20

Estimation of Effective Population Size and Migration Parameters from Genetic Data

JOSEPH E. NEIGEL

In recent years, the arrival of molecular genetic data has brought forth a renaissance in the field of population genetics. Classical population genetic models defined the allele as the unit of variation, and were formulated in terms of the frequencies of alleles within populations. Now DNA sequence data has redefined the description of genetic variation in terms of the nucleotide sequence differences that distinguish alleles. In those cases in which these differences are the result of an orderly stepwise process of change, they can be used to infer relationships of descent among sequences. New theoretical models, often referred to as "coalescent models" (Hudson, 1990) have been developed to incorporate this richer, more historical view of genetic variation. Other developments include methods that can detect highly variable portions of the genome. While these methods have gained most attention in applications to forensics and paternity analysis, they are now being applied to population genetics as well (see other chapters in this volume, and references therein).

Conservation biology is also a rapidly developing field, with a new emphasis on genetic variation in both natural and managed populations (see, for example, the volume edited by Soulé, 1987). The combination of molecular data, new theoretical models, and computerized methods of analysis holds the promise of new and powerful tools that can be applied to conservation genetics. There are essentially two ways that molecular genetic markers can be used in conservation biology. The first is motivated by the concern that genetic variation is an important determinant of population viability and adaptability. The aim is simply to use genetic markers to indicate current levels and distributions of genetic variation. The main assumption is that a given set of genetic markers is representative of the variation that is relevant to conservation. The second use is the measurement of processes, such as migration, which may be important for ecological as well as genetic reasons, but are expected to produce measurable effects on patterns of genetic variation. Genetic marker data can thus provide an "indirect" view of these processes, but its analysis and interpretation must be based upon theoretical population genetic models. A major assumption of indirect uses of genetic markers is the suitability of the models upon which these analyses are based.

The purpose of this chapter is to consider how molecular population genetic

data can be analyzed to provide inferences about two parameters that are regarded as important to conservation genetics: effective population size and migration. The term effective population size may refer to any of several quantities, but all are concerned with the rate at which some measure of genetic variation changes in a finite population (Ewens, 1982). Migration may also be defined in several ways, with a principal distinction between the migration *rate*, which is the proportion of individuals that move between populations, and the migration *distance*, which is how far a typical individual moves in one generation (Slatkin, 1985). Because these parameters represent the rates of processes, they must be estimated "indirectly" from genetic marker data. This chapter is largely concerned with the theoretical models that relate these parameters to molecular population genetic data.

There are many ways to approach the question of what is the "best way" to estimate either effective population size or migration parameters. The question that must be answered first is: What use will be made of these estimates? The method of choice from an evolutionary perspective may not be the best for conservation biology. In particular, the time-scales appropriate for evolutionary studies may be inappropriate for addressing the more immediate concerns of conservation biology. Therefore, along with offering some definitions of the quantities to be estimated, I will attempt to summarize some of the major reasons why they are of importance in conservation biology. These will hopefully serve as criteria for evaluating alternative approaches to estimating these parameters.

EFFECTIVE POPULATION SIZE

Much of the early development of population genetics theory incorporated a set of simplifying assumptions that together defined an "ideal population" (see, for example, Wright 1969). In an ideal population of N diploid individuals, each generation is constituted from a sample of $2N$ gametes drawn randomly from the individuals of the preceding generation. Thus every individual in the population is considered equal in reproductive capacity; there is no overlap in the individuals from one generation to the next; mating occurs at random; and the population remains at a constant size over time. Because gametes are sampled randomly, some alleles carried by gametes will, by chance, be represented more than other alleles. These sampling accidents are expected to produce a range of related effects that accumulate over time (Figure 20-1). These include random fluctuations of allele frequencies (a process referred to as genetic drift), loss of alleles from the population (allele extinction), and a decrease in heterozygosity (considered a form of "inbreeding"). Application of probability theory to the ideal population model allows predictions to be made about the expected rates of these processes that occur as consequences of gamete sampling.

There is variety of ways in which real populations can depart from the ideal population described above. These include unequal sex ratios, overlapping generations, differential reproductive success, and changes in population size. These deviations cause departures from predictions based on the ideal population model (a good introduction to this topic is provided by Hartl and Clark, 1989).

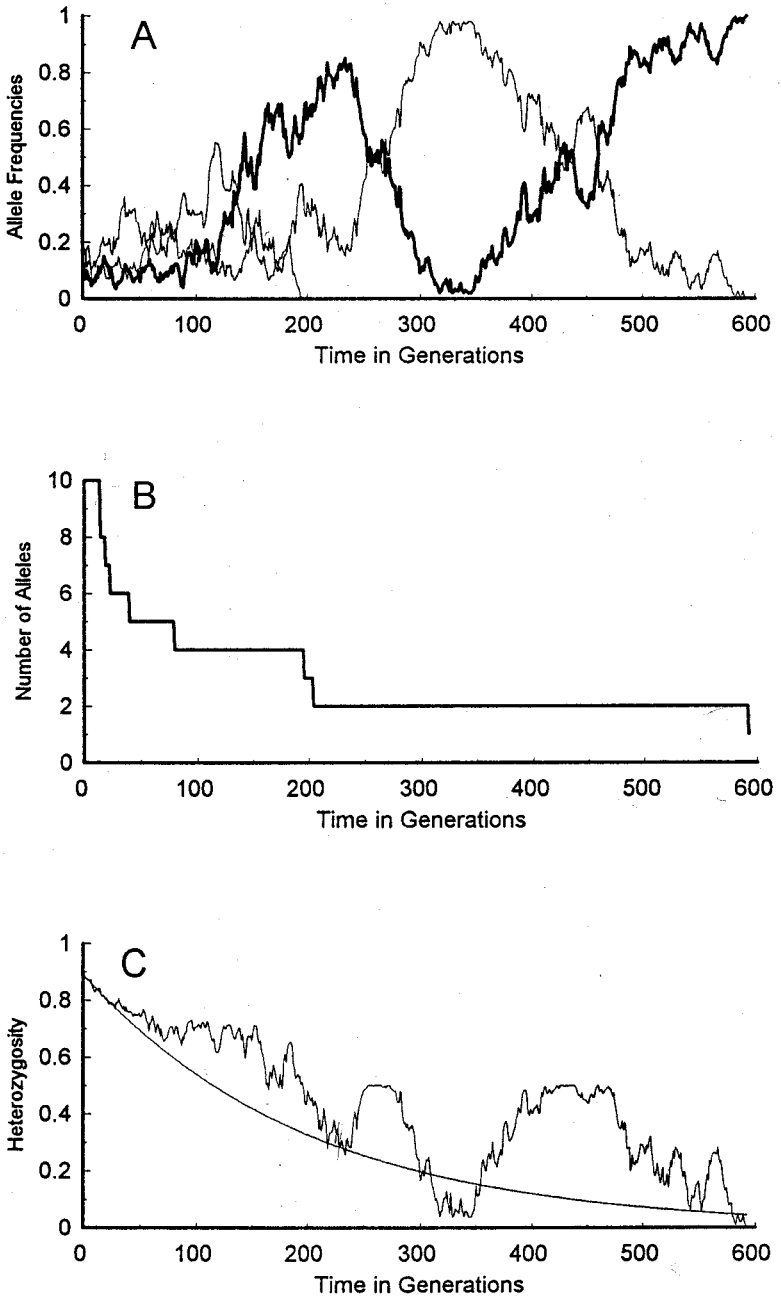

Figure 20-1 Computer simulation of the effects of genetic drift. An ideal population was initialized with 10 alleles at one locus. In panel (A) the frequencies of three alleles are monitored, with the trajectory of the allele that eventually reaches fixation in bold. In panel (B) the decline in the number of alleles over time is shown. In panel (C) the smooth curve represents the expected decline in heterozygosity, which can be contrasted with the more erratic trajectory of the actual changes in heterozygosity at a single locus.

Depending on the species, effective population sizes may be nearly the same as the "census size" (e.g., vertebrate examples in Nunney and Elam, 1994) or orders of magnitude lower (e.g., plant and invertebrate examples in Orive, 1993). For any nonideal population, its effective size, represented as N_e, is the size that an ideal population would be that behaved in the same way. More precise definitions of effective population size must be made specifically with respect to one of the process that is dependent on population size (Ewens, 1982). Thus, the inbreeding effective population size is defined with respect to the probability of homozygosity due to common ancestry; the variance effective population size is defined with respect to the variance of gene frequencies over time; and the extinction effective population size is defined with respect to the loss of alleles from the population. The distinctions among these definitions of N_e, which are in some cases numerically equivalent, are often considered trivial. However, these distinctions may assume considerable importance in conservation biology, where a specific effect of finite population size is considered detrimental. Management plans or breeding programs may affect each of these population size effects differently (see review by Simberloff, 1988).

Inbreeding Effective Population Size

Inbreeding depression is a reduction in fitness that occurs in the progeny of matings between related individuals. At least two mechanisms may account for inbreeding depression (reviewed by Charlesworth and Charlesworth, 1987). The one that is most easily demonstrated is an increase in the proportion of genetic loci that are homozygous for rare, recessive alleles with deleterious effects. These alleles are generated by mutation, and can persist as rare alleles in large outbreeding populations because they will generally occur as heterozygotes, which are not subject to selection. In small populations, there is a higher probability that matings will occur between related individuals that carry the same alleles. The increase in the occurrence of homozygotes for recessive deleterious alleles will result in inbreeding depression. Over time, selection may eliminate these rare alleles in small populations, and so inbreeding depression may be reduced in populations that have been small for many generations. In effect, the combination of inbreeding and selection can "purge" these deleterious alleles. A second potential cause of inbreeding depression is a decrease in the proportion of genetic loci that are heterozygous for pairs of alleles that confer greater fitness as heterozygotes than they would as either of the alternative homozygotes. The classic example of heterozygote superiority is the human hemoglobin locus, for which there is an allele that causes sickle cell anemia as a homozygote, but resistance to malaria (without severe anemia) as a heterozygote (Allison, 1955). Few other examples of heterozygote advantage are so well documented but, if it were a general phenomenon a reduction in the proportion of heterozygous loci would reduce fitness. Unlike the effect of deleterious recessive alleles, selection could not temper a population against the effects of loss of heterozygosity. Furthermore, the potential to restore heterozygosity will be reduced if alleles are lost from a population (mutation can replenish alleles, but only over relatively long timescales). Additional mechanisms of inbreeding depression may involve multilocus

interactions, such as epistasis, although these complicated mechanisms are seldom considered.

There are two occasions when it would be desirable to obtain an estimate of inbreeding effective population size. First, an estimate of inbreeding N_e from an undisturbed population could serve as a reference point. Inbreeding depression would be expected to occur if N_e suddenly became much smaller. The ideal estimate of N_e for this purpose would represent a long-term average. Thus, a measure of N_e that was relatively insensitive to short-term changes would be desirable. The second occasion would be in a threatened or managed population. An estimate of N_e could be used to evaluate the effectiveness of a management strategy in reducing inbreeding, especially if N_e estimates from undisturbed populations were available for comparison. In this case, the ideal estimate of N_e should reflect the current status of the population, and not its history prior to disturbance.

Variance Effective Population Size

Genetic drift has been considered to be an important force in evolution because it provides opportunities for evolutionary changes beyond those that would occur by natural selection acting alone, including genetic divergence between isolated populations. In small populations, random changes in gene frequency that occur from genetic drift may reduce the effects of selection. In the "shifting-balance theory of evolution" proposed by Sewall Wright (1932), evolution within populations can become highly constrained when genetic variation becomes caught in a balance of selective forces. This balance of selective forces may prevent the formation of new, and potentially advantageous, combinations of genes. Genetic drift within small populations can loosen this constraint, and thereby allow new combinations of genes to form, which may in turn precipitate sudden shifts to a new balance of selective forces. In a species divided into many separate populations, a sudden shift to a new and better adapted combination of genes in one population could be propagated to others. In some cases, this may involve extinction and recolonization of local populations, In somewhat anthropomorphic terms, each population is an "experiment" in finding new adaptations, and the results of a successful experiment can be shared by the entire species. The validity of Wright's theory has been extensively debated, and is still an open question. It is certainly reasonable to expect that the theory will apply more to some species than to others. An important implication for conservation biology is that processes that may be viewed as detrimental to an individual population may enhance the adaptive potential of the species as a whole.

Extinction Effective Size

The extinction of alleles from a population is irreversible if there are effectively no sources to replenish them. Thus, for a species that consists of essentially a single small population, loss of alleles is a special concern. The upper limit for heterozygosity is set by the number of alleles at each locus, as is the number of possible combinations of genes that can be formed. However, it is not clear at

which point the loss of alleles begins to impact the viability or adaptive potential of a population. As indicated above, a general benefit from heterozygosity per se has been difficult to demonstrate. Furthermore, variation in quantitative traits may be rather resistant to the loss of alleles (Robertson, 1966). These concerns become important in the design of management or captive breeding programs, because the best strategies for maintaining a high extinction N_e will reduce the inbreeding and variance N_e values. For example, allelic diversity can be preserved by maintaining a large number of small populations. Within each population, high levels of genetic drift and inbreeding will occur, and alleles will be lost. But because different populations will not lose the same alleles, a large number of alleles can be maintained indefinitely among all the populations as a whole. Furthermore, remixing separately maintained populations will restore much of the original heterozygosity. Thus, if the goal is to eventually reestablish a species into its former range, this strategy has merit.

METHODS FOR THE ESTIMATION OF EFFECTIVE POPULATION SIZE

Temporal Method

It would seem relatively straightforward to estimate variance effective population size by determining allele frequencies from a single population in successive generations, and calculating the variance, F, directly from these frequencies. Development of this approach has been based on a theoretical population in which mating is completely random, there is no migration from other populations, and generations are discrete and nonoverlapping. Waples (1989) has provided a general synthesis of the temporal method, and tested its performance with simulations. The expected variance in allele frequency, $E(F)$ over t generations in a population with variance effective size N_e is

$$E(F) \approx 1 - \left(1 - \frac{1}{2N_e}\right)^t$$

Note that this equation relates N_e to the *expected* variance in allele frequency. Because genetic drift is random, a specific prediction cannot be made about the net change in frequency of a particular allele over a period of time (see Figure 20-1). Thus, observations for multiple time points and/or multiple independent alleles are essential for the estimation of N_e. In addition, differences in allele frequencies between samples drawn from a population will occur from random sampling error as well as any actual temporal changes in the population. Estimation of N_e from temporal samples thus requires both a method for combining data for multiple alleles and/or multiple time points, and a correction for sampling error. Several formulas have been proposed for combining data from several alleles at a single locus. Although no one method appears to work best in all situations, Nei and Tajima's (1981) estimate, $\hat{F}_c$ appears to be generally reliable

$$\hat{F}_c = \frac{1}{K} \sum_{i=1}^{K} \frac{(x_i - y_i)^2}{\frac{1}{2}(x_i + y_i) - x_i y_i}$$

where K is the number of alleles at the locus, x_i is the frequency of the ith allele in the sample collected at generation 0, and y_i is the frequency of the ith allele in the sample collected at generation t. A weighted average, $\hat{F}_c$, for multiple loci can be calculated as

$$\hat{F}_c = \frac{\sum K_j F_{cj}}{\sum K_j}$$

where the index j specifies the locus.

If the effective population size N_e is not too small and the number of generations t is not too large, the exponential increase in the expected population allele frequency variance, $E(F)$ can be approximated as a linear function of t

$$E(F) \approx \frac{t}{2N_e}$$

With the addition of a correction for sampling variance (suggested by Krimbas and Tsakas, 1971), the expected variance in the estimator, $\hat{F}_c$, can be expressed in terms of effective population size, number of generations, and sample sizes

$$E(\hat{F}_c) \approx \frac{1}{2S_0} + \frac{1}{2S_t} + \frac{t}{2N_e}$$

where S_0 is the size of the sample taken at generation 0, and S_t is the size of the sample taken at generation t. This has led to the following as an estimate of variance effective population size

$$\hat{N}_e = \frac{t-2}{2\left[\hat{F}_c - \frac{1}{2S_0} - \frac{1}{2S_t}\right]}$$

The distribution of this estimator is approximately χ^2 with n degrees of freedom equal to one less than the number of alleles used to estimate $\hat{F}_c$

$$\chi^2 = \frac{n\hat{F}}{E(\hat{F})}$$

This estimator involves several mathematical approximations that appear valid in most situations but need not be made for those situations in which they are questionable (see Waples, 1989). In general, the method is most effective for populations with relatively small values of N_e. A typical application might involve 10 independent genetic markers, with samples of 100 individuals taken five generations apart. The theoretical accuracy of the estimate can be increased by increasing the number of genetic markers used, the length of the time interval (measured in generations) between samples, or the number of individuals sampled. However, the practical limitations of this approach may depend on how well the population under study fits the simple model upon which this estimator is based. Migration, in particular, has the potential to profoundly impact allele frequencies. If migrants are continuously received from a larger population or metapopulation

with relatively stable allele frequencies, temporal variation in the recipient population will be reduced, and temporal estimates of N_e will be biased upward (Nei and Tajima, 1981). On the other hand, episodic migration from sources with different allele frequencies may cause rapid, although transient, shifts in allele frequencies, which would bias temporal estimates of N_e downward. Selection on the genetic marker, or at loci linked to the markers may also introduce inaccuracies. This is likely to occur precisely under the conditions where estimates of N_e would be sought: small populations subject to inbreeding depression.

Estimates of N_e from DNA Sequences

The description of genetic variation in terms of DNA sequence differences rather than allele frequencies has led to the development of corresponding "gene genealogical" models in population genetics (reviewed by Hudson, 1990). Within a pedigree that traces the descent of individuals from their ancestors, the pedigrees of individual genes can also be traced. These genealogies are subject to the effects of finite population sizes. In particular there is an expected relationship between inbreeding effective population size and parameters that can be determined from gene genealogies.

At present, an estimate of N_e based on DNA sequence data must begin with some rather restrictive assumptions about DNA sequence evolution. The approximate mutation rate must be known, so that it is possible to convert measures based on DNA sequence differences into estimates of time. This mutation rate must also be high enough to provide sufficient variation for statistical analyses. The majority of these mutations should be simple single nucleotide substitutions, rather than length changes or rearrangements. Otherwise, it would be difficult to determine genealogical relationships among the sequences. Finally, there can be no recombination among DNA sequences. Recombination would create sequences with ambiguous genealogical relationships. In practice, these restrictions have limited most attempts at estimates of N_e from DNA sequence data to the use of animal mitochondrial DNA (mtDNA). In addition to coming closest to meeting the requirements listed above, animal mtDNA is relatively easy to isolate and analyze. However an important caveat is that, as a general rule, mtDNA appears to be inherited maternally in animals (Avise et al., 1987). Thus, only the effective number of females in a population can be estimated.

There are basically two steps to converting DNA sequence data to an estimate of effective population size. First, the DNA sequence data is used either to infer an mtDNA genealogy (also referred to as a "phylogeny") from which statistics are derived, or statistics may be estimated directly from the sequence data. The second step is to convert these statistics into an estimate of effective population size. In general, these methods do not estimate N_e directly, but rather the product $4N_e\mu$, often represented simply as θ, where μ is the mutation rate. An estimate of μ is therefore needed to convert an estimate of θ into an estimate of N_e.

A simple approach is suggested by Watterson's (1975) analysis of the number of segregating (variable) nucleotide sites expected for a sample of sequences. In a sample of n sequences that have a mutation rate μ for the entire sequence and are

drawn from a population of effective size N_e, the expected number of segregating sites $E(K)$ is approximated by

$$E(K) \approx \theta \sum_{i=1}^{n-1} \frac{1}{i}$$

Rearrangement of this expression provides an estimate of N_e that is based on a simple count of the number of segregating sites in a sample of DNA sequences and an estimate of μ. This approach avoids the problem of converting sequence data to estimates of divergence times. However, it still assumes that the occurrence of mutations follows a Poisson distribution, and that the parameter μ is known.

Other approaches have been based on the expected average divergence time for a pair of randomly chosen sequences, which is $2N_e$. Nei and Tajima (1981) suggested first calculating the nucleotide diversity index, π, defined as the average number of differences between two DNA sequences *at each nucleotide site* within a sequence (Nei and Li, 1979). Assuming the probability of multiple mutations at a single site is low enough to ignore, and the mutation rate μ (in this case, per nucleotide site) is known, effective population size can be estimated as

$$\hat{N}_e = \frac{\pi}{4\mu}$$

A constant of 4 rather than 2 appears in this expression because mutations can occur in both lines leading from two sequences back to a common ancestor. Avise et al. (1998) proceeded in a slightly different way, using the average estimated divergence time between mtDNA sequences as a direct estimate of $2N_e$. In place of a single mutation rate parameter, this approach can incorporate any method to convert sequence data to estimates of divergence time, and does not necessarily involve an explicit estimate of the mutation rate.

Felsenstein (1992) criticized the above approaches because they are based on statistics that represent only a fraction of the information potentially available in sequence data. This criticism is justified because the variance of these estimates tends to be rather high. To demonstrate this point, he developed a method that could be used if the gene genealogy for a random sample of sequences were known with complete accuracy. In this case, the distribution of time intervals between successive branch events in the genealogy contains all of the information that can be used to estimate N_e. A maximum-likelihood estimation procedure based on this distribution proved to have a much lower variance than methods based either on the number of segregating sites or on pairwise sequence comparisons. Unfortunately, gene genealogies cannot be determined with complete accuracy, and so this method could not be directly applied to real data sets.

More recently, Fu (1994) developed a method for estimating N_e from DNA sequence data that uses all of the information in an estimated gene genealogy. This method is illustrated for the hypothetical gene genealogy in Figure 20-2. Each branching event is numbered starting from the root of the genealogical tree. Coalescent theory predicts that as a consequence of the extinction of individual gene lineages over time, the time intervals between the deeper branch points in a

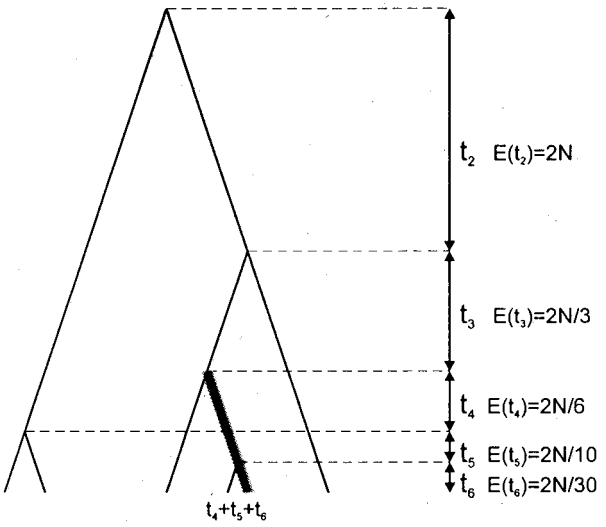

Figure 20-2 Expected branch lengths in a gene genealogy. Each expected branch length is the sum of expected time intervals between ordered branch points, and is proportional to effective population size.

genealogy will tend to be greater than those between the more shallow branches. More precisely, the expected time between the $(k-1)$th and kth branch points is

$$E(t_k) = \frac{4N_e}{k(k-1)}$$

From Figure 20-2 it can be seen that the length of each branch in the genealogy is the sum of one or more of these time intervals between branch points. Thus, expected time intervals between selected branch points can be summed to provide the expected time traversed by any given branch. For example, the expected length of the branch indicated by shading is

$$t_4 + t_5 + t_6$$

The number of mutations expected to occur in the branch is the product of its length and the mutation rate

$$E(n_i) = \mu l_i$$

By comparing the number of mutations observed on each branch with its expected length, each branch can furnish an estimate of θ. An overall estimate of θ can be based on a weighted average of estimates from each branch. The problem is to assign values to weighting factors to yield an estimate that is unbiased and has the minimum variance. This **b**est **l**inear **u**nbiased **e**stimator (BLUE) can be determined by solving a system of equations that include the weighting factors and the covariances of the estimates. A complication arises because, although the covariances can be determined from a coalescent model, they are functions of the parameter θ, which is itself being estimated. However, an iterative procedure can

be applied, starting with an initial estimate of θ (i.e., based on some other method) to yield a new estimate of θ, which is then fed back into the procedure.

A potential weakness of this method is that the development of the BLUE estimator is based on the topology of the gene genealogy, which itself may be inaccurate. In a series of simulations, it was shown that the combination of genealogies inferred by the UPGMA procedure (one of the simplest methods of inferring a genealogy) together with the BLUE procedure should provide good estimates of θ. Estimates with this combination, UPBLUE, had low variance and a small bias that can be easily corrected (Fu, 1994). A FORTRAN program that takes a distance matrix as its input and outputs the UPBLUE estimate of θ, along with other estimates of θ is available from Fu.

Unlike an estimate of N_e that has been based on temporal variation in allele frequencies observed over a set period of time, an estimate from a DNA sequence genealogy reflects the end results of a process that may have occurred over a very long period of time. For this reason, such methods are sometimes said to estimate "evolutionary effective population size" (Ball et al., 1990). The expected time for two DNA sequences to trace back to a single point in their genealogy is $2N_e$ generations. For species with large effective population sizes, this could be many thousands of generations. For many species, changes in population size that occurred during the Pleistocene (12 000 to 1.8 million years ago) could easily be reflected in sequence-based estimates of N_e.

As with temporal-based estimates of N_e, a basic assumption of DNA sequence-based estimates is that the population being sampled has been closed with respect to migration. Whether or not this assumption holds may be difficult to determine directly, because, even if migration is not presently occurring, it may have occurred recently enough in the past to influence the present genealogical structure of the population. At present, there are no methods available for the estimation of N_e from genetic data for populations with migration. However, the problem of migration has been studied in its own right, and there are methods for the estimation of migration parameters that are analogous to those used for estimation of N_e. A rational approach to the estimation of N_e should probably begin by defining the population to be studied with respect to its history and possible sources of migration. Methods appropriate for this are discussed in the remainder of this chapter.

MIGRATION

Migration is important as both an ecological and as a population genetic process. It has the immediate ecological effect of altering population size or density. Colonization of new habitats and recolonization of formerly occupied habitats are also direct consequences of migration. A major population genetic effect of migration is to reduce genetic divergence between populations, which might otherwise occur as a result of either genetic drift or natural selection. Theoretical results indicate that surprisingly small levels of migration, a few individuals per generation, can reduce divergence due to genetic drift. Thus migration is sometimes viewed as a constraining force in evolution that prevents both adaptation to local

conditions and the formation of new species (see for example, Mayr, 1963). However, migration also reduces inbreeding, and as proposed by Wright (1932), may allow the spread of favorable adaptations among populations.

The terms migration and gene flow are often used interchangeably, although the latter refers strictly to movements of genes within or between populations. Equating these terms assumes that migrants are as successful at reproduction as residents. A distinction also arises for species in which dispersal of gametes may occur independently of the movement of individuals. Thus it should be kept in mind that genetic measures of migration are actually measures of gene flow. In population genetics, the term migration is generally applied to any movement of individuals that affects the mating structure of populations. Most theoretical treatments assume that populations consist of discrete units, within which mating is random. In this case, movements of individuals between these populations constitute migration, and can be quantified by the migration rate m, the proportion of individuals that enter a population by migration. A different case is presented by individuals that are distributed continuously (rather than in discrete units) but tend to mate more often with nearby individuals. Here any movement may alter mating interactions, and so can be considered migration. A useful way of quantifying these movements in continuous populations is with the standard dispersal distance, σ_d, the standard deviation of the distances moved from the site of birth to the site of reproduction.

In the management of natural populations, it would be useful to use migration parameters to predict how demographic changes in one population would affect others. For example, through migration, one population (a source) could prevent a decline in size of another population (a sink). If this were known, it could be predicted that loss of the source population would also cause the loss of the sink population (Pulliam, 1988). However genetically based measurements of migration generally do not specify directionality. A single parameter, the migration rate, is estimated, and assumed to represent migration in both directions. Furthermore, there is no reason to expect that migration rates will remain constant as populations undergo demographic changes. Thus, it is important to consider the time-frame over which a migration estimate is relevant.

F_{ST} Based Methods of Estimating Migration Rate

Genetic drift and migration have opposite effects on the distribution of genetic variation among populations. Over time, genetic drift will result in the divergence of allele and genotype frequencies between isolated populations. Because migrants introduce alleles at frequencies that reflect their source populations, they reduce any genetic differences between populations. The basic theory that relates migration to the distribution of genetic variation was developed by Wright (1951, 1965). He defined a set of correlation coefficients ("F-statistics") in terms of the correlations between gametes. If gamete are represented as random variables that reflect their ancestry, then a correlation coefficient can be defined for the pairs of gametes that combine to form zygotes. Positive correlation coefficients result when gametes of common ancestry combine more frequently than expected. F_{ST} was defined as "the correlation between random gametes within a population, relative

to gametes of the total population" (Wright, 1965). If gametes drawn from the same population are more likely to have a common ancestor than gametes drawn from different populations, F_{ST} is positive. When allele frequencies vary among populations, this implies that gametes within an individual population are correlated, and F_{ST} has a positive value. Considering just one allele, F_{ST} can be equated with the standardized variance of the allele's frequency

$$F_{ST} = \frac{V(q)}{\bar{q}(1 - \bar{q})}$$

where $V(q)$ is the variance in frequency among populations, and $\bar{q}$ is the average frequency over populations.

F_{ST} is often used to estimate migration rate because a variety of theoretical models indicate a robust relationship between F_{ST} and the product of the migration rate and effective population size: $N_e m$. For example, for a model in which an infinite number of subpopulations exchange migrants at rate m, this relationship is

$$F_{ST} = \frac{1}{4Nm + 1}$$

Figure 20-3 shows diagrammatic representations of several other models, and

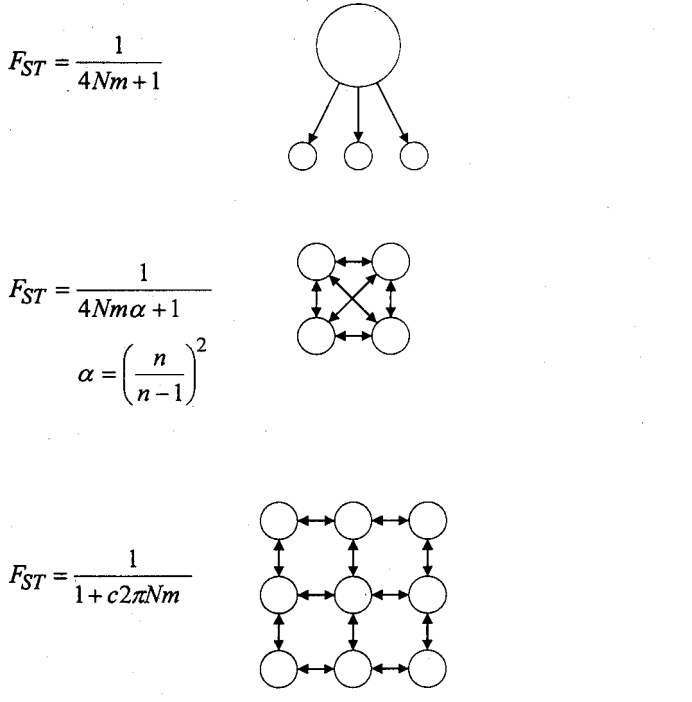

Figure 20-3 Several models of migration between populations, and the expected relationship between F_{ST} and Nm under each. (See Slatkin and Barton, 1989.)

the expected relationship between N_em and F_{ST} (see Slatkin and Barton, 1989). These models suggest that N_em can be determined by applying an equation of the form

$$N_em = \frac{1 - F_{ST}}{\alpha F_{ST}}$$

where α is a constant that depends on the model. However, there are some limitations to this approach. First, the expected value of F_{ST} is inconveniently small for values of N_em much over 10. Thus, in general, only relatively low rates of migration can be estimated with much precision. A second potential problem with the interpretation of F_{ST} is the possibility that the populations have not reached an equilibrium between genetic drift and migration. Crow and Aoki (1984) showed that the number of generations, t, for F_{ST} to be near an equilibrium value is

$$t \approx \frac{1}{2m + \dfrac{1}{2N_e}}$$

If the migration rate m is relatively large, then an equilibrium will rapidly be approached. However, if both m and $1/2N_e$ are small, the approach to equilibrium will occur slowly, on the order of $1/2N_e$ generations. Thus, while F_{ST} can be rapidly lowered by a high migration rate, the elevation of F_{ST} by a reduction in migration may be much slower. As a result, observations based on F_{ST} cannot distinguish between the immediate effects of ongoing migration and the residual effects of past migration.

There has also been some confusion over how F_{ST} should be estimated from actual data, and how it should be interpreted. Part of this confusion is due to the distinction between F_{ST} as a demographic parameter which can be defined without reference to genetic variation, and F_{ST} as a statistic which is calculated from genetic data (see Weir and Cockerham, 1984). This distinction becomes important when considering multiple alleles or multiple loci. A demographic definition of F_{ST} implies that there is only a single value, although F_{ST} statistics are expected to vary among alleles and loci. Weir and Cockerham (1984) developed a method to estimate F_{ST} as a single parametric value, which they defined as θ, to distinguish it from other definitions of F_{ST}. (Note that this use of the symbol θ differs from its use earlier in this chapter, where it represented $4N_e\mu$). Nei (1973) introduced the statistic G_{ST}, which is widely used because it can be calculated for multiple alleles and multiple loci. However, G_{ST} cannot be considered an estimate of a single parameter because it its expected to vary with the number of populations observed.

It is important to consider two components of variance in allele frequencies among population samples. One is the actual variance in allele frequencies among the populations, the other is the variance due to sampling a limited number of individuals from each population. If corrections are not applied for the latter component of variance, the estimate of F_{ST} will be upwardly biased. Methods for obtaining unbiased estimates of F_{ST} and related quantities have been developed by Nei and Chesser (1983) and Weir and Cockerham (1984). The latter method also corrects for the sampling error associated with a small number of populations.

Estimating Migration Parameters from DNA Data

As with estimates of effective population size, there are two basic kinds of genetic data that can be used to estimate migration parameters: allele frequencies in population samples and sets of DNA sequences sampled from populations. Some forms of DNA sequence variation do not reflect orderly changes of character states (e.g., microsatellite length variation), and so cannot be used to infer genealogical relationships among sequences. Such variation can be analyzed in essentially the same ways that allozyme variation is analyzed. However, estimates of F_{ST} that are based upon the variance in allele frequencies among populations are sensitive to mutation rates. For example, G_{ST} is generally expressed as the ratio of two quantities, D_{ST}/H_T (Nei, 1973). D_{ST} represents the covariance of genes within subpopulations, and is called the average gene diversity between subpopulations. H_T represents the variance of genes in the total population, and is called the gene diversity in the total population. Higher mutation rates increase both D_{ST} and H_T. For low mutation rates, the relationships between these gene diversity indices and mutation rates are approximately linear, so that the proportionate effects on both D_{ST} and H_T are nearly the same, and their ratio is nearly independent of mutation rate. However, as gene diversity approaches its maximum value of one, the rate at which gene diversity increases with mutation rate declines. In effect, gene diversity becomes "saturated" as a higher proportion of mutations occur in alleles that have already been differentiated by previous mutations. Because D_{ST} is generally higher than H_T, the saturation effect is greater for D_{ST}, and so D_{ST}/H_T becomes smaller. In practice, this effect is negligible if mutation rates are at least several orders of magnitude lower than migration rates, as is generally assumed for allozyme loci. However hypervariable makers, which may have mutation rates that approach or exceed migration rates, may be expected to exhibit lower values of G_{ST}. This suggests that extremely variable markers may be of limited usefulness in detecting and measuring genetic divergence between populations.

The basic theory represented by the use of F_{ST} as a population structure parameter has now been extended to include maternally transmitted mtDNA (Takahata and Palumbi, 1985) as well as genealogical relationships among DNA sequences (Slatkin, 1991). However, this approach does not represent a very effective use of mtDNA sequence data, since the entire mitochondrial genome is analyzed as if it were a single polymorphic locus. The information represented by the genealogical relationships among mtDNA sequences is essentially discarded. For an alternative, more appropriate use of mtDNA data, Slatkin and Maddison (1989) developed a method for estimating the product of effective population and migration rate (N_em) from gene genealogies. This method treats the geographic location of each individual as a character, and uses a parsimony analysis to determine the minimum number of migration events that could reconcile this location "character" with the mtDNA phylogeny. Computer simulations provide a simple function to convert the minimum number of migration events to an estimate of N_em.

Neigel et al. (1991) introduced a method for estimating σ_d, the standard deviation of the distances moved from the site of birth to the site of reproduction, from mitochondrial DNA data. The premise of the method is that for many species

the mutation rate of mtDNA is high relative to the rate at which new mutations are dispersed geographically. As a result, mtDNA lineages (the branches on a genealogical tree) should develop a hierarchical geographic structure that reflects their genealogical structure. This process can be modeled as a random walk, in which each mtDNA lineage originates at a unique geographic location and spreads outward at a rate dependent on σ_d. A prediction of this model is that the variance of the geographic locations of the individuals within a lineage should be roughly proportional to both the age of the lineage and σ_d. Geographical surveys of mtDNA variation, along with an estimate of the rate at which mtDNA sequences mutate, are needed for this method. A computer program, PHYFORM, is available to convert these data into estimates of σ_d. On the basis of simulation studies, Neigel and Avise (1993) found this approach to be fairly robust under various types of population density regulation and dynamic change. One advantage of this approach is that it allows separate estimates of σ_d to be made from lineages of different ages. It is thus possible to examine the possibility of historical changes in this parameter.

CONCLUSIONS

Genetic drift and migration operate over a range of time-scales, and are important determinants of both the immediate viability and the long-term adaptive potential of species. Genetic drift in small populations may result in inbreeding depression and loss of adaptive potential. However, genetic drift may also play a positive role in creating novel opportunities for evolutionary change. Migration can relieve inbreeding depression, but may also act as a constraining force in evolution by preventing local adaptation. Management decisions should be based on an awareness of both the positive and negative roles that genetic drift and migration may play.

Rates of genetic drift and migration can be estimated by the analysis of molecular genetic markers, but these are indirect estimates, based on the predictions of theoretical models. Methodological choices should be based on appropriate models and a clear understanding of the parameters they embody. Recent developments in both molecular genetics technology and population genetics theory have provided new methods for the study of genetic drift and migration in natural populations. In particular, methods based on DNA sequences and coalescence models have the potential to avoid the equilibrium assumptions of classical models, and to resolve historical patterns of genetic drift and migration.

ACKNOWLEDGMENTS

This chapter has benefited from collaborative work with J.C. Avise and R.M. Ball, and discussions with M. Slatkin. Supported by NSF/LEQSF grant (1992–1996)-ADP-02.

REFERENCES

Allison, A.C. 1955. Aspects of polymorphism in man. Cold Spring Harbor Symp. Quant. Biol. 20: 239–255.

Avise, J.C., J. Arnold, R.M. Ball, E. Bermingham, T. Lamb, J.E. Neigel, C.A. Reed, and N.C. Saunders, 1987. Intraspecific phylogeography: The mitochondrial DNA bridge between population genetics and systematics. Annu. Rev. Ecol. Syst. 18: 489–522.

Avise, J.C., R.M. Ball, and J. Arnold. 1988. Current versus historical population sizes in vertebrate species with high gene flow: A comparison based on mitochondrial DNA lineages and inbreeding theory for neutral mutations. Mol. Biol. Evol. 5: 331–344.

Ball, R.M., J.E. Neigel, and J.C. Avise. 1990. Gene genealogies within the organismal pedigrees of random mating populations. Evolution 44: 360–370.

Charlesworth, D. and B. Charlesworth. 1987. Inbreeding depression and its evolutionary consequences. Ann. Rev. Ecol. Syst. 18: 237–268.

Crow, J.F. and K. Aoki. 1984. Group selection for a polygenic behavioral trait: Estimating the degree of population subdivision. Proc. Natl. Acad. Sci. U.S.A. 81: 6073–6077.

Ewens, W.J. 1982. On the concept of effective population size. Theor. Pop. Biol. 21: 373–378.

Felsenstein, J. 1992. Estimating effective population size from samples of sequences: inefficiency of pairwise and segregation sites as compared to phylogenetic estimates. Genet. Res. 59: 139–147.

Fu, Y. 1994. A phylogenetic estimator of effective population size or mutation rate. Genetics 136: 685–692.

Hartl, D.L. and A.G. Clark. 1989. Principles of population genetics, 2d ed. Sunderland, Mass.: Sinauer Associates.

Hudson, R.R. 1990. Gene genealogies and the coalescent process. In: D. Futuyma and J. Antonovics, eds. Oxford Surveys in Evolutionary Biology, Vol. 7, pp. 1–44. New York: Oxford University Press.

Krimbas, C.B. and S. Tsakas. 1971. The genetics of *Dacus oleae*. V. Changes of esterase polymorphism in a natural population following insecticide control—selection or drift? Evolution 25: 454–460.

Mayr, E. 1963. Animal species and evolution. Cambridge, Mass.: Harvard University Press.

Nei, M. 1973. Analysis of gene diversity in subdivided populations. Proc. Natl. Acad. Sci. U.S.A. 70: 3321–3323.

Nei, M. and R.K. Chesser. 1983. Estimation of fixation indices and gene diversities. Ann. Hum. Genet. 47: 253–259.

Nei, M. and W.H. Li. 1979. Mathematical model for studying genetic variation in terms of restriction endonuclease. Proc. Natl. Acad. Sci. U.S.A. 76: 5269–5273.

Nei, M. and F. Tajima. 1981. Genetic drift and estimation of effective population size. Genetics 98: 625–640.

Neigel, J.E. and J.C. Avise. 1993. Application of a random walk model to geographic distributions of animal mitochondrial DNA variation. Genetics 135: 1209–1220.

Neigel, J.E., R.M. Ball, and J.C. Avise. 1991. Estimation of single generation migration distances from geographic variation in animal mitochondrial DNA. Evolution 45: 423–432.

Nunney, L. and D.R. Elam. 1994. Estimating the effective population size of conserved populations. Conserv. Biol. 8: 175–184.

Orive, M.E. 1993. Effective population size in organisms with complex life histories. Theor. Pop. Biol. 44: 316–340.

Pulliam, H.R. 1998. Sources, sinks and population regulation. Am. Nat. 132: 652.

Robertson, A. 1966. Artificial selection in plants and animals. Proc. R. Soc. Lond., Ser. B 164: 341–349.

Simberloff, D. 1988. The contribution of population and community biology to conservation science. Ann. Rev. Ecol. Syst. 19: 473–511.

Slatkin, W.M. 1985. Gene flow in natural populations. Ann. Rev. Ecol. Syst. 16: 393–430.

Slatkin, M. 1991. Inbreeding coefficients and coalescence times. Genet. Res. Camb. 58: 167–175.

Slatkin, M. and N.H. Barton. 1989. A comparison of three methods for estimating average levels of gene flow. Evolution 43: 1349–1368.

Slatkin, M. and W.P. Maddison. 1989. A cladistic measure of gene flow inferred from the phylogenies of alleles. Genetics 123: 603–613.

Soulé, M.E., ed. 1987. Viable populations for conservation. Cambridge, U.K.: Cambridge University Press.

Takahata, N. and S.R. Palumbi. 1985. Extranuclear differentiation and gene flow in the finite island model. Genetics 109: 441–457.

Waples, R.S. 1989. A generalized approach for estimating effective population size from temporal changes in allele frequency. Genetics 121: 379–391.

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

Weir, B.S. and C.C. Cockerham. 1984. Estimating F-statistics for the analysis of population structure. Evolution 38: 1358–1370.

Wright, S. 1932. The roles of mutation, inbreeding, crossbreeding and selection in evolution. Proc. 6th Int. Cong. Genet. 1: 356–366.

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

Wright, S. 1965. The interpretation of population structure by F-statistics with special regard to systems of mating. Evolution 19: 395–420.

Wright, S. 1969. Evolution and the genetics of populations, vol. 2. The theory of gene frequencies. Chicago: University of Chicago Press.