Evolution, 40(6), 1986, pp. 1312–1327

MIGRATION AND GENETIC DRIFT IN HUMAN POPULATIONS

ALAN R. ROGERS
Department of Anthropology, 3H01 Forbes Quad, University of Pittsburgh, Pittsburgh, PA 15260

AND

HENRY C. HARPENDING
Department of Anthropology, 409 Carpenter Building, Pennsylvania State University, University Park, PA 16802

Abstract.—In humans and many other species, mortality is concentrated early in the life cycle, and is low during the ages of dispersal and reproduction. Yet precisely the opposite is assumed by classical population-genetics models of migration and genetic drift. We introduce a model in which population regulation occurs before migration. In contrast to the conventional model, our model implies that geographic variation in the allele frequencies of newborns should exceed that of adults. Thus, it is important to distinguish genetic variation of adults from that of newborns in species with human-like life cycles.

Classical models deal with the variance of group allele frequencies about the allele frequency of a hypothetical "continent" or "foundation stock." Empirical studies, however, can only measure "reduced" variance, i.e., variance about the current population mean. Our model deals with reduced variance, and should therefore be more relevant to field studies. We show that reduced variance converges faster, which implies that populations are more likely to be at equilibrium with respect to reduced than unreduced variance.

To summarize the effect of migration on genetic population structure, we introduce a new parameter, the effective migration rate. Unlike most population structure statistics, it does not confound the effects of mobility and population size, and it should therefore be useful for comparisons between populations. Finally, we show that the difference between geographic variation of newborn and adult allele frequencies contains information about both effective population size and effective migration rate.

Received June 3, 1985. Accepted July 17, 1986

The causes and consequences of genetic variation among local groups have been central concerns of population genetics for many years. There is an extensive theoretical literature and an even more extensive empirical literature. However, it often seems that the two have little to do with each other. Theoretical work generally seeks qualitative insights rather than quantitative predictions, whereas empirical work has often been based on ad hoc measures of genetic distance or similarity that have no connection with theory. These measures have proved useful as guides to intuition but provide no basis for inference.

The major exception to these generalizations is the family of models, collectively called "migration matrix models," that were introduced by Malécot (1950), Bodmer and Cavalli-Sforza (1968), and Smith (1969). They are appealing because they deal gracefully with patterns of mobility that are nearly as complex as those of real populations. Their generality, however, is also their principal failing. Their use in theoretical work has been limited by the difficulty of obtaining explicit, general formulas (Felsenstein, 1976). They remain popular, however, among empiricists. Many authors have compared observed genetic variation among a set of local populations with that predicted by a migration matrix model.

Such studies may one day allow us to evaluate ideas about the effects of drift and migration on genetic variation, but so far they seem to have taught us little. Sometimes the variance observed is reasonably close to that predicted, and sometimes it is not (Bodmer and Cavalli-Sforza, 1974; Jorde, 1980). In either case, little can be inferred—observations may differ from predictions for so many reasons that it is impossible to interpret the concordance between theory and data.

When observed and predicted variances differ, the discrepancy is often attributed either to failure of the assumption of equilibrium or to some kind of non-Markovian migration. For example, dispersal may involve kin groups rather than individuals

(Fix, 1978; Neel and Salzano, 1967; Smouse et al., 1981), or the tendency to move may be inherited culturally (Hiorns et al., 1977). These factors are surely important, but they are not the only sources of discrepancy between theory and observation.

In this paper, we argue that classical population-genetics models of migration and genetic drift, as developed by Wright, Malécot, and others involve assumptions that are inappropriate for humans and other species with similar life cycles. We introduce a model that is more appropriate for such species. In addition, we introduce a maximum-likelihood estimator that is compatible with the assumptions of our theory. Finally, we compare the predictions of our theory with published genetic statistics for several human populations.

Models of the Life Cycle and Their Effects

Assumptions.—Models of migration and genetic drift usually incorporate assumptions about the life cycle of the organisms studied. Some of the conclusions of population-genetics theory are robust with respect to these assumptions, while others are quite sensitive. In this section we discuss the assumptions embodied in some models that have been used as a basis for analysis of human genetic data.

The set of individuals that disperses from the ith to the jth local group will be referred to as the "ijth migrant set." Regardless of a species' life cycle, we can write

$$p_j = \sum_i m_{ij} q_{ij}, \qquad (1)$$

where

q_{ij} = the frequency of allele A in the ijth migrant set,
m_{ij} = the proportion of group j after migration derived from group i, and
p_j = the frequency of A in group j after migration.

Clearly, q_{ij} is the allele frequency of a sample of individuals obtained from group i. Unless the propensity to migrate depends on genotype, the expectation of q_{ij} is equal to the allele frequency in group i prior to migration. The variance of q_{ij} depends on the size of the ijth migrant set, and this depends on the life cycle of the species being studied. Consider, for example, model A below.

Model A

Newborns (∞) —migration→ Adults (∞) —population regulation→
→ Adults (n) —reproduction→ Newborns (∞)

Here, the q_{ij} refer to infinite samples and are therefore equal to their expectations. Migration can be described by the deterministic equation

$$p_j^* = \sum_i m_{ij} p_i', \qquad (2)$$

where p_j^* is the allele frequency in group j after migration but before population regulation, p_i' is the allele frequency in group i before migration. Genetic drift occurs when density regulation reduces the population to size n_j, adding a random increment with mean zero and variance $p_j^*(1 - p_j^*)/2n_j$. Wright (1931, 1943) pioneered this approach to the problem, and it is central to most theoretical work on migration and genetic drift (for example, Bodmer and Cavalli-Sforza, 1968; Smith, 1969; Courgeau, 1974; Carmelli and Cavalli-Sforza, 1976).

Model A is a reasonable description of the life cycle of species in which dispersal occurs at the gamete stage, as in most plants, or in which large numbers of juveniles are involved, as in many other species. However, it is a poor description of species such as our own, in which most mortality occurs before dispersal (Coale, 1972). Nonetheless, results derived using model A have often been used to interpret human genetic data (see for example, Bodmer and Cavalli-Sforza, 1974). As an alternative, consider:

Model B

Zygotes (∞) —population regulation→ Newborns (n) →
—migration→ Adults (n) —reproduction→ Zygotes (∞)

This model assumes that no mortality occurs during migration and reproduction, which is probably more realistic than model A for humans and other species with low

mortality after infancy. It is particularly appropriate in human genetics, when the birthplaces of adult subjects are routinely recorded, and individuals may be classified either by adult residence or by birthplace. Such data refer not to the birth cohort, but to that portion of it that survives to maturity. With this life cycle, the q_{ij} of equation (1) are based on small samples so their variances are no longer negligible. Thus, genetic drift occurs during migration as well as during population regulation.

Some authors view (2) as a deterministic approximation to dynamics under life cycle B rather than as a model of dynamics under A (Sved and Latter, 1977; Latter and Sved, 1981; Harpending and Ward, 1982). Latter and Sved use the term "stochastic migration" to refer to models incorporating the stochastic effects on allele frequencies produced by migration under life cycle B, and also to models in which the m_{ij} themselves are random variables (see also Nagylaki, 1979, 1980, 1983). The geographic variation predicted by all these models is substantially greater than that predicted by analogous deterministic models. These stochastic models are of limited value for comparisons with natural populations, however, because of restrictive assumptions. All assume either that migration follows some simple symmetric pattern with equal group sizes, or else that the rate of migration is extremely high.

All of the models discussed above refer to allele frequencies of adults after migration and population regulation. Malécot (1948, 1969), on the other hand, attempts to deal with newborns, but his analysis contains a subtle error. On page 67 of the (1969) English translation of his book, he says that if genes are drawn from individuals born in generation $n + 1$ in locations C and D, whose parents are both known to come from location E, then "they will have conditional probability $1/[2\delta(E)dS_E]$ of coming from the same locus [gene copy] of the same parent and probability $1 - 1/[2\delta(E)dS_E]$ of coming from loci infinitely close but distinct," that is, of being copies of distinct genes from individuals born at location E in generation n. Here $\delta(E)dS_E$ is the number of individuals born at location E and must be greater than zero (see Felsenstein, 1975). But, since Malécot's model does not allow for the possibility that individuals may breed in more than one location, genes from individuals born in C and D cannot possibly be copies of the same parental gene unless $C = D$. If $C = D$, the probability that they are copies of the same parental gene is $1/2n_{EC}$, where n_{EC} is the number of individuals that disperse from E to C, and will be smaller than the number born at E.

Malécot's theory can be rescued by redefining his terms so that individuals are identified with the locations in which they breed instead of with their birthplaces as Lalouel (1977) has done. For example, Malécot's $g(E,C)dS_E$ becomes the probability that an individual breeding in C was born in E. His analysis then rests on the assumption that two distinct individuals breeding in C both derive from E with probability $[g(E,C)dS_E]^2$. This assumption is incompatible with life cycle B since

$$g(E,C)dS_E = E\left\{\frac{n_{EC}}{\sum_X n_{XC}}\right\},$$

whereas the probability that two distinct adults breeding in C were both born in E is

$$[g(E,C)dS_E]E\left\{\frac{n_{EC} - 1}{\sum_X n_{XC} - 1}\right\},$$

which is not the same as Malécot's formula unless the numbers of individuals migrating (the n_{XE}) are large. Thus, Malécot's theory refers to adults under life cycle A and applications of this theory to humans should be regarded with some skepticism.

In summary, no theory has been developed describing the dynamics of migration and genetic drift under life cycle B. Consequently, studies of the population structure of humans and similar species have been based on a theory that may be inappropriate.

The Distinction between Adults and Newborns.—Under life cycle A, no changes in allele frequency occur at reproduction, so the allele frequencies of newborns should equal those of their parents. Under life cycle B, on the other hand, allele frequencies of newborns differ from those of their parents

because of the effect of the population regulation component of genetic drift. Since drift tends to increase variation among local groups, variation of newborns should exceed that of their parents. Similarly, since migration tends to reduce variation, the variation of a single cohort should be smaller after migration than before. For both reasons, variation of newborns should exceed that of adults. At equilibrium these effects are in balance so that variation among adults of adjacent generations is the same, yet the difference between newborns and adults persists. This effect has also been noted by Long (1986).

Model

This section introduces a revised version of the migration matrix models developed by Malécot (1950, 1973), Bodmer and Cavalli-Sforza (1968), and Smith (1969). This revised model is used in the appendix to derive the expectations of several measures of local genetic variation.

Migration and Genetic Drift

Let n_{ij} denote the size of the ijth migrant set. We assume that migration within a generation follows a discrete Markov process. The allele frequencies of migrants are treated as random variables as life cycle B implies, but the n_{ij} are assumed constant in time. In the real world, of course, the numbers of migrants may be far from constant. Latter and Sved (1981) have investigated the effect of this assumption under Wright's (1943) "island model" of population structure. They find that the variance among groups implied by our assumption is smaller than that implied by the assumption that individuals migrate independently. The model of independent migration, however, may not be more realistic. If there is density regulation within groups, the probability that an individual moves into a group may depend on the number there already. It is not clear which model is the better approximation to reality.

We assume that each local group is panmictic so that its effective size (Wright, 1969) equals its actual size. If, in addition, individuals migrating from group i to j are a random subset of group i, then each of their genes can be treated as an independent, random draw from the gene pool of group i in the previous generation. Hence, the number of copies of allele A in migrants from i to j in generation $t + 1$ is a binomial random variable with parameters $2n_{ij}$ and $p_i^{(t)}$.

Systematic Pressure

To ensure that the process will have internal equilibria, we assume that, in addition to migration among groups, a fraction s of the residents of each group are immigrants from a "continent" with unchanging allele frequency π. This linear systematic pressure could also be interpreted as mutation or weak selection. Without it, the process would have no equilibria short of fixation. Continental migration is assumed to occur after population regulation so that it also contributes a component to genetic drift.

We estimate s as the fraction of immigrants from outside the study area, although this is almost certainly an overestimate. Most external immigrants derive from neighboring populations with similar allele frequencies, so their impact on local genetic structure will be smaller than their numbers imply. Predictions obtained by setting $s = 0$ may often be better approximations to reality.

Our assumptions that the fraction of external immigrants in each local group is the same and that external immigrants are all drawn at random from the same population are also unrealistic, and reduce the variance predicted among local groups (Wagener, 1973; Harpending and Ward, 1982). This effect is negligible when local genetic structure is dominated by the effects of local migration, but it may be important when systematic pressure is strong relative to local migration. Thus, the model we are building is most appropriate for populations that are relatively isolated from the outside world.

Measures of Local Variation

Our basic definitions are not of parameters but of functions of allele frequencies. These functions are random variables, and we attempt to characterize their first and second moments under the model described above. In what follows, the term "expectation" refers to an average over a hypo-

thetical infinite ensemble of populations representing different realizations of the same stochastic process. We neglect the sampling problem entirely, assuming that allele frequencies of local groups are known without error.

Let

$n.. = \sum_{ij} n_{ij}$, the total population size,

$w_j = \sum_{i} n_{ij}/n..$, the relative size of the jth group,

$\bar{p} = \sum_{i} w_i p_i$, the mean allele frequency, and

g = the number of local groups.

Following Harpending and Jenkins (1974), we define the genetic correlation of allele frequencies in groups i and j as

$$r_{ij} = \frac{(p_i - \bar{p})(p_j - \bar{p})}{\bar{p}(1 - \bar{p})} .$$

The genetic correlation matrix for adults is $\mathbf{R} = [r_{ij}]$. The analogous quantities for newborn allele frequencies are r_{ij}' and $\mathbf{R}'$.

A useful measure of variation among groups is (for adults)

$$r_0 = \frac{\sum_{i=1}^{g} w_i(p_i - \bar{p})^2}{\bar{p}(1 - \bar{p})} = \sum_{i=1}^{g} w_i r_{ii} . \qquad (3)$$

r_0 is a random variable, and we denote its expectation by ρ. An unbiased estimate of ρ can be obtained by inserting estimates of group allele frequencies into (3), since the conditional expectation of r_0 given $\bar{p}$ is independent of $\bar{p}$ (see Appendix). Our r_0 is equivalent to the "Wahlund variance" (Wahlund, 1928, 1975), and to one of the several meanings that have been attached to Wright's (1951) F_{ST} (see Wood, 1986). The analogous quantities for allele frequencies of newborns are denoted by r_0' and ρ'.

Note that ρ is defined in terms of variation about the current population mean $\bar{p}$. A related parameter describing the expected variation about the continental allele frequency, π, is

$$\phi = \frac{E\left\{\sum_{j=1}^{g} w_j(p_j - \pi)^2\right\}}{\pi(1 - \pi)} .$$

Following Cavalli-Sforza and Piazza (1975) and Felsenstein (1982), we refer to variances about $\bar{p}$ as "reduced variances." Most theoretical results refer to ϕ, while data analysts generally work with estimates of r_0. The distinction between these measures is often ignored, and has produced a good deal of confusion. This confusion can be avoided either by rewriting the theory in terms of reduced variances (Harpending and Jenkins, 1974) or by attempting to estimate unreduced variances from genetic data (Morton et al., 1968, 1971; Morton, 1975; Weir and Cockerham, 1984).

Theoretical Results

Exact Formulas

Exact formulas for the expectations of $\mathbf{R}$ and $\mathbf{R}'$ are derived in the appendix, and ρ and ρ' can be obtained from these using (3). These formulas are unwieldy, but do provide a method for predicting genetic variation from demographic data, and a computer program to do this is available. Analogous formulas are derived by Malécot (1950, 1973; Bodmer and Cavalli-Sforza, 1968; Smith, 1969; Courgeau, 1974).

Approximations

A variety of approximations and simplifying assumptions have been used in theoretical work on population structure. For example, Malécot (1973) assumes that $\mathbf{M}$ is symmetric, that group sizes are equal, and that (in expectation) $r_{ii} = r_{jj}$, for all i and j. The last assumption should often hold approximately, but the others are unfortunately restrictive. We assume instead that the number of individuals moving from group i to group j in a generation is the same as the number moving from j to i. This seems reasonable if the sizes of local groups are stable, and in other situations an equilibrium theory is of little interest anyway. In addition, we use the approximation $r_{ii} \simeq r_{jj} \simeq r_0$. In the appendix, we derive expressions for $\mathbf{R}$ and $\mathbf{R}'$, and show that

$$\rho \simeq \frac{1 - \rho}{2n..} \sum_{i=2}^{g} \frac{1}{1 - (1 - s)^2\lambda_i^2} , \qquad (4)$$

and

$$\rho' \simeq \frac{1-\rho}{2n..} \sum_{i=2}^{g} \frac{2-(1-s)^2\lambda_i^2}{1-(1-s)^2\lambda_i^2}, \quad (5)$$

where λ_i is the ith eigenvalue of **M**. Unless local groups are equally isolated from each other and exchange between each pair of groups is symmetric, these formulas are only approximate. Their accuracy is investigated in the applications section below.

Variances of Newborns and Adults

In this section, we investigate the magnitude of the difference between newborn and adult variances and define an "effective migration rate," which summarizes the effect of migration on ρ at equilibrium. Equations (4) and (5) imply that adult and newborn Wahlund variances are related by

$$\rho' = \rho + \frac{1-\rho}{2ng/(g-1)}, \quad (6)$$

where $n = n../g$ is the average group size. Except for the factor $g/(g-1)$, the increase in ρ at reproduction is identical to the increase in inbreeding between generations in a finite population (Crow and Kimura, 1970 p. 320). The effective population size is inflated by this amount since we are dealing with reduced variances. Thus, we refer to $n_e = ng/(g-1)$ as the "reduced variance effective group size."

Using (4) we can also write

$$\rho = \frac{1}{4m_e n_e + 1}, \quad (7)$$

where m_e is the effective migration rate, defined by

$$\frac{1}{2m_e} = \frac{1}{g-1} \sum_{i=2}^{g} \frac{1}{1-(1-s)^2\lambda_i^2}. \quad (8)$$

Equation (7) is a generalization of Wright's (1943) formula for the inbreeding coefficient. However, Wright's formula assumes that the number of groups is large, that the rate of mobility between each pair of groups (m) is the same, and that m is small. Equation (7), on the other hand, relies only on the assumptions that $n_{ij} = n_{ji}$ and that s is much smaller than ρ. Unlike Wright's formula, it is valid for large m_e. It can be shown that, under Wright's assumptions, m_e approaches m as m approaches 0, so (7) reduces to Wright's formula when m_e is small.

Our m_e is simply a number that summarizes the effect of mobility on ρ, and has no connection with the effective migration rate defined by Wright (1969). As (8) shows, m_e depends on the eigenvalues of **M** and on systematic pressure, but does not depend on population size. The effective migration rate ranges from 0 to ½. The maximal value is reached when all the λ_i equal zero, which occurs under "random dispersal," i.e., when community of residence is independent of community of origin. Using (6) and (8), a little algebra produces

$$\rho' = \rho(1 + 2m_e). \quad (9)$$

Thus, the ratio of adult and newborn Wahlund variances depends only on mobility.

Inferring m_e *and* $n..$ *from Genetic Data*

Solving equations (6) and (9) for $n..$ and m_e produces

$$n.. = \frac{1-\rho}{\rho'-\rho} \cdot \frac{g-1}{2}, \quad (10)$$

and

$$m_e = \frac{\rho'-\rho}{2\rho}. \quad (11)$$

These equations express $n..$ and m_e in terms of quantities that are readily estimated from genetic data, and they may prove useful as estimators. Their statistical properties, however, are as yet unknown.

A Symmetric Estimator of the Migration Matrix

Before this theory can be used to predict genetic variation, one must estimate $\mathbf{M} = [m_{ij}]$. The simplest estimator of m_{ij} is the proportion of the adult residents in group j that originated in group i. Some authors, however, prefer to impose some kind of symmetry constraint (Bodmer and Cavalli-Sforza, 1968; Morton, 1973). These differing approaches apparently arise from slightly different evolutionary models. All methods of predicting genetic variation assume that local group sizes are unchanging. For some authors, this invariance is a consequence of population regulation (Lalouel, 1977). This approach imposes no constraints on the pattern of migration, so we

refer to it as the "unrestricted" model. Others view unchanging group sizes as an approximation that is reasonable provided that migration is conservative (i.e., has no tendency to change the sizes of local groups). This, the "conservative" model, does impose constraints since it implies that $\mathbf{1}^T\mathbf{N} = \mathbf{N}\mathbf{1}$, where $\mathbf{N} = [n_{ij}]$ is the matrix of migrant set sizes, and $\mathbf{1}$ is a column vector with each entry equal to unity.

The approximations discussed above rely on the somewhat stronger assumption that exchange between each pair of groups is symmetric, i.e., that $\mathbf{N} = \mathbf{N}^T$. We refer to this as the "symmetric" model. Note that it does not imply that $\mathbf{M}$ is symmetric unless local group sizes are equal. Let $a_{ij} = w_j m_{ij}$ denote the joint probability of being in group i before migration, and in group j after. The symmetric model imposes $g(g-1)/2$ constraints of the form

$$a_{ij} = a_{ji}. \tag{12}$$

Since w_j is the marginal probability of being in the jth group, the a_{ij} must also satisfy g constraints of the form

$$w_i = \sum_j a_{ij}. \tag{13}$$

None of the estimators that have previously been proposed satisfy all of these constraints.

Let Y_{ij} denote the observed number of individuals that move from group i to group j. If mobility among groups follows a Markov process with transition matrix $\mathbf{M} = [m_{ij}]$, it can be shown that the unconstrained maximum likelihood estimate of m_{ij} is $\hat{m}_{ij} = Y_{ij}/\sum_j Y_{ij}$ (Smouse and Wood, unpubl.). This estimate employs g constraints, since the sum of each column must equal 1. It is appropriate under the unrestricted model, but it does not satisfy the constraints imposed by the conservative or symmetric models. The easiest way to obtain a symmetric estimate of $\mathbf{M}$ is to replace Y_{ij} and Y_{ji} by their average, and then compute $\mathbf{M}$ as above (Bodmer and Cavalli-Sforza, 1968; Morton, 1973). Although this estimator satisfies (12), it does not satisfy (13), and it is not consistent with any model that we know of.

To obtain an estimator that obeys both (12) and (13), we estimate a_{ij} first, and then use the relation $m_{ij} = a_{ij}/w_j$ to obtain $\mathbf{M}$. The log likelihood of our observations is

$$L = \sum_{i=1}^{g}\sum_{j=1}^{g} Y_{ij}\log a_{ij} + \sum_i \psi_{ii}\left(w_i - \sum_j a_{ij}\right) + \sum_{i\neq j}\psi_{ij}(a_{ij} - a_{ji}),$$

where the ψ_{ij}'s are Lagrange multipliers (Chiang, 1974). Maximum likelihood estimators of a_{ij} can be obtained by setting the partial derivatives with respect to a_{ij} and ψ_{ij} equal to zero. We find that

$$\hat{a}_{ij} = \frac{Y_{ij} + Y_{ji}}{2(\psi_{ii} - \psi_{ij} + \psi_{ji})},$$

and

$$\psi_{ji} = \psi_{ij} + \frac{\psi_{jj} - \psi_{ii}}{2}.$$

Hence, the maximum likelihood estimate of m_{ij} is

$$\hat{m}_{ij} = \frac{Y_{ij} + Y_{ji}}{w_j(\hat{\psi}_{ii} + \hat{\psi}_{jj})}.$$

The ψ_{ii} obey

$$\psi_{ii} = \frac{1}{w_i}\sum_j \frac{Y_{ij} + Y_{ji}}{1 + \frac{\psi_{jj}}{\psi_{ii}}}.$$

They can be estimated by starting with a trial value (say $\psi_{ii} = \psi_{jj} = \sum_{ij} Y_{ij}$), and applying this relation iteratively. In our experience, this algorithm usually converges rapidly, but we have no guarantee that it will always do so. A computer program for doing these calculations is available from the authors.

Application to Human Migration Data

The theory developed here avoids the effects of several factors that have reduced the realism of previous versions of the migration matrix model. To assess the magnitude of these effects, we analyzed published migration data for several human populations, listed in Table 1. In most cases,

the raw migration data were cross tabulations of birthplace against residence, or of parent's birthplace against offspring's birthplace. For the Bundi, however, the raw data comprise a matrix whose ijth entry is the number of husbands in clan j whose wife is of clan i. This matrix was converted to the origin-residence format by adding the column sums to the main diagonal. Local group sizes for the Gidra and Oxfordshire could not be found in published literature, so these were approximated by the harmonic means of the row and column sums of the raw migration matrices (**N**). Consequently, results for these populations are tentative.

The migration matrix for Bougainville Island is not irreducible, as our theory requires; four of the villages are completely isolated from the others. We applied our formulas to these data anyway, but also analyzed the largest irreducible subset of the Bougainville villages.

To obtain numerical results, we replace the symbol $n_{..}$ in our formulas with estimates of the variance-effective population size (Crow and Kimura, 1970), taken as one third the actual population size. This procedure is only an approximation both because the estimates of variance-effective size are crude and because our analysis assumes that the variance-effective size is equal to the actual population size.

Effect of the Life Cycle.—Equation (19) (see Appendix), which refers to adults under life cycle B is nearly identical to Carmelli and Cavalli-Sforza's (1976) equation 1.17, which refers to adults under life cycle A. Consequently, life cycles A and B have similar implications for the differentiation of local groups at the adult stage. Conclusions about the genetic statistics of adults are little affected by assumptions about the life cycle. They do, however, affect conclusions about newborns.

The Difference between Newborns and Adults.—Equation (9) indicates that variation of newborns exceeds that of adults by a factor of $1 + 2m_e$. Values of m_e calculated from the symmetric estimate of **M** can be found in Table 1, along with g (the number of local groups), s (estimated as the fraction of external migrants), and n_e (estimated as one-third the census size of the total population divided by $g - 1$). Effective migration rates for the human populations we studied range between 0.10 and 0.33, so Wahlund variances of newborns should exceed those of adults by 20 to 70%. Where mobility is great, m_e can be as large as ½, so ρ' may be twice as large as ρ. Variation of newborns will be substantially greater than that of adults unless m_e is small. These remarks also apply to metric characters since the variance among groups is proportional to ρ for any neutral metric character with an additive genetic basis (Wright, 1951; Rogers and Harpending, 1983). Thus, unless only an order of magnitude estimate is wanted, the genetic variance of newborns should not be confused with that of adults in populations whose life cycles resemble model B. The discrepancy between newborns and adults is undoubtedly smaller in studies of variation at larger scales of distance, since there is less mobility between continents than between neighboring villages.

When genetic data are collected, some investigators sample only adults, while others sample individuals of all ages. Since there can be a substantial difference between the

TABLE 1. Number of groups (g), systematic pressure (s), effective group size (n_e), and effective migration rate (m_e) in several human populations.

Population	g	s	n_e	m_e $s \neq 0$	m_e $s = 0$
Åland[a]					
All periods	12	0.028	581	0.160	0.130
Pre-1900	11	0.021	607	0.102	0.069
Bedik[b]	6	0.018	113	0.227	0.213
Bougainville[c]					
All groups	17	0.250	52	0.278	—
Subset[d]	14	0.250	54	0.294	0.045
Bundi[e]	15	0.122	114	0.325	0.247
Gainj[f]	11	0.119	32	0.284	0.201
Gidra[g]	13	0.050	50	0.143	0.078
!Kung[h]					
All groups	9	0.010	125	0.270	0.262
Subset[i]	6	0.010	140	0.294	0.288
Makiritare[j]	6	0.278	88	0.279	0.033
Oxfordshire[k]	8	0.353	133	0.326	0.051
Papago[l]	10	0.080	189	0.247	0.184

[a] Jorde, 1979; Jorde et al., 1982.
[b] Jacquard, 1974; Langaney and Gomila, 1973.
[c] Friedlaender, 1975.
[d] The largest irreducible set of groups on Bougainville Island. Villages Nupatoro, Turungum, Moronei, and Old Siwai are excluded.
[e] Malcolm, 1970; Malcolm et al., 1971.
[f] Wood et al., 1982; Wood, 1986.
[g] Ohtsuka et al., 1985.
[h] Harpending and Jenkins, 1974.
[i] Six relatively "pure" !Kung local groups.
[j] Ward and Neel, 1970.
[k] Hiorns et al., 1969.
[l] Workman and Niswander, 1970; Workman et al., 1973.

TABLE 2. A comparison of ρ and ρ' using the unrestricted (U), symmetric (S), and approximate (A) formulas. For each population, systematic pressure (s) is set to zero for one run and to the fraction of external migrants for the other.

Population	s	Newborns			Adults		
		U	S	A	U	S	A
Åland							
All periods**	0.000	0.006	0.004	0.004	0.005	0.003	0.003
	0.028	0.004	0.003	0.004	0.003	0.003	0.003
Pre-1900	0.000	0.007	0.007	0.007	0.006	0.006	0.006
	0.021	0.005	0.005	0.005	0.004	0.004	0.004
Bedik*	0.000	0.016	0.015	0.015	0.012	0.010	0.010
	0.018	0.015	0.014	0.014	0.011	0.009	0.010
Bougainville							
All groups†	0.000	0.335	0.279	—	0.328	0.272	—
	0.250	0.024	0.023	0.026	0.014	0.014	0.017
Subset	0.000	0.164	0.099	0.102	0.156	0.091	0.094
	0.250	0.022	0.022	0.025	0.013	0.013	0.016
Bundi**	0.000	0.032	0.013	0.013	0.028	0.009	0.009
	0.122	0.013	0.010	0.011	0.009	0.006	0.007
Gainj**	0.000	0.053	0.053	0.053	0.038	0.037	0.038
	0.119	0.040	0.040	0.042	0.024	0.024	0.027
Gidra††	0.000	0.076	0.068	0.070	0.067	0.059	0.061
	0.050	0.043	0.042	0.044	0.033	0.032	0.034
!Kung							
All groups**	0.000	0.011	0.012	0.012	0.007	0.008	0.008
	0.010	0.011	0.011	0.011	0.007	0.007	0.007
Six groups**	0.000	0.009	0.010	0.010	0.006	0.006	0.006
	0.010	0.009	0.010	0.010	0.006	0.006	0.006
Makiritare*	0.000	0.073	0.081	0.083	0.068	0.076	0.078
	0.278	0.014	0.014	0.016	0.008	0.008	0.010
Oxfordshire††	0.000	0.044	0.039	0.039	0.040	0.035	0.036
	0.353	0.008	0.008	0.009	0.004	0.004	0.006
Papago**	0.000	0.012	0.010	0.010	0.009	0.007	0.007
	0.080	0.008	0.008	0.008	0.006	0.005	0.005

* Hypothesis that underlying pattern of migration is symmetric and consistent with census sizes of local groups can be rejected at the 0.05 significance level.
** The hypothesis above can be rejected at the 0.005 level.
† Algorithm did not converge, and was stopped after 700 iterations.
†† Local group sizes were approximated by the harmonic mean of row and column sums of the raw migration matrix (M) since independent data were unavailable. No significance test is possible.

genetic variances of adults and newborns, the age structure of the sample from which genetic data are obtained is an important confounding influence.

Effect of the Approximations Used.—Table 2 presents predictions of ρ and ρ' for several human populations. For each population, there are two rows, one for the values expected in the absence of systematic pressure ($s = 0$), and one for the values expected if s is equal to the frequency of external immigrants. The "unrestricted" and "symmetric" columns were computed iteratively using equations (17) and (18), and the "approximate" columns were computed using (4) and (5). The "unrestricted" columns are based on the restricted estimate of the migration matrix, and the "symmetric" and "approximate" columns are based on the symmetric estimate of **M**.

The approximate predictions involve the additional assumption that s is much smaller than ρ. This assumption fails, however, for all groups except the !Kung, as indicated by column "s" of Table 2. Nonetheless, the symmetric and approximate predictions are, with one exception, in close agreement, indicating that our approximate formula is remarkably robust to failures of this assumption.

TABLE 3. The second eigenvalue (λ_2) of the symmetric estimate of **M**, half-life (HL) of convergence in the absence of external migration, ρ as predicted by migration data for newborns (N), adults (A), and by previous studies (P), and published estimates of ρ from genetic data.

			ρ			
			Migration data			
Population	λ_2	HL	N	A	P	Genetic data
!Kung (subset)	0.829	2	0.010	0.006	0.004	0.0067
Bedik	0.878	3	0.014	0.010	0.022	0.012
Gainj	0.899	3	0.042	0.027	0.0201	0.0332[a]
!Kung (all)	0.890	3	0.011	0.007	—	—
Papago	0.911	4	0.008	0.005	0.0077	0.0208
Bundi	0.916	4	0.011	0.007	0.002	0.008
Åland (all)	0.944	6	0.004	0.003	0.013	0.0097
Gidra	0.981	18	—	—	—	—
Åland (pre-1900)	0.982	19	0.005	0.004	—	—
Makiritare	0.986	24	0.016	0.010	—	—
Bougainville (subset)	0.992	43	0.025	0.016	—	—
Bougainville (all)	1.000	∞	0.026	0.017	0.0337	0.0477

[a] J. W. Wood. pers. comm.

The only substantial difference between the symmetric and approximate predictions occurs with the full Bougainville data set, which violates our assumption that the migration matrix is irreducible. The numerical results for this data set should be regarded with profound suspicion, and we provide them only as an example of how our formulas behave when applied to inappropriate data.

Effects of Different Estimators of ***M***.— The only appreciable discrepancies in Table 2 are between the unrestricted and symmetric columns, and are especially pronounced when systematic pressure is ignored (i.e., when $s = 0$). They may be due either to asymmetric migration or to discrepancies between the census sizes of local groups and the group sizes implied by the unrestricted estimate of **M**. Since both estimators are maximum likelihood estimators, a likelihood ratio test (Rao, 1973) with $g(g-1)/2$ degrees of freedom can be used to test the hypothesis that the underlying pattern of migration is symmetric and compatible with census sizes of local groups. It was possible to reject this hypothesis ($P < 0.05$) in most of the matrices studied (see Table 2). Thus, discrepancies between the symmetric and unrestricted columns are not surprising.

The difference between symmetric and unrestricted predictions is often small, even when migration does not conform to our assumptions. Thus, our results are reasonably robust to minor asymmetries in the pattern of migration. In some populations, however, the difference between symmetric and unrestricted predictions is large. For the Bundi, the unrestricted model yields a prediction that is three times as large as that of the symmetric model. For the irreducible subset of the Bougainville villages the prediction of the unrestricted model is larger by about 65%. Curiously, this is one of the cases in which it was not possible to reject the hypothesis that migration conforms to our model ($P = 0.40$). Thus, although the unrestricted and symmetric estimates of **M** lead to substantially different predictions of ρ, there is little reason to prefer one to the other. The symmetric estimate is in much better agreement with that obtained from genetic data for the full set of villages ($\hat{\rho} = 0.0477$, Friedlaender, 1975).

Even when the symmetric model can be rejected, it is not clear that the unrestricted model is more realistic. Although it makes no assumption about the pattern of migration, it does assume that group sizes are prevented from changing by population regulation. The changes in group size produced by migration each generation must be reversed by differential mortality and fertility of local groups. When migration is highly asymmetric, this requires high fecundity, a questionable assumption in species that reproduce slowly, such as our own. Thus, the unrestricted model may not be more realistic. The symmetric estimate of **M** is prob-

ably more reliable since, involving fewer parameters, its sampling variance must be smaller. We suspect that the symmetric estimate is less sensitive to "noise" in the data and will often be closer to reality even when the underlying pattern of migration is moderately asymmetric.

Convergence of Reduced Variances.—Unless the environment is extremely stable, natural populations are unlikely to be close to equilibria that take dozens of generations to reach. As shown in the appendix, ρ converges half way to its equilibrium value in $\log(\frac{1}{2})/2\log[(1-s)\lambda_2]$ generations. The half-life of convergence for several human populations is tabulated in Table 3, assuming s to be zero and using the eigenvalues of the symmetric estimate of **M**. Even without systematic pressure, most of these populations approach equilibrium rapidly. The half-life of convergence is often less than five generations, and convergence is even faster for realistic values of s. Thus, there is reason to be optimistic about the relevance of this theory to human populations. Similar results have been reported by Wood (1986).

Comparison with Genetic Data.—In Table 3, predictions of ρ and ρ' obtained from (7) and (9) are tabulated along with the predictions of previous studies and some estimates of ρ from genetic data. Except for the genetic value for the Bedik, which was computed from data in Jacquard (1974), the genetic values are also from previous studies. The figure for Bougainville in the "previous studies" column was calculated, using (3), from Friedlaender's (1975) published **R** matrix.

Except for Åland and Bougainville, our results fit the genetic data better than do the predictions of previous studies, especially in those populations where convergence is fastest. In the populations with fastest convergence, the estimate of ρ obtained from genetic data falls between the values predicted for newborns and adults. Where convergence is slower, the fit between observations and predictions is less impressive. This is to be expected for several reasons. Obviously, populations that converge slowly are less likely to be at equilibrium. Even at equilibrium, moreover, our model is probably a poor description of such populations. As discussed above, our treatment of external migration is most appropriate in populations that are relatively isolated from the outside world. When the effect of external migration is large compared with that of local migration, local genetic structure is strongly influenced by factors that are ignored by our model. Notice that, for most populations in Table 2, there is little difference between the predictions of ρ and ρ' obtained by setting $s = 0$, and those obtained with realistic values of s. In these populations, local genetic structure is dominated by the effects of local mobility and is little affected by assumptions concerning s. For the Papago, Åland (all periods), the Gidra, the Makiritare, and Bougainville (subset), the values of ρ predicted without systematic pressure exceed those with systematic pressure by 33%, 23%, 79%, 678%, and 507%, respectively. The populations in which the effect of systematic pressure is largest are also those that fit our theory most poorly.

Inferring m_e *and* $n_{..}$ *from Genetic Data.*—We know of only one population for which published data allow estimates of $n_{..}$ and m_e to be made from (10) and (11). Workman et al. (1973) published two estimates of ρ for the Papago, one computed by assigning individuals to village of origin, and the other by assigning them to village of residence. These are probably fair approximations to ρ' and ρ. Their figures are $\hat{\rho}' = 0.0208$ and $\hat{\rho} = 0.0162$, and there are 10 groups in their sample. The estimate obtained from (10) is $\hat{n}_{..} = 962$. The census size of this population is 5,102 and Workman et al. (1973) used one-third the census size as a rough estimate of effective size. Equation (10) suggests that their estimate may have been too large. The estimate of effective migration rate obtained from (11) is $\hat{m}_e = 0.142$, in only rough agreement with that obtained, using (8), from the migration matrix (see Table 1). This analysis should not be taken too seriously, since the statistical properties of (10) and (11) are as yet unknown.

This method is most likely to be useful with alleles of intermediate frequency since rare alleles provide little information about ρ. Slatkin (1981, 1985) has developed a method for estimating mobility from genetic data that should work better with rare

alleles and distantly related populations. An alternative method has been developed by Morton (1982).

Discussion and Conclusions

In humans and many other species, mortality is concentrated early in the life cycle and is low during the ages of dispersal and reproduction. Yet precisely the opposite is assumed by classical population-genetics models of migration and genetic drift. We have developed a theory incorporating a life cycle that is more appropriate for humans and species with similar life cycles.

These differences in the life cycle turn out to have a substantial effect on genetic statistics referring to newborns (i.e., individuals before migration). The expected value of r_0 (our synonym for Wright's F_{ST}) of newborns can be twice as great as that of adults. On the other hand, these differences in the life cycle have little effect on genetic statistics of adults.

These results imply that, in species like our own, geographic variation will appear larger if newborns are sampled for genetic data than if only adults are sampled. Thus, the age structure of the sample is a potent confounding influence in studies of population structure. This remark also applies to quantitative characters with additive genetic bases, since among-group variance in such characters is proportional to ρ (Wright, 1951; Rogers and Harpending, 1983).

We define a new measure of mobility, the effective migration rate, and show that Wright's (1931, 1943) formula, $\rho = (4n_e m_e + 1)^{-1}$, applies more generally than has been appreciated. An estimate of m_e can be obtained either from **M** or from genetic data, and this statistic should be useful for comparisons between populations. Unlike most statistics used in studies of population structure, it does not confound the effects of mobility and group size.

Some authors predict genetic variation using symmetrized estimates of the migration matrix and others use unconstrained estimates. These differences arise from differences in the evolutionary models being used, and can lead to substantially different answers, even when the data provide no basis for preferring one to the other. Hence, some of the differences between empirical studies may arise from the statistical methods that are used. To resolve this problem, we introduce a maximum likelihood estimator of the migration matrix that is compatible with the assumptions of our model.

In most of the populations studied, the predictions of our model are closer to estimates of ρ obtained from genetic data than are those of previous versions of the migration matrix model. The genetic estimates frequently fall between the values predicted for adults and newborns. The fit is not as good in populations where local migration is weak compared to external migration, but this is as it should be since our model is most appropriate for populations that are relatively isolated from the outside world.

Although the age structure of the sample has been a confounding influence, it is potentially rich in information. If genetic statistics are computed separately for individuals before and after migration, the difference between $\hat{\rho}$ and $\hat{\rho}'$ can be used to estimate both effective population size and effective migration rate directly from genetic data.

Acknowledgments

We thank Joseph Felsenstein, Lynn Jorde, and Elizabeth Cashdan for their comments, and Deirdre Sanders for helping with the data analysis.

Literature Cited

Bodmer, W. F., and L. L. Cavalli-Sforza. 1968. A migration matrix model for the study of random genetic drift. Genetics 59:565–592.

———. 1974. The analysis of genetic variation using migration matrices, pp. 45–61. *In* J. F. Crow and C. Denniston (eds.), Genetic Distance. Plenum, N.Y.

Carmelli, D., and L. L. Cavalli-Sforza. 1976. Some models of population structure and evolution. Theoret. Popul. Biol. 9:329–359.

Cavalli-Sforza, L. L., and A. Piazza. 1975. Analysis of evolution: Evolutionary rates, independence and treeness. Theoret. Popul. Biol. 8:127–165.

Chiang, A. C. 1974. Fundamental Methods of Mathematical Economics, 2nd Ed. McGraw-Hill, N.Y.

Coale, A. J. 1972. The Growth and Structure of Human Populations: A Mathematical Investigation. Princeton Univ. Press, Princeton, NJ.

Courgeau, D. 1974. Migration, pp. 351–387. *In* A. Jacquard (ed.), The Genetic Structure of Populations. Springer-Verlag, N.Y.

Crow, J. F., and M. Kimura. 1970. An Introduction to Population Genetics Theory. Harper & Row, N.Y.

Felsenstein, J. 1975. A pain in the torus: Some dif-

ficulties with models of isolation by distance. Amer. Natur. 109:359–368.

———. 1976. The theoretical population genetics of variable selection and migration. Ann. Rev. Genet. 10:253–280.

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

Fix, A. G. 1978. The role of kin-structured migration in genetic microdifferentiation. Ann. Hum. Genet. 41:329–339.

Friedlaender, J. S. 1975. Patterns of Human Variation: The Demography, Genetics, and Phenetics of the Bougainville Islanders. Harvard Univ. Press, Cambridge, MA.

Harpending, H. C., and T. Jenkins. 1974. !Kung population structure, pp. 137–164. *In* J. F. Crow and C. Denniston (eds.), Genetic Distance. Plenum, N.Y.

Harpending, H. C., and R. Ward. 1982. Chemical systematics and human populations, pp. 213–256. *In* M. Nitecki (ed.), Biochemical Aspects of Evolutionary Biology. Univ. Chicago Press, Chicago, IL.

Hiorns, R. W., G. A. Harrison, A. J. Boyce, and C. F. Kuchemann. 1969. A mathematical analysis of the effects of movement on the relatedness between populations. Ann. Hum. Genet. 32:237–250.

Hiorns, R. W., G. A. Harrison, and J. B. Gibson. 1977. Genetic variation in some Oxfordshire villages. Ann. Hum. Biol. 4:197–210.

Jacquard, A. 1974. The Genetic Structure of Populations. Springer-Verlag, N.Y.

Jorde, L. B. 1979. The genetic structure of the Åland Islands, Finland. Ph.D. Diss., Univ. New Mexico, Albuquerque.

———. 1980. The genetic structure of subdivided human populations: A review, pp. 135–208. *In* J. H. Mielke and M. H. Crawford (eds.), Current Developments in Anthropological Genetics, Vol. 1. Plenum, N.Y.

Jorde, L. B., P. L. Workman, and A. W. Eriksson. 1982. Genetic microevolution in the Åland Islands, Finland, pp. 333–365. *In* M. H. Crawford and J. H. Mielke (eds.), Current Developments in Anthropological Genetics, Vol. 2. Plenum, N.Y.

Lalouel, J. M. 1977. The conceptual framework of Malécot's model of isolation by distance. Ann. Hum. Genet. 40:355–360.

Langaney, A., and J. Gomila. 1973. Bedik and Niokholonko: Intra and inter-ethnic migration. Hum. Biol. 45:137–150.

Latter, B. D. H., and J. A. Sved. 1981. Migration and mutation in stochastic models of gene frequency change. II. Stochastic migration with a finite number of islands. J. Math. Biol. 13:95–104.

Long, J. 1986. The allelic correlation structure of Gainj- and Kalam-speaking people. I. The estimation and interpretation of Wright's F-statistics. Genetics 112:629–647.

Malcolm, L. A. 1970. Growth and development in New Guinea—A study of the Bundi people of the Madang District. Monog. Ser. No. 1, Institute of Human Biology, Madang, Papua New Guinea.

Malcolm, L. A., P. B. Booth, and L. L. Cavalli-Sforza. 1971. Intermarriage patterns and blood group gene frequencies of the Bundi people of the New Guinea Highlands. Hum. Biol. 43:187–199.

Malécot, G. 1948. Les Mathématiques de l'Hérédité. Masson & Cie, Paris, France.

———. 1950. Quelques schemas probabilistes sur la variabilité des populations naturelles. Ann. l'Univ. Lyon, Sci. 13:37–60.

———. 1969. The Mathematics of Heredity. Freeman, San Francisco, CA.

———. 1973. Isolation by distance, pp. 72–75. *In* N. E. Morton (ed.), Genetic Structure of Populations. Univ. Hawaii Press, Honolulu.

Morton, N. E. 1973. Prediction of kinship from a migration matrix, pp. 119–123. *In* N. E. Morton (ed.), Genetic Structure of Populations. Univ. Hawaii Press, Honolulu.

———. 1975. Kinship, information and biological distance. Theoret. Popul. Biol. 7:246–255.

———. 1982. Estimation of demographic parameters from isolation by distance. Hum. Hered. 32:37–41.

Morton, N. E., C. Miki, and S. Yee. 1968. Bioassay of population structure under isolation by distance. Amer. J. Hum. Genet. 20:411–419.

Morton, N. E., S. Yee, D. E. Harris, and R. Lew. 1971. Bioassay of kinship. Theoret. Popul. Biol. 2:507–524.

Nagylaki, T. 1979. The island model with stochastic migration. Genetics 91:163–176.

———. 1980. The strong-migration limit in geographically structured populations. J. Math. Biol. 9:101–114.

———. 1983. The robustness of neutral models of geographical variation. Theoret. Popul. Biol. 24: 268–294.

Neel, J. V., and F. Salzano. 1967. Further studies on the Xavante Indians. X. Some hypotheses-generalizations resulting from these studies. Amer. J. Hum. Genet. 19:554–574.

Ohtsuka, R., T. Kawabe, T. Inaoka, T. Akimichi, and T. Suzuki. 1985. Inter- and intra-population migration of the Gidra in lowland Papua: A population-ecological analysis. Hum. Biol. 57:33–45.

Rao, C. R. 1973. Linear Statistical Inference and its Applications, 2nd Ed. Wiley & Sons, N.Y.

Rogers, A. R., and H. C. Harpending. 1983. Population structure and quantitative characters. Genetics 105:985–1002.

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

———. 1985. Rare alleles as indicators of gene flow. Evolution 39:53–65.

Smith, C. A. B. 1969. Local fluctuations in gene frequencies. Ann. Hum. Genet. 32:251–260.

Smouse, P. E., V. J. Vitzthum, and J. V. Neel. 1981. The impact of random and lineal fission of the genetic divergence of small human groups: A case study among the Yanomama. Genetics 98:179–197.

Strang, G. 1976. Linear Algebra and its Applications. Academic Press, N.Y.

Sved, J. A., and B. D. H. Latter. 1977. Migration and mutation in stochastic models of gene frequency change. J. Math. Biol. 5:61–73.

Wagener, D. K. 1973. An extension of migration matrix analysis to account for differential immigration from the outside world. Amer. J. Hum. Genet. 25:47–56.

WAHLUND, S. 1928. The combination of populations and the appearance of correlation examined from the standpoint of the study of heredity (in German). Hereditas 11:65–106.

———. 1975. Composition of populations and of genotypic correlations from the viewpoint of population genetics, pp. 224–263. *In* K. Weiss and P. Ballonoff (eds.), Demographic Genetics. Dowden, Hutchinson & Ross, Stroudsburg, PA. Translated from Hereditas 11:65–106 (1928).

WARD, R. H., AND J. V. NEEL. 1970. Gene frequencies and microdifferentiation among the Makiritare Indians. IV. A comparison of a genetic network with ethnohistory and migration matrices; a new index of genetic isolation. Amer. J. Hum. Genet. 22:538–561.

WEIR, B. S., AND C. C. COCKERHAM. 1984. Estimating F-statistics for the analysis of population structure. Evolution 38:1358–1370.

WOOD, J. W. 1986. Convergence of genetic distances in a migration matrix model. Amer. J. Phys. Anthro. *In press.*

WOOD, J. W., P. L. JOHNSON, R. L. KIRK, K. MCLOUGHLIN, N. M. BLAKE, AND F. A. MATHESON. 1982. The genetic demography of the Gainj of Papua New Guinea. I. Local differentiation of blood group, red cell enzyme, and serum protein allele frequencies. Amer. J. Phys. Anthro. 57:15–25.

WORKMAN, P. L., H. C. HARPENDING, J. M. LALOUEL, C. LYNCH, J. D. NISWANDER, AND R. SINGLETON. 1973. Population studies on southwestern Indian tribes. VI. Papago population structure: A comparison of genetic and migration analyses, pp. 166–194. *In* N. E. Morton (ed.), Genetic Structure of Populations. Univ. Hawaii Press, Honolulu.

WORKMAN, P. L., AND J. D. NISWANDER. 1970. Population studies on southwestern Indian tribes. II. Local genetic differentiation in the Papago. Amer. J. Hum. Genet. 22:24–49.

WRIGHT, S. 1931. Evolution in Mendelian populations. Genetics 16:97–159.

———. 1943. Isolation by distance. Genetics 28:114–138.

———. 1951. The genetical structure of populations. Ann. Eugen. 15:323–354.

———. 1969. Evolution and the Genetics of Populations. II. The Theory of Gene Frequencies. Univ. Chicago Press, Chicago, IL.

Corresponding Editor: M. K. Uyenoyama

APPENDIX

In this appendix we modify the recurrence equations of Malécot (1950), Bodmer and Cavalli-Sforza (1968), and Smith (1969) to deal with: a) the stochastic effects of migration implied by model B of the life cycle (see above), b) variance about $\bar{p}$ rather than about π, and c) both newborn and adult allele frequencies. We derive expressions for the equilibrium matrix of genetic covariances and for the Wahlund variances of newborns and adults.

Definitions

The expectation of a random variable will be denoted by a superposed tilde, i.e., $\tilde{x} = \mathrm{E}\{x\}$. The set of individuals that disperses from the ith to the jth local group will be referred to as the "ijth migrant set." Let

n_{ij} = the size of the ijth migrant set,
$n_{\cdot j} = \sum_i n_{ij}$, the size of group j after migration,
$n_{i\cdot} = \sum_j n_{ij}$, the size of group i before migration,
$n_{\cdot\cdot} = \sum_{ij} n_{ij}$, the total population size,
$\mathbf{N} = [n_{ij}]$, a matrix of migrant set sizes, and
$m_{ij} = n_{ij}/n_{\cdot j}$, the fraction of the jth local group comprising immigrants from group i.

As is customary, we assume that the m_{ij} and n_{ij} are constant from generation to generation, and that the stochastic matrix $\mathbf{M} = [m_{ij}]$ is ergotic and aperiodic (so that it has exactly one eigenvalue equal to unity). We also define

q_{ij} = the frequency of allele A in the ijth migrant set,
$q_{i\cdot}$ = the frequency of A in local group i prior to migration,
$q_{\cdot j} = \sum_i m_{ij} q_{ij}$, the frequency of A in the jth local group among adults after migration among local groups,
s = the proportion of each local group exchanged each generation with a "continent" with unchanging allele frequency π,
u_j = the frequency of allele A among external migrants to group j, and
p_j = the allele frequency in group j after both local and continental migration.

Where it is necessary to distinguish quantities referring to particular generations, we write $p_j^{(t)}$, $q_{ij}^{(t)}$, etc.

These definitions and assumptions imply that

$$p_j^{(t+1)} = (1 - s) \sum_i m_{ij} q_{ij}^{(t+1)} + s u_j^{(t+1)}. \qquad (14)$$

The moments of p_j depend on those of the $q_{i\cdot}$. If selection is absent and migratory propensities are independent of genotype, then the conditional expectation of $q_{ij}^{(t-1)}$, given the genetic structure of the previous generation, is $p_i^{(t)}$, the frequency of A among adults in group i in the previous generation. Consequently, equation (14) can be rewritten in matrix notation as

$$p^{(t+1)} = (1 - s)\mathbf{M}^{\mathrm{T}} p^{(t)} + s\pi\mathbf{1} + \epsilon^{(t+1)} \qquad (15)$$

where

p = a column vector of group allele frequencies after local and continental migration,
$\epsilon^{(t+1)} = p^{(t+1)} - \mathrm{E}\{p^{(t+1)} | p^{(t)}\}$, a column vector of deviations produced by genetic drift, and
$\mathbf{1}$ = a column vector with each element equal to unity.

This recursion is fundamental to the migration matrix model as developed by Bodmer and Cavalli-Sforza (1968) and Smith (1969). These authors and others after them have used this equation to study dispersion of group allele frequencies about π, the continental allele frequency. Since π is ordinarily unknown and unknowable, however, empirical studies of population

structure necessarily deal with dispersion about $\bar{p}$, the current population mean allele frequency.

Let

w_i = $n_{\cdot i}/n_{\cdot\cdot}$, relative population size,
w = $[w_i]$, a column vector of relative population sizes,
$\bar{p}$ = $w^{\mathrm{T}}p$, the weighted mean of group allele frequencies,
$\mathbf{C^*}$ = the conditional expectation of $\epsilon\epsilon^{\mathrm{T}}$, given p', that is, the matrix of variances and covariances of the effects of one generation of genetic drift, and
$\mathbf{C}$ = the expectation of $\epsilon\epsilon^{\mathrm{T}}/\bar{p}(1-\bar{p})$, a matrix of normalized variances and covariances.

Note that $(\mathbf{I} - \mathbf{1}w^{\mathrm{T}})p$ is a vector of deviations from $\bar{p}$, where the superscript "T" denotes matrix transpose, and $\mathbf{I}$ is the identity matrix. We define a vector of normalized deviations,

$$z = \frac{(\mathbf{I} - \mathbf{1}w^{\mathrm{T}})p}{[\bar{p}(1-\bar{p})]^{1/2}}.$$

Applying this transformation to both sides of (15) produces

$$z^{(t+1)} = (1-s)\mathbf{L}^{\mathrm{T}}z^{(t)} + (\mathbf{I} - \mathbf{1}w^{\mathrm{T}})\epsilon^{(t)}[\bar{p}(1-\bar{p})]^{-1/2},$$

where $\mathbf{L} = \mathbf{M}(\mathbf{I} - w\mathbf{1}^{\mathrm{T}})$. Our assumption that $\mathbf{M}$ is ergotic and aperiodic implies that it has exactly one eigenvalue equal to unity and that its associated right and left eigenvectors are w and $\mathbf{1}^{\mathrm{T}}$, respectively. Consequently, $\mathbf{L} = \mathbf{M} - w\mathbf{1}^{\mathrm{T}}$. Note that $(\mathbf{I} - w\mathbf{1}^{\mathrm{T}})^2 = (\mathbf{I} - w\mathbf{1}^{\mathrm{T}})$, implying that $\mathbf{L}^t = \mathbf{M}^t(\mathbf{I} - w\mathbf{1}^{\mathrm{T}})$. It is convenient to define the zero'th power of $\mathbf{L}$ as $\mathbf{L}^0 \equiv \mathbf{M}^0(\mathbf{I} - w\mathbf{1}^{\mathrm{T}}) = \mathbf{I} - w\mathbf{1}^{\mathrm{T}}$.

The normalized dispersion matrix is $\mathbf{R} = [r_{ij}] = zz^{\mathrm{T}}$, and follows, in expectation,

$$\tilde{\mathbf{R}}^{(t+1)} = (1-s)^2\mathbf{L}^{\mathrm{T}}\tilde{\mathbf{R}}^{(t)}\mathbf{L} + (\mathbf{L}^0)^{\mathrm{T}}\mathbf{C}\mathbf{L}^0. \tag{16}$$

Equation (16) can be applied iteratively to obtain, for large t,

$$\tilde{\mathbf{R}}^{(t+1)} = \sum_{l=0}^{t}(1-s)^{2l}(\mathbf{L}^l)^{\mathrm{T}}\mathbf{C}\mathbf{L}^l. \tag{17}$$

This equation differs from the analogous formulas of Malécot (1973), Carmelli and Cavalli-Sforza (1976), and Smith (1969) only in the definitions of $\mathbf{L}$ and $\mathbf{C}$. If the form of $\mathbf{C}$ is known, successive terms in (17) can be added in a computer program until the result no longer changes. The result is a prediction of the normalized genetic dispersion matrix at equilibrium between migration and genetic drift.

The Effect of One Generation of Genetic Drift

Under life cycle B, genetic drift occurs during migration as well as during population regulation, and the variance introduced depends on $\mathbf{N}$, the matrix of sizes of migrant sets. We assume that mating within local groups is random so that the effective size of local groups (Wright, 1969) is the same as their actual size, and also that individuals migrating from i to j are a random subset of group i. Each gene can therefore be treated as an independent, random draw from the gene pool of group i in the previous generation, and the number of copies of allele A among migrants from i to j in generation $t+1$ is binomially distributed with parameters $2n_{ij}$ and $p_i^{(t)}$. Thus, $\mathrm{E}\{q_{ij}^{(t+1)}\} = p_i^{(t)}$, and

$$\mathrm{Var}\{q_{ij}^{(t+1)}\} = p_i^{(t)}(1-p_i^{(t)})/2n_{ij}.$$

We are interested in the dispersion of p, which depends on the first and second moments of

$$\epsilon_j^{(t+1)} = (1-s)\sum_i m_{ij}(q_{ij}^{(t+1)} - p_i^{(t)}) + s(u_j^{(t+1)} - \pi).$$

The expectation of ϵ_j is zero, as are the offdiagonal entries of $\mathbf{C^*}$. The diagonal entries are

$$\mathbf{C}_{jj}^* = (1-s)^2\sum_i m_{ij}^2\mathrm{Var}\{q_{ij}|p_i\} + s^2\mathrm{Var}\{u_j\},$$
$$= \frac{(1-s)^2}{2n_{\cdot j}}\sum_i m_{ij}p_i(1-p_i) + \frac{s\pi(1-\pi)}{2n_{\cdot j}},$$

since $m_{ij} = n_{ij}/n_{\cdot j}$. The substitution $s\bar{p}(1-\bar{p}) \simeq s\pi(1-\pi)$ should have little effect on the answer because when s is small the contribution of this term is negligible and when it is large $\bar{p} \simeq \pi$. We also substitute $\bar{p}(1-\bar{p})(1-\tilde{r}_{ii}) = p_i(1-p_i)$, where $\tilde{r}_{ii}$ is the ith diagonal entry of $\tilde{\mathbf{R}}$. With these substitutions,

$$\mathbf{C}_{jj}^* \simeq \frac{\bar{p}(1-\bar{p})}{w_j}\left[\frac{(1-s)^2\sum_i m_{ij}(1-\tilde{r}_{ii}) + s}{2n_{\cdot\cdot}}\right],$$
$$= \frac{\bar{p}(1-\bar{p})h_j}{w_j},$$

where h_j is the term in brackets above. Thus, the matrix of variances and covariances of the increments due to genetic drift is $\mathbf{C^*} = \bar{p}(1-\bar{p})\mathbf{W}^{-1}\mathbf{H}$, where $\mathbf{H} = \mathrm{Diag}\{h_i\}$ and $\mathbf{W} = \mathrm{Diag}\{w_i\}$.

$\mathbf{C}$ is obtained from $\mathbf{C^*}$ as follows: $\mathbf{C} = \mathrm{E}\{\epsilon\epsilon^{\mathrm{T}}/p(1-p)\} = \mathrm{E}\{\mathrm{E}[\epsilon\epsilon^{\mathrm{T}}|\bar{p}]/\bar{p}(1-\bar{p})\} = \mathrm{E}\{\mathbf{C^*}/\bar{p}(1-\bar{p})\} = \mathbf{C^*}/\bar{p}(1-\bar{p}) = \mathbf{W}^{-1}\mathbf{H}$, since $\mathbf{C^*}/\bar{p}(1-\bar{p})$ is a constant. Note that although $\mathbf{C}$ is conditioned on $\bar{p}$, it is not a function of $\bar{p}$. This implies that ρ, which is also conditioned on $\bar{p}$, is also independent of $\bar{p}$. Our "unrestricted" and "symmetric" predictions of ρ for adults were obtained using this formula and (17), as discussed above.

Dispersion of Newborn Allele Frequencies

Let $p'^{(t)}$ denote the vector of allele frequencies among newborns in generation t. Our assumptions about the life cycle imply that the population regulation component of genetic drift occurs at reproduction. Hence, newborn allele frequencies are related to those of their parents by $p' = p + e$, where e is a vector of deviations due to that component of drift. Under random mating, the expectation of e_i is zero and its variance is $p_i(1-p_i)/2n_{i\cdot}$. The conditional expectation of $e^{(t+1)}(e^{(t+1)})^{\mathrm{T}}$, given $p'^{(t)}$, is $\bar{p}^{(t)}(1-\bar{p}^{(t)})\mathbf{W}^{-1}\mathbf{D}$, where $\mathbf{D} = \mathrm{Diag}\{(1-\tilde{r}_{ii})/2n_{\cdot\cdot}\}$. The newborn dispersion matrix is

$$\tilde{\mathbf{R}}' = \tilde{\mathbf{R}} + (\mathbf{L}^0)^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{D}\mathbf{L}^0. \tag{18}$$

Approximate Results

These results can be greatly simplified if **N** is assumed symmetric, that is, if the number of migrants traveling from group i to j in a generation is the same as the number traveling from j to i. This should often hold approximately where migration is conservative. Let $\mathbf{A} = \mathbf{N}/n_{..}$. The ijth element of **A** is the proportion of the entire population that moves from group i to group j in a generation, and our assumption implies that **A** is symmetric. It is related to **M** by $\mathbf{M} = \mathbf{AW}^{-1}$.

Consider the matrix $\mathbf{X} = \mathbf{W}^{-\frac{1}{2}}\mathbf{A}\mathbf{W}^{-\frac{1}{2}}$. Since **A** is symmetric, so is **X**, and the spectral theorem (Strang, 1976) ensures that it can be written as $\mathbf{X} = \mathbf{S}\mathbf{\Lambda}^*\mathbf{S}^T$, where **S** is an orthogonal matrix of eigenvectors, and $\mathbf{\Lambda}^*$ is a diagonal matrix of eigenvalues. This implies that the diagonal form of **M** is $\mathbf{M} = [\mathbf{W}^{\frac{1}{2}}\mathbf{S}]\mathbf{\Lambda}^*[\mathbf{S}^T\mathbf{W}^{-\frac{1}{2}}] = \mathbf{U}\mathbf{\Lambda}^*\mathbf{V}^T$.

We denote the eigenvalues of **L** and **M** by λ_i and λ_i^*, respectively, and assume that those of **M** are indexed in descending order. **1** and w^T are eigenvectors of $\mathbf{M}^T$ with eigenvalue $\lambda_1^* = 1$, and the definition of $\mathbf{L}^1$ implies that they are also eigenvectors of that matrix, but with eigenvalue $\lambda_1 = 0$. For $i > 1$, $\lambda_i^* = \lambda_i$, and the corresponding eigenvectors of **L** and **M** are identical. Thus, $\mathbf{L} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$, where $\mathbf{\Lambda} = \mathrm{Diag}\{\lambda_i\}$.

In addition to the symmetry assumption, we use the approximation

$$\sum_i m_{ij}\tilde{r}_{ii} \simeq \sum_i w_i\tilde{r}_{ii} = \rho.$$

Consequently, $h_i = h = [2n_{..}]^{-1}[(1 - s)^2(1 - \rho) + s]$, and $\mathbf{C} = h\mathbf{W}^{-1}$. Note that $(\mathbf{L}^t)^T\mathbf{W}^{-1}\mathbf{L}^t = \mathbf{V}\mathbf{\Lambda}^{2t}\mathbf{V}^T$. Using (17) and the formula for the sum of a geometric series, we have, as $t \to \infty$

$$\tilde{\mathbf{R}} = \mathbf{V}\mathbf{B}\mathbf{V}^T, \tag{19}$$

where **B** is diagonal with diagonal entries $\beta_1 = 0$, and

$$\beta_i = \frac{(1 - \rho)(1 - s)^2 + s}{2n_{..}[1 - (1 - s)^2\lambda_i^2]}$$

$$\simeq \frac{1 - \rho}{2n_{..}[1 - (1 - s)^2\lambda_i^2]},$$

if $i \neq 1$ and $s \ll \rho$. Equation (19) is not necessarily the diagonal form of $\tilde{\mathbf{R}}$ because, unless group sizes are equal, the matrix $\mathbf{V} = \mathbf{W}^{-\frac{1}{2}}\mathbf{S}$ is not orthogonal. On the other hand, we have found the diagonal form of the weighted R-matrix, $\tilde{\mathbf{R}}^* = \mathbf{W}^{\frac{1}{2}}\tilde{\mathbf{R}}\mathbf{W}^{\frac{1}{2}} = \mathbf{S}\mathbf{\Delta}\mathbf{S}^T$.

Using the same approximations we obtain, for newborns,

$$\tilde{\mathbf{R}}' = \mathbf{V}\mathbf{B}'\mathbf{V}^T, \tag{20}$$

where the diagonal entries of $\mathbf{B}'$ are $\beta_1' = 0$, and

$$\beta_i' \simeq \left(\frac{1 - \rho}{2n_{..}}\right)\frac{2 - (1 - s)^2\lambda_i^2}{1 - (1 - s)^2\lambda_i^2} \quad \text{if } i \neq 1.$$

These matrices are related to ρ and ρ' by $\rho = \sum_i w_i\tilde{r}_{ii} = \mathrm{Trace}\{\tilde{\mathbf{R}}^*\}$, and $\rho' = \sum_i w_i\tilde{r}_{ii} = \mathrm{Trace}\{\tilde{\mathbf{R}}'^*\}$.

The Rate of Convergence

The convergence of $\tilde{\mathbf{R}}$ is governed by the convergence of the β_j, which is measured by

$$\frac{\beta_j^{(\infty)} - \beta_j^{(t)}}{\beta_j^{(\infty)}} = 1 - \{1 - [(1 - s)\lambda_j]^2\} \cdot \sum_{\tau=0}^{t-1}[(1 - s)\lambda_j]^{2\tau},$$

$$= [(1 - s)\lambda_j]^{2t}.$$

Since λ^2 is the largest eigenvalue of **L**, the convergence of $\tilde{\mathbf{R}}$ is asymptotically determined by $[(1 - s)\lambda_2]^{2t}$. The process converges halfway to equilibrium in $\log(0.5)/\{2\log[(1 - s)\lambda_2]\}$ generations. If $(1 - s)\lambda_2 = 0.9$, this half-life is about three generations, and it is only 34 generations when $(1 - s)\lambda_2 = 0.99$. The second eigenvalue is often a good deal less than 1.0 in human data, so convergence is rapid.