t)}_{mn}/\mathrm{d}t$, we obtain the following differential equation:

$$\frac{\mathrm{d}\mu'^{(t)}_{mn}}{\mathrm{d}t} = -\frac{m+n}{4N}(A+m+n-1)\,\mu'^{(t)}_{mn} + \frac{m}{4N}(B+m-1)\,\mu'^{(t)}_{(m-1)n} + \frac{n}{4N}(B+n-1)\,\mu'^{(t)}_{m(n-1)}, \tag{5}$$

where the terms involving $1/N^2$ and higher-order terms are neglected. In the absence of mutation $A = B = 0$, and (5) reduces to that of Kimura (1955).

It is not easy to obtain a general solution of (5), but all the moments can be obtained step by step, starting from $\mu'^{(t)}_{10}$ and $\mu'^{(t)}_{01}$. The complete expressions of the moments that are required for our purpose are given in the Appendix. The mean and variance of $j(t)$ can be obtained by using these moments together with the single moments given by (3). Since, however, the general formulae are very complicated, we shall present simplified formulae for the case of $k \to \infty$. In practice the following formulae are sufficiently accurate if $k \geqslant 20$. We also note that at the molecular level k is practically infinite.

$$\bar{j}(t) = \{j(0)-\hat{j}\}\exp\left\{-\left(2v+\frac{1}{2N}\right)t\right\}+\hat{j}, \tag{6}$$

$$V(j(t)) = C\exp\left\{-\left(4v+\frac{6}{2N}\right)t\right\} + D\exp\left\{-\left(3v+\frac{3}{2N}\right)t\right\} + G\exp\left\{-\left(2v+\frac{1}{2N}\right)t\right\} + \overline{j^2}(\infty)-\overline{j^2}(t), \tag{7}$$

where

$$\begin{aligned}
\hat{j} &\equiv \bar{j}(\infty) = 1/(M+1),\\
\overline{j^2}(\infty) &= (M+6)/\{(M+1)\,(M+2)\,(M+3)\},\\
G &= 2\{j(0)-\hat{j}\}\,(M+16)/\{(M+4)\,(M+5)\},\\
D &= 4\{j(0)+2\Sigma x_i^3(0)\}/(M+6)-2G(M+5)/(M+6)-4\hat{j}/(M+2),\\
C &= j^2(0)-D-G-\overline{j^2}(\infty),
\end{aligned}$$

in which $M = 4Nv$. Expression (6) is equivalent to Malécot's (1948) formula for the inbreeding coefficient for the case of $k = 2$. Noting that $h(t) = 1-j(t)$ and

$$V(h(t)) = V(j(t)),$$

it is clear that at equilibrium

$$\bar{h}(\infty) = M/(M+1), \tag{8}$$

$$\hat{V} \equiv V(h(\infty)) = 2M/\{(M+1)^2\,(M+2)\,(M+3)\}. \tag{9}$$

Formula (8) is identical with Kimura's (1968) and (9) with Stewart's (1974).

When $(4v+3/N)t$ is much smaller than 1, formulae (6) and (7) can further be simplified. The mean and variance of heterozygosity are then given by

$$\bar{h}(t) = h(0)+\frac{(M+1)j(0)-1}{2N}\,t, \tag{10}$$

https://doi.org/10.1017/S0016672300015664 Published online by Cambridge University Press

$$V(h(t)) = k_1 t + k_2 t^2, \tag{11}$$

approximately, where

$$Nk_1 = 2(\Sigma x_i^3(0) - j^2(0)),$$

$$8N^2 k_2 = 4C(M+3)^2 + 9D(M+2)^2/4 + G(M+1)^2 - 2j(0)\{2j(0) - 3\hat{j}\}(M+1)^2 - 2.$$

We note that if all alleles in the initial population are equally frequent,

$$\Sigma x_i^3(0) = j^2(0) \quad \text{and thus} \quad k_1 = 0,$$

but otherwise $k_1 \neq 0$. We also note that k_1 is much larger than k_2 unless $\Sigma x_i^3(0) - j^2(0)$ is very small – say less than $1/N$. Therefore, we have two different situations. Namely, if $\Sigma x_i^3(0) = j^2(0)$,

$$V(h(t)) = k_2 t^2. \tag{12}$$

In this case $V(h(t))$ is very small and the standard deviation $(\sqrt{\{V(h(t))\}})$ of $h(t)$ increases linearly with time. Since the mutation rate is generally very small, this linearity is expected to hold for a long period of time, if population size is large. On the other hand, if $\Sigma x_i^3(0) - j^2(0)$ is not small, then

$$V(h(t)) = k_1 t \tag{13}$$

approximately. In this case the variance is much larger than that for the first case and increases linearly with time.

In the derivation of formulae (6) and (7), we have neglected terms involving v/N, $1/N^2$, and higher orders. This approximation is satisfactory as long as there are one or more variant alleles in the population and $2Nv < 1$. However, if the initial population is completely homozygous for an allele with $j(0) = 1$, then formula (7) is not very accurate in the very early generations. For example, in the first generation (7) gives $V(h(1)) = v/N$ while the exact value is $2v/N$ if $2N < 1/v$. (Note that the variance is of the order of v/N, which has been assumed to be negligible in our formulation.) Numerical computations, however, have shown that the difference between the approximate and exact values relative to the approximate value decreases very rapidly. For example, if $N = 250{,}000$ and $v = 10^{-7}$, formula (7) gives a good approximation for $t \geqslant 100$.

The formulae developed above may be applied to two different situations. One is the case where a large number of independent populations are derived simultaneously from a common ancestral population and the variation of heterozygosity at a particular locus among populations is to be studied. In this case, if the effective size is the same for all populations, then formulae (6) and (7) are directly applicable. The other case is that where the variation of heterozygosity among loci in a single population is to be studied. If the initial gene frequencies and mutation rate are the same for all loci, formulae (6) and (7) are again applicable. In practice, however, the initial gene frequencies vary from locus to locus except in some special cases. In this case the initial conditions $j(0)$, $j^2(0)$, and $\Sigma x_i^3(0)$ in (6) and (7) should be replaced by their means over all loci, i.e. $\bar{j}(0)$, $\bar{j^2}(0)$, and $E\{\Sigma x_i^3(0)\}$, respectively. Therefore

$$\bar{j}(t) = \{\bar{j}(0) - \hat{j}\} \exp\left\{-\left(2v + \frac{1}{2N}\right)t\right\} + \hat{j}, \tag{14}$$

https://doi.org/10.1017/S0016672300015664 Published online by Cambridge University Press

$$V(j(t)) = \bar{C}\exp\left\{-\left(4v+\frac{6}{2N}\right)t\right\}+\bar{D}\exp\left\{-\left(3v+\frac{3}{2N}\right)t\right\} + \bar{G}\exp\left\{-\left(2v+\frac{1}{2N}\right)t\right\}+\bar{j}^2(\infty)-\bar{j}^2(t), \quad (15)$$

where $\bar{C}$, $\bar{D}$ and $\bar{G}$ are respectively the means of C, D and G over all loci with respect to the initial gene frequencies.

It is noted that $V(j(t))$ in (15) corresponds to the 'interlocus' variance in Nei & Roychoudhury (1974*b*). This interlocus variance can be decomposed into two components, i.e. the 'interclass' and 'intraclass' variances. The former refers to the variance due to the differences in initial gene frequencies among different loci, while the latter is the expected variance within gene frequency classes. The intraclass variance can be computed in the following way. We first classify the loci in the genome according to the initial gene frequencies, and let P_s be the proportion of the sth class of loci whose initial gene frequencies are identical with each other. The expected intraclass variance can then be computed by

$$V_w(h(t)) = \sum_s P_s V(h_s(t)), \quad (16)$$

where $V(h_s(t))$ is the variance of heterozygosity for the sth class of initial gene frequencies and given by (7). Therefore,

$$V_w(h(t)) = \bar{C}\exp\left\{-\left(4v+\frac{6}{2N}\right)t\right\}+\bar{D}\exp\left\{-\left(3v+\frac{3}{2N}\right)t\right\} + [\bar{G}-2\hat{j}\{\bar{j}(0)-\hat{j}\}]\exp\left\{-\left(2v+\frac{1}{2N}\right)t\right\} - \{\bar{j}^2(0)-2\bar{j}(0)\hat{j}+\hat{j}^2\}\exp\left\{-\left(4v+\frac{1}{N}\right)t\right\}+\hat{V}. \quad (17)$$

On the other hand, the interclass variance is the variance of $h(t)$ due to the variation of $h(0)$ in (6). Therefore,

$$V_b(h(t)) = V(h(0))\exp\left\{-\left(4v+\frac{1}{N}\right)t\right\}. \quad (18)$$

This indicates that the effect of initial gene frequencies declines at a rate of $4v+1/N$ in every generation. Furthermore, it can be shown that

$$V(h(t)) = V_w(h(t))+V_b(h(t)), \quad (19)$$

as it should be.

As noted earlier, $V(h(t))$ refers to the variation of heterozygosity *among loci*. Therefore, if we compute average heterozygosity from n loci which are randomly chosen from the genome, the expected variance of average heterozygosity is given by $V(h(t))/n$, neglecting the sampling variance at the time of gene frequency survey. If, however, one is interested in the variation of average heterozygosity *among different populations* which have been derived simultaneously from the same common ancestral population and average heterozygosity is computed from the same set of n random loci, then the expected variance of average heterozygosity is given by

$V_w(h(t))/n$. In nature, of course, most isolated populations or species would not have been derived at the same time. Furthermore, there is no assurance that the effective size has been the same or similar for all populations. Therefore, except in artificial populations it is difficult to get the observed value of $V_w(h(t))$.

Table 1. *Means* ($\bar{h}$) *and standard deviations* ($\sigma(h)$) *of heterozygosity in transient states*

(The mutation rate is assumed to be 10^{-7} per locus per generation.)

Generation	0	10	10^2	10^4	10^6	10^7
			$N = 250000$			
$\bar{h}$	0·0	2×10^{-6}	2×10^{-5}	0·0020	0·081	0·0909
$\sigma(h)$	0·0	$6{\cdot}3 \times 10^{-6}$	$6{\cdot}3 \times 10^{-5}$	0·0062	0·150	0·159
$\bar{h}$	0·5	0·49999	0·49991	0·491	0·136	0·0909
$\sigma(h)$	0·0	$1{\cdot}4 \times 10^{-5}$	$1{\cdot}4 \times 10^{-4}$	0·014	0·189	0·159
$\bar{h}$	0·48	0·47999	0·47991	0·472	0·134	0·0909
$V(h)$	0·0	$7{\cdot}7 \times 10^{-7}$	$7{\cdot}7 \times 10^{-6}$	0·0009	0·035	0·0253
$\sigma(h)$	0·0	$8{\cdot}7 \times 10^{-4}$	0·0028	0·03	0·188	0·159
			$N = 200$			
$\bar{h}$	0·0909	0·0887	0·071	0·00008	0·00008	0·00008
$\sigma(h)$	0·159	0·160	0·153	0·0052	0·0052	0·0052

In Table 1 four examples are given to illustrate how the mean ($\bar{h}$) and standard deviation ($\sigma(h) = \sqrt{[V(h(t))]}$) of $h(t) \equiv 1 - j(t)$ change with evolutionary time. It is assumed that $v = 10^{-7}$ and $k = \infty$ in all cases. The mutation rate of $v = 10^{-7}$ seems to be appropriate for an organism whose generation time is about one year (cf. Kimura & Ohta, 1971; Nei, 1975). The population size is assumed to be $N = 250000$ in all cases except in the last. In the first case of $h(0) = 0$, $\bar{h}$ and $\sigma(h)$ both increase almost linearly with increasing t up to about $t = 10000$ and then the rate of increase gradually declines. (As mentioned earlier, however, the value of $\sigma(h)$ is not very accurate for $t < 100$.) In the second case, where $x_1(0) = x_2(0) = 0{\cdot}5$, $\sigma(h)$ again increases linearly with time up to about $t = 10000$. In the third example of

$$x_1(0) = 0{\cdot}6 \quad \text{and} \quad x_2(0) = 0{\cdot}4,$$

$V(h)$ rather than $\sigma(h)$ increases linearly with time in the early generations, as mentioned earlier. Comparison of $\sigma(h)$ between the second and third cases shows that it is much larger in the latter case than in the former. It is also seen that in all of the above three examples both $\bar{h}$ and $\sigma(h)$ practically reach the equilibrium value by generation 10^7. The asymptotic rate of approach to the equilibrium value is the same for both $\bar{h}$ and $V(h)$, as is clear from (6) and (7). It is interesting to note that the variance of heterozygosity in the transient state may be larger than the equilibrium value (the second and third examples).

In the fourth example, it is assumed that the initial population was in equilibrium with $N = 250000$ but the population size was subsequently reduced to $N = 200$, and the mean and variance of heterozygosity among loci is computed. Therefore, $\bar{h}(0) = 0{\cdot}0909$ and $\sigma(h(0)) = 0{\cdot}159$ from (8) and (9), whereas $E\{\Sigma x_i^2(0)\} = 0{\cdot}8658$

https://doi.org/10.1017/S0016672300015664 Published online by Cambridge University Press

from (23′), which is given in the next paragraph. This example may simulate the evolution of a cave population in the characid fish *Astyanax mexicanus* (Avise & Selander, 1972). It is clear from Table 1 that average heterozygosity declines rather rapidly in the early generations and practically reaches the equilibrium value by generation 10000. On the other hand, the standard deviation of heterozygosity obtained from (15) first increases slightly and then starts to decrease. By generation 10000 it again reaches the equilibrium value.

Stewart (1974) derived a formula for the variance of heterozygosity at steady state for an arbitrary value of k by studying the equilibrium joint distribution of $x_1, \ldots, x_k$. If we use the present method, his result can be obtained very easily. Namely, at steady state the left-hand side of equation (5) is 0, so that the joint moment of gene frequencies becomes

$$\mu'^{(\infty)}_{mn} = \frac{\Gamma(B+m)\,\Gamma(B+n)\,\Gamma(A)}{\Gamma(A+m+n)\,\Gamma(B)\,\Gamma(B)}, \tag{20}$$

while the single moment is

$$\mu'^{(\infty)}_{n} = \mu'^{(\infty)}_{n0} = \frac{\Gamma(B+n)\,\Gamma(A)}{\Gamma(A+n)\,\Gamma(B)}, \tag{21}$$

which also follows from formula (3). Therefore,

$$\begin{aligned} V(h(\infty)) &= \sum_{i=1}^{k} \mu'^{(\infty)}_{4} + \sum_{i \neq j}^{k} \mu'^{(\infty)}_{22} - \left\{ \sum_{i=1}^{k} \mu^{(\infty)}_{2} \right\}^2 \\ &= \frac{2M(1+M/l)}{(1+M+M/l)^2\,(2+M+M/l)\,(3+M+M/l)}. \end{aligned} \tag{22}$$

This is identical with Stewart's formula. It can also be shown that

$$E\{\Sigma x_i^3(\infty)\} = \frac{(1+M/l)\,(2+M/l)}{(1+M+M/l)\,(2+M+M/l)}, \tag{23}$$

which reduces to

$$2/[(M+1)\,(M+2)] \tag{23′}$$

when k or $l \to \infty$.

3. VARIANCE OF GENETIC DISTANCE

In the last two decades, several different measures of genetic distance between populations have been proposed (e.g. Sanghvi, 1953; Cavalli-Sforza & Edwards, 1967; Rogers, 1972). However, most of these measures are constructed from the statistical point of view and it is not clear what biological unit they are going to measure (see Nei, 1973, for review). In contrast to these measures, the genetic distance proposed by Nei (1972) is intended to estimate the accumulated number of gene substitutions (net codon differences) per locus between populations. He has devised three different estimates of this number, i.e. the minimum (D_m), standard (D) and maximum (D') distances. For the biological meanings of these estimates or distances the reader may refer to Nei (1972, 1973) and Nei & Roychoudhury (1972).

Nei's genetic distance is based on the identities of genes within and between populations. Let x_i and y_i be the frequencies of the ith allele at a locus in populations

https://doi.org/10.1017/S0016672300015664 Published online by Cambridge University Press

1 and 2, respectively. The probability of identity of two randomly chosen genes from population 1 is $j_1 = \Sigma x_i^2$ and that from population 2 is $j_2 = \Sigma y_i^2$. The identity of two genes chosen at random, one from each population, is $j_{12} = \Sigma x_i y_i$. The three distance measures are then defined as:

$$\text{minimum:} \quad D_m = (J_1 + J_2)/2 - J_{12}, \tag{24}$$

$$\text{standard:} \quad D = -\log_e [J_{12}/\sqrt{(J_1 J_2)}], \tag{25}$$

$$\text{maximum:} \quad D' = -\log_e [J'_{12}/\sqrt{(J'_1 J'_2)}], \tag{26}$$

where J_1, J_2 and J_{12} are the arithmetic means of j_1, j_2 and j_{12} over all loci, respectively, while J'_1, J'_2 and J'_{12} are the geometric means. Since the genetic distances defined above are intended to measure the number of gene substitutions per locus, a large number of loci which are ideally a random sample of the genome should be used, including the polymorphic and monomorphic loci, as in the case of estimation of average heterozygosity. Note that J'_{12} is 0 if one of j_{12}'s is 0; then the maximum distance is meaningless. Actually, D' always tends to be an overestimate of the number of gene substitutions, and it is safe not to use this estimate if any one of j_{12}'s is small compared with unity (Nei, 1972). In the following, we shall not consider the maximum distance. As in the case of heterozygosity, the variance of genetic distance may be computed among a random set of loci between a given pair of populations as well as among independent pairs of populations at the same loci. In the present paper we shall first study the variance among loci and then show how to compute the variance appropriate for the latter case.

Now suppose that a population splits into two populations and thereafter no migration occurs between the two populations. In the absence of selection the differentiation of gene frequencies occurs due to mutation and random genetic drift. The genetic distance measures mentioned above were originally designed to be applied to the case where the sizes of the ancestral and the two descendant populations are more or less the same, so that in each population equilibrium between mutation and genetic drift is maintained throughout the process considered. Chakraborty & Nei (1974), however, showed that the distance measures are quite robust and applicable even when the size of one of the two descendant populations is 100 times larger or smaller than that of the other. Let N_1 and N_2 be the effective sizes of populations 1 and 2, respectively, and assume that they are so large that the terms involving $1/N_1^2$, $1/N_2^2$, $1/(N_1 N_2)$ and higher order terms are negligible. We first consider the mean and variance of the minimum genetic distance in generation t after reproductive isolation.

At a particular locus, the minimum genetic distance is defined as

$$d_m(t) = \tfrac{1}{2}[j_1(t) + j_2(t)] - j_{12}(t). \tag{27}$$

The mean $[\bar{d}_m(t)]$ and variance $[V(d_m(t))]$ of $d_m(t)$ over loci are given by

$$\bar{d}_m(t) = \tfrac{1}{2}[\bar{j}_1(t) + \bar{j}_2(t)] - \bar{j}_{12}(t), \tag{28}$$

$$V(d_m(t)) = \tfrac{1}{4}[V(j_1(t)) + V(j_2(t))] + \tfrac{1}{2}\mathrm{cov}\,(j_1(t), j_2(t)) + V(j_{12}(t)) - \mathrm{cov}\,(j_1(t), j_{12}(t)) - \mathrm{cov}\,(j_2(t), j_{12}(t)). \tag{29}$$

Clearly, D_m in (24) is an estimate of $\bar{d}_m(t)$. We have seen that $\bar{j}_1(t)$ and $\bar{j}_2(t)$ are given by (14), while $\bar{j}_{12}(t)$ is given by $\bar{j}_{12}(0)\,e^{-2vt}$ (Nei, 1972; Nei & Feldman, 1972). If $N_1 = N_2 = N$ and equilibrium between mutation and genetic drift is maintained throughout the process with $\overline{j_1}(0) = \overline{j_2}(0) = \overline{j_{12}}(0) = \hat{j}$, then (28) reduces to

$$\bar{d}_m(t) = \hat{j}(1 - \exp\{-2vt\}). \tag{30}$$

We also know that $V(j_1(t))$ and $V(j_2(t))$ are given by (15). On the other hand, $\mathrm{cov}\,(j_1(t), j_2(t))$ is given by

$$\sum_i E\{x_i^2(t)\,y_i^2(t)\} + \sum_{i \neq j} E\{x_i^2(t)\,y_j^2(t)\} - \bar{j}_1(t)\,\bar{j}_2(t).$$

Therefore, in order to know $\mathrm{cov}\,(j_1(t), j_2(t))$, we must evaluate the joint moments $\mu'_{20,20} = E\{x_i^2(t)\,y_i^2(t)\}$ and $\mu'_{20,20} = E\{x_i^2(t)\,y_j^2(t)\}$. These moments can be obtained by the method given in the Appendix. The results are, however, very complicated and we may consider only the case of $k = \infty$. In this case we have

$$\mathrm{cov}\,(j_1(t), j_2(t)) = \mathrm{cov}\,(j_1(0), j_2(0)) \exp\left\{-\left(4v + \frac{1}{2N_1} + \frac{1}{2N_2}\right)t\right\}. \tag{31}$$

The variance $V(d_m(t))$ includes three more quantities to be determined. They can be obtained in the same way as the above and are given by

$$\begin{aligned} V(j_{12}(t)) = {} & A_3 \exp\left\{-\left(4v + \frac{1}{2N_1} + \frac{1}{2N_2}\right)t\right\} + B_3 \exp\left\{-\left(3v + \frac{1}{2N_1}\right)t\right\} \\ & + C_3 \exp\left\{-\left(3v + \frac{1}{2N_2}\right)t\right\} + F_3 \exp\{-2vt\} - \bar{j}_{12}^2(0) \exp\{-4vt\}, \end{aligned} \tag{32}$$

$$\begin{aligned} \mathrm{cov}\,(j_1(t), j_{12}(t)) = {} & A_4 \exp\left\{-\left(4v + \frac{3}{2N_1}\right)t\right\} + B_4 \exp\left\{-\left(3v + \frac{1}{2N_1}\right)t\right\} \\ & + F_4 \exp\{-2vt\} - \bar{j}_1(t)\,\bar{j}_{12}(t), \end{aligned} \tag{33}$$

where

$$\begin{aligned} M_1 &= 4N_1 v, \quad M_2 = 4N_2 v, \quad A_3 = E\{j_{12}^2(0)\} - B_3 - C_3 - F_3, \\ B_3 &= 2[E\{\Sigma x_i^2(0)\,y_i(0)\} - 2\bar{j}_{12}(0)/(M_1+2)]/(M_2+2), \\ C_3 &= 2[E\{\Sigma x_i(0)\,y_i^2(0)\} - 2\bar{j}_{12}(0)/(M_2+2)]/(M_1+2), \\ F_3 &= 2\bar{j}_{12}(0)\,[N_1/(M_1+2) + N_2/(M_2+2)]/(N_1 M_2 + N_1 + N_2), \\ A_4 &= E\{j_1(0)\,j_{12}(0)\} - B_4 - F_4, \\ B_4 &= [4E\{\Sigma x_i^2(0)\,y_i(0)\} - (8 - 2M_1)\bar{j}_{12}(0)/(M_1+2)]/(M_1+4) \end{aligned}$$

and

$$F_4 = 6\bar{j}_{12}(0)/[(M_1+2)\,(M_1+3)].$$

The formula for $\mathrm{cov}\,(j_2(t), j_{12}(t))$ can be obtained by interchanging $x_i(0)$ and $y_i(0)$ and subscripts 1 and 2 in (33).

Therefore, putting the above variances and covariances of $j_1(t), j_2(t)$ and $j_{12}(t)$ into (29), $V(d_m(t))$ can be evaluated. One of the important cases to which (29) may be

applied is that where the sizes of the ancestral and the two descendant populations are more or less the same. In this case we may assume that

$$N_1 = N_2 = N, \quad M_1 = M_2 = M,$$

$$\bar{j}_1(t) = \bar{j}_2(t) = \bar{j}_{12}(0) = \hat{j} = 1/(M+1),$$

$$V(j_1(t)) = V(j_2(t)) = \operatorname{cov}(j_1(0), j_2(0))$$

$$= \hat{V} = 2M/[(M+1)^2(M+2)(M+3)],$$

$$E\{j_{12}^2(0)\} = E\{j_1(0)j_{12}(0)\} = \overline{j^2}(\infty) = (M+6)/\{(M+1)(M+2)(M+3)\}$$

and

$$E\{\Sigma x_i^2(0)\, y_i(0)\} = E\{\Sigma x_i(0)\, y_i^2(0)\} = E\{\Sigma x_i^3(0)\} = 2/[(M+1)(M+2)].$$

Therefore, (29) reduces to

$$V(d_m(t)) = \tfrac{1}{2}\hat{V} + 2\hat{j}\left\{\hat{j} - \frac{4M+6}{(M+2)^2(M+3)}\right\}\exp\{-2vt\} - \hat{j}^2\exp\{-4vt\}$$
$$- \frac{4M\hat{j}}{(M+2)(M+4)}\exp\left\{-\left(3v+\frac{1}{2N}\right)t\right\}$$
$$+ \left\{\tfrac{3}{2}\hat{V} + \frac{(M\hat{j})^2}{(M+2)^2}\right\}\exp\left\{-\left(4v+\frac{1}{N}\right)t\right\}$$
$$+ \frac{2M\hat{j}}{(M+3)(M+4)}\exp\left\{-\left(4v+\frac{3}{2N}\right)t\right\}. \tag{34}$$

When $(4v + 3/2N)\, t$ is much smaller than 1, the above formula can be approximated as follows:

$$V(d_m(t)) = k_1 t^2, \tag{35}$$

where

$$k_1 = \frac{M\hat{j}}{(M+3)(M+4)}\left(4v+\frac{3}{2N}\right)^2 - 4\hat{j}v^2\left[\hat{j} + \frac{4M+6}{(M+2)^2(M+3)}\right]$$
$$- \frac{2M\hat{j}}{(M+2)(M+4)}\left(3v+\frac{1}{2N}\right)^2 + \frac{1}{2}\left[\tfrac{3}{2}\hat{V} + \frac{(M\hat{j})^2}{(M+2)^2}\right]\left(4v+\frac{1}{N}\right)^2.$$

Thus, both the mean (30) and standard deviation of minimum genetic distance increase linearly with time in the early generations. This linearity is expected to hold for a long period of time, if N is large.

As in the case of heterozygosity, the variance of genetic distance among loci can be decomposed into the intraclass $[V_w(d_m(t))]$ and interclass $[V_b(d_m(t))]$ variances. Namely,

$$V(d_m(t)) = V_w(d_m(t)) + V_b(d(t)). \tag{36}$$

The interclass variance is the variance of $d_m(t)$ in (27) due to the variation of $j_1(0)$, $j_2(0)$ and $j_{12}(0)$ among loci. Therefore, we have

$$V_b(d_m(t)) = [V_b(j_1(t)) + V_b(j_2(t))]/4$$
$$+ \tfrac{1}{2}\operatorname{cov}(j_1(0), j_2(0))\exp\left\{-\left(4v+\frac{1}{2N_1}+\frac{1}{2N_2}\right)t\right\} + V(j_{12}(0))\exp\{-4vt\}.$$

https://doi.org/10.1017/S0016672300015664 Published online by Cambridge University Press

$$-\operatorname{cov}(j_1(0), j_{12}(0)) \exp\left\{-\left(4v + \frac{1}{2N_1}\right)t\right\}$$

$$-\operatorname{cov}(j_2(0), j_{12}(0)) \exp\left\{-\left(4v + \frac{1}{2N_2}\right)t\right\}. \tag{37}$$

If $j_1(0) = j_2(0) = j_{12}(0) = j(0)$ at each locus and $N_1 = N_2 = N$, the above formula reduces to

$$V_b(d_m(t)) = V(h(0))\left\{\exp\left\{-\left(2v + \frac{1}{2N}\right)t\right\} - \exp\{-2vt\}\right\}^2. \tag{38}$$

Furthermore, if $(4v + 1/N)\,t \ll 1$,

$$V_b(d_m(t)) = V(h(0))\,(t/2N)^2 \tag{39}$$

approximately. Formulae (38) and (39) indicate that the interclass variance initially increases with time but eventually becomes 0.

It is noted that (39) can be directly obtained from the following approximate formula for $d_m(t)$ for $(2v + 1/2N)\,t \ll 1$:

$$d_m(t) = h(0)\,(t/2N). \tag{40}$$

Note also that if $\hat{j}_1 = \hat{j}_2 = \bar{j}(0) = \hat{j}$, then the mean of (40) over all loci is

$$\begin{aligned} \bar{d}_m(t) &= \frac{4Nv}{4Nv+1}\frac{t}{2N} \\ &= 2v\hat{j}t, \end{aligned} \tag{41}$$

as expected from (30).

The intraclass variance of genetic distance can be obtained by the same method as (16), but the result is somewhat complicated. However, if

$$N_1 = N_2 = N, \quad \bar{j}_1(t) = \bar{j}_2(t) = \hat{j}_{12}(0) = \hat{j},$$

and

$$V(j_1(t)) = V(j_2(t)) = \operatorname{cov}(j_1(0), j_2(0)) = \hat{V},$$

then it is given by the difference between (34) and (38).

The intraclass variance gives the expected variance of genetic distance among random pairs of populations when the same set of random loci are used for all populations. For a given value of $M = 4Nv$ the proportion of intraclass variance among the total variance remains constant in the early generations and is given by $1 - \hat{V}/(4N^2k_1)$. This proportion is 0·78 when M is close to 0, and increases with increasing M. For example, it is 0·80 for $M = 0{\cdot}1$ and 0·83 for $M = 0{\cdot}2$. Therefore, in all cases the intraclass variance accounts for a major part of the variance of genetic distance, as long as the balance between mutation and genetic drift is maintained in the process of gene differentiation.

Let us now consider the standard genetic distance, D. This distance is designed to be applied to a set of loci and it is not meaningful to compute the distance for each locus separately. Furthermore, since it involves the logarithm, it is not easy to obtain the exact mean and variance of D computed from a finite number of loci. However, approximate formulae may be obtained by using the method of Taylor

expansion, assuming that the probabilities of J_1, J_2 and J_{12} deviating far from their means are negligibly small. This assumption seems to be satisfactory as long as a large number of loci are used to estimate J_1, J_2 and J_{12} and the populations to be compared belong to the same species or different species in the same genus (Nei, 1973, 1975). The number of loci used (r) is generally larger than 20 in practice.

The expectation of D based on r loci is given by

$$\begin{aligned}\bar{D} &= -E\log_e(J_{12}/\sqrt{(J_1J_2)}) \\ &= -\log_e(\bar{J}_{12}/\sqrt{(\bar{J}_1\bar{J}_2)}) + E(J_1-\bar{J}_1)\frac{\partial D}{\partial J_1}\Big|_{\bar{J}_1} + E(J_2-\bar{J}_2)\frac{\partial D}{\partial J_2}\Big|_{\bar{J}_2} + \dots \\ &\approx -\log_e(\bar{J}_{12}/\sqrt{(\bar{J}_1\bar{J}_2)}),\end{aligned} \tag{42}$$

where $\bar{J}_1 = E(J_1)$, $\bar{J}_2 = E(J_2)$ and $\bar{J}_{12} = E(J_{12})$, and the second and higher order terms of $(J_1-\bar{J}_1)$, $(J_2-\bar{J}_2)$ and $(J_{12}-\bar{J}_{12})$ are neglected. The value $(\bar{D}(t))$ of $\bar{D}$ at the tth generation is given by (42) replacing $\bar{J}$'s by $\bar{J}(t)$'s where $\bar{J}(t) = \Sigma\bar{j}(t)/r$ in which the summation is over all loci. We note that if $\bar{J}_1(t) = \bar{J}_2(t) = \bar{J}_{12}(0) = \hat{j}$, then

$$\bar{D} = 2vt \tag{43}$$

(Nei & Feldman, 1972). Therefore, $\bar{D}$ is proportional to the divergence time.

Similarly, neglecting the third and higher order terms of $(J_1-\bar{J}_1)$, $(J_2-\bar{J}_2)$ and $(J_{12}-\bar{J}_{12})$, the variance of D is given by

$$\begin{aligned}V(D) = \frac{1}{r^2}\Bigg[&\frac{\Sigma V(j_1)}{4\bar{J}_1^{\,2}} + \frac{\Sigma V(j_2)}{4\bar{J}_2^{\,2}} + \frac{\Sigma\,\mathrm{cov}(j_1,j_2)}{2\bar{J}_1\bar{J}_2} + \frac{\Sigma V(j_{12})}{\bar{J}_{12}^{\,2}} \\ &- \frac{\Sigma\,\mathrm{cov}(j_1,j_{12})}{\bar{J}_1\bar{J}_{12}} - \frac{\Sigma\,\mathrm{cov}(j_2,j_{12})}{\bar{J}_2\bar{J}_{12}}\Bigg]\end{aligned} \tag{44}$$

approximately. Since the denominator of each term in the bracket in the above formula is the same for all loci, the contribution of a locus to the total variance is

$$V(d) = \frac{V(j_1)}{4\bar{J}_1^{\,2}} + \frac{V(j_2)}{4\bar{J}_2^{\,2}} + \frac{\mathrm{cov}(j_1,j_2)}{2\bar{J}_1\bar{J}_2} + \frac{V(j_{12})}{\bar{J}_{12}^{\,2}} - \frac{\mathrm{cov}(j_1,j_{12})}{\bar{J}_1\bar{J}_{12}} - \frac{\mathrm{cov}(j_2,j_{12})}{\bar{J}_2\bar{J}_{12}}, \tag{45}$$

if the mutation rate is the same for all loci and each locus behaves independently. $V(d)$ in the tth generation can be evaluated by replacing $\bar{j}$'s by $\bar{j}(t)$'s. Decomposition of $V(d)$ into the intraclass and interclass variances can be made by the same method as that for the minimum genetic distance, and we shall not repeat it here.

Chakraborty & Nei (1974) showed that when the size (N_2) of one of the two populations is drastically reduced after divergence a better measure of genetic distance is

$$D_a = -\log_e(J_{12}/J_1), \tag{46}$$

since this is porportional to the divergence time. The expectation and variance of this measure can be obtained in the same way as the above. Namely,

$$\bar{D}_a \approx -\log_e(\bar{J}_{12}/\bar{J}_1), \tag{47}$$

$$V(D_a) \approx \frac{1}{r^2}\left[\frac{\Sigma V(j_1)}{\bar{J}_1^{\,2}} + \frac{\Sigma V(j_{12})}{\bar{J}_{12}^{\,2}} - \frac{2\Sigma\,\mathrm{cov}(j_1,j_{12})}{\bar{J}_1\bar{J}_{12}}\right]. \tag{48}$$

https://doi.org/10.1017/S0016672300015664 Published online by Cambridge University Press

Therefore, the contribution of a locus to the total variance is

$$V(d_a) = \frac{V(j_1)}{\bar{J}_1^2} + \frac{V(j_{12})}{\bar{J}_{21}^2} - \frac{2\,\mathrm{cov}\,(j_1, j_{12})}{\bar{J}_1 \bar{J}_{12}}. \tag{49}$$

When $\bar{J}_1(r) = \bar{J}_{12}(0) = \hat{\jmath}$, $\bar{D}_a = 2vt$, as shown by Chakraborty & Nei (1974).

In table 2 the means and standard deviations of genetic distances are given for two different situations. In Case 1 a population splits into two completely isolated populations of equal size with equilibrium between mutation and genetic drift maintained. We assume that $N_1 = N_2 = 250\,000$ and $\bar{\jmath}_1(t) = \bar{\jmath}_2(t) = \bar{\jmath}_{12}(0) = \hat{\jmath} = 1/(M+1)$, etc., as mentioned earlier. In Case 2 a small population of effective sizes $N_2 = 200$ is

Table 2. *Means and standard deviations of genetic distances in various generations after divergence of two populations*

(The mutation rate is assumed to be 10^{-7} per locus per generation.)

Generation	0	10	10^3	10^5	10^6	10^7
			$N_1 = N_2 = 250\,000$			
$\bar{d}_m$	0	$1{\cdot}8 \times 10^{-6}$	$1{\cdot}8 \times 10^{-4}$	0·018	0·165	0·786
$\sigma(d_m)$	0	$7{\cdot}2 \times 10^{-6}$	$7{\cdot}2 \times 10^{-4}$	0·064	0·334	0·326
$\bar{D}$	0	2×10^{-6}	2×10^{-4}	0·020	0·20	2·00
$\sigma(d)$	0	$8{\cdot}1 \times 10^{-6}$	$8{\cdot}1 \times 10^{-4}$	0·074	0·448	2·52
			$N_1 = 250\,000$, $N_2 = 200$			
$\bar{D}$	0	0·0012	0·044	0·067	0·24	2·05
$\sigma(d)$	0	0·0041	0·147	0·216	0·52	2·60
$\bar{D}_a$	0	2×10^{-6}	0·0002	0·02	0·20	2·00
$\sigma(d_a)$	0	0·020	0·125	0·20	0·52	2·60

derived from a large population whose effective size (N_1) is 250 000. We assume that $\bar{\jmath}_1(t) = \bar{\jmath}_{12}(0) = \hat{\jmath} = 1/(M+1)$. In both cases the mutation rate is assumed to be 10^{-7} per locus per generation.

In case 1 both $\bar{d}_m$ and $\bar{D}$ first increase linearly with increasing t, but after about 10^6 generations the linearity for $\bar{d}_m$ is destroyed. The standard deviations

$$\sigma(d_m) = \sqrt{\{V(d_m)\}} \quad \text{and} \quad \sigma(d) = \sqrt{\{V(d)\}}$$

of d_m and D for a locus increase linearly until about 10^5 generations, and then the rate of increase declines. The standard deviations are both about four times larger than their respective means until about 10^5 generations and then $\sigma(d_m)/\bar{d}_m$ or $\sigma(d)/\bar{D}$ gradually declines. This indicates that in order to have a reliable estimate of genetic distance a large number of loci must be used. If $\sigma(d)/\bar{D}$ is 4 and one wants to make $\sigma(D)/\bar{D}$ to be $\frac{1}{2}$ or less, then at least 64 loci should be used, where

$$\sigma(D) = \sqrt{\{V(D)\}}.$$

If $\sigma(d)/\bar{D} = 2{\cdot}2$ (the value at generation 10^6), then the number of loci to be used will be 19. The square root of the intraclass variance is smaller than $\sigma(d_m)$ but not much. For example, the value for the minimum distance is $6{\cdot}5 \times 10^{-6}$ for $t = 10$ compared with $\sigma(d_m) = 7{\cdot}2 \times 10^{-6}$.

https://doi.org/10.1017/S0016672300015664 Published online by Cambridge University Press

In case 2 the means and single-locus standard deviations of D and D_a are given. As expected, $\bar{D}$ is not linear with divergence time and increases rapidly in the early generations. On the other hand, $\bar{D}_a$ increases linearly with increasing number of generations. The standard deviation ($\sigma(d_a) = \sqrt{\{V(d_a)\}}$) of D_a for a locus is, however, very large in the early generations. Therefore, to obtain a reliable estimate of D_a a large number of loci must be studied. In the later generations there is not much difference between the two distance measures in both the mean and standard deviation

4. DISCUSSION

In the present study we have assumed that mutation rate is the same for all loci. This assumption is certainly unrealistic and mutation rate would vary from locus to locus. If this is the case, the variances of heterozygosity and genetic distance among loci will be larger than those given in this paper. In practice, however, we do not know the magnitude of the variation of mutation rate. It is worth noting that if mutation rate varies with locus, the expected heterozygosity at steady state and genetic distance become smaller than those obtained by replacing ν in (8) and (43) by the average mutation rate (Nei, 1975). For example, the expectation of average heterozygosity at steady state can be shown to be equal to

$$\frac{\bar{M}}{\bar{M}+1} - \frac{\sigma_M^2}{(\bar{M}+1)^3}$$

approximately, where $\bar{M}$ and σ_M^2 are the mean and variance of M respectively.

We have also assumed that the number of possible allelic states at a locus is so large, that whenever a mutation occurs in a population it represents a new allele (the model of infinite number of alleles). This model seems to be appropriate if allelic variants are identified at the nucleotide or codon level. In practice, however, genetic variation is often studied by electrophoresis. Under the model of stepwise change of electrophoretic mobility of protein, Ohta & Kimura (1973) and Nei & Chakraborty (1973) studied the expectations of heterozygosity and genetic distance, respectively. Ohta & Kimura showed that the expected homozygosity at steady state is given by $\bar{J} = 1/\sqrt{(1+8Nv)}$, where v is the rate of mutation that induces electrophoretic charge change. This value is much larger than $1/(1+4Nv)$, if $4Nv$ is large, In practice, however, $4Nv$ is generally of the order of 0·15 or less (cf. Nei, 1975), and the difference between the two formulae is very small. The formula for genetic distance in (43) is also approximately applicable for electrophoretic data as long as $\bar{D}$ is smaller than 1 (Nei & Chakraborty, 1973). It is not easy to determine the variance of heterozygosity under the model of stepwise mutation, but the computer simulation conducted by Ohta & Kimura (1974) suggests that it is slightly smaller than that given by our formula. The variance of genetic distance is also expected to be slightly smaller. However, if such a technique as heat denaturation treatment (Bernstein, Throckmorton & Hubby, 1973) is used in combination with electrophoresis to detect protein variation, both the means and variances of heterozygosity and genetic distance become closer to those for the model of infinite number of alleles.

In practice, average heterozygosity and genetic distance are measured by taking samples from populations, and this sampling process introduces another variance. Nei & Roychoudhury (1974*b*) presented a method for decomposing the variances of average heterozygosity and genetic distance into the interlocus and intralocus (sampling) variances. Empirical data have shown that the interlocus variance is much larger than the sampling variance even when only about 40 genes (20 individuals) per locus are sampled. This large interlocus variance is of course expected to occur due to random genetic drift. In fact, in a number of organisms there is a good agreement between the observed interlocus variance of heterozygosity and the theoretical variance computed from (9) (Nei, 1975). The present study indicates that in transient states the drift variance of heterozygosity relative to the mean may be larger than that at steady state. This is particularly so when population size has recently been reduced. The drift variance of genetic distance is also very large, and this large value of the expected variance of genetic distance is in agreement with the empirical observations by Nei & Roychoudhury (1974*a*) and Chakraborty & Nei (1974). These results again emphasize the importance of studying many loci for estimating average heterozygosity and genetic distance. The strategy of determining the number of loci and sample size per locus in estimating these quantities has been discussed by Nei & Roychoudhury (1974*b*).

As mentioned earlier, a number of authors proposed different measures of genetic distance. It is known that there is a strong positive correlation between these measures when they are applied to closely related populations. One might, therefore, suspect that the results obtained here are also applicable approximately to other distance measures. We are not sure about this. Actually, there are a number of problems in evaluating the variances of other distance measures. For example, Cavalli-Sforza's distance involves the square roots of gene frequencies, so that both the mean and variance of this distance are expected to be a complicated function of evolutionary time. (Under certain circumstances with no mutation, the expectation of his f_θ (Cavalli-Sforza, 1969) is *approximately* linear with time for $t \ll 2N$.) Furthermore, his distance is intended to be applied not to a random set of loci but to polymorphic loci alone. Therefore, in order to evaluate the variance, we must know the initial gene frequencies, which are not obtainable in natural populations. The same comment applies to all distance measures which make use of polymorphic (selected) loci only.

In addition to our distance measures, those devised by Latter (1972), Rogers (1972), and Hedrick (1971) are clearly intended to be applied to a random set of loci. Latter's distance is defined as $\gamma = 2D_m/(J_X + J_Y)$ in our terminology. Thus, the expectation of γ under mutation-drift balance is linear with evolutionary time in the early generations. The variance of γ is somewhat complicated, but if the number of loci used is large, the ratio of the standard deviation to the mean of this distance would not be far from that of D_m. Rogers' coefficient of dissimilarity is a function of the square root of our $d_m(t)$. Therefore, the expectation of this coefficient cannot be linear with time even in the early generations. Its variance is also expected to be a complicated function of evolutionary time. Hedrick's coefficient of similarity (rather

than dissimilarity) measures the degree of similarity of genotype frequencies rather than gene frequencies between two populations and is a function of fourth moments of gene frequencies in diploid organisms. Therefore, the variance of his coefficient relative to the mean is expected to be large, since it becomes a function of eighth moments of gene frequencies. This is true also with the sampling variance at the time of gene frequency survey. For the comparison of other properties of various distance measures, see Nei (1973).

We thank Dr Alan Robertson for his valuable comments on the manuscript. This study was supported by U.S. Public Health Service Grant GM 20293.

APPENDIX

Joint Moments of the Gene Frequencies

(i) *One population*

From equation (5) we obtain the following formulae:

$$\mu_{11}^{\prime(t)} = [p_i p_j - AG'(A+2)^{-1} - \mu_{11}^{\prime(\infty)}] \exp\left\{-\left(2c+\frac{1}{2N}\right)t\right\} + AG'(A+2)^{-1}\{\exp -ct\} + \mu_{11}^{\prime(\infty)}, \quad \text{(A 1)}$$

$$\mu_{21}^{\prime(t)} = Q\exp\left\{-\left(3c+\frac{3}{2N}\right)t\right\} + R\exp\left\{-\left(2c+\frac{1}{2N}\right)t\right\} + T\exp\{-ct\} + \mu_{21}^{\prime(\infty)}, \quad \text{(A2)}$$

$$\mu_{22}^{\prime(t)} = A'\exp\left\{-\left(4c+\frac{6}{2N}\right)t\right\} + B'\exp\left\{-\left(3c+\frac{3}{2N}\right)t\right\} + C'\exp\left\{-\left(2c+\frac{1}{2N}\right)t\right\} + D'\exp\{-ct\} + \mu_{22}^{\prime(\infty)}, \quad \text{(A 3)}$$

where $\mu_{mn}^{\prime(\infty)}$ is given by (12), and

$$G' = (p_i + p_j - 2/k)/k,$$

$$Q = p_i^2 p_j - R - T - \mu_{21}^{\prime(\infty)},$$

$$R = [2(B+1)\{p_i p_j - AG'(A+2)^{-1} - \mu_{11}^{\prime(\infty)}\} + B\{p_i^2 - 2(B+1)(p_i - 1/k)(A+2)^{-1} - \mu_2^{\prime(\infty)}\}](A+4)^{-1},$$

$$T = (B+1)\{AG' + B(p_i - 1/k)\}\{(A+2)(A+3)\}^{-1},$$

$$A' = p_i^2 p_j^2 - B' - C' - D' - \mu_{22}^{\prime(\infty)},$$

$$B' = 2(B+1)[p_i p_j^2 + p_i^2 p_j - 4(B+1)\{p_i p_j - AG'(A+2)^{-1} - \mu_{11}^{\prime(\infty)}\}(A+4)^{-1} - B\{p_i^2 + p_j^2 - 2(B+1)G'k(A+2)^{-1} - 2\mu_2^{\prime(\infty)}\}(A+4)^{-1} - 2(B+1)AG'\{(A+2)(A+3)\}^{-1} - (B+1)AG'\{(A+2)(A+3)\}^{-1} - 2\mu_{21}^{\prime(\infty)}](A+6)^{-1},$$

$$C' = (B+1)[4(B+1)\{p_i p_j - AG'(A+2)^{-1} - \mu_{11}^{\prime(\infty)}\}(A+4)^{-1} + B\{p_i^2 + p_j^2 - 2(B+1)G'k(A+2)^{-1} - 2\mu_2^{\prime(\infty)}\}(A+4)^{-1}](A+5)^{-1},$$

$$D' = 2(B+1)^2 AG'\{(A+2)(A+3)(A+4)\}^{-1}.$$

By virtue of symmetry, $\mu'^{(t)}_{12}$ can be obtained simply by interchanging p_i and p_j in formula (A 2). In the absence of mutation, $A = B = 0$ and the above formlae reduce to those of Kimura (1955).

(ii) *Two populations*

Let $x_i(t)$, $x_j(t)$, $y_i(t)$ and $y_j(t)$ be the frequencies of the ith and jth alleles at generation t in populations 1 and 2, respectively. Since we assume that there is no migration between the two populations, these gene frequencies satisfy the recurrence equation (4). Note also that $\delta x_i(t)$ and $\delta x_j(t)$ are independent of $\delta y_i(t)$ and $\delta y_j(t)$. Let

$$\mu'^{(t)}_{mn,pq} = E\{x_i^m(t)\, x_j^n(t)\, y_i^p(t)\, y_j^q(t)\}$$

be the m, n, p, q-th moment of gene frequencies at generation t. Then, a differential equation equivalent to (5) can be derived. Using this differential equation all the moments required can be obtained successively starting from $\mu'_{10,01}$. However, the results obtained are so complicated, that we present only the following typical moments:

$$\mu'^{(t)}_{10,01} = (p_i q_j - G_2 - 1/k^2)\exp\{-2ct\} + G_2\exp\{-ct\} + 1/k^2, \tag{A 4}$$

$$\begin{aligned}\mu'^{(t)}_{30,01} = {} & A_5\exp\left\{-\left(4c+\frac{3}{2N_1}\right)t\right\} + B_5\exp\left\{-\left(3c+\frac{3}{2N_1}\right)t\right\} + C_5\exp\left\{-\left(3c+\frac{1}{2N_1}\right)t\right\} \\ & + D_5\exp\left\{-\left(2c+\frac{1}{2N_1}\right)t\right\} + E_5\exp\{-2ct\} + F_5\exp\{-ct\} + \mu'^{(\infty)}_{30,01},\end{aligned} \tag{A 5}$$

$$\begin{aligned}\mu'^{(t)}_{20,02} = {} & A_6\exp\left\{-\left(4c+\frac{1}{2N_1}+\frac{1}{2N_2}\right)t\right\} + B_6\exp\left\{-\left(3c+\frac{1}{2N_1}\right)t\right\} \\ & + C_6\exp\left\{-\left(3c+\frac{1}{2N_1}\right)t\right\} + D_6\exp\left\{-\left(2c+\frac{1}{2N_1}\right)t\right\} + E_6\exp\left\{-\left(2c+\frac{1}{2N_2}\right)t\right\} \\ & + F_6\exp\{-2ct\} + G_6\exp\{-ct\} + \mu'^{(\infty)}_{20,02},\end{aligned} \tag{A 6}$$

where

$$p_i = x_i(0), \quad q_i = y_i(0), \quad A_1 = 4N_1c, \quad A_2 = 4N_2c,$$

$$B_1 = 4N_1d, \quad B_2 = 4N_2d,$$

$$\mu'^{(\infty)}_{30,01} = 3W_1(A_1+3)^{-1} + B_1\Gamma(B_1+3)\,\Gamma(A_1)/\{2\Gamma(A_1+4)\,\Gamma(B_1)\},$$

$$W_\lambda = [2 + A_\lambda(B_\lambda+1)(A_\lambda+1)^{-1}](3A_\lambda+2)^{-1}k^{-2} \quad (\lambda = 1 \text{ or } 2),$$

$$\mu'^{(\infty)}_{20,02} = (N_1W_1 + N_2W_2)(2N_1A_2 + N_1 + N_2)^{-1}, \quad G_2 = (p_i + q_j - 2/k)/k,$$

$$A_5 = p_i^3q_j - B_5 - C_5 - D_5 - E_5 - F_5 - \mu'^{(\infty)}_{30,01}, \quad B_5 = U_1/k,$$

$$C_5 = 6O_1/(A_1+4), \quad D_5 = (6P_1 + B_1U_2)/(2A_1+4),$$

$$E_5 = 3S_1/(A_1+3), \quad F_5 = (6T_1 + B_1U_3)/(3A_1+6),$$

$$U_1 = p_i^3 - U_2 - U_3 - (B_1+2)(B_1+1)\{(A_1+2)(A_1+1)k\}^{-1},$$

$$U_2 = 3(B_1+2)[p_i^2 - 2(B_1+1)(p_i - 1/k)(A_1+2)^{-1} - (B_1+1)\{(A_1+1)k\}^{-1}]/(A_1+4),$$

$$U_3 = 3(B_1+2)(B_1+1)(p_i - 1/k)\{(A_1+2)(A_1+3)\}^{-1},$$

$$O_1 = p_i^2 q_j - P_1 - S_1 - T_1 - W_1, \quad O_2 = p_i q_j^2 - P_2 - S_2 - T_2 - W_2,$$

$$P_1 = [p_i^2 - 2(B_1+1)(p_i - 1/k)(A_1+2)^{-1} - (B_1+1)\{(A_1+1)k\}^{-1}]/k,$$

$$P_2 = [q_j^2 - 2(B_2+1)(q_j - 1/k)(A_2+2)^{-1} - (B_2+1)\{(A_2+1)k\}^{-1}]/k,$$

$$S_\lambda = 2(p_i q_j - G_2 - 1/k^2)/(A_\lambda + 2), \quad \lambda = 1 \quad \text{or} \quad 2,$$

$$T_1 = G_2(A_1+1)^{-1} + B_1(B_1+1)(p_i - 1/k)\{(A_1+1)(A_1+2)\}^{-1},$$

$$T_2 = G_2(A_2+1)^{-1} + B_2(B_2+1)(q_j - 1/k)\{(A_2+1)(A_2+2)\}^{-1},$$

$$A_6 = p_i^2 p_j^2 - B_6 - C_6 - D_6 - E_6 - F_6 - G_6 - \mu'^{(\infty)}_{20,02}, \quad B_6 = 2O_1/(A_2+2),$$

$$C_6 = 2O_2/(A_1+2), \quad D_6 = P_1/(A_2+1), \quad E_6 = P_2/(A_1+1),$$

$$F_6 = (N_1 S_1 + N_2 S_2)/(N_1 A_2 + N_1 + N_2), \quad G_6 = 2(N_1 T_1 + N_2 T_2)/(3N_1 A_2 + 2N_1 + 2N_2).$$

REFERENCES

AVISE, J. C. & SELANDER, R. K. (1972). Evolutionary genetics of cave-dwelling fishes of the genus *Astyanax*. *Evolution* **26**, 1–19.

BERNSTEIN, S. C., THROCKMORTON, L. H. & HUBBY, J. L. (1973). Still more genetic variability in natural populations. *Proceedings of the National Academy of Sciences* (*U.S.A.*) **70**, 3928–3931.

CAVALLI-SFORZA, L. L. (1969). Human diversity. *Proceedings of the XIIth International Congress of Genetics* (*Tokyo*) **3**, 405–416.

CAVALLI-SFORZA, L. L. & EDWARDS, A. W. F. (1967). Phylogenetic analysis: models and estimation procedures. *American Journal of Human Genetics* **19**, 233–257.

CHAKRABORTY, R. & NEI, M. (1974). Dynamics of gene differentiation between incompletely isolated populations of unequal sizes. *Theoretical Population Biology* **5**, 460–469.

CROW, J. F. & KIMURA, M. (1956). Some genetic problems in natural populations. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* **4**, 1–22.

HEDRICK, P. W. (1971). A new approach to measuring genetic similarity. *Evolution* **25**, 276–280.

KIMURA, M. (1955). Random genetic drift in a multi-allelic locus. *Evolution* **9**, 419–435.

KIMURA, M. (1968). Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. *Genetical Research* **11**, 247–269.

KIMURA, M. & CROW, J. F. (1964). The number of alleles that can be maintained in a finite population. *Genetics* **49**, 725–738.

KIMURA, M. & OHTA, T. (1971). Protein polymorphism as a phase of molecular evolution. *Nature* **229**, 467–469.

LATTER, B. D. H. (1972). Selection in finite populations with multiple alleles. III. Genetic divergence with centripetal selection and mutation. *Genetics* **70**, 475–490.

MALÉCOT, G. (1948). *Les Mathématiques de l'hérédité*. Paris: Masson et Cie.

NEI, M. (1972). Genetic distance between populations. *American Naturalist* **106**, 283–292.

NEI, M. (1973). The theory and estimation of genetic distance. *Genetic Structure of Populations* (ed. N. E. Morton), pp. 45–54. Honolulu: University of Hawaii Press.

NEI, M. (1975). *Molecular Population Genetics and Evolution*. Amsterdam: Elsevier, Excerpta Medica and North-Holland.

NEI, M. & CHAKRABORTY, R. (1973). Genetic distance and electrophoretic identity of proteins between taxa. *Journal of Molecular Evolution* **2**, 323–328.

NEI, M. & FELDMAN, M. W. (1972). Identity of genes by descent within and between populations under mutation and migration pressures. *Theoretical Population Biology* **3**, 460–465.

NEI, M. & ROYCHOUDHURY, A. K. (1972). Gene differences between Caucasian, Negro, and Japanese populations. *Science* **177**, 434–436.

NEI, M. & ROYCHOUDHURY, A. K. (1974*a*). Genic variation within and between the three major races of man, Caucasoids, Negroids, and Mongoloids. *American Journal of Human Genetics* **26**, 421–443.

NEI, M. & ROYCHOUDHURY, A. K. (1974*b*). Sampling variances of heterozygosity and genetic distance. *Genetics* **76**, 379–390.

OHTA, T. & KIMURA, M. (1973). A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genetical Research* **22**, 201–204.

OHTA, T. & KIMURA, M. (1974). Simulation studies on electrophoretically detectable genetic variability in a finite population. *Genetics* **76**, 615–624.

ROGERS, J. S. (1972). Measures of genetic similarity and genetic distance. *Studies in Genetics*: VII. University of Texas Publication no. 7213, pp. 145–153.

SANGHVI, L. D. (1953). Comparison of genetic and morphological methods for a study of biological differences. *American Journal of Physical Anthropology* **11**, 385–404.

STEWART, F. M. (1974). Variability in the amount of heterozygosity maintained by neutral mutations. *Theoretical Population Biology* (in press).