GENE IDENTITY AND GENETIC DIFFERENTIATION OF POPULATIONS IN THE FINITE ISLAND MODEL[1]

NAOYUKI TAKAHATA

National Institute of Genetics, Mishima, Shizuoka-ken, 411 Japan

Manuscript received August 3, 1982
Revised copy accepted February 9, 1983

ABSTRACT

A formula for the variance of gene identity (homozygosity) was derived for the case of neutral mutations using diffusion approximations for the changes of gene frequencies in a subdivided population. It is shown that when gene flow is extremely small, the variance of gene identity for the entire population at equilibrium is smaller than that of the panmictic population with the same mean gene identity. On the other hand, although a large amount of gene flow makes a subdivided population equivalent to a panmictic population, there is an intermediate range of gene flow in which population subdivision can increase the variance. This increase results from the increased variance between colonies. In such a case, each colony has a predominant allele, but the predominant type may differ from colony to colony. The formula for obtaining the variance allows us to study such statistics as the coefficient of gene differentiation and the correlation of heterozygosity. Computer simulations were conducted to study the distribution of gene identity as well as to check the validity of the analytical formulas. Effects of selection were also studied by simulations.

NATURAL populations are generally subdivided into a number of subpopulations or demes, and there is often significant genetic differentiation among subpopulations. To measure the degree of genetic differentiation of structured populations, WRIGHT (1943) introduced a statistic called the fixation index. NEI (1973) extended it to the case of multiple alleles, proposing an index called the coefficient of gene differentiation. He also proposed a quantity appropritate to measure the genetic distance between two related populations (NEI 1972). For measuring genetic differentiation, there are many other quantities, and the reader may refer to FELSENSTEIN (1976) for them. Although they are diverse, one quantity common to them is gene identity, *i.e.*, the probability of identity of two randomly chosen alleles, which has been intensively studied in relation to geographic distance (WRIGHT 1943, 1946, 1951; MALECOT 1951, 1955; KIMURA and WEISS 1964; WEISS and KIMURA 1965; MARUYAMA 1969, 1970a,b,c; and others). However, the theoretical study of gene identity is generally restricted to the mean value, except for (1) only two populations (NEI and FELDMAN 1972; LI and NEI 1975, 1977), (2) a finite number of populations that are completely isolated (NEI and CHAKRAVARTI 1977), or (3) diallelic systems without mutation (NEI, CHAKRAVARTI and TATENO 1977).

[1] Contribution no. 1450 from the National Institute of Genetics, Mishima, Shizuoka-ken 411, Japan.

In this paper I derive a formula for the variance of gene identity for a finite number of incompletely isolated populations and study the variance of the coefficient of gene differentiation and the correlation of heterozygosity. These formulas are derived for neutral mutations at a single locus with K possible allelic states (KIMURA 1968a). Computer simulations have been conducted to check the validity of the formulas and examine the distribution of gene identity. Simulations were also extended to the case of multiallelic mutations with selection.

GENE IDENTITY IN THE ISLAND MODEL

We consider the finite island model in which the entire population is subdivided into L colonies, each with effective size N, and each colony exchanges individuals at the rate m with equal likelihood with the remaining colonies. Suppose that the organism is diploid and migration is independent of genotype. Let K be a fixed number of potential alleles at a locus and $v/(K-1)$ be the mutation rate from one to any of the other $K-1$ alleles, the total rate being v. We denote by $A_k(i)$ the kth allele in the ith colony and by $x_k(i)$ the frequency of $A_k(i)$.

We make use of the diffusion approximation method for describing stochastic changes of gene frequencies (KIMURA 1964). Hence, the formulas and results obtained are valid so long as the higher order terms of m, v and N^{-1} can be ignored. The mean $M[\delta x_k(i)]$ and covariance $V[\delta x_k(i)\delta x_{k'}(j)]$ of the change of gene frequencies per generation are given by

$$M[\delta x_k(i)] = v^* - (Lm^* + Kv^*)x_k(i) + m^* \sum_{j=1}^{L} x_k(j) \tag{1}$$

and

$$V[\delta x_k(i)\delta x_{k'}(j)] = \frac{1}{2N} x_k(i)[\delta_{kk'} - x_{k'}(j)]\delta_{ij} \tag{2}$$

where $\sum_{k=1}^{K} x_k(i) = 1$, $v^* = v/(K-1)$ and $m^* = m/(L-1)$.
In (2), δ_{ij} stands for the Kronecker's delta function, and it is assumed that random sampling of gametes takes place independently in each colony. The diffusion operator, B, for the Kolmogorov backward equation is

$$B = \sum_{i=1}^{L} \left[\sum_{k=1}^{K-1} M[\delta x_k(i)] \frac{\partial}{\partial x_k(i)} + \frac{1}{2} \sum_{k=1}^{K-1} \sum_{k'=1}^{K-1} V[\delta x_k(i)\delta x_{k'}(i)] \frac{\partial^2}{\partial x_k(i)\partial x_{k'}(i)} \right] \tag{3}$$

and the expectation, $E\{f\}$, of any function of $x_k(i)$'s satisfies

$$\frac{dE\{f\}}{dt} = E\{Bf\} \tag{4}$$

in which time, t, is measured in generations.

For simplicity, we will restrict our study to the case in which the equilibrium is reached or in which the initial condition of f is independent of the geography of colonies when we want to study nonequilibrium solution of (4); otherwise,

we must formulate an intractable number of moment equations. Twelve moments are required to obtain the variance of gene identity at equilibrium. We define the gene identities within and between colonies, j_0 and j_1,

$$j_0 = <j_0(i)> = <\sum_{k=1}^{K} x_k^2(i)> = \frac{1}{L}\sum_{i=1}^{L}\sum_{k=1}^{K} x_k^2(i)$$
$$j_1 = <\sum_{k=1}^{K} x_k(i_1)x_k(i_2)> \qquad \text{for} \quad i_1 \neq i_2 \tag{5}$$

and define the gene identity for the entire population

$$j_T = \sum_{k=1}^{K} y_k^2, \qquad y_k = <x_k(i)> \tag{6}$$

Thus, $j_T = \frac{1}{L} j_0 + \left(1 - \frac{1}{L}\right) j_1$ and in (5) and (6) a symbol $< >$ denotes the expectation taken over the appropriate set of colonies.

The mean values of j_0 and j_1, denoted by J_0 and J_1, respectively, satisfies

$$\begin{pmatrix} \dot{J}_0 \\ \dot{J}_1 \end{pmatrix} = \begin{pmatrix} -a_1 & 4M \\ 4M^* & -a_2 \end{pmatrix}\begin{pmatrix} J_0 \\ J_1 \end{pmatrix} + \begin{pmatrix} 1 + 4\theta^* \\ \theta^* \end{pmatrix} \tag{7}$$

where the time scale has been changed to a unit of $2N$ generations ($\tau = t/2N$) and the dot over J_0 and J_1 indicates the differentiation with respect to τ. The parameters in (7) are

$$\begin{aligned} M^* &= M/(L-1) = Nm/(L-1), \\ \theta^* &= \theta/(K-1) = Nv/(K-1), \\ a_1 &= 1 + 4K\theta^* + 4M, \\ a_2 &= 4K\theta^* + 4M^*. \end{aligned} \tag{8}$$

Equations in (7) are equivalent to those studied by MAYNARD SMITH (1970) for small values of m and v (see also MARUYAMA 1970a; LATTER 1973; NEI 1975).

The third moments concern identity probabilities for which we choose three genes randomly from one, two and three different colonies,

$$\begin{aligned} T_1 &= <\sum_{k=1}^{K} E\{x_k^3(i)\}> \\ T_2 &= <\sum_{k=1}^{K} E\{x_k^2(i_1)x_k(i_2)\}> \end{aligned} \tag{9}$$

and

$$T_1 = <\sum_{k=1}^{K} E\{x_k(i_1)x_k(i_2)x_k(i_3)\}>$$

where the subscripts of i indicate different colonies. For the fourth moments we must know the quantities concerning identity probabilities when we sample four genes randomly from one, two, three and four different colonies,

$$
\begin{aligned}
F_1 &= <\sum E\{x_{k_1}^2(i)x_{k_2}^2(i)\}> \\
F_2 &= <\sum E\{x_{k_1}^2(i_1)x_{k_2}(i_1)x_{k_2}(i_2)\}> \\
F_3 &= <\sum E\{x_{k_1}(i_1)x_{k_1}(i_2)x_{k_2}(i_1)x_{k_2}(i_2)\}> \\
F_4 &= <\sum E\{x_{k_1}^2(i_1)x_{k_2}^2(i_2)\}> \\
F_5 &= <\sum E\{x_{k_1}^2(i_1)x_{k_2}(i_2)x_{k_2}(i_3)\}> \\
F_6 &= <\sum E\{x_{k_1}(i_1)x_{k_1}(i_2)x_{k_2}(i_1)x_{k_2}(i_3)\}> \\
F_7 &= <\sum E\{x_{k_1}(i_1)x_{k_1}(i_2)x_{k_2}(i_3)x_{k_2}(i_4)\}>
\end{aligned} \tag{10}
$$

in which the sum is taken over all k_1 and k_2. Substitution of (9) and (10) into (4) gives the moment equations. The third moment equations are

$$
\begin{pmatrix} \dot{T}_1 \\ \dot{T}_2 \\ \dot{T}_3 \end{pmatrix} = \begin{pmatrix} -b_1 & 6M & 0 \\ 2M^* & -b_2 & 4(L-2)M^* \\ 0 & 12M^* & -b_3 \end{pmatrix} \begin{pmatrix} T_1 \\ T_2 \\ T_3 \end{pmatrix} + \begin{pmatrix} (3+6\theta^*)J_0 \\ 2\theta^* J_0 + (1+4\theta^*)J_1 \\ 6\theta^* J_1 \end{pmatrix} \tag{11}
$$

where $b_1 = 3(1 + 2K\theta^* + 2M)$, $b_1 = 1 + 2(3K\theta^* + (2L - 3)M^*)$ and $b_3 = 6(K\theta^* + 2M^*)$. The column vector $\mathbf{F} = (F_1, F_2, \ldots, F_7)^t$ of the fourth moments satisfies

$$
\dot{\mathbf{F}} = C\mathbf{F} + \mathbf{d} \tag{12}
$$

in which C is the 7×7 matrix given in APPENDIX, and $\mathbf{d}$ is the column vector whose elements d_i $(i = 1, 2, \ldots, 7)$ are

$$
\begin{aligned}
d_1 &= (2 + 8\theta^*)J_0 + 4T_1 \\
d_2 &= 4\theta^* J_0 + (1 + 4\theta^*)J_1 + 2T_2 \\
d_3 &= 8\theta^* J_1 + 2T_2 \\
d_4 &= 2(1 + 4\theta^*)J_0 \\
d_5 &= 4\theta^* J_0 + (1 + 4\theta^*)J_1 \\
d_6 &= 8\theta^* J_0 + T_3 \\
d_7 &= 8\theta^* J_1.
\end{aligned} \tag{13}
$$

When the population consists of only two colonies, (11) and (12) should be read for the first two Ts and four Fs. The remaining variables cannot be defined in this case and are not required to obtain the variance of gene identities as will be seen. The same note applies to the population with three colonies, in which case F_7 is ignored. Also, we note that the maximum eigenvalue of the matrix in (7) is not greater than $-4K\theta^*$ and those in (11) and (12) are at most $-6K\theta^*$ and $-8K\theta^*$, respectively. Thus, the rate at which the equilibrium state is reached does not exceed $4K\theta^* = 4NvK/(K - 1)$ in unit of $2N$ generations. This fact provides us with a rough estimate of the number of generations required to study nonequilibrium properties by using the previous formulas. On the other hand, the equilibrium solutions of (7), (11) and (12) can be directly obtained by equating the left-hand sides to 0 and solving these equations in a standard way.

Actual calculation of such equations except for (7) is, however, often tedious so that it was done numerically.

In the following, we define several quantities related to gene identity and express them in terms of J_i, T_i and F_i. The variance of gene identity within a subpopulation, V_w, is

$$V_w = <E\{j_0^2(i)\}> - E\{j_0\}^2 = F_1 - J_0^2 \tag{14}$$

and the variance between subpopulations, V_b, is

$$\begin{aligned} V_b &= E\{j_1^2\} - E\{j_1\}^2 \\ &= \frac{1}{L(L-1)}\{2F_3 + 4(L-2)F_6 + (L-2)(L-3)F_7\} - J_1^2. \end{aligned} \tag{15}$$

Likewise, the variance of j_T, denoted by V_T, is

$$\begin{aligned} V_T &= E\{j_T^2\} - E\{j_T\}^2 \\ &= \frac{1}{L^2}\left\{\frac{1}{L}V_w + \left(1-\frac{1}{L}\right)(F_4 - J_0^2)\right\} + \left(1-\frac{1}{L}\right)^2 V_b \\ &\quad + 2\frac{1}{L}\left(1-\frac{1}{L}\right)\mathrm{cov}(j_0, j_1) \end{aligned} \tag{16}$$

where

$$\mathrm{cov}(j_0, j_1) = \frac{1}{L}\{2F_2 + (L-2)F_5\} - J_0 J_1.$$

NEI (1973) extended WRIGHT'S F_{st} statistic to the case of multiple alleles and called it the G_{st} statistic. In the present notation, it is given by

$$G_{st} = \frac{J_0 - J_T}{1 - J_T}.$$

In addition, to study the coefficient of gene differentiation for a large number of loci, NEI'S group considered the mean and variance of

$$g_{st} = (j_0 - j_T)/(1 - j_T),$$

which henceforth are denoted by G_{st}^* and $V_{g_{st}}$, respectively. Exact analysis of G_{st}^* and $V_{g_{st}}$ is difficult, so that truncated Taylor expansions are used to examine the behavior of these variables. Using this approach NEI and CHAKRAVARTI (1977) found that

$$\begin{aligned} G_{st}^* &\approx 1 - \frac{1-J_0}{1-J_T} + \frac{\mathrm{cov}(j_0, j_T)}{(1-J_T)^2} - \frac{(1-J_0)V_T}{(1-J_T)^2} \\ V_{g_{st}} &\approx \left(\frac{1-J_0}{1-J_T}\right)^2\left[\frac{V_w}{(1-J_0)^2} + \frac{V_T}{(1-J_T)^2} - \frac{2\mathrm{cov}(j_0, j_T)}{(1-J_0)(1-J_T)}\right]. \end{aligned} \tag{17}$$

In (17), J_T denotes the mean of j_T, i.e.,

$$J_T = \frac{1}{L}J_0 + \left(1-\frac{1}{L}\right)J_1$$

and (18)

$$cov(j_0, j_T) = \frac{1}{L^2}\{F_1 + (L-1)(2F_2 + F_4 + (L-2)F_5)\} - J_0 J_T.$$

It is obvious that (17) is a poor approximation when the amount of polymorphism in the entire population is low, *i.e.*, for a large value of J_T and, thus, (17) should not be used in such a case. As will be discussed later, a more accurate formula particularly for $V_{g_{st}}$ is needed that includes higher moments of j_0 and j_T.

Finally, we define the correlation of heterozygosity between colonies, R, as

$$R = (F_4 - J_0^2)/V_w. \tag{19}$$

This is equivalent to the correlation of heterozygosities from two randomly chosen colonies among different loci when the mutation rate is the same for all loci. Based on the infinite allele model, LI and NEI (1975) showed that in completely isolated populations R decreases exponentially as time increases and eventually becomes 0. In a subdivided population with gene flow, however, the equilibrium value does not equal 0 even for $K = \infty$ and takes a value depending heavily on levels of gene flow. Thus, R may be a useful statistic to measure the degree of genetic differentiation of subpopulations.

Before going to the next section, I would like to give some results concerning the equilibrium solution of (7). In particular, when $K = \infty$ (KIMURA and CROW 1964) the solution is simple (MAYNARD SMITH 1970; MARUYAMA 1970a; CROW and MARUYAMA 1971; LATTER 1973; NEI 1975). The mean genetic identity for the entire population J_T and G_{st} are then given by

$$J_T = \left[1 + 4L\theta + \frac{(L-1)\theta}{LM^* + \theta}\right]^{-1}$$

and (20)

$$G_{st} = \left[1 + \frac{4L}{L-1}(\theta + LM^*)\right]^{-1}$$

where, and subsequently, a symbol indicating the equilibrium state is suppressed. Formulas in (20) are equivalent to those given in MAYNARD SMITH (1970) and LATTER (1973), provided m, v and N^{-1} are all small. Note that $J_T = [L(1 + 4\theta)]^{-1}$ and $G_{st} = \left[1 + \frac{4L\theta}{L-1}\right]^{-1}$ for $m = 0$ so that the population is very polymorphic regardless of the value of $\theta (J_T < 1/L)$, and that for small θ the value of G_{st} is close to 1 because of random genetic drift. On the other hand, when $m \gg v$, $J_T = [1 + 4L\theta]^{-1}$ and $G_{st} = [1 + 4\alpha M]^{-1}$, where $\alpha = \left(\frac{L}{L-1}\right)^2$. Namely, the population can be regarded as panmictic in the sense that J_T is equivalent to the mean genetic identity in a panmictic population with effective size NL. Even under the situation, however, G_{st} is different from 0 if Nm is small and the finiteness of L affects G_{st} through α. The formula of WRIGHT (1943), $F_{st} = 1/(1 + 4Nm)$, corresponds to G_{st} with $L = \infty$.

COMPUTER SIMULATION

To check the validity of my formula, I conducted computer simulation keeping K finite (K = 4), examining the distribution of gene identity, as well as the mean and variance of g_{st}. I also examined the effect of selection on these parameters, considering two selection schemes. Both schemes assume that there exists a normal type allele in each colony and that the other alleles are all selectively disadvantageous. One model assumes that the normal type allele varied randomly from colony to colony, whereas the other model assumes that the same allele is favored in all colonies. Let $1 - s_k(i)$ be the relative fitness of the kth allele in the ith colony and assume that fitness is multiplicative and that fitnesses do not vary with time.

The mean change of $x_k(i)$ per generation due to mutation and migration is given by the right-hand side of (1), and the change due to selection is

$$\Delta x_k(i) = \{w(i) - s_k(i)\}x_k(i)/(1 - w(i)) \tag{21}$$

where $w(i) = \sum_{l=1}^{K} s_l(i)x_l(i)$. In (21), $s_k(i)$ 0 for some k depending on the model used and is a positive constant for all other alleles. For preassigned values of $s_l(i)$'s, v and m, the mean changes of gene frequencies were calculated using (1) and (21), followed by random sampling of gametes using multinomial pseudo-random variables (as described by KIMURA and TAKAHATA 1983).

To establish the equilibrium state, the first v^{-1} generations for each set of parameter values were discarded and, thereafter, 5000 observations were made at every 100 generations. Choosing v as 10^{-3} in the simulations, the total of 5×10^5 generations in each run was used to study the equilibrium properties. Table 1 gives such simulation results for K = 4 and L = 10. Comparison of these results with the theoretical values shows that the formulas except that for $V_{g_{st}}$ provide good approximations. As pointed out by NEI and CHAKRAVARTI (1977), the formula of $V_{g_{st}}$ omits the second- and higher-order terms of j_0 and j_T in the Taylor expansion. Although these terms could not be evaluated analytically, the approximation (17) seems so poor that we cannot use it, particularly for the case of large Nm and small Nv.

MEAN AND VARIANCE OF GENE IDENTITY

Many statistical analyses have been proposed to test the neutrality of polymorphic genes (KIMURA 1968b), among which methods using the relationship between the mean and variance of heterozygosity (STEWART 1976; LI AND NEI 1975) have received much attention (NEI, CHAKRABORTY and FUERST 1976a,b; FUERST, CHAKRABORTY and NEI 1977; YAMAZAKI 1976; GOJOBORI 1982). However, the theoretical relationships used in these tests are obtained based on the assumption of random mating, so that it is interesting to examine them in the case of a subdivided population. In the following, I will consider only the equilibrium relationships. We keep in mind that the rate at which the equilibrium state is reached is given approximately by mutation rate when gene flow among colonies seldom occurs.

First, we note from (20) that the total genetic diversity, $H_T = 1 - J_T$, increases as m/v decreases, and that the range of H_T becomes $[1 - (1/L), 1]$ if m/v is less

TABLE 1

Simulation results for K = 4 and L = 10

Nm		0.01	0.1	1	10
			Nv = 0.1		
J_0	exp	0.727 (0.0432)	0.644 (0.0403)	0.452 (0.0190)	0.379 (0.0123)
	obs	0.726 (0.0434)	0.640 (0.0409)	0.448 (0.0177)	0.377 (0.0092)
J_1	exp	0.254 (0.0766)	0.280 (0.0561)	0.342 (0.0192)	0.365 (0.0122)
	obs	0.263 (0.0776)	0.277 (0.0549)	0.339 (0.0162)	0.366 (0.0087)
J_T	exp	0.301 (0.0017)	0.317 (0.0034)	0.353 (0.0087)	0.366 (0.0112)
	obs	0.309 (0.0020)	0.313 (0.0026)	0.349 (0.0061)	0.367 (0.0079)
G_{st}	exp	0.609	0.479	0.153	0.196
	obs	0.605	0.476	0.151	0.164
G_{st}^*	exp	0.609 (0.0882)	0.478 (0.0856)	0.151 (0.0337)	0.0193 (0.0044)
	obs	0.606 (0.0086)	0.476 (0.0083)	0.150 (0.0018)	0.0164 (0.00002)
			Nv = 0.01		
J_0	exp	0.938 (0.0180)	0.840 (0.0361)	0.756 (0.0476)	0.741 (0.0496)
	obs	0.935 (0.0179)	0.813 (0.0397)	0.718 (0.0514)	0.736 (0.0433)
J_1	exp	0.303 (0.1775)	0.518 (0.1444)	0.702 (0.0636)	0.735 (0.0509)
	obs	0.321 (0.1799)	0.434 (0.1386)	0.657 (0.0669)	0.731 (0.0439)
J_T	exp	0.366 (0.0081)	0.550 (0.0360)	0.707 (0.0491)	0.736 (0.0496)
	obs	0.381 (0.0079)	0.472 (0.0280)	0.663 (0.0516)	0.732 (0.0429)
G_{st}	exp	0.902	0.644	0.167	0.0198
	obs	0.894	0.645	0.163	0.0162
G_{st}^*	exp	0.901 (0.0448)	0.635 (0.1623)	0.156 (0.1753)	0.183 (0.0312)
	obs	0.899 (0.0052)	0.622 (0.0250)	0.136 (0.0056)	0.0148 (6×10^{-5})

The value of the variance for each quantity is presented in parentheses: exp = theoretical value, obs = observed value in simulation.

than about 0.1. This means that in a small subdivided population with extremely limited migration, different alleles are quasifixed in different colonies, and the total amount of genetic variability can be quite high. Such a situation is theoretically conceivable but does not agree with the observation of $0 \leq H_T \leq 0.3$ for most species studied so far (FUERST, CHAKRABORTY and NEI 1977). There are many assumptions that may be responsible for the discrepancy between observed and theoretical values of H_T, among which the assumed value of m/v may be important. For the lowest value of H_T to be close to 0, say ϵ, the ratio m/v has to be as large as L/ϵ. This indicates that gene flow between colonies should be very high compared with the mutation rate, e.g., if we take $\epsilon = 0.01$, $L = 100$ and $v = 10^{-6}$, the migration rate required is 1% per generation. There is, however, an interesting analysis that reveals low levels of gene flow among colonies. Using the *conditional average frequency* (the average frequency

of an allele conditioned on the number of colonies it appears in), SLATKIN (1981, 1982) estimated the level of gene flow in a subdivided population and showed that some species such as salamanders apparently have low levels of gene flow.

STEWART'S formula for the variance of heterozygosity allows us to calculate a theoretical variance, V_p, when the population with a given H_T is assumed to be panmictic. When K is sufficiently large, we have that

$$V_p = \frac{2H_T(1 - H_T)^3}{(2 - H_T)(3 - 2H_T)}. \tag{22}$$

It is important to note that at equilibrium and for sufficiently large K, all variables J_i, T_i and F_i (except J_0, T_1, F_1 and F_4) are negligibly small when m is much smaller than v and N^{-1}. This indicates that the genetic constitution differs from colony to colony and that V_b and $cov(j_0, j_1)$ are greatly reduced. Actually we can show that for $m = 0$ and $K = \infty$, $V_b = 0$, $cov(j_0, j_1) = 0$ and

$$V_T = \frac{2(1/L - 1 + H_T)(1 - H_T)^3}{(1/L + 1 - H_T)(1/L + 2 - 2H_T)}$$

for the range of $1 - 1/L \leq H_T \leq 1$. Clearly, $V_p > V_T$ for $L \geq 2$ and the value of V_T decreases at the rate of $1/L^2$ as L increases. It is interesting to examine whether the whole population of the salamander mentioned before has lower than expected variance of gene diversity.

On the other hand, if m is sufficiently large, population subdivision should not affect the relationship of (22). Therefore, STEWART'S formula (22) is expected to hold for large m. However, there are intermediate values of m for which population subdivision results in a variance larger than expected from that in the panmictic case.

Figure 1 shows the result of comparison between V_p in (22) and V_T for the entire population. The shaded regions indicate that the inequality $V_p > V_T$ holds, i.e., population subdivision reduces the variance of gene diversity. It

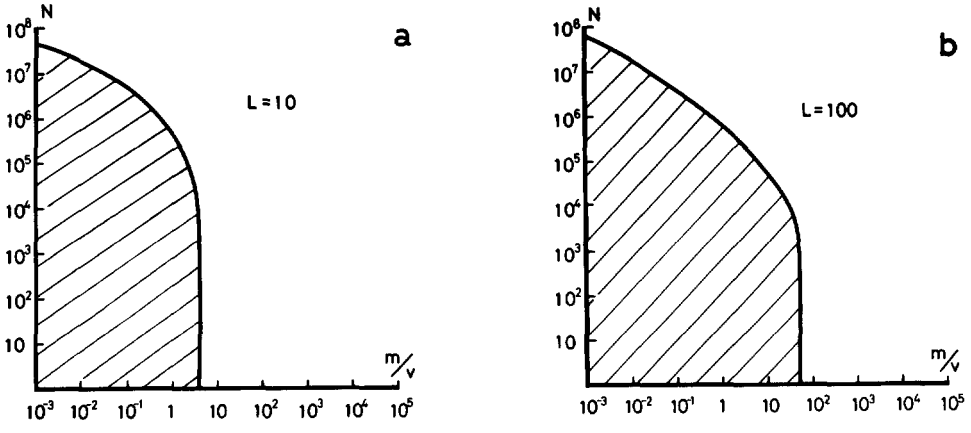

FIGURE 1.—Parameter region where the variance of gene identity in the entire population is smaller than that expected in a panmictic population with the same mean gene identity. The ordinate represents the local population size, and the abscissa represents the ratio of migration rate to mutation rate. The infinite allele model is used. The number of colonies in a whole population is 10 (a) and 100 (b). $v = 10^{-6}$.

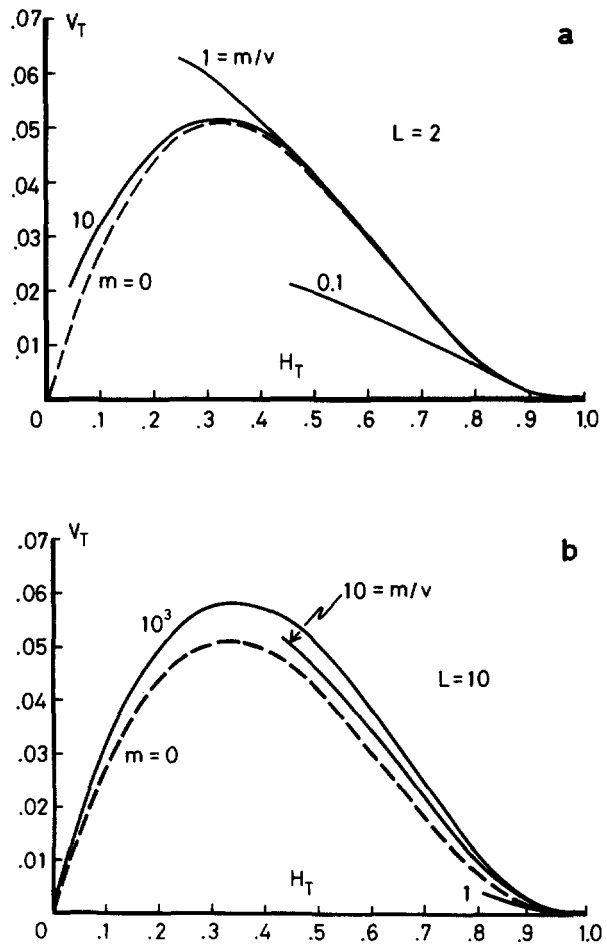

FIGURE 2.—Relationship between the mean (H_T) and variance (V_T) of gene identity in the entire population. The infinite allele model is used. The number beside each curve represents the ratio of migration rate to mutation rate which is kept constant in drawing each curve. a, Corresponds to the case of $L = 2$; b, corresponds to the case of $L = 10$. The case of $L = 100$ is not presented here since the relationship does not differ significantly from that for a panmictic population (dashed lines). $v = 10^{-6}$.

should be noted that the region depends not only on the ratio m/v but also on the number of colonies L. For the theoretical value of H_T to be within the observed range, gene flow must be frequent as mentioned before. Under this condition, population subdivision increases the variance of gene diversity (Figure 2). The large variance of V_T as is found in the case of intermediate gene flow and small values of Nv results from a large value of V_b. The situation is most remarkable when the number of colonies is small. Under intermediate gene flow and small Nv and L, either similar or dissimilar allele can be quasifixed in different colonies from time to time so that F_3 and J_1 tend to be 0.5. In other words, the probability density of j_1 is U-shaped. This is shown in the case of $m/v = 1$, $Nv = 10^{-4}$ and $L = 2$ in Table 2. As L increases, the probability

TABLE 2

Mean and variance (in parentheses) of gene identity; K = ∞

	m/v		$Nv = 10^{-4}$	$Nv = 10^{-3}$	$Nv = 10^{-2}$	$Nv = 10^{-1}$
L = 2	0.1	j_0	1.00 (0.0001)	0.996 (0.0014)	0.958 (0.0129)	0.696 (0.0504)
		j_1	0.091 (0.0825)*	0.091 (0.0816)*	0.087 (0.0730)*	0.063 (0.0293)*
		j_T	0.545 (0.0207)	0.543 (0.0206)	0.523 (0.0200)	0.380 (0.0144)
	1	j_0	0.999 (0.0002)	0.994 (0.0020)	0.943 (0.0169)	0.625 (0.0488)
		j_1	0.500 (0.2496)*	0.497 (0.2460)*	0.472 (0.2144)*	0.313 (0.0768)*
		j_T	0.750 (0.0625)*	0.746 (0.0622)*	0.708 (0.0599)*	0.469 (0.0373)
	10	j_0	0.999 (0.0003)	0.992 (0.0025)	0.929 (0.0211)*	0.567 (0.0477)*
		j_1	0.908 (0.0825)*	0.902 (0.0812)*	0.845 (0.0753)*	0.515 (0.0524)*
		j_T	0.954 (0.0208)*	0.947 (0.0222)*	0.887 (0.0340)*	0.541 (0.0455)*
L = 10	10	j_0	0.998 (0.0008)	0.978 (0.0072)	0.813 (0.0410)	0.304 (0.0164)
		j_1	0.525 (0.0634)*	0.515 (0.0619)*	0.428 (0.0487)*	0.160 (0.0078)*
		j_T	0.572 (0.0514)*	0.561 (0.0506)*	0.467 (0.0419)*	0.174 (0.0074)*
	10^2	j_0	0.996 (0.0012)	0.964 (0.0115)	0.730 (0.0540)*	0.213 (0.0112)*
		j_1	0.914 (0.0300)*	0.885 (0.0369)*	0.670 (0.0596)*	0.195 (0.0105)*
		j_T	0.922 (0.0245)*	0.893 (0.318)*	0.676 (0.0572)*	0.197 (0.0104)*
	10^3	j_0	0.996 (0.0014)	0.962 (0.0132)	0.716 (0.0569)*	0.201 (0.0109)*
		j_1	0.987 (0.0049)*	0.953 (0.0160)*	0.710 (0.0574)*	0.199 (0.0108)*
		j_T	0.988 (0.0043)*	0.954 (0.0154)*	0.710 (0.0571)*	0.200 (0.0108)*

* The variance is greater than that expected from STEWART's formula for a given mean value of gene identity.

that two randomly chosen colonies are genetically identical decreases, reducing the variance of j_1 but still $V_p > V_b$. Thus, large variances of V_b and V_T are expected under intermediate gene flow between a small number of colonies. For a subdivided population consisting of a large number of colonies such as $L = 100$, we cannot expect a large variance of gene identities. This is because under the situation, the probability density of j_1 tends to be *J*-shaped, *i.e.*, it is more likely that any pair of colonies is genetically dissimilar.

COEFFICIENT OF GENE DIFFERENTIATION AND CORRELATION OF HETEROZYGOSITY

Table 3 shows some numerical results of the coefficient of gene differentiation G_{st} (or G_{st}^*) and the correlation of heterozygosity, R. These two inversely correlated quantities measure the degree of genetic differentiation or similarity between colonies. It is obvious from the table and (20) that large values of m/v and Nv increase the genetic similarity and prevent colonies from local differentiation. Although the larger the number of colonies the larger the gene flow required for panmixia of the entire population, it is interesting to examine the L dependence of these statistics keeping m/v constant.

G_{st} is rather insensitive to the change of the number of L and becomes $G_{st} \approx 1/(1 + 4N(m + v))$ for large K and L. Actually, the value for $L = 10$ does not differ much from that for $L = \infty$. On the other hand, the value of R depends

TABLE 3

Coefficient of gene differentiation and correlation of heterozygosity; $K = \infty$

	m/v		$Nv = 10^{-4}$	$Nv = 10^{-3}$	$Nv = 10^{-2}$	$Nv = 10^{-1}$
$L = 2$	1	G_{st}	0.998	0.977	0.807	0.294
		G_{st}^*	0.997	0.969	0770	0.288
		R	0.0002	0.002	0.018	0.103
	10	G_{st}	0.984	0.856	0.373	0.056
		G_{st}^*	0.929	0.513	0.163	0.054
		R	0.005	0.049	0.309	0.694
	10^2	G_{st}	0.862	0.383	0.059	0.006
		G_{st}^*	†	†	0.031	0.006
		R	0.055	0.353	0.806	0.957
$L = 10$	10	G_{st}	0.995	0.949	0.650	0.157
		G_{st}^*	0.994	0.947	0.644	0.156
		R	0.001	0.012	0.086	0.249
	10^2	G_{st}	0.953	0.667	0.167	0.020
		G_{st}^*	0.935	0.604	0.158	0.020
		R	0.035	0.251	0.688	0.870
	10^3	G_{st}	0.669	0.168	0.020	0.002
		G_{st}^*	†	†	0.019	0.002
		R	0.285	0.766	0.956	0.986
$L = 100$	10^2	G_{st}	0.960	0.708	0.195	0.024
		G_{st}^*	0.960	0.708	0.195	0.024
		R	0.010	0.071	0.185	0.255
	10^3	G_{st}	0.710	0.197	0.024	0.002
		G_{st}^*	0.705	0.196	0.024	0.002
		R	0.234	0.634	0.808	0.863
	10^4	G_{st}	0.197	0.024	0.002	0.0002
		G_{st}^*	0.191	0.024	0.002	0.0002
		R	0.725	0.940	0.977	0.985

† Approximation of (17) is invalid.

markedly on L. For instance, when $Nv = 0.01$ and $Nm = 1$, $R = 0.688$ for $L = 10$ and equals 0.185 for $L = 100$. This L dependence of R is caused by a significant change of F_4, which in turn depends heavily on L. However, when we want to estimate the degree of local differentiation from observations without knowledge of L, a statistic sensitive to L may not give a correct estimate.

DISTRIBUTION OF GENE IDENTITY AND EFFECTS OF SELECTION

The distribution of gene identity and the effect of selection were studied by computer simulation for the case of $K = 4$ and $L = 10$. The probability density of j_T subject to intermediate gene flow is shown in Figure 3. Let us first examine the case of neutral mutations. When migration occurs rather frequently ($Nm =$

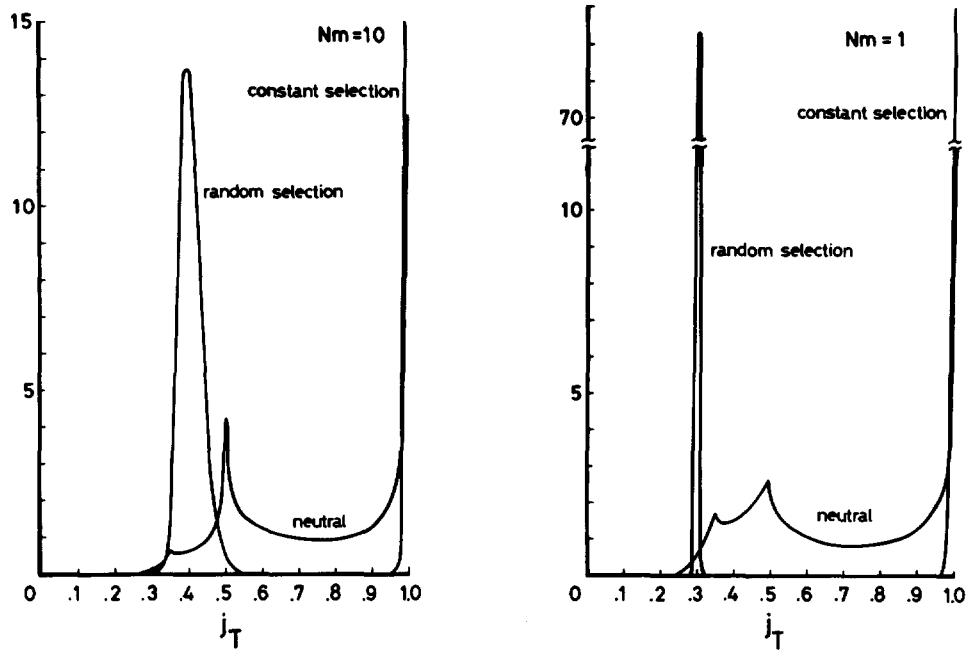

FIGURE 3.—Distribution of gene identity in the entire population in the case of $K = 4$, $Nv = 0.01$, $Ns = 10$ and $L = 10$. The number of migrants per generation between colonies is 10 (a) and 1 (b). Three curves are plotted in each figure, corresponding to the neutral, random selection and constant selection models. The abscissa is gene identity for the entire population and the ordinate is the corresponding probability density. $v = 10^{-3}$. See the text for details.

10), the pattern of the distribution is qualitatively the same as that for a panmictic population with the same parameters (see STEWART 1976). As Nm decreases, however, the spikes of the distribution at $j_T = \frac{1}{2}$ and $\frac{1}{3}$ become moderate (Figure 3b) and eventually disappear. Instead, a new peak emerges around the mean value of j_1 due to the similarity between colonies. Note, however, that this does not necessarily mean that the distribution of gene identity between colonies has a peak near its mean value. Actually, when $Nm = 0.1$ and $Nv = 0.01$, this distribution is U-shaped with the mean and variance being 0.434 and 0.139, respectively. Under these circumstances, two different colonies can take both genetically similar and dissimilar states to each other as time goes on (because of intermediate gene flow and small K relative to L). The proportion of genetically similar colonies to the total colonies at any given time and it's time average determine the position of a new peak. Thus, in contrast to the distribution of j_1, the j_T distribution can be unimodal in the intermediate range of j_T even though NLv is smaller than 1. This pattern forms a contrast to that in a panmictic population.

The effect of selection on the distribution of j_T for the case if $K = 4$ is conspicuous in both selection models. The random selection model is where the normal type allele in each colony is determined at random, and the constant selection model is where the normal type allele is the same in all colonies. The

distribution in either model has a single sharp peak; the position, however, depends on the selection model used. In the constant selection model, the position is always near 1 because of the high frequency of the common advantageous allele in any colony. On the other hand, in the random selection model, there is the possibility that only a few colonies have a common advantageous allele. In the present simulation, the number of combinations of a pair of colonies which have a common advantageous allele was 10. As the number of all different combinations of two colonies out of $L = 10$ is 45, the proportion of the combinations was 2/9. And this proportion in turn mainly determines the position of a peak of the distribution. Thus, the position shifts toward 0 as K/L increases. Another interesting feature is the width of the distribution which depends not only on the magnitude of Ns (where s is selection coefficient) but also on Nm. As shown by SLATKIN (1973), a population cannot respond to local selection when gene flow is large relative to the strength of selection (see also SLATKIN and MARUYAMA 1975; FELSENSTEIN 1975; WALSH 1983). In such a case, the distribution will be broad.

We can confirm a well-known effect of random selection on the maintenance of polymorphism (LEVENE 1953 and see pages 258–262 in FELSENSTEIN 1976); in our simulation J_T reduces to 0.399 and 0.297 from 0.732 and 0.663, respectively (Figure 3). In fact, the random selection model is an efficient mechanism for maintaining genetic polymorphism. At the same time the variance of j_T is greatly decreased. On the other hand, the inbreeding coefficient or the coefficient of gene differentiation G_{st}^* is increased, although the variance $V_{g_{st}}$ is decreased compared with the case of neutral mutations. Our simulation result is that $G_{st}^* = 0.812$ for $Nm = 1$ and is 0.117 for $Nm = 10$ which are 6 and 10 times larger than those for neutral mutations. The increased mean value, G_{st}^*, comes entirely from the occurrence of genetically similar colonies in a population.

SLATKIN (1977) and MARUYAMA and KIMURA (1980) studied the effect of local extinction and recolonization of colonies on genetic variation and showed that the effective population size is greatly reduced compared with the case of the absence of this effect. As this effect reduces between-colony differentiation, G_{st} is also reduced. In other words, extinction and subsequent recolonization of colonies is a mechanism equivalent to that of mass migration, counterbalancing the reduction of the effective size. In terms of the variance of gene identity, this process makes not only j_1 but also j_T approach STEWART'S relationship.

I thank BRUCE WALSH, YOSHIO TATENO, MASATOSHI NEI and an anonymous reviewer for their suggestions and comments which greatly improved the manuscript. I am also grateful to MONTGOMERY SLATKIN for his critical reading of the manuscript and CURTIS STROBECK for his interest and unpublished paper with G. B. GOLDING.

LITERATURE CITED

CROW, J. F. and T. MARUYAMA, 1971 The number of neutral alleles maintained in a finite, geographically structured population. Theor. Pop. Biol. **2**: 437–453.

FELSENSTEIN, J., 1975 Genetic drift in clines which are maintained by migration and natural selection. Genetics **81**: 191–207.

FELSENSTEIN, J., 1976 The theoretical population genetics of variable selection and migration. Annu. Rev. Genet. **10**: 253–280.

FUERST, P. A., R. CHAKRABORTY and M. NEI, 1977 Statistical studies on protein polymorphism in natural populations. I. Distribution of single locus heterozygosity. Genetics **86:** 455–483.

GOJOBORI, T., 1982 Means and variances of heterozygosity and protein polymorphism. pp. 137–148. In: *Molecular Evolution, Protein Polymorphism and the Neutral Theory*, Edited by M. KIMURA. Japan Scientific Societies Press, Tokyo, and Springer-Verlag, Berlin.

KIMURA, M., 1964 Diffusion models in population genetics. J. Appl. Probab. **1:** 177–232.

KIMURA, M., 1968a Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. Genet. Res. **11:** 247–269.

KIMURA, M., 1968b Evolutionary rate at the molecular level. Nature **217:** 624–626.

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

KIMURA, M. and N. TAKAHATA, 1983 Selective constraint in protein polymorphism: study of the effectively neutral mutation model by using an improved pseudosampling method. Proc. Natl. Acad. Sci. USA **80:** 1048–1052.

KIMURA, M. and G. H. WEISS, 1964 The stepping stone model of population structure and the decrease of genetic correlation with distance. Genetics **49:** 561–576.

LATTER, B. D. H., 1973 The island model of population differentiation: a general solution. Genetics **73:** 147–157.

LEVENE, H., 1953 Genetic equilibrium when more than one ecological niche is available. Am. Nat. **87:** 331–333.

LI, W.-H. and M. NEI, 1975 Drift variances of heterozygosity and genetic distance in transient states. Genet. Res. **25:** 229–248.

LI, W.-H. and M. NEI, 1977 Persistence of common alleles in two related populations or species. Genetics **86:** 901–914.

MALECOT, G., 1951 A stochastic treatment of linear problems (mutation, linkage, migration) in population genetics. Ann. Univ. Lyon Sci. (Sect. A) **14:** 79–117.

MALECOT, G., 1955 Remarks on decrease of relationship with distance. Cold Spring Harbor Symp. Quant. Biol. **20:** 52–53.

MARUYAMA, T., 1969 Genetic correlation in the stepping stone model with non-symmetrical migration rate. J. Appl. Probab. **6:** 463–477.

MARUYAMA, T., 1970a Effective number of alleles in a subdivided population. Theor. Pop. Biol. **1:** 273–306.

MARUYAMA, T., 1979b Stepping stone models of finite length. Adv. Appl. Probab. **2:** 229–258.

MARUYAMA, T., 1979c Analysis of population structure. I. One dimensional stepping stone models of finite length. Ann. Hum. Genet. **34:** 201–219.

MARUYAMA, T. and M. KIMURA, 1980 Genetic variability and effective population size when local extinction and recolonization of subpopulations are frequent. Proc. Natl. Acad. Sci. USA **77:** 6710–6714.

MAYNARD SMITH, J., 1970 Population size, polymorphism, and rate of non-Darwinian evolution. Am. Nat. **104:** 231–236.

NEI, M., 1972 Genetic distance between populations. Am. Nat. **106:** 283–292.

NEI, M., 1973 Analysis of gene diversity in subdivided populations. Proc. Natl. Acad. Sci. USA **70:** 3321–3323.

NEI, M., 1975 *Molecular Population Genetics and Evolution*. American Elsevier, North Holland, New York.

NEI, M., R. CHAKRABORTY and P. A. FUERST, 1976a Testing the neutral mutation hypothesis by distribution of single locus heterozygosity. Nature **262:** 491–493.

NEI, M., R. CHAKRABORTY and P. A. FUERST, 1976b Infinite allele model with varying mutation rate. Proc. Natl. Acad. Sci. USA **73:** 4164–4168.

NEI, M. and A. CHAKRAVARTI, 1977 Drift variances of F_{st} and G_{st} statistics obtained from a finite number of isolated populations. Theor. Pop. Biol. **11:** 307–325.

NEI, M., A. CHAKRAVARTI and Y. TATENO, 1977 Mean and variance of F_{st} in a finite number of incompletely isolated populations. Theor. Pop. Biol. **11:** 291–306.

NEI, M. and M. FELDMAN, 1972 Identity of genes by descent within and between populations under mutation and migration pressures. Theor. Pop. Biol. **3:** 469–465.

SLATKIN, M., 1973 Gene flow and selection in a cline. Genetics **75:** 733–756.

SLATKIN, M., 1977 Gene flow and genetic drift in a species subject to frequent local extinctions. Theor. Pop. Biol. **12:** 253–262.

SLATKIN, M., 1981 Estimating levels of gene flow in natural populations. Genetics **99:** 323–335.

SLATKIN, M., 1982 Testing neutrality in subdivided populations. Genetics **100:** 533–545.

SLATKIN, M. and T. MARUYAMA, 1975 Genetic drift in a cline. Genetics **81:** 209–222.

STEWART, F. M., 1976 Variability in the amount of heterozygosity maintained by neutral mutations. Theor. Pop. Biol. **9:** 188–201.

WALSH, J. B., 1983 Conditions for protection of an allele in linear homogeneous stepping stone models. Theor. Pop. Biol. In press.

WEISS, G. H. and M. KIMURA, 1965 A mathematical analysis of the stepping stone model of genetic correlation. J. Appl. Probab. **2:** 129–149.

WRIGHT, S., 1943 Isolation by distance. Genetics **28:** 114–138.

WRIGHT, S., 1946 Isolation by distance under diverse systems of mating. Genetics **31:** 39–59.

WRIGHT, S., 1951 The genetical structure of populations. Ann. Eugen. **15:** 323–354.

YAMAZAKI, T., 1976 Enzyme polymorphism and functional difference: mean, variance, and distribution of heterozygosity. pp. 189–225. In: *Proceedings of the Second Taniguchi International Symposium on Biophysics, Molecular Evolution, and Polymorphism*, Edited by M. KIMURA. Academic Press, New York.

Corresponding editor: M. NEI

APPENDIX

The matrix C in (12) giving the fourth moments is given by

$$C = \begin{pmatrix} -c_1 & 8M & 0 & 0 & 0 & 0 & 0 \\ 2M^* & -c_2 & 4M^* & 2M^* & 2(L-2)M^* & 4(L-2)M^* & 0 \\ 0 & 8M^* & -c_3 & 0 & 0 & 8(L-2)M^* & 0 \\ 0 & 8M^* & 0 & -c_4 & 8(L-2)M^* & 0 & 0 \\ 0 & 4M^* & 0 & 4M^* & -c_5 & 8M^* & 4(L-3)M^* \\ 0 & 8M^* & 4M^* & 0 & 4M^* & -c_6 & 4(L-3)M^* \\ 0 & 0 & 0 & 0 & 8M^* & 16M^* & -c_7 \end{pmatrix}$$

where $c_1 = 6 + 8K\theta^* + 8M$, $c_2 = 3 + 8K\theta^* + (6L-4)M^*$, $c_3 = c_4 = 2 + 8K\theta^* + 8M$, $c_5 = c_6 = 1 + 8K\theta^* + 4(L+1)M^*$ and $c_7 = 8K\theta^* + 24M^*$.