Copyright © 1996 by the Genetics Society of America

Inferring Patterns of Migration From Gene Frequencies Under Equilibrium Conditions

Jarle Tufto,* Steinar Engen† and Kjetil Hindar*

** Norwegian Institute for Nature Research, 7005 Trondheim, Norway and † Department of Mathematics and Statistics, Norwegian University of Science and Technology, 7034 Trondheim, Norway*

Manuscript received May 27, 1996
Accepted for publication August 23, 1996

ABSTRACT

A new maximum likelihood method to simultaneously estimate the parameters of any migration pattern from gene frequencies in stochastic equilibrium is developed, based on a model of multivariate genetic drift in a subdivided population. Motivated by simulations of this process in the simplified case of two subpopulations, problems related to the nuisance parameter q, the equilibrium gene frequency, are eliminated by conditioning on the observed mean gene frequency. The covariance matrix of this conditional distribution is calculated by constructing an abstract process that mimics the behavior of the original process in the subspace of interest. The approximation holds as long as there is limited differentiation between subpopulations. The bias and variance of estimates of long-range and short-range migration in a finite stepping stone model are evaluated by fitting the model to simulated data with known values of the parameters. Possible ecological extensions of the model are discussed.

THE pattern of migration (or gene flow) between a set of geographically separated populations reflects a large number of ecological and genetic processes. First, migration is often restricted to relatively short distances (*e.g.*, LEVIN and KERSTER 1974). Second, if carrying capacities vary between populations, optimality models predict that individuals should adopt a conditional dispersal strategy responding to the differences in local carrying capacities (HOLT 1985; JOHNSON and GAINES 1990). Finally, effective rates of migration may be modified by the breeding system (ANDERSSON 1994) and by geographic variation in selection (ENDLER 1986). Actual estimates of migration are important because they suggest how important migration is for the species, for example in limiting the development of local adaptations (*e.g.*, SLATKIN 1973; NAGYLAKI 1975).

In subdivided populations, when rates of migration are high, migration interacts with genetic drift occurring in each subpopulation, and the gene frequencies will then, after a number of generations, reach a stationary equilibrium distribution. Under the island model this distribution is the beta (WRIGHT 1931). For more complicated migration patterns such as stepping stone models (KIMURA and WEISS 1964) and models of spatially continuous populations (MALÉCOT 1975), only the variances around the equilibrium gene frequencies and the correlations between populations at different distances have been found analytically. If there is a limited amount of migration, and if effective population sizes are large, the number of generations necessary to reach equilibrium distribution may be very large, and the observed genetic structure may then reflect the initial historic genetic composition of the populations.

Evolutionary forces such as mutation and selection can also be important in determining geographic genetic variation. For many loci, and for many problems in population genetics, however, it is reasonable to assume that these forces are small compared to migration and drift, and that they therefore can be neglected (CROW 1985; AVISE 1994).

Given the large amount of already existing data on geographic genetic variation, it should be of great interest to develop models that make inferences about general migration patterns possible. Previous approaches to this problem have mostly been based on the expected amount of genetic differentiation under the island model as measured by the parameter F_{st} (WRIGHT 1951), estimated for all pairs of subpopulations or for all subpopulations taken together. Using the theory of the island model, some overall measure of gene flow is then calculated. While this approach has verified that the genetic correlations decreases with distance as predicted by *e.g.*, stepping stone models (SLATKIN 1993), it is clear that the assumptions of the island model are not valid in general.

The dependencies between the gene frequencies between the subpopulations are an important feature of the data that must be incorporated in a general model for the problem. Except under the island model, it is these dependencies that contain most of the information about the unknown migration pattern. Here, using some ideas in FELSENSTEIN (1982), we develop a model that can be used to estimate the parameters of any migration pattern by maximum likelihood. The model

Corresponding author: Jarle Tufto, Norwegian Institute for Nature Research, Tungasletta 2, 7005 Trondheim, Norway.
E-mail: jarle.tufto@nina.nina.no

is based on the underlying multivariate genetic drift process that generates the data. An attempt is made to eliminate the unknown equilibrium gene frequencies from the model by considering the distribution generated by the drift process, conditioned on the sufficient statistics for these nuisance parameters. Using simulations, we show that it is reasonable to approximate this conditional distribution by the multivariate normal. A method for calculating the covariance matrix of the distribution similar to COURGEAU (1974) is suggested, based on insights gained from further simulations. Finally, to evaluate the properties such as bias and efficiency of parameter estimates obtained using the model, we estimate the parameters of an example migration pattern, a finite stepping stone model, from simulated data.

THE GENERAL MODEL

Consider $n + 1$ populations indexed $i = 1, 2, \ldots, n + 1$. Let N_i be the variance effective size (EWENS 1979, eq. 3.96) of population i. We will only consider the simplest case of a single diallelic locus with two alleles A_1 and A_2. Let the elements of the column vector $\mathbf{p}_t$ represent the allele frequencies of A_1 in the populations in generation t, and let m_{ij} be the probability that an individual born in population i received a gene from a parent in population j. Each row of the $(n + 1) \times (n + 1)$ migration matrix $\mathbf{M}^* = [m_{ij}]$ therefore sums to one. The gene frequency in the $(n + 1)$th population remains constant and equal to q, that is, this population is of infinite effective size, and can thus be thought of as a large outside world population. The $(n + 1)$th column of $\mathbf{M}^*$ represents the immigration rates from this outside world into each subpopulation. These immigration rates can in general be different. We will let these immigration rates also include mutations, since the effects of mutations are indistinguishable from the effects of immigration from the outside world.

Apart from these assumptions, $\mathbf{M}^*$ can take any form depending on what assumptions we make about the underlying migration pattern. The migration matrix is generally a function of the parameters of some migration pattern model.

If we include genetic drift, the gene frequencies in generation $t + 1$ may be expressed as

$$\mathbf{p}_{t+1} = \mathbf{M}^*\mathbf{p}_t + \mathbf{e}, \qquad (1)$$

where the elements of $\mathbf{e}$ represent the stochastic changes in the process. The elements of $\mathbf{e}$ are binomial variables rescaled to have zero expectations and variances equal to $p_{t,i}(1 - p_{t,i})/2N_i$, except e_{n+1} that always equals zero. Also note that we assume that the changes in the gene frequencies due to migration and drift are small so that the sequential order of the events in the life cycle can be ignored.

Substituting the gene frequencies with their deviation $x_i = p_i - q$ from the equilibrium gene frequency, we see that (1) can be rewritten as

$$\mathbf{x}_{t+1} = \mathbf{M}\mathbf{x}_t + \mathbf{e}, \qquad (2)$$

where $\mathbf{x}_t$ is an n-dimensional column vector and $\mathbf{M}$ is a $n \times n$ matrix that equals $\mathbf{M}^*$ except that the $(n + 1)$th row and column have been dropped. The sum of each row of $\mathbf{M}$ is consequently equal to or less than one.

As noted by BODMER and CAVALLI-SFORZA (1968) and FELSENSTEIN (1982), the variances of the elements of $\mathbf{e}$ depend on the gene frequencies. To make the variances constant, one might use the arcsine square root transformation, but this also changes the expectations in the process. In fact, no transformation exists that will make both the variances constant and the expectations linear in p, and a more careful analysis will show that such a transformation is not necessary, at least to derive the covariance matrix of the stationary distribution of this multivariate process.

This derivation can be done as follows. We first want to find the recursion relation between the variances and covariances from one generation to the next. Multiplying each side of (2) with their own transposed yields

$$\mathbf{x}_{t+1}\mathbf{x}_{t+1}^T = \mathbf{M}\mathbf{x}_t\mathbf{x}_t^T\mathbf{M}^T + \mathbf{M}\mathbf{x}_t\mathbf{e}^T + (\mathbf{M}\mathbf{x}_t)^T\mathbf{e} + \mathbf{e}\mathbf{e}^T. \qquad (3)$$

The elements of the matrices in the second and third term on the right hand side, formed by taking products of row and column vectors, involve products of the stochastic variables e_i and $x_{j,t}$. Even though these stochastic variables are dependent for $i = j$, we always have $E(e_i x_{j,t}) = 0$ since $E(e_i \mid x_{j,t}) = 0$. If we take expectations of (3), these terms therefore vanish and we get

$$\mathbf{C}_{t+1} = \mathbf{M}\mathbf{C}_t\mathbf{M}^T + E(\mathbf{e}\mathbf{e}^T), \qquad (4)$$

where $\mathbf{C}_t$ is the covariance matrix of the distribution at time t.

It remains to evaluate the expectation of the last term in (4). Because the elements of $\mathbf{e}$ are independent and with zero expectations, only the elements of the matrix $\mathbf{e}\mathbf{e}^T$ along the diagonal have nonzero expectations. If we first write each element of the column vector $\mathbf{e}$ as

$$e_i = e_{0,i}\sqrt{\frac{p_{t,i}(1 - p_{t,i})}{2N_i}}, \qquad (5)$$

where $e_{0,i}$ is stochastic with variance equal to one and expectation equal to zero, then we see that

$$E(e_i e_i) = E\left(\frac{e_{0,i}^2}{2N_i} p_{i,t}(1 - p_{i,t})\right) = \frac{1}{2N_i}(Ep_{i,t} - E(p_{i,t}^2))$$

$$= \frac{1}{2N_i}(Ep_{i,t} - (Ep_{i,t})^2 - E(p_{i,t}^2) + (Ep_{i,t})^2)$$

$$= \frac{1}{2N_i}(q(1 - q) - c_{ii,t}), \qquad (6)$$

since $E(p_{t,i}) = q$ and since $E(p_{i,t}^2) - (Ep_{i,t})^2 = \mathrm{Var}(p_{i,t}) = c_{ii,t}$. The average genetic drift in the process is thus reduced by an amount proportional to the variance c_{ii} around the equilibrium gene frequency q, as also shown by COURGEAU (1974) p. 365.

Substituting (6) into (4) and noting that the covariance matrix $\mathbf{C}_{t+1}$ must equal $\mathbf{C}_t$ as t tends to infinity and a stationary distribution is attained, we now know that the covariance matrix $\mathbf{C}$ of this stationary distribution must satisfy the equation

$$\mathbf{C} = \mathbf{M}\mathbf{C}\mathbf{M}^T + \mathbf{E}, \tag{7}$$

where the matrix $\mathbf{E} = E(\mathbf{e}\mathbf{e}^T)$. Note that $\mathbf{E}$ depends on $\mathbf{C}$. Equation 7 can be rewritten to a system of linear equations in the $n(n+1)/2$ unknown covariances c_{ij} and solved for $\mathbf{C}$.

A more formal proof of the existence of this limit is given in COURGEAU (1974). It should be noted that the solution of (7) is exact to the extent that the order of the events in the life cycle can be ignored.

FITTING THE MODEL TO A GENETIC SAMPLE

Some introductory remarks: Typically, in studies on genetic differentiation, a large number of individuals have been sampled from a set of subpopulations $i = 1, 2, \ldots, n$, and the frequencies in each subpopulation $p_1, p_2, \ldots, p_n$ of different alleles have been determined by *e.g.*, protein electrophoresis or restriction fragment length polymorphism (RFLP) analysis. Our interest is to make inferences about the parameters of the underlying migration pattern from these gene frequencies. Formally, this can be done by assuming that the population system has reached its stationary distribution, then use this stationary distribution of the process as the probability distribution for the data, and finally estimate the parameters of the model by maximizing the probability of the observations.

An additional difficulty is however introduced by the parameter q, the frequency of the long-range migrants, which in general will be unknown. We have no interest in making inferences about q, that is, q is a nuisance parameter in the model. Sufficient statistics are important in models containing such nuisance parameters because they contain all information in the data about the unknown parameter. The sampling distribution of the model, conditioned on some sufficient statistics t for the nuisance parameter θ, is the relevant distribution to consider, because this distribution, by definition, is independent of the true value of the nuisance parameter. By conditioning on t, we restrict our attention to only a small part of the sample space, and ignore other possible outcomes that are irrelevant for the problem. In the context of maximum likelihood estimation, this principle is called conditional likelihood (MCCULLAGH and NELDER 1989, ch. 7).

In the present case, with n subpopulations in the system, the relevant distribution that we seek, on which calculations of the likelihood must be based, is the $(n-1)$-dimensional stationary distribution of the process, conditioned on some sufficient statistic for q. Because of the complexity of the generated distribution, finding a sufficient statistic for q and finding the corresponding conditional distribution, will necessarily have to be based on some approximations.

Approximations for small fluctuations: If the fluctuations around q are small, for example if there is a high rate of immigration from the outside world into each subpopulation, then the genetic drift will be nearly constant, and the process can then be approximated by a multivariate autoregressive process. This process is known to have the multivariate normal as its stationary distribution. Since $E(p_i) = q$ for all populations, we know from the properties of the multivariate normal, that the weighted mean gene frequency

$$\bar{p} = \sum_{i=1}^{n} w_i p_i \tag{8}$$

is sufficient for q, provided that the weights $w_1, \ldots, w_n$ are chosen to minimize the variance of $\bar{p}$ (APPENDIX B).

It can also be shown that the multivariate normal distribution conditioned on the linear combination $\bar{p} = \sum_{i=1}^{n} w_i p_i$ is also multivariate normal (KENDALL *et al.* 1983, exercise 15.1). Also, if we as suggested by FELSENSTEIN (1982), work with the deviations of each gene frequency from the weighted sample mean, that is, an $(n-1)$-dimensional vector $\mathbf{y}$ where $y_i = p_i - \bar{p}$, then the distribution of $\mathbf{y}$ conditional on $\bar{p}$ is independent, not only of q, but also of $\bar{p}$ (APPENDIX C).

The vector $\mathbf{y}$ may be expressed as a linear matrix transformation of $\mathbf{p}$,

$$\mathbf{y} = \mathbf{K}\mathbf{p}, \tag{9}$$

where the $(n-1) \times (n)$ matrix

$$\mathbf{K} = \begin{bmatrix} 1-w_1 & -w_2 & \cdots & -w_{n-1} & -w_n \\ -w_1 & 1-w_2 & \cdots & -w_{n-1} & -w_n \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -w_1 & -w_2 & \cdots & 1-w_{n-1} & -w_n \end{bmatrix}. \tag{10}$$

The unconditional covariance matrix of $\mathbf{y}$ is then

$$\mathbf{C}_y = \mathbf{K}\mathbf{C}\mathbf{K}^T. \tag{11}$$

Again, since $\mathbf{y} \mid \bar{p}$ is independent of $\bar{p}$ (APPENDIX C) it follows that the conditional covariance matrix $\mathbf{C}_{y|\bar{p}}$ that we seek equals $\mathbf{C}_y$ given by (11).

We can therefore consider (11), when $\mathbf{C}$ is calculated from (7), to be a naive approximation of $\mathbf{C}_{y|\bar{p}}$. By making some further distributional assumptions, the likelihood can be calculated. Our main concern at this stage, however, is how well (11) approximates $\mathbf{C}_{y|\bar{p}}$ when the fluctuations around the equilibrium gene frequency q become large. This will be investigated in the next subsection. It still seems reasonable, however, to rely on the assumption that the distribution of $\mathbf{p}$, when conditioned on (8), is approximately independent of q, also more generally.

Simulations of the two population case: To get an impression of the behavior of the process, we will first simulate its stationary distribution. We will look at the

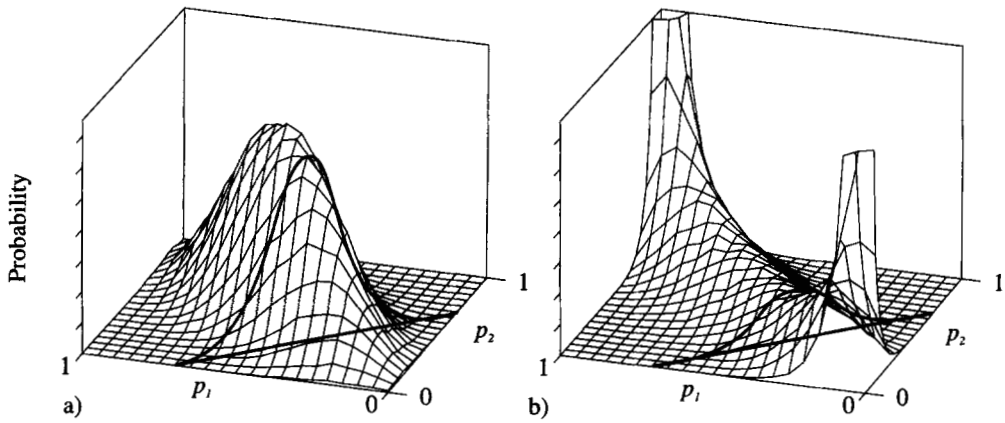

FIGURE 1.—Simulations of the stationary distribution in the two population case with $N = 10$, $m = 0.3$ and $q = 0.5$. (A) When the long-range migration rate is high ($u = 0.1$), the stationary distribution is close to binormal. (B) As the long-range migration rate becomes small ($u = 0.01$), the gene frequencies are close to fixation most of the time and the total distribution is far from binormal. Note that the $(n - 1)$-dimensional conditional distribution indicated by boldface lines is still close to normal.

special case of two populations, each of equal effective size $N = 10$, both receiving long range migrants with gene frequency $q = 0.5$ at rate u. In addition, we will let the two between-population migration rates be equal to m. The migration matrix is then

$$\mathbf{M} = (1 - u)\begin{bmatrix} 1 - m & m \\ m & 1 - m \end{bmatrix}. \qquad (12)$$

We also assume, in the simulations, that genetic drift occurs after migration. Each of the components of the vector $\mathbf{p} = (p_1, p_2)$ can then take any value of the rational numbers 0, $1/20$, $2/20$, ..., 1, and there are $21 \cdot 21 = 441$ possible discrete states in the process. By simulating, say 200,000 generations, and counting the number of generations the population spends in each state, we can find the stationary bivariate distribution of the process. Figure 1 shows this distribution for two different long-range migration rates. We see that the distribution in fact is close to multivariate normal as long as the long-range migration rate u is high. However, as soon as this migration rate becomes small the populations are close to fixation for a large part of the time. Interestingly, the relevant conditional distribution does however still appear to be close to normally distributed, which suggests that the multivariate normal may still be a good approximation of the $(n - 1)$-dimensional conditional distribution that we seek.

We will now look at how well (11), which is based on multivariate normality, approximates the covariance matrix of the conditional distribution generated by the original process for large fluctuations around the equilibrium gene frequency q. We will check this by doing some further simulations, but we will still only consider the simplified case of two subpopulations. There is then just one contrast $y_1 = p_1 - \bar{p} = 1/2(p_1 - p_2)$ and the covariance matrix is the variance of this single contrast. The exact conditional variance of y_1 can be found numerically by doing simulations as described above, if the calculation of this variance is based on the values of y_1 at the points in time when the process is in one of the states

$$\mathbf{p} = (0/20, 20/20), (1/20, 19/20), \ldots, (20/20, 0/20)$$

if the observed mean gene frequency is say, $\bar{p} = 1/2$.

For the unconditional variance we can get an explicit solution. With the symmetry assumptions above it is clear that the variances of p_1 and p_2 are equal, that is, $c_{11} = c_{22}$. Solving (7) for c_{11} and c_{12} we then find, after some algebra, that

$$\begin{aligned}\mathrm{Var}(y_1) = 1/4(c_{11} + c_{22} - c_{12}) &= 1/2\, qu \\ &\times (2 - u - 2q + uq)/(2m - 2m^2 + 2u - u^2 \\ &- 4um + 4um^2 + 2u^2m - 2u^2m^2 - 16Num^2 \\ &- 40Nu^2m + 40Nu^2m^2 + 16Num - 32Nu^3m^2 \\ &+ 8Nu^4m^2 + 8Nu^2 - 8Nu^3 \\ &+ 2Nu^4 + 32Nu^3m - 8Nu^4m). \qquad (13)\end{aligned}$$

Note that this expression tends to zero as the long range migration rate u becomes small (provided that $m > 0$).

Figure 2, A and B, shows the conditional variance (calculated using simulations) and the unconditional variance (calculated using Equation 13). We see that $\mathrm{Var}(y_1)$ only approximates $\mathrm{Var}(y_1 \mid \bar{p})$ when u and m are large. This shows that our naive approximation (11) of the conditional covariance matrix does not hold in general. An alternative method is therefore needed to calculate the conditional covariance matrix of the $n - 1$ contrasts.

An alternative, similarly behaving process: It is interesting to notice (Figure 1B) that the conditional distribution appears to be only weakly dependent on the value of $\bar{p}$. In fact, if the stationary distribution was multivariate normal, then this conditional distribution would be completely independent of $\bar{p}$ (APPENDIX C). This suggests that if we instead look at an alternative process $\tilde{\mathbf{p}}_t$, forced to move only within the subspace

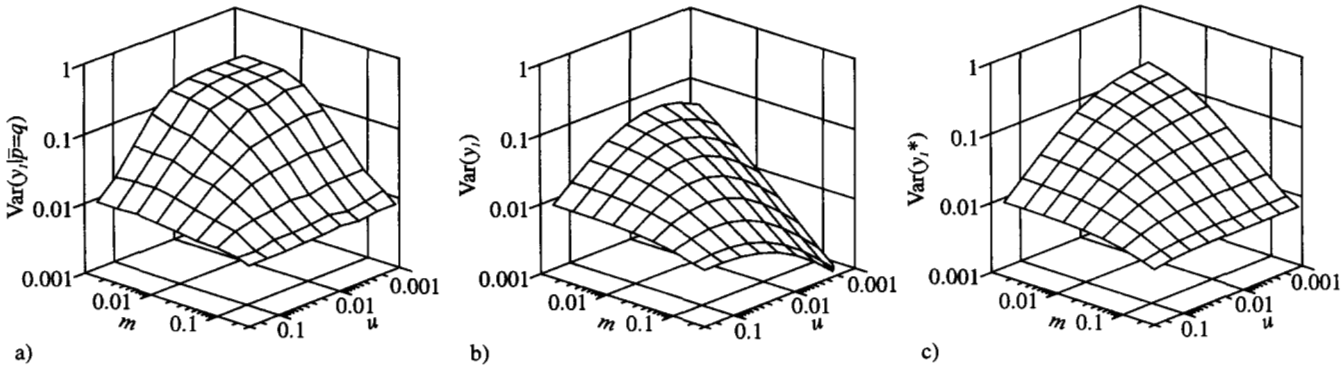

FIGURE 2.—Comparison of $\mathrm{Var}(y_1 \mid \bar{p})$ (calculated using simulations) (A), $\mathrm{Var}(y_1)$ (Equation 13) (B), and $\mathrm{Var}(\tilde{y}_1)$ (C) of the alternative process described in the text (Equation 19).

where $\bar{p}$ equals the observed mean gene frequency, but otherwise change the process as little as possible, the stationary distribution of this *alternative* process may be more similar to the conditional distribution of the original process $\mathbf{p}_t$ that we seek. This modification of the process can be done as follows.

Let $\tilde{\mathbf{x}}^*_{t+1}$ designate the temporary state of the process after migration and drift have occurred. Conditional on that the covariance matrix is $\tilde{\mathbf{C}}_t$ in the previous generation, the covariance matrix of $\tilde{\mathbf{x}}^*_{t+1}$ is, as before,

$$\tilde{\mathbf{C}}^*_{t+1} = \mathbf{M}\tilde{\mathbf{C}}_t\mathbf{M}^T + \mathbf{E}. \tag{14}$$

Remember that $\mathbf{E}$ depends on $\tilde{\mathbf{C}}_t$. To keep the process within the $n-1$-dimensional subspace, we now project $\tilde{\mathbf{x}}^*_{t+1}$ down into the $(n-1)$-dimensional subspace where $\bar{\tilde{x}} = 0$ by subtracting the weighted mean $\overline{\tilde{x}^*_{t+1}}$ from each element of $\tilde{\mathbf{x}}^*_{t+1}$ (Figure 3). This can be done using the transformation

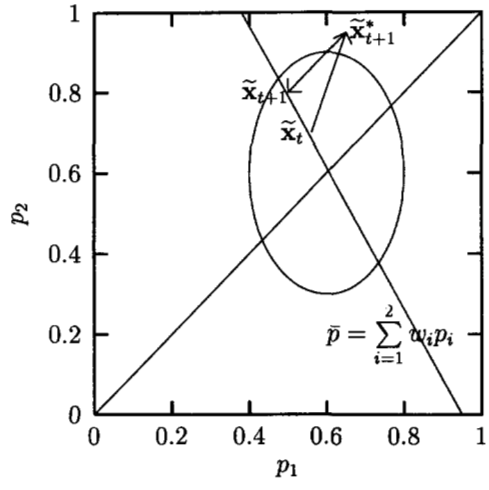

FIGURE 3.—In the multivariate normal case, the distribution conditional on $\bar{p}$ is independent of q, provided that $\bar{p}$ is sufficient for q. The conditional distribution of $\mathbf{y}$ also happens to be independent of $\bar{p}$ itself. This suggests that a process $\tilde{p}_t$ forced to move only within the subspace of interest will generate the relevant conditional distribution of the original process p_t, also when the overall distribution is far from multinormal, provided that the conditional distribution is only weakly dependent on $\bar{p}$.

$$\tilde{\mathbf{x}}_{t+1} = \mathbf{D}\tilde{\mathbf{x}}^*_{t+1}, \tag{15}$$

where element ij of the $n \times n$ matrix $\mathbf{D}$ is

$$d_{ij} = \begin{cases} 1 - w_j & \text{for } i = j \\ -w_j & \text{for } i \neq j. \end{cases} \tag{16}$$

We now know that the covariance matrix of $\tilde{\mathbf{x}}_{t+1}$ is

$$\tilde{\mathbf{C}}_{t+1} = \mathbf{D}\tilde{\mathbf{C}}^*_{t+1}\mathbf{D}^T. \tag{17}$$

Substituting (14) into (17) and noting that $\tilde{\mathbf{C}}_{t+1}$ must equal $\tilde{\mathbf{C}}_t$ as t tends to infinity, we know that the covariance matrix of the stationary distribution satisfies

$$\tilde{\mathbf{C}} = (\mathbf{DM})\,\tilde{\mathbf{C}}\,(\mathbf{DM})^T + \mathbf{DED}^T. \tag{18}$$

This equation is, like (7) for the original process, still a linear system of $n(n+1)/2$ unknown covariance. It must in general be solved numerically (APPENDIX A).

It is interesting to look at the symmetrical two-population case. The solution of (18) is then straightforward, and the variance of the contrast is

$$\mathrm{Var}(\tilde{y}_1) = \tfrac{1}{4}(c_{11} + c_{22} - 2c_{12}) = \bar{p}(1-\bar{p})/ \\ ((16m + 8u - 16m^2 - 4u^2 - 32um + 32um^2 \\ + 16u^2m + 16u^2m^2)N + 1). \tag{19}$$

Figure 2C shows the behavior of this approximation for different values of u and m in the two-population case. Compared to the conditional variance of the original process (Figure 2A), we see that the approximation works reasonably well, at least when there is a high rate of between-population migration m.

Some special cases are of interest. As u tends to zero, and for small m, (19) tends to the limit

$$\mathrm{Var}(\tilde{y}_1) = \frac{\bar{p}(1-\bar{p})}{16Nm + 1}, \tag{20}$$

in contrast to approximation (13) that tends to zero. The simulations of the original process also suggest that this limit is larger than zero. Notice that the variance given the mean (Figure 2A) seems to be independent of u when u is much smaller than the between-population migration rate m.

Another check of result (19) is the special case of m equal to zero, in which the model is equivalent to the island model, with our parameter u representing the parameter usually called m in the island model. For small u, (19) then reduces to

$$\mathrm{Var}(\tilde{y}_1) = \frac{\bar{p}(1-\bar{p})}{8Nu+1}. \qquad (21)$$

The island model predicts that $\mathrm{Var}(p_1) = \mathrm{Var}(p_2) = q(1-q)/(4Nu+1)$. Since p_1 and p_2 are independent, it then follows, from the island model, that

$$\mathrm{Var}(y_1) = \mathrm{Var}(\tfrac{1}{2}(p_1 - p_2)) = \frac{q(1-q)}{8Nu+2}, \qquad (22)$$

which is approximately equal to (21), unless Nu is small.

A third special case is the limit obtained when both u and m tend to zero, that is, when there is large between population differentiation. For the original process all the probability is now concentrated on the "edges" of the unit square, which constitutes the state space of the process. Because we are considering the conditional distribution given $\bar{p}$ there is, in the limit, for $\bar{p} < \frac{1}{2}$, only two equally probable states $\mathbf{p} = (2\bar{p}, 0)$ and $\mathbf{p} = (0, 2\bar{p})$ and it is easy to show that

$$\mathrm{Var}(y_1 \mid \bar{p}) = \begin{cases} \bar{p}^2 & \text{for } \bar{p} < \frac{1}{2} \\ (1-\bar{p})^2 & \text{for } \bar{p} > \frac{1}{2}. \end{cases} \qquad (23)$$

However, in this limit, approximation (19) becomes

$$\mathrm{Var}(\tilde{y}_1) = \bar{p}(1-\bar{p}), \qquad (24)$$

which only equals (23) when $\bar{p} = \frac{1}{2}$.

In conclusion, the approximation may mimic the behavior of the original process quite well, at least in the two-population case, as long as there is not too much between-population differentiation, which is a fortunate feature of many real data sets.

Maximizing the likelihood: As suggested by the simulation, if there is little genetic differentiation *between* subpopulations, and if we have large sample sizes, it is reasonable to use the multinormal distribution (see *e.g.*, KENDALL *et al.* 1983, p. 479) as an approximation of the conditional stationary distribution of the process given $\bar{p}$. The likelihood function for the observed gene frequencies at n_k loci is then

$$L(\boldsymbol{\beta} \mid N_e, \mathbf{p}_1, \ldots, \mathbf{p}_{n_k}) = \prod_{k=1}^{n_k} \frac{1}{(2\pi)^{n/2} |\tilde{\mathbf{C}}_k|^{1/2}} \exp\left\{ -\tfrac{1}{2}\, \mathbf{y}_k^T \tilde{\mathbf{C}}_k^{-1} \mathbf{y}_k \right\}, \qquad (25)$$

where $\boldsymbol{\beta}$ is the parameter vector of the assumed migration pattern $\mathbf{M}$.

In summary, to find the maximum likelihood estimates of the parameters of the migration pattern being assumed (see *e.g.*, the next section), we start with some initial parameter values and calculate the migration matrix corresponding to the underlying migration pattern, using *e.g.*, Equation 26. We then calculate the unconditional covariance matrix $\mathbf{C}$ of the whole distribution generated by the process by solving (7) (APPENDIX A). The appropriate weights are then given by (B.5) and the transformation matrix $\mathbf{D}$ by (15). Solving (18) for the alternative process using this transformation matrix, we have the covariance matrix $\mathbf{C}_{y|\bar{p}} \approx \tilde{\mathbf{C}}$ of the relevant conditional distribution. Since the solution of (18) depends on the equilibrium gene frequencies, which may differ between loci, it is adviseable first to solve the system for the standardized covariance matrix $\mathbf{C}_0 = [1/\bar{p}(1-\bar{p})]\tilde{\mathbf{C}}$ once, and then calculate covariance matrices from $\mathbf{C}_0$ for each locus separately. Having calculated $\tilde{\mathbf{C}}$, the likelihood is given by (25). Going back, adjusting the parameters repeatedly, using some numerical algorithm for maximizing functions of several variables, *e.g.*, the AMOEBA routine (PRESS *et al.* 1986), we obtain maximum likelihood estimates of the parameters.

AN EXAMPLE MIGRATION PATTERN

A finite stepping stone model: So far, no assumptions have been made about the form of the migration matrix $\mathbf{M}$. In general, any parametric migration model should be possible to incorporate. Here, to test the performance of the model, we will simulate data from a stepping stone model, which has theoretically well known behavior, and use the general model to estimate its parameters. We will use a finite stepping stone model with $n = 10$ subpopulations, along a single dimension, each of effective size N_e. Each subpopulation exchanges individuals at a rate $m_0/2$ with each of its neighbors, and receives long range migrants from the outside world at rate u. The subpopulations at the edges have emigration and immigration at rate $m_0/2$ only in one direction. With these assumptions, the migration matrix is

$$\mathbf{M} = (1-u)\begin{bmatrix} 1-\frac{1}{2}m_0 & \frac{1}{2}m_0 & 0 & & \\ \frac{1}{2}m_0 & 1-m_0 & \frac{1}{2}m_0 & 0 & \\ \ddots & \ddots & \ddots & \ddots & \\ & 0 & \frac{1}{2}m_0 & 1-\frac{1}{2}m_0 & \end{bmatrix}. \qquad (26)$$

Simulation of data from this model, for a given set of parameters, is again done by assuming that genetic drift occurs after immigration. Data can be generated by initializing all gene frequencies to say $p_i = 0.5$ and then simulating 500 generations. Repeating this procedure 10 times, ignoring linkage disequilibrium, we have data at 10 different loci. For small values of u it may happen that one allele becomes fixed in all populations at the end of a simulation. In such cases, we simulated 500 additional generations, one or more times until the mean gene frequency was between 0.1 and 0.9.

Data generated from the stepping stone model de-

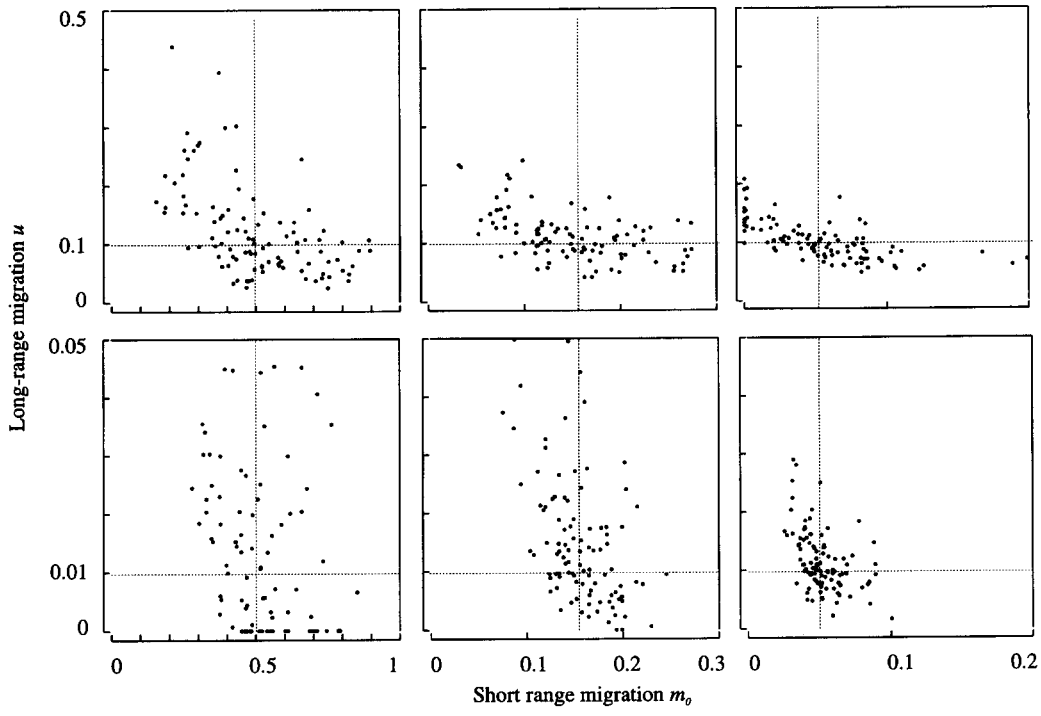

FIGURE 4.—Sampling distributions of $\hat{u}$ and $\hat{m}_0$, for various parameter combinations for 100 simulated data sets from a finite stepping stone model with 10 populations and 10 loci. The effective population size $N_e = 100$ in all simulations and is treated as a known parameter. The true values of u and m_0 are indicated by dotted lines.

scribed above do potentially contain information about all the three parameters N_e, u, and m_0 (in addition to the equilibrium gene frequencies), that is, each parameter combination potentially generates its own unique genetic covariance structure. However, from the theory of infinite one-dimensional stepping stone models, we know that the effective number of parameters are reduced to two in the limit when u becomes much smaller than m. One parameterization of the model is then the rate of exponential decrease $\sqrt{2u/m_0}$ in the correlation between populations with distance, and the variance $q(1-q)/(1+4N_e\sqrt{2um_0})$ of the fluctuations of each subpopulation around the equilibrium gene frequency q (CROW and KIMURA 1970, Equations 9.9.28 and 9.9.29). We can therefore in this situation only expect to obtain estimates of, in our parameterization, *e.g.*, u and m_0 as functions of N_e. We will therefore concentrate on estimating u and m_0 and treat N_e as a known parameter. When working with real data, prior information about N_e could be used if this is available in the literature.

Bias and variance of estimates: Figure 4 shows the sampling distributions of the estimators $\hat{u}$ and $\hat{m}_0$ for 100 simulated data sets, with $N_e = 100$, and various combinations of u and m_0. The bias and standard deviations of the estimators are summarized in Table 1.

There are several interesting points to notice. First, the sampling distribution is concentrated around the true values, indicating that the method works. Both u and m_0 appear to become slightly overestimated, but the bias is small, and is anyhow dependent on the more or less arbitrary parameterization of the model.

Second, as u becomes small relative to m_0 we also see that the uncertainty in the estimate of u becomes very large. This is consistent with our simulations of the two-population case that indicated that the conditional distribution of the contrasts becomes independent of u when u tends to zero. Interestingly, the uncertainty in $\hat{m}_0$ seems, at the same time, to decrease.

Third, there is also negative covariance between the estimates, which makes intuitive sense. Genetic similarity can be due to two causes: long-range and short-range migration. If there is less long-range migration, there has to be more short-range migration. This negative covariance seems to disappear when u becomes much smaller than m_0 (upper right plot, Figure 4).

Finally, note that the maximum likelihood estimate of u for some data sets becomes effectively equal to zero (Figure 4, lower left plot) indicating that long-range migration has no detectable effect on the genetic structure for these particular data sets. Figure 5 shows a contour plot of the likelihood function for one such data set, plotted on log scale for u. We see that the likelihood, given this particular data set, tends to a limit

TABLE 1

Bias and standard deviation of $\hat{u}$ and $\hat{m}_0$ based on the simulations shown in Figure 4 and on untransformed estimates

u	Bias ($\hat{u}$)	SD ($\hat{u}$)	m_0	Bias ($\hat{m}_0$)	SD ($\hat{m}_0$)
0.10	0.022	0.079	0.50	0.000	0.18
0.10	0.012	0.042	0.16	−0.002	0.08
0.10	0.003	0.033	0.05	0.002	0.04
0.01	0.012	0.020	0.50	0.003	0.13
0.01	0.005	0.011	0.16	−0.003	0.03
0.01	0.002	0.005	0.05	0.002	0.02

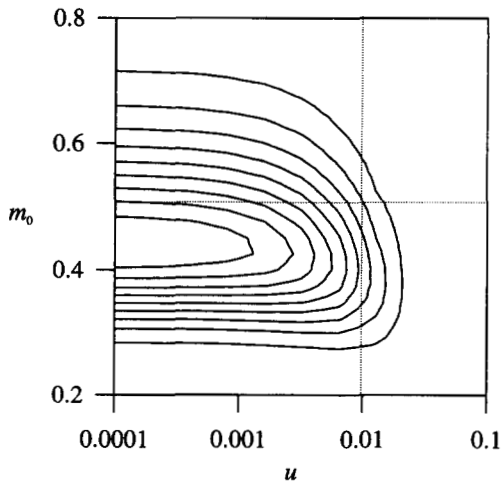

FIGURE 5.—Contour plot of the likelihood function $L(u, m_0)$ for one particular simulated data set from the example stepping stone model with long-range migration $u = 0.01$ and short-range migration $m_0 = 0.5$ (indicated by dotted lines on the graph).

as log u tends to $-\infty$. We can only interpret the contours in Figure 5 as parameter combinations given equal support by the data. The only inference that can be made about u is that it must be smaller than a certain value. Like in all interval estimation, it is however difficult to make any probabilistic statements as to whether the true value of u and m_0 lies within a certain contour, without assuming that a prior probability distribution for u exists.

Also, for small values of m_0 (Figure 4, upper right plot), the maximum likelihood estimate of m_0 sometimes becomes equal to zero, that is, the genetic structure is sometimes indistinguishable from the genetic structure generated by the island model.

DISCUSSION

Assumptions: The approach used here is based on several approximations, and it would be desirable to more formally analyze the behavior of the alternative process $\tilde{\mathbf{p}}_t$, not just for the two-population case, and not just using simulations. The fact that the alternative process starts to behave poorly as the between-population differentiation increases is disturbing. Even so, the model appears to give good estimates for both the short- and long-range migration rates for data generated with our example stepping stone model, and this is encouraging. For real data and for more complex migration patterns, we expect the model to at least give a good approximation of the *relative* likelihood of competing migration patterns.

Another problem is that real data in general are generated by a multivariate process in a much larger number of populations than those sampled. The state of the whole system is therefore no longer completely characterized by the n-dimensional vector $\mathbf{p}_t$, and the process is therefore no longer Markovian because these extra neighboring populations will act as "memory" of the subprocess in the populations under consideration. In our example we avoided this problem by using exactly the same model for simulations and model fitting.

Although we initially assume the existence of an outside world population, the simulations indicate that the model also can be used in situations where the subpopulations are completely isolated. This is because the distribution of the gene frequencies, when conditioning on $\bar{p}$, tends to a limit as the rate of immigration from the outside world becomes small. The model is in fact perhaps most suitable in such situations because the problems with neighboring populations mentioned above are then eliminated.

The current version of the model is also limited in that it handles only two alleles at each locus, and thereby ignores much almost independent information available in multiallelic data. Combining the least frequent allelic classes should however only lead to less efficient estimates and not to additional bias. We have also ignored sampling error, but the additional variance due to finite sample size is easily incorporated by adding the binomial sampling variance to the calculated covariance matrix before each calculation of the likelihood. Since additional sampling error will tend to mask any underlying patterns in the data, large sample sizes will of course be essential for the model to work.

Practical considerations: Even though the method presented here is computationally intensive, it is reasonably fast for analysis of a small number of populations. Finding maximum likelihood estimates of the parameters of the example migration pattern, typically took ~30 sec for single data sets on a Pentium 90 MHz processor, when starting in the true parameter values. Almost all of the computer time is consumed in solving (7) and (18), for each computation of the likelihood. However, the number of populations that in general can be analyzed is quite limited as the computer time increases very rapidly with the number of populations n (APPENDIX A). For idealized migration patterns the number of unknown covariances can potentially be greatly reduced if one takes advantage of possible symmetries. For example, for the migration pattern used here, one could use the fact that c_{ij} equals $c_{(n+1-i)(n+1-j)}$. FELSENSTEIN (1982) discusses explicit solutions for the covariance matrix of another similar approximation based on arcsine transformations and symmetric migration matrices, which only involves taking the inverse of $n \times n$ matrices.

Comparison with other approaches: The method used here differs from recent alternative approaches to the problem based on association tests between genetic and geographic distance matrices (*e.g.*, MANLY 1991; SLATKIN 1993; RAYBOULD *et al.* 1996). Although these approaches have proven useful in practical data analysis, it is not clear if the assumptions underlying the estimators of genetic distance on which these methods are based are valid in general. One such estimator, WEIR

and COCKERHAM's (1984) $\hat{\theta}$, does, for example, assume that all populations have descended from a common ancestral population and that all populations then are maintained under the same conditions, and the expected amount of genetic differentiation in these subpopulations is then estimated. In contrast to this, NEI and CHESSER (1983) estimate the actual amount of genetic differentiation between populations at the time of sampling as defined by WRIGHT's (1943) F_{st} parameter. These estimates are then used in further exploratory data analysis. The problem with these approaches, however, is that there, in general, is no simple relationship between rates of migration and genetic differentiation, and this makes the interpretation of the estimated parameters difficult.

SLATKIN and BARTON (1989) suggested using the beta distribution to estimate gene flow by maximum likelihood, also for stepping stone like models, based on the result of MARUYAMA (1972) who found that the distribution of gene frequencies in finite stepping stone models closely resembles the beta. However, this result refers to the marginal distribution in single subpopulations. Most of the information in the data, except under the island model, lies in the dependencies between the gene frequencies in each subpopulation, and to do any likelihood based inference the full multivariate distribution is therefore needed.

New maximum likelihood methods based on Monte-Carlo Markov-Chain methods and coalescent theory (KUHNER *et al.* 1995) for making inferences from sequence data are perhaps more promising in that they can be extended to incorporate more complex patterns of migration.

Future directions: The motivation behind this work has been to develop a model that makes it possible to calculate the approximate likelihood of more general ecological models of the migration pattern, including effects of *e.g.*, differences in population size, geographic distance, and other factors that possibly influence rates of migration, and that therefore potentially are reflected in the genetic structure generated by the drift process. We are currently working with such models. One could alternatively use all the n^2 elements of the migration matrix as unknown parameters in the model. However, apart from the computational difficulties with so many parameters, we believe that simpler models based on the biology of the species in question are likely to produce more insights.

This work puts estimation of migration patterns and evolutionary trees into a common framework, making more formal likelihood ratio simulation tests between alternative geographical and historical hypothesis possible, a problem discussed in FELSENSTEIN (1982). Whether such tests are computationally feasible and whether they will have the necessary statistical power remains to be seen.

It must be stressed that there are several other ways to approximate the distribution generated by the drift process, that need to be explored. Here an attempt was made to incorporate the fact that the variances of the stochastic changes in the process depend on the gene frequencies, also when approximating the conditional distribution. Alternative solutions, perhaps based on transformations of the data, may produce better results in a larger part of the parameter space.

We thank PETER BEERLI and BERNT-ERIK SÆTHER for valuable comments on the manuscript. The Norwegian Research Council and the Norwegian Institute for Nature Research provided financial support for the study.

LITERATURE CITED

ANDERSSON, M., 1994 *Sexual Selection.* Princeton University Press, Princeton, NJ.

AVISE, J. C., 1994 *Molecular Markers, Natural History and Evolution.* Chapman & Hall, New York.

BODMER, W. F., and L. L. CAVALLI-SFORZA, 1968 A migration matrix model for the study of random genetic drift. Genetics **59:** 565–592.

COURGEAU, D., 1974 Migration, pp. 351–387 in *The Genetic Structure of Populations,* edited by A. JACQUARD. Springer-Verlag, Berlin.

CROW, J. F., 1985 The neutrality-selection controversy in the history of evolution and population genetics, pp. 1–18 in *Population Genetics and Molecular Genetics,* edited by T. OHTA and K. AOKI. Springer-Verlag, Berlin.

CROW, J. F., and M. KIMURA, 1970 *An Introduction to Population Genetics Theory.* Harper & Row, New York.

ENDLER, J. A., 1986 *Natural Selection in the Wild.* Princeton University Press, Princeton, NJ.

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

FELSENSTEIN, J., 1982 How can we infer geography and history from gene frequencies? J. Theoret. Biol. **96:** 9–20.

HOLT, R. D., 1985 Population dynamics in two-patch environments: some anomalous consequences of an optimal habitat distribution. Theoret. Popul. Biol. **28:** 181–208.

JOHNSON, M. L., and M. S. GAINES, 1990 Evolution of dispersal: theoretical models and empirical tests using birds and mammals. Annu. Rev. Ecol. Syst. **21:** 449–80.

KENDALL, M., A. STUART and J. K. ORD, 1983 *Kendall's Advanced Theory of Statistics,* Vol. 1. Charles Griffin and Company Ltd., London.

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.

KUHNER, M. K., J. YAMATO and J. FELSENSTEIN, 1995 Estimating effective population size and mutation rate from sequence data using metropolis-hastings sampling. Genetics **140:** 1421–1430.

LEVIN, D. A., and H. W. KERSTER, 1974 Gene flow in seed plants. Evol. Biol. **7:** 139–220.

MALÉCOT, G., 1975 Heterozygosity and relationship in regularly subdivided populations. Theoret. Popul. Biol. **8:** 212–241.

MANLY, B. F. J., 1991 *Randomization and Monte Carlo Methods in Biology.* Chapman and Hall, London.

MARUYAMA, T., 1972 Distribution of gene frequencies in a geographically structured finite population. Ann. Hum. Genet. **35:** 411–423.

MCCULLAGH, P., and J. A. NELDER, 1989 *Generalized Linear Models.* Chapman & Hall, London.

NAGYLAKI, T., 1975 Conditions for the existence of clines. Genetics **80:** 595–615.

NEI, M., and R. K. CHESSER, 1983 Estimation of fixation indices and gene diversities. Ann. Hum. Genet. **47:** 253–359.

PRESS, W. H., B. P. FLANNERY, S. A. TEUKOLSKY and W. T. VETTERLING, 1986 *Numerical Recipes in Pascal.* Cambridge University Press, Cambridge.

RAYBOULD, A. F., J. GOUDET, R. J. MOGG, C. J. GLIDDON and A. J. GRAY, 1996 Genetic structure of a linear population of *Beta vulgaris* ssp. *maritima* (sea beet) revealed by isozyme and RFLP analysis. Heredity **76:** 111–117.

SLATKIN, M., 1973 Gene flow and selection in a cline. Genetics **75:** 733–756.
SLATKIN, M., 1993 Isolation by distance in equilibrium and non-equilibrium populations. Evolution **47:** 264–479.
SLATKIN, M., and N. H. BARTON, 1989 A comparison of three indirect methods for estimating average levels of gene flow. Evolution **43:** 1349–1368.
WEIR, B. S., and C. C. COCKERHAM, 1984 Estimating *F*-statistics for the analysis of population structure. Evolution **38:** 1358–1370.
WRIGHT, S., 1931 Evolution in mendelian populations. Genetics **16:** 97–159.
WRIGHT, S., 1943 Isolation by distance. Genetics **28:** 114–138.
WRIGHT, S., 1951 The genetical structure of populations. Ann. Eugenics **15:** 323–354.

Communicating editor: W. J. EWENS

APPENDIX A

Comments on the numerical solution of (7) and (18): COURGEAU (1974) showed how (7) can be rewritten into a system of n^2 linear equation. However, for computations, it is important to take advantage of the fact that $c_{ij} = c_{ji}$. It is then only necessary to work with $m = n(n+1)/2$ equations and unknown covariances. First, define

$$\delta_{ij} = \begin{cases} 0 & \text{for } i \neq j \\ 1 & \text{for } i = j. \end{cases} \tag{A.1}$$

For each element of **C**, there is a corresponding linear equation that now can be written

$$c_{ij} = \sum_{k=1}^{n}\sum_{l=1}^{n} m_{ik}m_{jl}c_{kl} + \delta_{ij}\frac{1 - c_{ij}}{2N_i}. \tag{A.2}$$

If the first double sum is split in two in addition to the terms along the diagonal, then the covariance terms above the diagonal can be combined with those below, and so

$$\sum_{k=1}^{n}\sum_{l=1}^{k-1}(m_{ik}m_{jl} + m_{il}m_{jk})c_{kl} + \sum_{k=1}^{n} m_{ik}m_{jk}c_{kk} - \left(1 + \delta_{ij}\frac{1}{2N_i}\right)c_{ij} = -\delta_{ij}\frac{1}{2N_i}. \tag{A.3}$$

This system of equation can be reindexed to the system

$$\mathbf{Ac} = \mathbf{b}, \tag{A.4}$$

where **A** is a $m \times m$ matrix, and **c** and **b** are m-dimensional column vectors, indexed by the new indexes r and s that are uniquely determined by ij and kl, respectively. From (A.3) we see that element rs of **A** now is

$$a_{rs} = \begin{cases} m_{ik}m_{jl} + m_{il}m_{jk} & \text{for } r \neq s \wedge k \neq l \\ m_{ik}m_{jk} & \text{for } r \neq s \wedge k = l \\ m_{ik}m_{jl} + m_{il}m_{jk} - 1 - \delta_{ij}\dfrac{1}{2N_i} & \text{for } r = s \wedge k \neq l \\ m_{ik}m_{jl} - 1 - \delta_{ij}\dfrac{1}{2N_i} & \text{for } r = s \wedge k = l. \end{cases} \tag{A.5}$$

The elements of **b** in (A.4) is

$$b_r = -\delta_{ij}\frac{1}{2N_i}, \tag{A.6}$$

and the vector **c** has elements

$$c_r = c_{ij}. \tag{A.7}$$

Equation A.4 can then be solved using some numerical algorithm, *e.g.*, LU decomposition (PRESS *et al.* 1986, p. 39).

The rewriting of (18) is similar. First, calculate the $n \times n$ matrix $\mathbf{U} = \mathbf{DM}$. The equation corresponding to element ij of **C** is then

$$c_{ij} = \sum_{l=1}^{n}\sum_{k=1}^{n} u_{ik}u_{jl}c_{kl} + \sum_{k=1}^{n} d_{ik}d_{jk}\frac{1 - c_{kk}}{2N_k}, \tag{A.8}$$

which can be rearranged to

$$\sum_{k=1}^{n}\sum_{l=1}^{k-1}(u_{ik}u_{jl} + u_{il}u_{jk})c_{kl} + \sum_{k=1}^{n}\left(u_{ik}u_{jk} - d_{ik}d_{jk}\frac{1}{2N_k}\right)c_{kk} - c_{ij} = -\sum_{k=1}^{n} d_{ik}d_{jk}\frac{1}{2N_k} \tag{A.9}$$

and reindexed into a system similar to (A.4).

Note that since the computer time needed to solve a system of m equations is of order m^3, the time needed to solve (18) with n subpopulations, is of order n^6. The computer memory requirements (storage of matrices) are of order n^4. If there are many subpopulations, it may therefore be better to solve (7) and (18) iteratively, initializing all the elements of $\mathbf{C}_0$ to *e.g.*, zero or to some previous solution, and then compute subsequent matrices $\mathbf{C}_1, \mathbf{C}_2, \cdots$ until **C** converges. However, for small values of long-range migration rate, we found the convergence of **C** to be very slow.

APPENDIX B

Choosing the weights of $\bar{p}$: Given the covariance matrix **C** of the gene frequencies the weights should be chosen to minimize the variance of $\bar{p}$ given by

$$f(w_1, \ldots, w_n) = \mathrm{Var}(\bar{p}) = \mathrm{Var}\left(\sum_{i=1}^{n} w_i p_i\right) = \sum_{i=1}^{n} w_i^2 c_{ii} + 2\sum_{i=1}^{n}\sum_{j=1}^{i-1} w_i w_j c_{ij}, \tag{B.1}$$

subject to the constraint

$$g(w_1, \ldots, w_n) = \sum_{i=1}^{n} w_i - 1 = 0. \tag{B.2}$$

Using the method of Lagrange multipliers, the optimal weights must satisfy a set of n equations, the kth equation being

$$\frac{\partial f}{\partial w_k} = \lambda\frac{\partial g}{\partial w_k}. \tag{B.3}$$

Substituting the partial derivatives of (B.1) and (B.2) into (B.3), we get

$$\sum_{i=1}^{n} c_{ki} w_i = \lambda, \tag{B.4}$$

which is equivalent to $\mathbf{Cw} = \boldsymbol{\lambda}$, implying $\mathbf{w} = \mathbf{C}^{-1}\boldsymbol{\lambda}$. Choosing λ to satisfy (B.2) it follows that

$$w_k = \frac{\sum_{i=1}^{n} c_{ki}^{-1}}{\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}^{-1}}. \tag{B.5}$$

These weights happen to be independent of the unknown q, since $\mathbf{C}$ is proportional to a standardized covariance matrix $\mathbf{C}_0$ rescaled by the factor $q(1-q)$. Note that the c_{ij}^{-1}'s in (B.5) are elements of the inverse matrix $\mathbf{C}^{-1}$.

APPENDIX C

Independence between y and $\bar{p}$: Let $\mathbf{t}$ be a column vector with all elements equal to the weighted mean gene frequency, $\bar{p} = t(\mathbf{p}) = \sum_{i=1}^{n} w_i p_i$. We want to establish that the distribution of the vector $\mathbf{y} = \mathbf{p} - \mathbf{t}$, in the multinormal case, is independent of the statistic t itself, when conditioned on t.

The problem can be simplified if we first transform $\mathbf{p}$ to $\mathbf{p}^*$ using the linear transformation

$$\mathbf{p}^* = \mathbf{Ap}, \tag{C.1}$$

where the matrix $\mathbf{A}$ has the property that the elements of $\mathbf{p}^*$ become independent and with expectations equal to q, that is, we require that

$$E(\mathbf{p}^*) = \mathbf{q}, \tag{C.2}$$

where all elements of the column vector $\mathbf{q}$ are equal to q. Since we also have $E(\mathbf{Ap}) = \mathbf{A}E(\mathbf{p}) = \mathbf{Aq}$, it follows that

$$\mathbf{q} = \mathbf{Aq}, \tag{C.3}$$

implying that the rows of $\mathbf{A}$ and $\mathbf{A}^{-1}$ sum to one.

By the same argument as in APPENDIX B, the sufficient statistic t for the unknown q as a function of $\mathbf{p}^*$ in this new model is

$$t(\mathbf{p}^*) = \sum_{i=1}^{n} w_i^* p_i^*, \tag{C.4}$$

where

$$w_i^* = \frac{1/\sigma_i^2}{\sum_{j=1}^{n} 1/\sigma_j^2}, \tag{C.5}$$

and $\sigma_i^2 = \mathrm{Var}(p_i^*)$. Defining the vector

$$\mathbf{y}^* = \mathbf{p}^* - \mathbf{t}, \tag{C.6}$$

it follows from the properties of $\mathbf{A}$ that

$$\begin{aligned}\mathrm{Cov}(y_i^*, t) &= \mathrm{Cov}(p_i^* - t, t) = \mathrm{Cov}(p_i^*, t) - \mathrm{Var}(t)\\ &= w_i^* \mathrm{Var}(p_i^*) - \sum_{j=1}^{n} (w_j^*)^2 \mathrm{Var}(p_j^*) = \frac{1/\sigma_i^2}{\sum_{k=1}^{n} 1/\sigma_k^2}\\ &\sigma_i^2 - \sum_{j=1}^{n} \left(\frac{1/\sigma_j^2}{\sum_{k=1}^{n} 1/\sigma_k^2}\right)^2 \sigma_j^2 = 0,\end{aligned} \tag{C.7}$$

implying that the vector $\mathbf{y}^*$ is independent of t.

We now know that

$$\begin{aligned}\mathbf{y} = \mathbf{p} - \mathbf{t} &= \mathbf{A}^{-1}\mathbf{p}^* - \mathbf{t} = \mathbf{A}^{-1}(\mathbf{y}^* + \mathbf{t}) - \mathbf{t}\\ &= \mathbf{A}^{-1}\mathbf{y}^* + (\mathbf{A}^{-1} - \mathbf{I})\mathbf{t}.\end{aligned} \tag{C.8}$$

Because each row of $\mathbf{A}^{-1}$ sums to one, it follows that $(\mathbf{A}^{-1} - \mathbf{I})\mathbf{t} = 0$ and hence

$$\mathbf{y} = \mathbf{A}^{-1}\mathbf{y}^*. \tag{C.9}$$

Since the distribution of $\mathbf{y}^*$ conditional on t is independent of t, so is the distribution of $\mathbf{y}$.