
CAUSES OF EVOLUTION 2

Populations are extremely complex entities. The Hardy–Weinberg model discussed in Chapter 1, useful as it is as a first approximation, ignores most of the complexities of actual populations. Populations are not infinitely large, and their sizes are rarely constant, so fluctuations in allele frequency can occur by chance. Populations are also subject to the systematic evolutionary forces of migration, mutation, and natural selection, all of which cause non-random or directional changes in allele frequency. Changes in allele frequency are changes in the genetic makeup of a population. Since **evolution** may be defined as cumulative change in the genetic makeup of a population resulting in increased adaptation to the environment, the fundamental process in evolution is change in allele frequency. In this chapter we focus on the various forces that change allele frequency. These forces hold the key to understanding the origin and maintenance of genetic variation.

RANDOM GENETIC DRIFT

Random genetic drift refers to chance fluctuations in allele frequency which occur, particularly in small populations, as a result of random sampling among gametes. To be specific, a population of nine diploid organisms arises

from a sample of just 18 gametes out of an essentially infinite pool of gametes. Because small samples are frequently not representative, an allele frequency in the sample may differ from that in the entire pool of gametes. In fact, if the number of gametes in a sample is represented as $2N$ (in the above example, $2N = 18$), the probability that the sample contains exactly 0, 1, 2, . . . , $2N$ alleles of type A is given by the successive terms in the expansions of $(pA + qa)^{2N}$, where p and q are the respective allele frequencies of A and a in the parental generation, and $p + q = 1$. The probability that the sample contains exactly i alleles of type A can be shown to equal

$$\frac{(2N)!}{i!(2N - i)!} p^i q^{2N-i} \quad (2.1)$$

Equation 2.1 is often called the **binomial probability**, and the symbol $i!$ (i **factorial**) represents the product of all integers up to and including i ; that is, $i! = i(i - 1)(i - 2) \dots (3)(2)(1)$. Similarly, $(2N)! = (2N)(2N - 1)(2N - 2) \dots (3)(2)(1)$. (For the sake of generality, the number $0!$ is defined to equal 1.)

To consider a concrete example, suppose a monoecious population of size $N = 9$ contains exactly 16 A alleles. Then $p = 16/(2N) = 16/18 = 8/9$, and $q = 1 - p = 1/9$. In the next generation, the probability that A becomes fixed (i.e., $p = 1$) is given by Equation 2.1 with $i = 18$, namely

$$\frac{18!}{18!0!} (8/9)^{18} (1/9)^0 = 0.12$$

Similarly, the probability that A is **lost** (i.e., that a becomes fixed) is given by Equation 2.1 with $i = 0$. This probability is very much smaller, namely

$$\frac{18!}{0!18!} (8/9)^0 (1/9)^{18} = 6.7 \times 10^{-18}$$

Thus, the probability that the population remains **unfixed or segregating** (i.e., neither A nor a becomes fixed) in the next generation equals $1 - 0.12 - (6.7 \times 10^{-18}) = 0.88$.

PROBLEM

1

A college student admires *Phlox cuspidata* and always keeps two plants in her dormitory room, which are grown from two seeds chosen at random from the plants of the previous year. (Because she is careful to ensure that mating is random, including an equal likelihood of cross- and self-fertilization, her method of perpetuation is equivalent to choosing four gametes at random and uniting them to form the two plants of the next generation.) Suppose that in one year her plants contain two Adh^a and two Adh^b alleles (Adh = alcohol dehydrogenase gene). Use Equation 2.1 to calculate the various probabilities that the next year's population contains exactly 0, 1, 2, 3, or 4 Adh^a alleles.

ANSWER Use Equation 2.1 with $N = 2$ (so $2N = 4$) and p the allele frequency of Adh^a . In the initial population, $p = 2/4 = 1/2$, so $q = 1 - p = 1/2$. In Equation 2.1, i represents the number of A (i.e., Adh^a) alleles in the next generation. The probability that $i = 0$ is thus $(4!/0!4!)(1/2)^0(1/2)^4 = 1/16 = 0.0625$; the probability that $i = 1$ is $(4!/1!3!)(1/2)^1(1/2)^3 = 4/16 = 0.25$; the probability that $i = 2$ is $(4!/2!2!)(1/2)^2(1/2)^2 = 6/16 = 0.375$; similarly, the probability that $i = 3$ is 0.25, and the probability that $i = 4$ is 0.0625.

The hypothetical population of *Phlox cuspidata* with $N = 2$ in Problem 1 is very small indeed, yet it is instructive because it typifies many of the important features of random genetic drift. Suppose there were many such populations, each of size $N = 2$. The calculations in Problem 1 imply that, in only one generation of random genetic drift, 6.25 percent of the populations would lose the Adh^a allele, 6.25 percent would become fixed for the Adh^a allele, and the other 87.5 percent would remain segregating. However, among the segregating populations, there is a spread in allele frequency: some populations have $p = 1/4$, some have $p = 1/2$, and some have $p = 3/4$. Of course, random genetic drift continues generation after generation. Each of the unfixed populations produces a spread of allele frequencies in the following generation, including some cases in which the Adh^a allele is newly fixed or newly lost. The fixed populations, whether fixed for Adh^a or Adh^b , remain fixed except in the unlikely event of mutation. Thus, as time goes on, a smaller and smaller fraction of the populations continues to segregate both alleles.

For the population in Problem 1 in which $p = 1/2$, what is the probability that the population is still segregating after two generations of random genetic drift? (Hint: Use the proportions calculated in Problem 1, along with Equation 2.1, to calculate the probability that A is fixed or lost in two generations; then determine the probability that the population is still segregating by subtraction.)

PROBLEM 2

ANSWER The allele frequency of A after one generation of random genetic drift is either 0, $1/4$, $1/2$, $3/4$, or 1 with respective probabilities 0.0625, 0.25, 0.375, 0.25 and 0.0625, as determined in Problem 1. When $p = 0$ or $p = 1$ the allele frequency remains 0 or 1, respectively. The probability of loss when $p = 1/4$, $1/2$, or $3/4$ is $(4!/0!4!) \times (1/4)^0(3/4)^4 = 0.31641$, $(4!/0!4!)(1/2)^0(1/2)^4 = 0.0625$, and $(4!/0!4!)(3/4)^0(1/4)^4 = 0.00391$, respectively. The overall probability of loss in two generations is therefore $(0.0625)(1) + (0.25)(0.31641) + (0.375)(0.0625) + (0.25)(0.00391) + (0.0625)(0) = 0.16602$. In a similar manner, it can be calculated that the probability of fixation in two generations is also 0.16602. Thus, the proportion of populations still expected to be segregating after two generations is $1 - 0.16602 - 0.16602 = 66.8$ percent. This compares with the 87.5 percent still segregating after one generation.

After random genetic drift has continued for a sufficient number of generations, most populations become fixed for one allele or the other. After only four generations of random genetic drift in the population with $N = 2$, for example, the proportion of unfixed populations is down to 38 percent. Moreover, the populations that are still segregating are nearly as likely to have any one allele frequency as any other; in the $N = 2$ example, the ratio of unfixed populations having $p = 1/4$ to those with $p = 1/2$ to those with $p = 3/4$ after four generations is 0.32:0.36:0.32. Fixation continues as time goes on. After 19 generations with $N = 2$, fully 99.6 percent of the populations have become fixed; among the remaining 0.4 percent of unfixed populations, however, the ratio of those having 1, 2, or 3 A alleles is still 0.32:0.36:0.32.

The progressive spreading out of allele frequencies due to random genetic drift and the accompanying accumulation of populations in which one or another allele has become fixed also occurs in larger populations, but in larger populations a correspondingly longer time is required for the effects to become pronounced. A computer-generated example based on random numbers is shown in Figure 1. Each small graph gives the number of A alleles in 19 successive generations of random genetic drift in a population of size $N = 9$ (so $2N = 18$). As is apparent from the figure, individual populations behave erratically. In seven populations (29 percent), the A allele became fixed; in six populations (25 percent), the A allele was lost. The remaining 11 populations (46 percent) were still segregating after 19 generations. (Compare this 46 percent with the corresponding value of 0.4 percent for populations with $N = 2$.) Among the segregating populations, however, the final allele frequency was as likely to be one value as any other. The principal conclusion from the example with $N = 2$ is that the overall pattern of allele frequency change due to random genetic drift is predictable—the

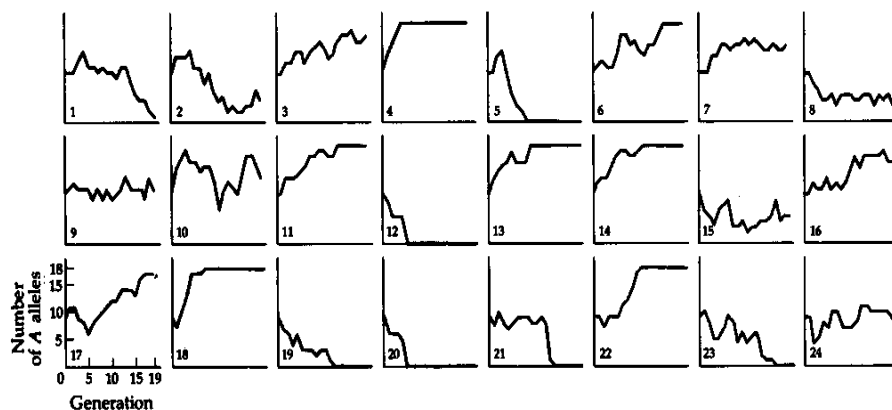


Figure 1. Change of allele frequency by random genetic drift over 19 generations in 24 hypothetical populations of size $N = 9$.

allele frequencies progressively spread out as the proportion of fixed populations steadily increases. The overall conclusion from Figure 1 is that allele frequencies in any single population behave so erratically that prediction of allele frequency change in any one population is virtually impossible.

In view of the unpredictability of changes in allele frequency in any single population, the effects of random genetic drift on allele frequency are often discussed in terms of the expected changes in allele frequency when averaged over a very large number of populations. A framework for thinking about the problem from this perspective is illustrated in the model in Figure 2. Here an infinitely large initial population is imagined to be split up into a large number of local populations or subpopulations, each of size N , which

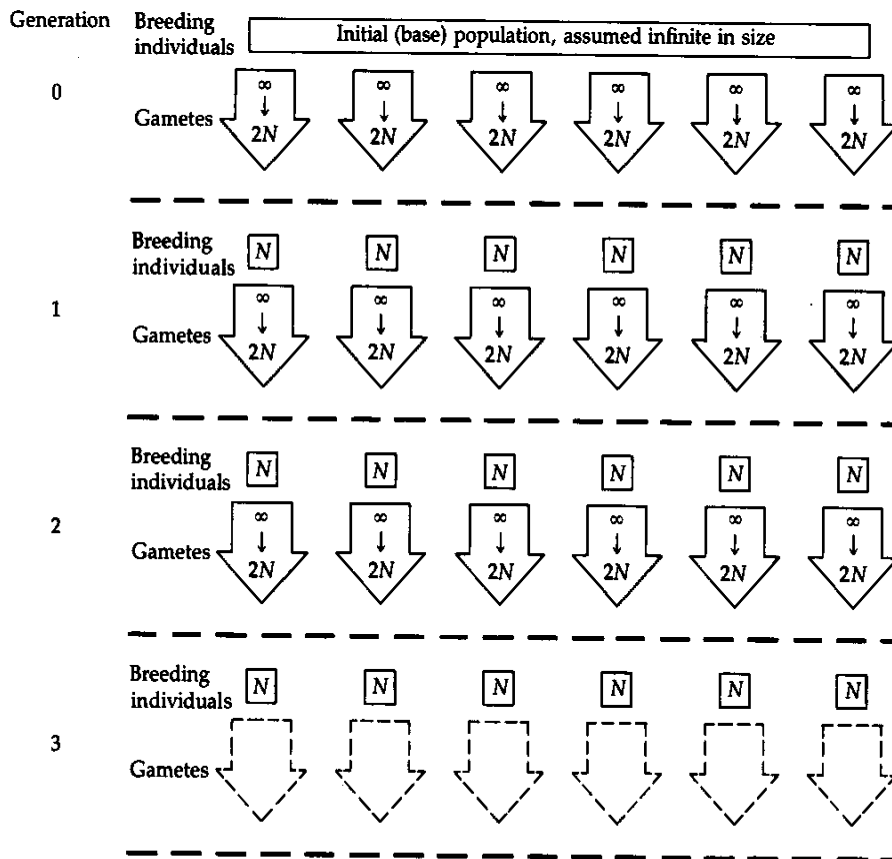


Figure 2. Model for analyzing effects of random genetic drift. Each of the subpopulations founded from the initial large population (shown by vertical columns of boxes and arrows) is assumed to be genetically isolated from the others. Each subpopulation produces an infinite number of gametes, of which $2N$ are chosen at random to form the next generation's breeding population. Random genetic drift results from sampling error in this process.

correspond to the populations in Figure 1. (As will be explained shortly, N is technically the **effective population number**.) The subpopulations are assumed to satisfy the Hardy–Weinberg assumptions of diploid organism, sexual reproduction, nonoverlapping generations, no mutation, and no selection. In addition, we assume that

1. Many independent subpopulations exist, each of constant size N .
2. Random mating occurs within each subpopulation.
3. No migration occurs among the subpopulations.
4. Each subpopulation has an equal number of males and females.
5. Every individual has an equal chance of contributing successful gametes to the next generation.
6. The number of successful gametes per individual conforms to a Poisson (random) distribution. [With this assumption, the probability that an individual contributes i successful gametes, when the mean number per individual is λ , is given by $(\lambda^i/i!)\exp(-\lambda)$.]

In short, the subpopulations in Figure 2 are theoretically uncomplicated or *ideal*, and reference will be made to such ideal populations throughout this chapter. The importance of the concept of ideal populations is that they provide a standard of comparison for other populations that violate the above assumptions.

Results of the kind of population structure in Figure 2 are shown for an actual case in Figure 3, which records the history of 19 generations of random genetic drift in 107 subpopulations of *Drosophila melanogaster* (Buri 1956). Each population was initiated with 16 bw^{75}/bw heterozygotes ($bw = brown\ eyes$) and maintained at a constant size of 16 by randomly choosing 8 males and 8 females as the breeding population in each generation. The horizontal axis in Figure 3 gives the number of populations having 0, 1, 2, . . . , 32 bw^{75} alleles. The overall pattern of change in allele frequency in Figure 3 is apparent: the initially humped distribution of allele frequency (in the background) gradually becomes flat (foreground) as populations fixed for bw^{75} or bw begin to pile up. By about 18 generations, half the populations are fixed for one allele or the other, and among the unfixed populations, the distribution of allele frequencies is essentially flat.

Consequences of Random Genetic Drift

The observed pattern of change in allele frequency shown in Figure 3 is very nearly that expected theoretically for an ideal population. Although the full-blown theory of random genetic drift is conceptually simple (it involves mathematical methods for the repeated application of the binomial formula to each subpopulation, as was illustrated for the case $N = 2$ in Problems 1

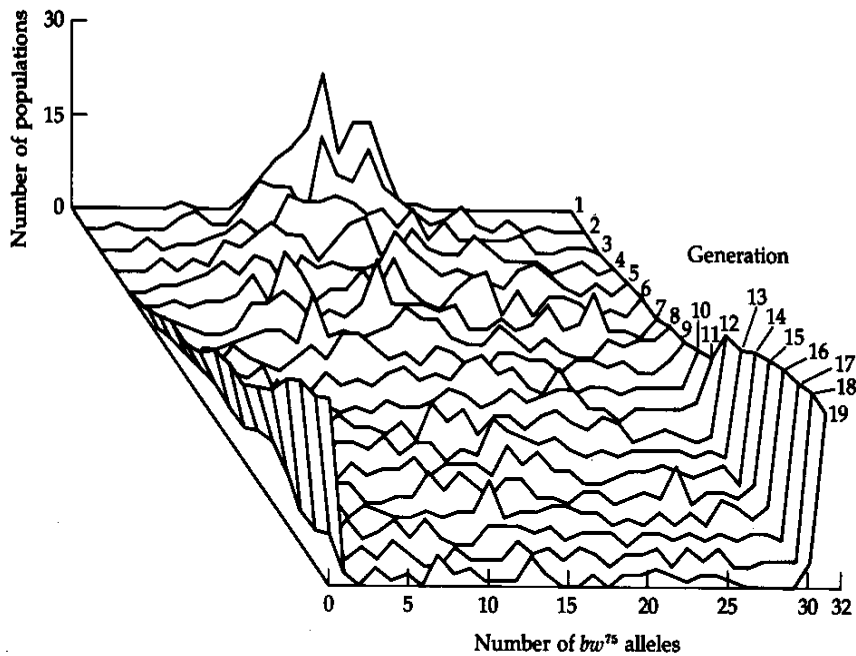


Figure 3. Random genetic drift in 107 actual populations of *Drosophila melanogaster*. Each of the initial 107 populations consisted of 16 bw^{75}/bw heterozygotes ($N = 16$; $bw = brown$ eyes). From among the progeny in each generation, eight males and eight females were chosen at random to be the parents of the next generation. The horizontal axis of each curve gives the number of bw^{75} alleles in the population, and the vertical axis gives the corresponding number of populations. (Data from Buri 1956.)

and 2), the actual calculations are formidable (see, for example, Crow and Kimura, 1970, and Kimura and Ohta, 1971). Some results of the theory are summarized in the two families of curves in Figure 4, which represent the theoretical distributions of allele frequency among unfixed (segregating) populations after various times (t) measured in units of N generations. In Figure 4(a), all populations have an initial allele frequency of 0.5 (as in the actual populations in Figure 3); after about $t = 2N$ generations, the distribution of allele frequency is essentially flat, and by this time about half the populations are still unfixed. The distributions in Figure 4 refer only to those populations that are unfixed. As time goes on, more and more of the populations do become fixed, and thus the distributions progressively pile up at 0 and 1, as demonstrated in Figure 3. Indeed, in Figure 4, the area under each curve is equal to the proportion of unfixed populations, and this area becomes progressively smaller as time goes on. Figure 4(b) shows what happens when the initial allele frequency is 0.1; here the distributions are highly asymmetrical, and the distribution of allele frequency does not become flat until about

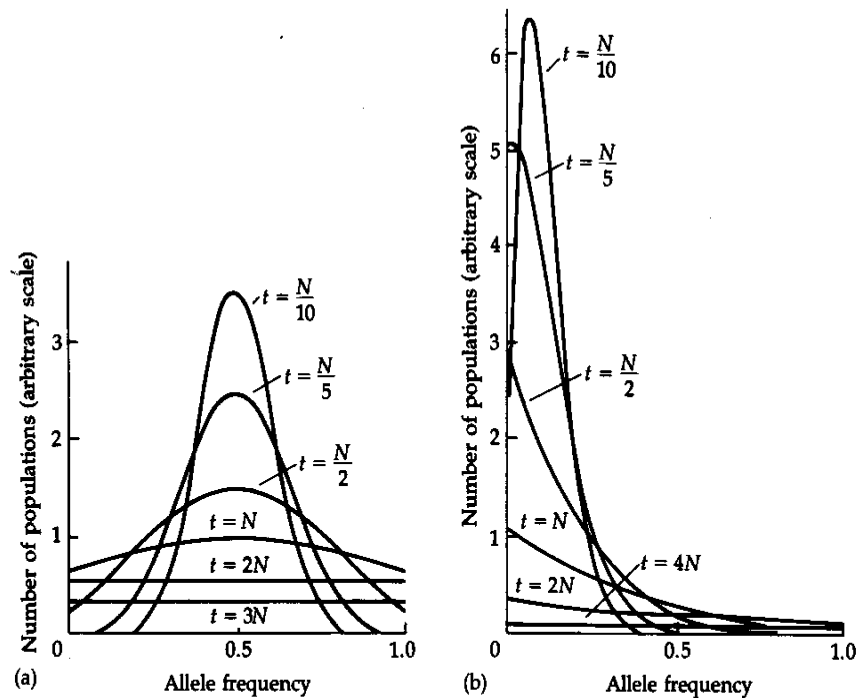


Figure 4. Theoretical results of random genetic drift. (a) Initial allele frequency = 0.5. (b) Initial allele frequency = 0.1. The curves have been scaled so that the area under each curve is equal to the proportion of populations in which fixation or loss has not yet occurred. The curves are therefore the distributions of allele frequencies among segregating populations. (From Kimura 1955.)

$t = 4N$ generations, by which time only about 10 percent of the populations remain unfixated. Once a flat distribution of allele frequency is reached, the distribution remains flat, but random drift continues until fixation or loss has occurred in all populations. The average time required for fixation or loss to occur depends on the initial allele frequency (p_0). When $p_0 = 0.5$, the average time that a population remains segregating is $2.8N$ generations, and when $p_0 = 0.1$, the average time is $1.3N$ generations. Thus, populations remain segregating longer when they are large or when the allele frequency is near 0.5.

Many of the important consequences of random genetic drift follow from the fact that the population structure in Figure 2 entails a peculiar sort of "inbreeding," but the nature of the inbreeding is very subtle because the population structure in Figure 2 can be considered on two distinct levels—the level of the individual subpopulation and the level of the total population (i.e., the aggregate of all subpopulations).

Consider first any one of the subpopulations in Figure 2. Within this subpopulation (call it subpopulation number i), mating is random because of assumption (2) in the earlier list. If the allele frequencies of A and a in the i th subpopulation are denoted p_i and q_i , then the genotype frequencies of AA , Aa , and aa are given by the familiar Hardy-Weinberg principle as p_i^2 , $2p_iq_i$, and q_i^2 , respectively. Furthermore, think about the situation in Figure 2 at a time so advanced that all subpopulations have become fixed for one allele or the other. When the i th subpopulation is fixed, either p_i equals 0 or p_i equals 1. The genotype frequencies of AA , Aa , and aa in the subpopulation are either 0, 0, and 1 (if $p_i = 0$), or 1, 0, and 0 (if $p_i = 1$). These genotype frequencies, though extreme, still satisfy the Hardy-Weinberg principle. Thus, within any one subpopulation in Figure 2, the frequency of heterozygotes equals the frequency expected with random mating.

The situation regarding the total population in Figure 2 is very different, however, as there is an overall deficiency of heterozygotes. The meaning of "total population" in this context can be made clear by considering a simple example. Suppose the subpopulations in question were colonies of mice in a large barn. (Were the mice real, we would have to worry about migration between subpopulations, but these are hypothetical mice which stay where they are put.) Suppose further that we are unaware of the existence of such subpopulations but instead think that the barn contains a single randomly mating population. To study the "total population" of the barn, we trap mice at random. If time is so advanced that a fraction p of the subpopulations are fixed for A and a fraction q are fixed for a (remember that $q = 1 - p$), then a fraction p of the time, we would trap an AA mouse and a fraction q of the time, an aa mouse. Since the overall allele frequency of A among the trapped animals is p , we would naively expect a fraction $2pq$ of the animals to be heterozygous. In fact, we would have caught no heterozygotes at all!

This rather paradoxical result—that there is a deficiency of heterozygotes in the total population even though random mating occurs within each subpopulation—is a consequence of the random genetic drift of allele frequencies among subpopulations due to their finite size. Were the subpopulations so large that random drift could be ignored, each subpopulation would have the same allele frequencies and Hardy-Weinberg genotype frequencies as any other. In such a case, to return to the mouse example, it would not affect the proportion of heterozygotes if mice were sampled from the total population or from any one of the subpopulations. For the model in Figure 2, however, random genetic drift is important. The total population has a deficiency of heterozygotes, much as if there were inbreeding. It is this inbreeding-like effect of population subdivision that we now set out to quantify.

Levels of Population Structure

Because population subdivision entails an inbreeding-like effect, it is convenient to measure the effect in terms of decrease in the proportion of heterozygous genotypes, much as the effects of nonrandom mating were measured in Chapter 1. A subdivided population has three distinct levels of complexity, however—individual organisms (*I*), subpopulations (*S*), and the total population (*T*). To avoid potential confusion of these various levels, it is useful to adopt the terminology below. We will also allow the possibility of nonrandom mating within subpopulations, so assumption (2) in the earlier list can be violated.

$$\begin{aligned} H_I &= \text{the heterozygosity of an } \textit{individual} \text{ in a subpopulation} \\ H_S &= \text{the expected heterozygosity of an individual in an} \\ &\quad \text{equivalent random mating } \textit{subpopulation} \\ H_T &= \text{the expected heterozygosity of an individual in an} \\ &\quad \text{equivalent random mating } \textit{total} \text{ population} \end{aligned} \tag{2.2}$$

As noted in Chapter 1, H_I can be interpreted as the average heterozygosity of all the genes in an individual or as the probability of heterozygosity of any one gene. H_S represents the level of heterozygosity that would be found in a subpopulation if the subpopulation were undergoing random mating; therefore, H_S always equals $2p_iq_i$ for a subpopulation with allele frequency p_i . The quantity H_T represents what the heterozygosity would be were all subpopulations pooled together and mated randomly; if the average allele frequency among subpopulations is denoted p_0 , then $H_T = 2p_0q_0$.

The **inbreeding coefficient** defined in Chapter 1 measures the reduction in heterozygosity of an individual due to nonrandom mating within its subpopulation. To avoid confusion, we will call this inbreeding coefficient F_{IS} (rather than simply F as in Chapter 1). Thus,

$$F_{IS} = \frac{H_S - H_I}{H_S} \tag{2.3}$$

The effects of population subdivision are measured by a quantity called the **fixation index** (symbolized F_{ST}), which is the reduction in heterozygosity of a subpopulation due to random genetic drift. Thus,

$$F_{ST} = \frac{H_T - H_S}{H_T} \tag{2.4}$$

The overall inbreeding coefficient of an individual (denoted F_{IT}) includes a contribution due to actual nonrandom mating within subpopulations (F_{IS}) and another contribution due to the subdivision itself (F_{ST}). The quantity F_{IT}

measures the reduction in heterozygosity of an individual relative to the total population, so

$$F_{IT} = \frac{H_T - H_I}{H_T} \quad (2.5)$$

The hierarchical *F*-statistics defined in Equations 2.3, 2.4, and 2.5 are all types of "inbreeding coefficients," but they differ in the reference populations. F_{IS} is concerned with inbreeding in individuals (*I*), relative to the subpopulation (*S*) to which they belong; F_{ST} is concerned with inbreeding in subpopulations (*S*), relative to the total population (*T*) of which they are a part; and F_{IT} is concerned with inbreeding in individuals (*I*), relative to the total population (*T*). F_{IT} is the most inclusive measure of inbreeding in that it takes into account both the effects of nonrandom mating within subpopulations (F_{IS}) and the effects of population subdivision (F_{ST}). The mathematical relation between the three types of inbreeding coefficient is demonstrated in the following problem.

Show that F_{IS} , F_{ST} , and F_{IT} are related by the equation $(1 - F_{IS})(1 - F_{ST}) = (1 - F_{IT})$.

PROBLEM

3

ANSWER From Equation 2.3, $F_{IS} = 1 - (H_I/H_S)$, or $1 - F_{IS} = H_I/H_S$. Equation 2.4 implies that $F_{ST} = 1 - (H_S/H_T)$, or $1 - F_{ST} = H_S/H_T$. And Equation 2.5 implies that $F_{IT} = 1 - (H_I/H_T)$, or $1 - F_{IT} = H_I/H_T$. Now multiply the expressions for $1 - F_{IS}$ and $1 - F_{ST}$ together to obtain $(1 - F_{IS})(1 - F_{ST}) = (H_I/H_S)(H_S/H_T) = H_I/H_T = (1 - F_{IT})$.

Equations 2.3–2.5 can be applied to data of Levin (1978) on natural subpopulations of *Phlox cuspidata* in Texas. Altogether, 43 subpopulations were studied with respect to the phosphoglucosyltransferase-2 (*Pgm-2*) gene, and two alleles (*Pgm-2^a* and *Pgm-2^b*) were found. In 40 subpopulations, the *Pgm-2^b* allele was fixed; in the remaining three subpopulations, the allele frequencies of *Pgm-2^b* were estimated as 0.49, 0.83, and 0.91, respectively. In the three polymorphic populations, the observed frequencies of heterozygotes were 0.17, 0.06, and 0.06, respectively.

The individual heterozygosity (H_I) can be estimated from the observed data as the average heterozygosity of individuals within subpopulations; thus $\langle H_I \rangle = [(40)(0) + 0.17 + 0.06 + 0.06]/43 = 0.0067$, where the first term comes from the 40 subpopulations in which the observed heterozygosity was 0. The subpopulation heterozygosity (H_S) can be estimated from the estimated allele frequencies within subpopulations and the expected frequency of heterozygotes under the supposition of random mating, thus $\langle H_S \rangle = [(40)(0) + 2(0.49)(0.51) + 2(0.83)(0.17) + 2(0.91)(0.09)]/43 = 0.0220$, where again the first term comes from the 40 fixed subpopulations. Finally, the total hetero-

zygosity (H_T) can be estimated from the average allele frequency in the combined subpopulations and the Hardy-Weinberg expectation of heterozygosity with random mating; the average allele frequency of the $Pgm-2^b$ allele is $[(40)(1) + 0.49 + 0.83 + 0.91]/43 = 0.9821$, so that $\langle H_T \rangle = 2(0.9821)(0.0179) = 0.0352$.

Now the various F -statistics can be estimated from Equations 2.3–2.5: $\langle F_{IS} \rangle = (\langle H_S \rangle - \langle H_I \rangle) / \langle H_S \rangle = (0.0220 - 0.0067) / (0.0220) = 0.70$; this is the estimated inbreeding coefficient due to nonrandom mating within subpopulations, and note that it is in reasonable agreement with the value calculated in Problem 21 in Chapter 1 based on the fraction of self-fertilization in this species. The fixation index F_{ST} is estimated as $\langle F_{ST} \rangle = (\langle H_T \rangle - \langle H_S \rangle) / \langle H_T \rangle = (0.0352 - 0.0220) / (0.0352) = 0.38$; this is the amount of inbreeding due solely to population subdivision. Finally, the total inbreeding coefficient is estimated as $\langle F_{IT} \rangle = (\langle H_T \rangle - \langle H_I \rangle) / \langle H_T \rangle = (0.0352 - 0.0067) / (0.0352) = 0.81$; this is the amount of inbreeding due to the combined effects of nonrandom mating within subpopulations and to random genetic drift among subpopulations. Note that the total inbreeding coefficient is quite large (as expected from the substantial fraction of self-fertilization), but that an appreciable fraction of the total inbreeding is attributable to subdivision and random genetic drift.

The method of estimating the F -statistics by replacing the parameters in Equations 2.3–2.5 with their observed or estimated values is not necessarily the best, particularly with small samples. Ideally, estimates of the F -statistics should correct for the effects of sampling a limited number of subpopulations, as well as for the effects of sampling a limited number of individuals per subpopulation. Methods for making these corrections have been developed but are complex and raise additional issues. For an excellent discussion, see Weir and Cockerham (1984). Important issues are also addressed in Wright (1978, p. 86ff.), Curie-Cohen (1982), Nei and Chesser (1983), and Nei (1986). We will use the uncorrected estimation procedure used above, which is adequate for purposes of illustration. When applied to the Pgm data from *Phlox cuspidata*, for example, the more sophisticated method of Weir and Cockerham gives $F_{IS} = 0.70$, $F_{ST} = 0.36$ and $F_{IT} = 0.81$.

PROBLEM 4 Levin (1978) also studied the glutamate oxalacetate transaminase-2 (*Got-2*) gene in the 43 subpopulations of *Phlox cuspidata* discussed earlier. Three alleles were found: *Got-2^a*, *Got-2^b* and *Got-2^c*. A total of 39 subpopulations were monomorphic for *Got-2^b*. One subpopulation contained *Got-2^a* and *Got-2^b* with respective allele frequencies 0.37 and 0.63, and the observed frequency of heterozygotes was 0.17. Three subpopulations contained only *Got-2^b* and *Got-2^c*; the allele frequencies of *Got-2^b* in these subpopulations were 0.87, 0.91, and 0.82, respectively, and the observed frequencies of heterozygotes were 0.09, 0.06, and 0.09, respectively. Estimate the various heterozygosities and F -statistics for this gene.

ANSWER $\langle H_i \rangle = [(39)(0) + 0.17 + 0.09 + 0.06 + 0.09]/43 = 0.00953$; $\langle H_s \rangle = [(39)(0) + 2(0.37)(0.63) + 2(0.87)(0.13) + 2(0.91)(0.09) + 2(0.82)(0.18)]/43 = 0.02678$. The overall estimated frequency of *Got-2^a* is $[(39)(0) + 0.37 + 3(0)]/43 = 0.0086$, that of *Got-2^b* is $[(39)(1) + 0.63 + 0.87 + 0.91 + 0.82]/43 = 0.9821$, so that of *Got-2^c* is $1 - 0.0086 - 0.9821 = 0.0093$. From the Hardy–Weinberg principle for multiple alleles (Chapter 1), $\langle H_T \rangle = 2(0.0086)(0.9821) + 2(0.0086)(0.0093) + 2(0.9821)(0.0093) = 0.03532$. Thus $\langle F_{IS} \rangle = 0.64$, $\langle F_{ST} \rangle = 0.24$, and $\langle F_{IT} \rangle = 0.73$. As you may have noted, the *F*-statistics vary somewhat from gene to gene; the range of $\langle F_{IS} \rangle$ for five genes studied in *Phlox cuspidata* was 0.60 to 0.77 (average 0.67), and that of $\langle F_{ST} \rangle$ was 0.17 to 0.86 (average 0.41).

Except for plants that have a high frequency of self-fertilization or for certain insects that regularly undergo parent–offspring or brother–sister mating, values of F_{IS} in most natural populations are typically close to zero, which indicates random mating within subpopulations. (Within very small subpopulations, there may be mating between relatives—but as long as the mating between relatives occurs no more often than would be expected by chance in a subpopulation of that size, F_{IS} remains zero.) For example, Crow and Mange (1965) studied a South Dakota population of Hutterites, a group that is relatively isolated from surrounding communities because of their religious beliefs. In this population, F_{ST} was estimated as 0.04 (a relatively large value for a human population), yet F_{IS} was essentially zero. For many natural populations, therefore, particularly in animals or outcrossing plants, it is reasonable to assume $F_{IS} = 0$. This assumption is equivalent to the assumption of random mating within subpopulations (assumption 2 in the earlier list), which puts us back in the context of the model outlined in Figure 2. Furthermore, when $F_{IS} = 0$, then $F_{ST} = F_{IT}$, and there is no longer need for the subscripts. For the sake of typographical convenience, therefore, the symbol *F* will hereafter be used in place of F_{ST} unless otherwise specified. In some contexts, F_{ST} is denoted G_{ST} (Nei 1975, p. 151; Crow and Aoki 1984).

Increase in Fixation Index from Random Genetic Drift

As was true of the inbreeding coefficient F_{IS} in Chapter 1, the fixation index F_{ST} (or, as we are now calling it, simply *F*) can be interpreted in terms of identity by descent. Indeed, *F* is the probability that two alleles chosen at random from within the same subpopulation are identical by descent. As random genetic drift continues generation after generation, the value of *F* will change, so it will be necessary to use the symbol F_t to represent the average value of the fixation index F_{ST} among subpopulations in generation *t*. The value of F_t can be calculated from the reasoning illustrated in Figure 5. This figure shows the $2N$ alleles in a breeding population of generation

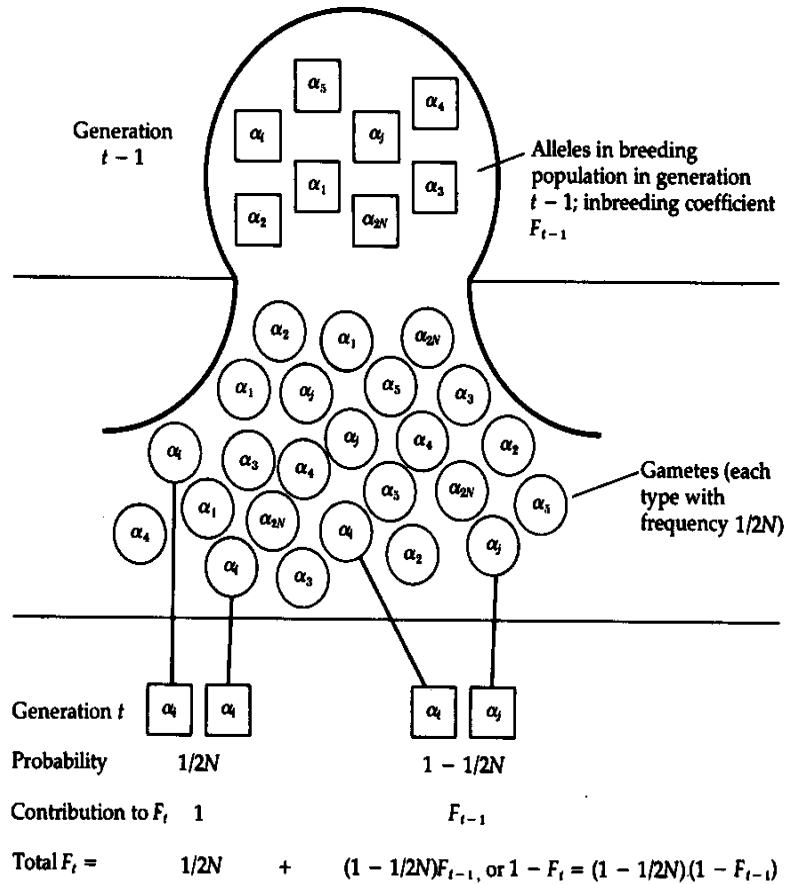


Figure 5. Random sampling in each of the subpopulations in Figure 2 implies the relation $F_t = (1/2N) + [1 - (1/2N)]F_{t-1}$.

$t - 1$ (labeled $\alpha_1, \alpha_2, \alpha_3, \dots, \alpha_i, \dots, \alpha_j, \dots, \alpha_{2N}$ to avoid confusion with the conventional allele symbols A and a), the gametes derived from generation $t - 1$ (each gametic type represents a fraction $1/2N$ of the entire pool of gametes), and two randomly chosen pairs of alleles in the breeding population of generation t . The probability that the second allele chosen is of the same type as the first is $1/2N$, because this is the frequency of each allelic type in the gametic pool; the probability that the second allele is of a different type from the first is accordingly $1 - 1/2N$. In the first case ($\alpha_i\alpha_i$), the probability of identity by descent is 1; in the second case ($\alpha_i\alpha_j$), it is F_{t-1} . Altogether $F_t = 1/2N + (1 - 1/2N)F_{t-1}$. Multiplying both sides by -1 and adding $+1$ leads to

$$1 - F_t = 1 - 1/2N - (1 - 1/2N)F_{t-1} \\ = (1 - 1/2N)(1 - F_{t-1})$$

the solution of which is

$$1 - F_t = (1 - 1/2N)^t(1 - F_0) \quad (2.6)$$

In the case when $F_0 = 0$, Equation 2.6 simplifies to

$$F_t = 1 - \left(1 - \frac{1}{2N}\right)^t \quad (2.7)$$

Strictly speaking, Equations 2.6 and 2.7 apply only to populations (many plants, for example) that can undergo self-fertilization, because the theoretical considerations in Figure 5 make no stipulation that half of the successful gametes in any generation must come from males and the other half from females. The correction for organisms such as those with separate sexes that cannot undergo self-fertilization is minor, however: if there are equal numbers of males and females, simply replace N in Equations 2.6 and 2.7 with $N + 1/2$ (see Crow and Kimura 1970). Since $1/2(N + 1/2)$ is very nearly equal to $1/2N$ for realistic values of N , Equations 2.6 and 2.7 as they stand (without the correction) may be taken as a satisfactory approximation for organisms with separate sexes.

As an example of the use of Equation 2.6, consider the following question: What is the inbreeding coefficient of a mouse in a colony that was established from a much larger colony 30 generations ago and since that time has maintained a constant size of 20 individuals each generation? For the sake of simplicity, assume that the colony is ideal in the sense discussed earlier, particularly with regard to random mating within the colony. Because random mating within the colony is assumed, the total inbreeding coefficient F_{IT} of an animal is equal to the fixation index F_{ST} of the colony, and $F_{ST} = F_t$ in Equation 2.6. Thus, F_{ST} after 30 generations is given by Equation 2.6 with $t = 30$, which is $1 - F_{30} = (1 - 1/2N)^{30}(1 - F_0)$, where we may take $N = 20$ (as noted earlier, use of $N = 20.5$ would be a bit more accurate), and $F_0 = 0$ (justified because the foundation population was said to be large). Substituting, $1 - F_{30} = (1 - 1/40)^{30} = 0.468$, or $F_t = 0.532$. Note that there is substantial inbreeding even though mating is random and genotypic frequencies within the colony are given by the Hardy-Weinberg principle. The inbreeding does not result from actual nonrandom mating but from the fact that the population is small in size.

From four large chicken barns near Ramona, California, Selander and Yang (1969) trapped wild mice (*Mus musculus*) and carried out an electrophoretic study of a large number of genes, including those for hexose-6-phosphate dehydrogenase, NADP-isocitrate dehydrogenase, and hemoglobin. Estimated values of F_{ST} for these genes were 0.10, 0.16, and 0.11, respectively, averaging $F_{ST} = 0.12$. Assuming that the mouse

PROBLEM
5

populations in each barn have about the same size (N), how long would it take for random genetic drift to result in a value of $F_{ST} = 0.12$ in the ideal (and undoubtedly oversimplified) case when migration between barns does not occur, assuming that $N = 20$? How long would it take when $N = 100$?

ANSWER In the case $N = 20$, we have $1 - F_t = (1 - 1/40)^t$ with $F_t = 0.12$, or $t = \ln(0.88)/\ln(0.975) = 5$ generations. For $N = 100$, the answer is $t = 26$ generations. As will be discussed later, migration has the effect of retarding the increase in F_t , so the times would be correspondingly longer with migration taken into account. On the other hand, founder effects (defined later) are probably significant in barn populations of mice, and these have the effect of accelerating genetic divergence due to the recurrent "bottlenecks" of population size.

Several important consequences of the population structure in Figure 2 are summarized in Table 1. First, although each subpopulation is finite in number, we can imagine so many of them that the total population number is effectively infinite. For an infinite population that obeys the assumptions of an ideal population discussed earlier, the allele frequencies remain constant. That is, even though the allele frequencies within the subpopulations may change willy-nilly due to random genetic drift, the overall average allele frequency of A among subpopulations always remains p_0 , where p_0 represents the allele frequency of A in the base population.

Secondly, when mating is random within subpopulations, F_t is the probability of autozygosity of a gene in an individual in generation t . For an individual chosen at random from the total population in generation t , therefore, the probability of allozygosity is $1 - F_t$. Allozygous alleles are those that have escaped the effects of inbreeding, and in this subset of cases the genotypes occur in Hardy-Weinberg proportions. Autozygous alleles are identical by descent because of the inbreeding, and in this subset of cases the genotype frequencies are the same as with complete inbreeding. Therefore, in a total population split into subpopulations as in Figure 2, in which p_0 is the overall average allele frequency of A , the genotype frequencies, averaged across all subpopulations, are given by

$$\begin{aligned} AA: & p_0^2(1 - F_t) + p_0F_t \\ Aa: & 2p_0q_0(1 - F_t) \\ aa: & q_0^2(1 - F_t) + q_0F_t \end{aligned} \tag{2.8}$$

Equation 2.8 implies that the genotype frequencies in the total population are given by the usual formula for inbreeding (see Table 5 in Chapter 1). However, *within any one subpopulation, the genotype frequencies still obey the Hardy-Weinberg principle because of random mating within the subpopulation.* When F_t is substituted from Equation 2.7 into the expression for the frequency of heterozygous genotypes in Equation 2.8, we see that the average hetero-

Table 1. Consequences of population subdivision and random genetic drift on genotype frequency.

		GENERATION		
		0	t	∞
Inbreeding coefficient (F_t) (Average over all populations)		0	$1 - (1 - 1/2N)^t$	1
Genotype frequency (Average over all populations)		$\left\{ \begin{array}{l} AA: p_0^2 \\ Aa: 2p_0q_0 \\ aa: q_0^2 \end{array} \right.$	$\left\{ \begin{array}{l} p_0^2(1 - F_t) + p_0F_t \\ 2p_0q_0(1 - F_t) \\ q_0^2(1 - F_t) + q_0F_t \end{array} \right.$	$\left\{ \begin{array}{l} p_0 \\ 0 \\ q_0 \end{array} \right.$
Allele frequency (Average over all populations)		$\left\{ \begin{array}{l} A: p_0 \\ a: q_0 \end{array} \right.$	$\left\{ \begin{array}{l} p_0 \\ q_0 \end{array} \right.$	$\left\{ \begin{array}{l} p_0 \\ q_0 \end{array} \right.$

zygosity among subpopulations at time t decreases according to the equation

$$2p_0q_0(1 - F_t) = 2p_0q_0(1 - 1/2N)^t \quad (2.9)$$

Third, and finally, since F_t eventually goes to 1, all subpopulations eventually become fixed for one allele or the other. Because the average allele frequency of A remains p_0 even when all subpopulations have become fixed, the proportion of subpopulations that eventually become fixed for A is p_0 (and the proportion that eventually become fixed for a is q_0). Stated another way, the probability of ultimate fixation of a designated allele in any ideal subpopulation is equal to the frequency of the allele in the initial population.

Effective Population Number

No real population can be expected to obey all of the assumptions pertaining to ideal populations. In any actual case, there must be corrections for age-related differences in reproductive rate, for unequal numbers of males and females, and for unequal family size (see Crow and Kimura, 1970, for a good review). To take a specific example, consider the experiment depicted in Figure 3, for which $p_0 = q_0 = 1/2$. The total heterozygosity is given by $H_T = 2(1/2)(1/2) = 0.5$, and this value remains constant throughout the experiment, as expected. Furthermore, the theory predicts that the average subpopulation heterozygosity (H_S) is given by Equation 2.9 as follows:

$$H_S = 2p_0q_0(1 - F_t) = 2p_0q_0(1 - 1/2N)^t = (0.5)(1 - 1/2N)^t$$

The actual size (N_a) of the subpopulations in Figure 3 is $N_a = 16$, and the predicted decrease in H_S with this population size is shown in Figure 6, along

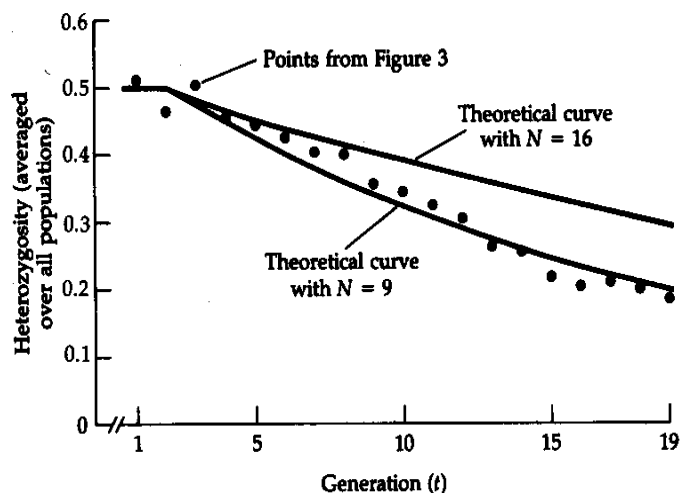


Figure 6. Theoretical curves for average heterozygosity with $N = 9$ or $N = 16$, along with actual values from the experiment in Figure 3. (Data from Buri 1956.)

with the observed values. It is evident that the fit is not very impressive, especially for the later generations. However, the pattern of decrease in H_S does follow the theoretical prediction if we assume that $N = 9$ (see Figure 6). That is, the observed decrease in average heterozygosity in the actual populations matches what would be expected in ideal populations of size $N = 9$. The number $N = 9$ is called the *inbreeding effective number* (or simply the *effective number*) of the populations in question.

To say the same thing somewhat more formally, the **inbreeding effective number** N_e of an actual (non-ideal) population is the population number of an ideal population which undergoes the same rate of increase in F (decrease in H_S) as the actual population. The use of Equation 2.9 to estimate the effective population number is illustrated in the following problem.

PROBLEM 6 The neotropical giant toad *Bufo marinus* became established in Hawaii and Australia through the release of a large number of individuals in the 1930s. Allele frequencies of ten polymorphic enzyme genes were studied in individuals from ten local populations in Hawaii and five local populations in Australia (Easteal 1985). Estimated values of F_{ST} were 0.056 in Hawaiian populations, 0.063 in Australian populations, and 0.059 combined. The species has one generation per year, and there have been approximately 45 years since the introductions. Use these data to estimate the effective population number in the Hawaiian, Australian, and combined populations.

ANSWER Equation 2.9 requires a little manipulation in order to obtain $1/2N = 1 - \exp[(1/t)\ln(1 - F_{ST})] = \omega$ (say), and then $N = (2\omega)^{-1}$. Substituting the estimates of $\langle F_{ST} \rangle$ and $\langle t \rangle = 45$ generations yields $\omega = 0.001280$ for Hawaii, 0.001445 for Australia and

0.001350 combined, for $\langle N \rangle = 391, 346$ and 370 , respectively. Eastel has estimated the 95 percent confidence intervals as (119, 812), (104, 719), and (112, 770), respectively. Estimates of population number based on dispersal rates and population densities have given estimates ranging from 4000 to 20,000, which emphasizes the principle that the long-term average effective population number can be very different than the actual population number.

In the consideration of random genetic drift, the effective population number (N_e) is usually of greater interest than the actual population number (N_a). There are three important instances in which the two quantities are related in a reasonably simple manner. The first case involves populations that fluctuate in size from generation to generation, the second case concerns populations with unequal numbers of males and females, and the third case pertains to populations that are spread out uniformly in two dimensions rather than being clustered into discrete colonies. These cases are considered in the following paragraphs.

Variable population number. Imagine a population that is ideal in all respects except that its number changes from generation to generation. Suppose, to be precise, that its effective number in generation 1 is N_1 , in generation 2 is N_2 , and in generation t is N_t . In this situation, the overall increase in F is given approximately by $1 - F_t = (1 - 1/2N_e)^t$, where

$$1/N_e = (1/t)(1/N_1 + 1/N_2 + \dots + 1/N_t) \quad (2.10)$$

Equation 2.10 says that the average effective population number is calculated as the reciprocal of the average of the reciprocals, which is a special sort of average called the **harmonic mean**. The harmonic mean tends to be dominated by the smallest terms. Suppose, for example, that $N_1 = 1000$, $N_2 = 10$, and $N_3 = 1000$ in a population that underwent a severe temporary reduction in population number (a **bottleneck**) in generation 2. Then, using Equation 2.10, $1/N_e = (1/3)(1/1000 + 1/10 + 1/1000) = 0.034$, or $N_e = 1/0.034 = 29.4$. The average effective number over the three-generation period is only 29.4, whereas the average actual number is $(1/3)(1000 + 10 + 1000) = 670$. A severe population bottleneck often occurs in nature when a small group of emigrants from an established subpopulation founds a new subpopulation. The random genetic drift accompanying such a founder event is known as a **founder effect**.

A certain highly isolated colony of the moth *Panaxia dominula* near Oxford, England, has been intensively studied by Ford and collaborators over the period 1928 to 1968 (see Ford and Sheppard 1969). This species has one generation per year, and estimates of population size were carried out yearly beginning in 1941. For the years 1950

PROBLEM

7

to 1961, inclusive, estimates of population size were as follows:

1950	4100	1956	11,000
1951	2250	1957	16,000
1952	6000	1958	15,000
1953	8000	1959	7000
1954	11,000	1960	2500
1955	2000	1961	1400

Assuming that the actual size of the population in any year equals the effective size in that year, use Equation 2.10 to estimate the average effective number over the entire 12-year period.

ANSWER $\langle 1/N_e \rangle = (1/12)[(1/4100) + (1/2250) + \dots + (1/2500) + (1/1400)] = 0.00025$, so $\langle N_e \rangle = 1/0.00025 = 4000$. Wright (1978) has argued that this is still a substantial overestimate of the true effective population number, being too large by a factor of about 10, because the effective number in any generation is only about one-tenth of the actual number owing to large differences in reproductive success among individuals.

Unequal sex ratio. A second important case in which the effective number of a nonideal population can readily be calculated concerns sexual populations in which the number of males and females is unequal. This inequality creates a peculiar sort of "bottleneck": because half of the alleles in any generation must come from each sex, any departure of the sex ratio from equality enhances the opportunity for random genetic drift. This situation is important in wildlife management, where for many game animals (pheasants and deer come immediately to mind) the legal bag limit for males is much larger than for females. Although some management considerations are served by such hunting regulations (for example, the species involved are usually polygamous, so a single male can fertilize many females and overall actual population size can be maintained), the resulting inequality in sex ratio reduces the effective population number. Specifically, if a sexual population consists of N_m males and N_f females, the actual size is $N_a = N_m + N_f$. However, the effective population number is given by

$$N_e = \frac{4N_mN_f}{N_m + N_f} \quad (2.11)$$

To take a realistic example, if hunting is permitted to a level at which the number of surviving males is one-tenth the number of females, then $N_m = 0.1N_f$ and $N_a = N_m + N_f = 1.1N_f$. From Equation 2.11, the effective number is $N_e = 4N_mN_f/(N_m + N_f) = (0.4/1.1)N_f = 0.33N_a$. In this case, the effective population number is a mere one-third of the actual number.

A dairy farmer has a herd consisting of 200 cows and 2 bulls. What is the effective size of the population?

PROBLEM

8

ANSWER $N_e = 4(2)(200)/(2 + 200) = 8$. Random genetic drift will have a marked effect on the allele frequencies in the population as time goes on, and with such a small effective population number may even result in an increase in the frequency of harmful alleles.

Uniform population dispersion. The remaining case of effective-size calculations to consider pertains to populations spread out uniformly in two dimensions, such as on a prairie, rather than being clumped into discrete colonies. In this case the effective size depends on two quantities: (1) the number of breeding individuals per unit area, usually denoted by the symbol δ , and (2) the amount of dispersion between an individual's own birthplace and that of its offspring. The latter quantity, usually denoted by the symbol σ^2 , is called the *one-way variance of distance between birth and breeding sites*. If dispersion follows a normal, bell-shaped curve in both dimensions (like an inverted salad bowl), then σ^2 has the following biological interpretation: 39 percent of all individuals have their offspring within a circle of radius σ centered at their own birthplace; 87 percent have their offspring within a circle of radius 2σ and 99 percent have their offspring within a circle of radius 3σ . In terms of δ and σ , the effective size of the population (often called **neighborhood size** in this context) is given by

$$N_e = 4 \pi \delta \sigma^2 \quad (2.12)$$

where $\pi = 3.14159$.

Equation 2.12 can be applied to data on the abundant prairie deer mouse *Peromyscus maniculatus*. In a large area in southern Michigan, Dice and Howard (1951) estimated the density of breeding individuals to be between $\langle \delta \rangle = 5/\text{ha}$ and $\langle \delta \rangle = 7/\text{ha}$ (ha = hectare = 10,000 m²—about 2.5 acres). By following the movement of marked animals from birth to breeding sites, they estimated $\langle \sigma^2 \rangle = 1.3$ ha (excluding woodlands, which this organism avoids). Applying Equation 2.12, the neighborhood size is estimated to be between $\langle N_e \rangle = 4(3.14159)(5)(1.3) = 82$ and $\langle N_e \rangle = 4(3.14159)(7)(1.3) = 114$, or approximately 80 to 110. The neighborhood size is surprisingly small for such an abundant animal.

Phlox pilosa and *Liatris cylindracea* are insect-pollinated, perennial plants that are very abundant in undisturbed prairies in Illinois. In plots with densities of 9 plants/m² (*P. pilosa*) and 5 plants/m² (*L. cylindracea*), the one-way variances for dispersal of gametes, averaged over both pollen and seed, were estimated as 3.9 m² and 2.6 m²,

PROBLEM

9

respectively (Levin and Kerster 1968, Schaal and Levin 1978). In the Texas bluebonnet (*Lupinus texensis*), estimates of δ and σ^2 were $\langle \delta \rangle = 15/\text{m}^2$ and $\langle \sigma^2 \rangle = 0.5 \text{ m}^2$ (Schaal 1980). Use Equation 2.12 to estimate the neighborhood size for each species.

ANSWER $\langle N_e \rangle = 4\pi\langle \delta \rangle\langle \sigma^2 \rangle = 4(3.14159)(9)(3.9) = 441$ individuals for *P. pilosa*, $\langle N_e \rangle = 163$ individuals for *L. cylindracea*, and $\langle N_e \rangle = 94$ individuals for *L. texensis*. The estimated effective sizes are all rather small considering the abundance of the organisms, and they give ample scope for the accumulation of local genetic differences.

Having discussed the relationship between effective population number and actual population number in the preceding special cases, we now drop the subscript e from N_e with the understanding that, hereafter, N always means N_e unless otherwise specified. Bear in mind, however, that the effective number of a population is usually smaller than the actual number, and sometimes much smaller.

A rare situation in which N_e can be greater than N occurs when the distribution of offspring among individuals is more even (smaller variance) than would be expected from the Poisson distribution. This possibility is sometimes of use to animal breeders, who, by deliberately evening out the offspring distribution among animals, can help maximize the effective size of the population.

Genetic Divergence among Subpopulations

The fixation index F_{ST} defined in Equation 2.4 serves as a convenient and widely used measure of genetic differences among populations. In an ideal population with no mutation, migration, or selection, the value of F_{ST} can easily be interpreted in terms of random genetic drift. Interpretations are not so easy for natural populations, because observed values of F_{ST} are influenced not only by random drift but also by mutation and especially by migration and natural selection. Difficulties in interpretation do not, however, invalidate the usefulness of F_{ST} as an index of genetic differentiation.

Although F_{ST} has a theoretical minimum of 0 (indicating no genetic divergence) and a theoretical maximum of 1 (indicating fixation for alternative alleles in the subpopulations), the observed maximum is usually much less than 1. Wright (1978) suggests the following qualitative guidelines for the interpretation of F_{ST} :

1. The range 0 to 0.05 may be considered as indicating *little* genetic differentiation.
2. The range 0.05 to 0.15 indicates *moderate* genetic differentiation.
3. The range 0.15 to 0.25 indicates *great* genetic differentiation.
4. Values of F_{ST} above 0.25 indicate *very great* genetic differentiation.

However, to quote Wright (1978), who developed the concept of F_{ST} , "Differentiation is by no means negligible if F_{ST} is as small as 0.05 or even less."

Amounts of genetic divergence among human subpopulations and among subpopulations of several other species are presented in Table 2. The values of F_{ST} imply that genetic divergence among human subpopulations is quite small. Of the total genetic variation found in three major races (Caucasoid, Negroid, and Mongoloid), only 0.07 (7 percent) is ascribable to genetic differences among races. That is, 93 percent of the total genetic variation is found within races. Again, of the total genetic variation found in the native Yanomama Indians of Venezuela and Brazil, only 0.077 (7.7 percent) is due to differences in allele frequency among villages, which implies that 92.3 percent of the total genetic variation is found within any single village. Values of F_{ST} for other organisms are quite variable, presumably because F_{ST} is influenced by the effective size of the subpopulations, by the amount and pattern of migration between subpopulations, and by other factors, including natural selection.

Table 2 provokes a brief discussion of the sensitive term *race* because the term is prone to misunderstanding or misuse. In population genetics, a *race* is a group of individuals in a species who are genetically more similar to each other than they are to the members of other such groups. Populations

Table 2. Total heterozygosity (H_T), average heterozygosity (H_S) among subpopulations, and fixation index (F_{ST}) for various organisms.

Organism	Number of populations	Number of loci	H_T	H_S	F_{ST}^a
Human (major races)	3	35	0.130	0.121	0.069
Human, Yanomama Indian villages	37	15	0.039	0.036	0.077
House mouse (<i>Mus musculus</i>)	4	40	0.097	0.086	0.113
Jumping rodent (<i>Dipodomys ordii</i>)	9	18	0.037	0.012	0.676
<i>Drosophila</i> <i>equinoxialis</i>	5	27	0.201	0.179	0.109
Horseshoe crab (<i>Limulus</i>)	4	25	0.066	0.061	0.076
Lycopod plant (<i>Lycopodium lucidulum</i>)	4	13	0.071	0.051	0.282

(Data from Nei 1975.)

^a F_{ST} is calculated as $(H_T - H_S)/H_T$.

PROBLEM

10

The data below are the allele frequencies of several genes in three human subpopulations: Blacks from West Africa; Blacks from Claxton, Georgia; and Whites from Claxton, Georgia (data from Adams and Ward 1973). Each gene has two predominant alleles and may, for purposes of this problem, be considered to have only two alleles. The genes control the MN blood group (alleles M and N), the Ss blood group (alleles S and s), the Duffy blood group (alleles Fy^a and Fy^b), the Kidd blood group (alleles JK^a and JK^b), the Kell blood group (alleles Js^a and Js^b), the enzyme glucose-6-phosphate dehydrogenase (alleles $G6PD^-$ and $G6PD^+$), and beta-hemoglobin (alleles β^S and β^+).

	Blacks (West Africa)	Blacks (Claxton)	Whites (Claxton)
M	0.474	0.484	0.507
S	0.172	0.157	0.279
Fy^a	0	0.045	0.422
JK^a	0.693	0.743	0.536
Js^a	0.117	0.123	0.002
$G6PD^-$	0.176	0.118	0
β^S	0.090	0.043	0

For each gene, use Equation 2.4 to estimate F_{ST} for the comparison West African Blacks versus Claxton Blacks and for the comparison West African Blacks versus Claxton Whites. Classify the F_{ST} for each gene and comparison according to the qualitative guidelines enumerated earlier.

ANSWER Let West African Blacks, Claxton Blacks and Claxton Whites be populations A, B and C, respectively. For the MN blood group in the A-B comparison, $\langle H_S \rangle$ for A is $2(0.474)(0.526) = 0.49865$, $\langle H_S \rangle$ for B is $2(0.484)(0.516) = 0.49949$, average $\langle H_S \rangle = (0.49865 + 0.49949)/2 = 0.49907$; average estimated allele frequency of M is $(0.474 + 0.484)/2 = 0.479$, so $\langle H_T \rangle = 2(0.479)(0.521) = 0.49912$. Replacing the parameters provided in Equation 2.4 by their estimated values results in $\langle F_{ST} \rangle = (0.49912 - 0.49907)/(0.49912) = 0.0001$. For the MN blood group in the A-C comparison, $\langle H_S \rangle$ for C is $2(0.507)(0.493) = 0.4999$, average $\langle H_S \rangle = (0.49865 + 0.4999)/2 = 0.49928$; the average estimated allele frequency of M is $(0.474 + 0.507)/2 = 0.4905$, so $\langle H_T \rangle = 2(0.4905)(0.5095) = 0.49982$. Thus $\langle F_{ST} \rangle = (0.49982 - 0.49928)/(0.49982) = 0.00108$. Calculations for the other genes are carried out similarly. Overall $\langle F_{ST} \rangle$ estimates and their qualitative interpretations are as follows:

	A versus B	A versus C
MN	0.0001 (little)	0.0011 (little)
Ss	0.0004 (little)	0.0164 (little)
Duffy	0.0230 (little)	0.2674 (very great)
Kidd	0.0031 (little)	0.0260 (little)
Kell	0.0001 (little)	0.0591 (moderate)
G6PD	0.0067 (little)	0.0965 (moderate)
β	0.0089 (little)	0.0471 (little)
Average	0.0060 (little)	0.0734 (moderate)

that have undergone some degree of genetic divergence as measured by, for example, F_{ST} , therefore qualify as races. Using this definition, the human population contains many races. Each Yanomama village represents, in a certain sense, a separate "race," and the Yanomama as a whole also form a distinct "race." Such fine distinctions are rarely useful, however. It is usually more convenient to group populations into larger units that still qualify as races in the definition given. These larger units often coincide with races based on physical characteristics such as skin color, hair color, hair texture, facial features, and body conformation, as defined by anthropologists (modern anthropologists also take cultural and linguistic similarities into account).

Here it must be pointed out that the data in Table 2, which indicate much more genetic variation within than among human races, may be misleading. The conclusion was based primarily on genes determining allozymes, and it certainly is not true for genes influencing skin color, hair color, hair texture, and other traits that most people think of in connection with the word "race." However, skin color and other prominent racial characteristics are used to delineate races precisely because racial differences for these traits are rather large, so the genes involved cannot be representative of the entire genome. On the other hand, allozyme loci may not be very representative of the genome either. See Nei and Roychoudhury (1982) for a review of the genetic relationship and evolution of human races.

MUTATION

Mutation is the ultimate source of genetic variation. This is an important statement, but hardly a profound one. New genetic variation is created by changes in the genetic material, and heritable changes in the genetic material are, by definition, **mutations**. The word *mutation* is used here in its widest sense to mean all genetic changes, including visible chromosome abnormalities, such as inversions and translocations, and polyploidy. **Polyploidy** refers to organisms or species having multiple sets of chromosomes. Stebbins (1976) has shown that polyploidy is of immense importance in the origin of new species of higher plants. However, the vast bulk of genetic variation within virtually all natural populations is presently thought to consist of changes in DNA which are invisible in the microscope. At the molecular level, one class of mutations results from the replacement of one pair of nucleotides with another, which includes the mutations responsible for allozymes. Another class results from the deletion or duplication of nucleotide sequences. Still another class results from the transposition of DNA sequences from one location in the genome to another. Transposition mutations often result from the movement of particular DNA sequences called **transposable elements**, most of which contain the genes that promote their own transposition. Mutations

resulting from transposable elements constitute an important source of genetic variation, and the implications of transposable elements in population genetics are discussed in Chapter 3.

Because spontaneous mutation rates are typically rather small (on the order of 10^{-4} to 10^{-6} mutations per gene per generation), the tendency for allele frequencies to change as a result of recurrent mutation (**mutation pressure**) is very small over the course of a few generations. On the other hand, the cumulative effects of mutation over long periods of time can become appreciable. A useful model for thinking about the problem is the Hardy-Weinberg model of Chapter 1, but with mutation allowed. For the moment we focus on mutations that are **selectively neutral**—mutations that have so little effect on the ability of the organism to survive and reproduce that natural selection does not appreciably influence their frequency.

Changes in Allele Frequency with Mutation Pressure

Consider a gene with two alleles A and a , and suppose that the mutation rate per generation from A to a is μ and that the rate from a to A is ν . Let p_t and q_t (with $q_t = 1 - p_t$) denote the allele frequencies of A and a in generation t . An A allele in generation t can originate in only two ways: it could have been an A allele in generation $t - 1$ that escaped mutation to a (probability $1 - \mu$), or it could have been an a allele that mutated to A (probability ν). Symbolically,

$$p_t = p_{t-1}(1 - \mu) + (1 - p_{t-1})\nu$$

The solution of this equation gives p_t in terms of p_0 (the initial frequency of A), and it is

$$p_t = \frac{\nu}{\mu + \nu} + \left(p_0 - \frac{\nu}{\mu + \nu}\right)(1 - \mu - \nu)^t \quad (2.13)$$

To understand the biological meaning of Equation 2.13, consider first the case when t is small, for example less than 100 generations. Because μ and ν are ordinarily very small, the term $(1 - \mu - \nu)^t$ is approximately equal to $1 - t(\mu + \nu)$. Suppose also that $p_0 = 0$ (i.e., that the initial population is fixed for a). With these provisions (small t , $p_0 = 0$), Equation 2.13 implies that $p_t = t\nu$. That is, the frequency of the A allele increases linearly with time, and the slope of the line equals ν . Because ν is small, however, the linear increase in p_t is difficult to detect experimentally except in very large populations. Large population sizes can be attained in bacterial **chemostats**, which are devices for maintaining populations of bacteria in a continuous state of growth and cell division (Figure 7). The linear increase in p_t that occurs in chemostats from mutation pressure is shown in Figure 8. Note the

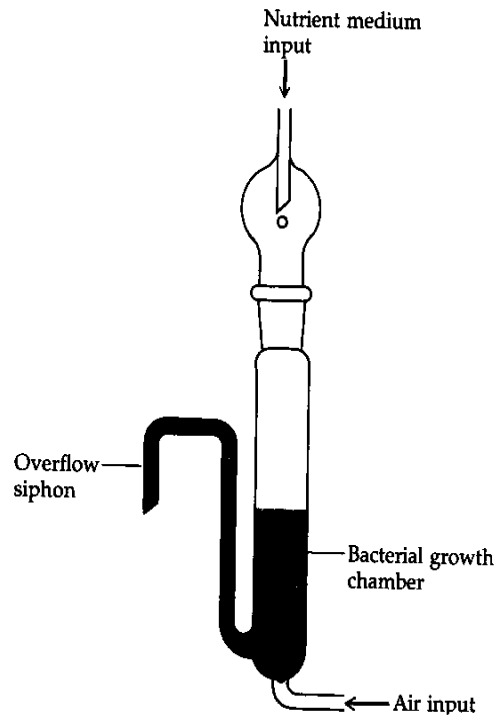


Figure 7. Diagram of a bacterial chemostat. Nutrient medium drips in at the top, but a constant volume is maintained by means of an overflow siphon. The air coming in at the bottom provides oxygen. At the steady state, the rate of nutrient inflow equals the rate of outflow. Cells within the chemostat are in a continuous state of division, but the population does not increase in size because, in any interval of time, the number of new cells produced by division is balanced by the number washing out through the siphon.

A genetic factor has been described in *Drosophila mauritiana* which results in the spontaneous deletion of the transposable genetic element *mariner* at a frequency of approximately one percent per generation for each copy (Bryan et al., 1987). In a population containing a locus at which a *mariner* insertion is fixed (homozygous), how many generations would be required for the frequency of individuals which are homozygous for a deletion of the element to exceed 5 percent? Assume that the population is large, that mating is random, that the excision factor is fixed, and that deletion of the element does not affect survival or reproduction.

ANSWER: Let p_t be the frequency of chromosomes in which the *mariner* element remains undeleted in generation t , and $\mu = 0.01$ be the probability of deletion of the element per generation. With $v = 0$ and $p_0 = 1$, Equation 2.13 implies that $p_t = (1 - \mu)^t$. The frequency of deletion homozygotes is greater than 5 percent when $(1 - p_t)^2 > 0.05$, or $p_t < 1 - (0.05)^{1/2} = 0.776$. Thus, t should be greater than $\ln(0.776)/\ln(0.99) = 25.2$ generations.

PROBLEM

11

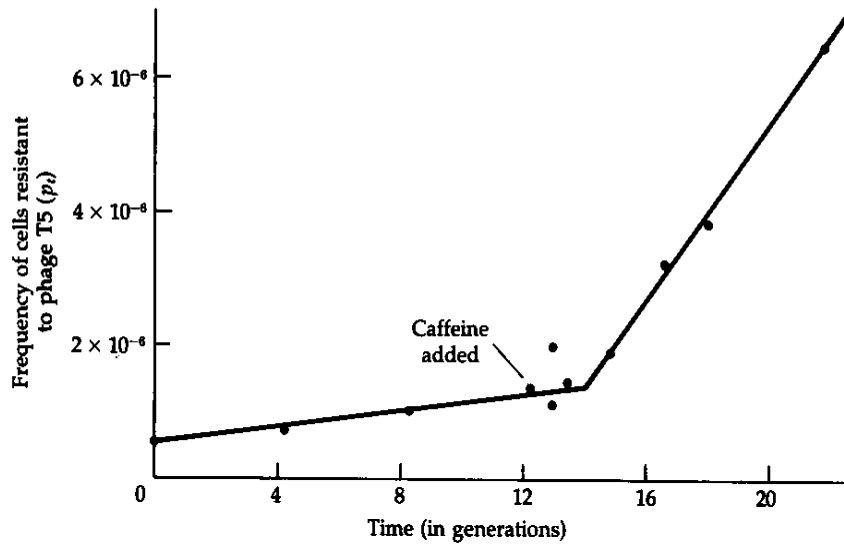


Figure 8. Measurement of mutation rates in bacterial chemostats. This example concerns the rate of mutation of a gene for resistance to bacteriophage T5. The mutation rate is estimated as the slope of the straight-line segments. Prior to addition of caffeine, the slope was $\langle \nu \rangle = 1.3 \times 10^{-8}$ per hour. After addition of caffeine at a concentration of 150 mg/l, the slope increased to $\langle \nu \rangle = 12 \times 10^{-8}$ per hour. Mutation rates per hour in bacterial chemostats are constant for generation times between 2 and 12 hours. The present experiment had a generation time of 5.5 hours. (From Novick 1955.)

abrupt increase in mutation rate (indicated by the increase in slope) shortly after the addition of caffeine, a bacterial mutagen.

To see what happens to allele frequency in the long run, consider Equation 2.13 in the case when t is very large, for example 10^5 or 10^6 generations. Even though $1 - \mu - \nu$ is ordinarily close to 1, the value of t eventually becomes so large that $(1 - \mu - \nu)^t$ becomes approximately 0. Thus, the whole right-hand term in Equation 2.13 drops out, and p_t attains a value that remains the same generation after generation. Such a value of p is called an **equilibrium** value, which we will denote by \hat{p} . In case of mutation, the equilibrium is given by

$$\hat{p} = \frac{\nu}{\mu + \nu} \quad (2.14)$$

The manner in which p_t converges to its equilibrium value is shown in Figure 9 for the case $\mu = 10^{-4}$ and $\nu = 10^{-5}$. Note that, whatever the initial frequency of A , the allele frequency of A eventually goes to \hat{p} , which in the example in the figure equals $0.00001/(0.0001 + 0.00001) = 1/11 = 0.091$. Figure 9 indicates that mutation pressure is usually very weak in changing allele frequency,

inasmuch as the population requires thousands or tens of thousands of generations to reach equilibrium.

**PROBLEM
12**

Stocker (1949) studied a case in the bacterium *Salmonella typhimurium* in which the mutation rates were sufficiently large that Equation 2.13 can be tested. The gene in question controls a protein component of the cellular flagella, and there are two alleles, which we can call A (for the "specific-phase" flagellar property) and a (for the "group-phase" flagellar property). The mutation rate from A to a was estimated as $\langle\mu\rangle = 8.6 \times 10^{-4}$ per generation, and that of a to A as $\langle\nu\rangle = 4.7 \times 10^{-3}$ per generation. These mutation rates are orders of magnitude larger than typically observed for other genes. The reason is that the change from A to a and back again does not involve mutation in the conventional sense, but results from intrachromosomal recombination (Simon et al., 1980). Formally, however, we can treat the system as one with reversible mutation. In cultures initially established with the frequency of A at $p_0 = 0$, Stocker found that the frequency increased to $p = 0.16$ after 30 generations and to $p = 0.85$ after 700 generations. In cultures initiated with $p_0 = 1$, the frequency decreased to 0.88 after 388 generations and to 0.86 after 700 generations. How do these values agree with those calculated from Equation 2.13 using the estimated mutation rates? What is the predicted equilibrium frequency of A?

ANSWER Note that $\langle\nu\rangle/(\langle\mu\rangle + \langle\nu\rangle) = 0.845$. This is the predicted equilibrium frequency (Equation 2.14). Also, $1 - (\langle\mu\rangle + \langle\nu\rangle) = 0.99444$, and this quantity determines the rate of approach to equilibrium. For the cultures with $p_0 = 0$, predicted values are $p_{30} = 0.845 - (0.845)(0.99444)^{30} = 0.13$ and $p_{700} = 0.845 - (0.845)(0.99444)^{700} = 0.83$. For the cultures with $p_0 = 1$, predicted values are $p_{388} = 0.845 + (0.155)(0.99444)^{388} = 0.86$ and $p_{700} = 0.845 + (0.155)(0.99444)^{700} = 0.85$. The predicted values are in good agreement with the observations.

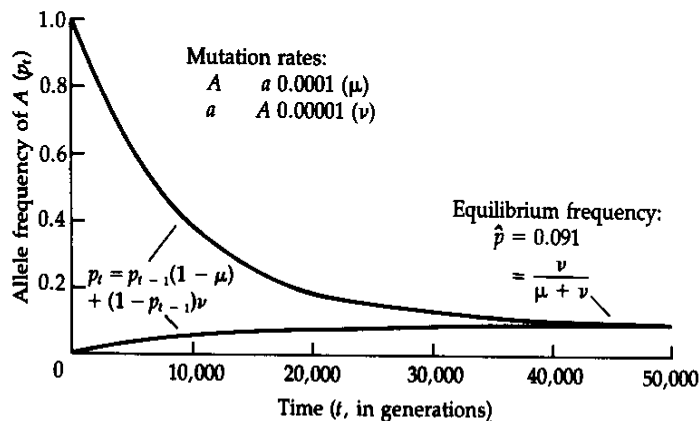


Figure 9. Theoretical change in allele frequency under pressure of reversible mutation. Note that attainment of near-equilibrium values requires tens of thousands of generations for realistic mutation rates.

Number of Alleles Maintained in Populations

Recall from Chapter 1 that genes controlling allozymes often have more than two alleles. It is therefore of some importance to determine how many alleles can be maintained by mutation pressure. If allozyme genes tend to have more alleles than would be expected from mutation pressure alone, then other forces that operate in nature must tend to accumulate alleles. On the other hand, if there are fewer alleles than expected, then other forces must tend to eliminate alleles. There is a technical problem in calculating the number of alleles that can be maintained by mutation. If the procedure for distinguishing among alleles has low resolving power, few alleles will be detected no matter how many may actually be present in the population. Electrophoresis augmented with other procedures may have satisfactory resolving power for some types of studies, but many alleles may still remain undetected. Nevertheless, in principle, we can imagine for the sake of argument that all alleles are distinguishable, as is the case when their DNA nucleotide sequences are determined. Since an average protein contains roughly 300 amino acids, the average length of the amino acid coding part of an average gene is roughly 900 nucleotides. The number of possible alleles is therefore very large— 4^{900} to be exact, which equals about 10^{542} . Hence, we can suppose that every time a mutation occurs, it creates a new allele that does not already exist in the population. This is called the **infinite-alleles model** of mutation. The infinite-alleles model is but one way to specify the characteristics of new mutations. Although it represents a somewhat simplified view of mutation, it nevertheless provides a useful standard of comparison for other models or for observed allele frequencies.

In the infinite-alleles model, we must also assume that the population in question is finite, because in an infinite population new alleles would continue to accumulate forever. Consequently, the model of interest is one in which both random genetic drift and mutation occur. In order to evaluate whether actual populations contain more or fewer alleles than would be expected from mutation pressure alone in the infinite-alleles model, we must determine how many alleles can be maintained in the infinite-alleles model. The question is, for what number of alleles is the creation of new alleles by mutation balanced exactly by the loss of old alleles due to random genetic drift? For the infinite-alleles model, the answer is surprisingly simple. It can be calculated by expressing the proportion of homozygous genotypes in two ways: in terms of the allele frequencies, and in terms of the fixation index (F_i in Equation 2.7). Setting these two proportions equal to each other yields the desired result. Accordingly, let the effective population number be N and the mutation rate be μ . Because of the infinite-alleles assumption, each allele

in the population arises only once, and homozygotes for any allele are therefore autozygous. For n alleles with frequencies $p_1, p_2, p_3, \dots, p_n$, the homozygosity in any subpopulation is

$$\Sigma p_i^2 = p_1^2 + p_2^2 + \dots + p_n^2 \quad (2.15)$$

because mating is random within each subpopulation. While Σp_i^2 equals the homozygosity in terms of allele frequency, the autozygosity in the infinite-alleles model also equals the fixation index F . The calculation of F_t in Figure 5 is still correct for this model providing neither allele (α_i or α_j) undergoes mutation in the passage of one generation. Therefore, $F_t = [1/2N + (1 - 1/2N)F_{t-1}](1 - \mu)^2$. The equilibrium value of F_t , call it \hat{F} , is found by solving $\hat{F} = [1/2N + (1 - 1/2N)\hat{F}](1 - \mu)^2$, from which

$$\hat{F} = \frac{1}{4N\mu + 1} \quad (2.16)$$

to an excellent approximation. Therefore, the number of selectively neutral alleles increases under mutation pressure until \hat{F} satisfies Equation 2.16.

We thus have two measures of homozygosity: Σp_i^2 in Equation 2.15 and \hat{F} in Equation 2.16. Since they measure the same thing, they must equal each other, so $\Sigma p_i^2 = \hat{F} = 1/(4N\mu + 1)$. (Alternative approaches leading to essentially the same result are discussed in Sved and Latter, 1977.) One complication is that any number of distributions of allele frequency can result in the same homozygosity. For example, an equilibrium population with four alleles at frequencies $p_1 = 0.7, p_2 = 0.1, p_3 = 0.1$, and $p_4 = 0.1$ has a homozygosity of $\hat{F} = p_1^2 + p_2^2 + p_3^2 + p_4^2 = (0.7)^2 + (0.1)^2 + (0.1)^2 + (0.1)^2 = 0.52$; likewise a population with two alleles at frequencies $p_1 = 0.6, p_2 = 0.4$ has a homozygosity of 0.52, because $p_1^2 + p_2^2 = (0.6)^2 + (0.4)^2 = 0.52$. The problem that many distributions of allele frequency can result in the same homozygosity can be sidestepped by assuming that all alleles are equally frequent. If the population contains n equally frequent alleles, then $p_1 = p_2 = p_3 = \dots = p_n = 1/n$, and the homozygosity is given in Equation 2.15 as $n(1/n)^2 = 1/n$. At equilibrium, therefore, $1/n = \hat{F} = 1/(4N\mu + 1)$, or $n = 4N\mu + 1$. The number n of equally frequent alleles is called the **effective number of alleles**, often symbolized as n_e . Diverse distributions of allele frequency can therefore be compared in terms of their effective number of alleles. The four-allele population and the two-allele population above that have identical homozygosities of 0.52 have the same effective number of alleles, namely $n_e = 1/0.52 = 1.92$. Biologically speaking, n_e is the number of equally frequent alleles in an ideal population that would be required to produce the same homozygosity as in an actual population. At equilibrium, $n_e = 4N\mu + 1$.

PROBLEM 13 In a large electrophoretic study of Caribbean populations of *Drosophila willistoni*, Ayala and Tracy (1974) found five alleles of the leucine amino peptidase-5 gene (*Lap-5*), with allele frequencies estimated as 0.494, 0.429, 0.057, 0.014, and 0.006 in the Santo Domingo population, and 0.801, 0.177, 0.014, 0.004, and 0.004 in the Yunque population. The Yunque population contained eight alleles of the xanthine dehydrogenase gene (*Xdh*), with frequencies estimated as 0.446, 0.406, 0.092, 0.034, 0.014, 0.004, 0.002, and 0.002. The Yunque population had estimated allele frequencies of 0.574, 0.309, 0.114, and 0.003 of alleles of the adenylate kinase-1 gene (*Adk-1*). Estimate the effective number of alleles in each of these cases.

ANSWER For *Lap-5* in Santo Domingo, $\langle n_e \rangle = 2.32$; for *Lap-5* in Yunque, $\langle n_e \rangle = 1.49$; for *Xdh*, $\langle n_e \rangle = 2.68$; for *Adk-1*, $\langle n_e \rangle = 2.28$. Note that the effective number of alleles is determined more by the uniformity of allele frequencies than by the actual number of alleles.

PROBLEM 14 Show that an infinite population having a gene with an infinite number of alleles at frequencies $1/2, 1/4, 1/8, 1/16, 1/32, \dots$ has an effective number of alleles equal to 3. (Hint: The infinite sum $x + x^2 + x^3 + x^4 + \dots$ equals $x/(1 - x)$ for any value of x less than 1.)

ANSWER Homozygosity equals $(1/2)^2 + (1/4)^2 + (1/8)^2 + \dots$, which is $1/4 + 1/16 + 1/64 + \dots$, or $x + x^2 + x^3 + \dots$ with $x = 1/4$. Thus, homozygosity equals $x/(1 - x) = (1/4)/(3/4) = 1/3$. Effective number of alleles is the reciprocal of the homozygosity, so $n_e = 3$.

The Neutrality Hypothesis

The hypothesis that observed allozyme polymorphisms result from selectively neutral alleles maintained in a balance between mutation and random genetic drift is known as the **neutrality hypothesis** or the hypothesis of **selective neutrality** (Kimura 1968; King and Jukes 1969). In essence, the neutrality hypothesis states that many mutations have so little effect on the organism that their influence on survival and reproduction is negligible. The frequencies of neutral alleles are not determined by the forces of natural selection, but are instead determined by other forces such as migration (discussed later) and random genetic drift. Because the neutrality hypothesis is of fundamental importance in population genetics and evolution, it has been (and still is) a subject of considerable discussion. If the hypothesis is true, or approximately true, then observed polymorphisms—of allozyme genes, for example—may have no particular significance in the adaptation of organisms to their present environments. From this perspective, selectively neutral polymorphisms are mere evolutionary “noise,” and regardless of how much their study may reveal about population structure and random genetic

drift, they tell us little or nothing about adaptive genetic changes in evolution. If, on the other hand, the neutrality hypothesis is false, the next important task would be to study in detail the actual adaptive significance of genetic polymorphisms.

To assess the plausibility of the neutrality hypothesis, many aspects of the hypothesis must be compared with the situation in actual populations. One aspect of the hypothesis developed in the preceding section concerns the homozygosity to be expected with the infinite-alleles model. Using observed homozygosities, we can estimate the effective number of alleles $\langle n_e \rangle$, and from the expression $n_e = 4N\mu + 1$, estimate the corresponding values of $\langle N\mu \rangle$. If the resulting values are grossly unreasonable, we can safely reject the infinite-alleles version of the neutrality hypothesis (or at least argue that actual populations cannot be in equilibrium).

Recall from Chapter 1 that observed values of heterozygosity of allozyme genes range from 0.04 to 0.14 in most organisms (see Figure 6 in Chapter 1). Observed homozygosities therefore range from $1 - 0.04 = 0.96$ to $1 - 0.14 = 0.86$, which correspond to estimated $\langle n_e \rangle$ in the range $1/0.96 = 1.04$ to $1/0.86 = 1.16$. Estimates of $\langle N\mu \rangle$, calculated as $(\langle n_e \rangle - 1)/4$, therefore range from 0.01 to 0.04. That the maximum estimated value of $\langle N\mu \rangle$ differs from the minimum by a factor of only about four is surprising inasmuch as the population number in different species ranges over a factor of 10^4 or more. The apparently too-uniform distribution of homozygosities of allozyme genes among diverse organisms has been interpreted as meaning that the neutrality hypothesis is wrong. This conclusion implies that natural selection is involved somehow in the maintenance of genetic polymorphisms. On the other hand, rejection of the neutrality hypothesis on grounds of the estimated range of $\langle n_e \rangle$ is probably premature because routine electrophoresis does not distinguish all alleles (see Chapter 1 for discussion). Beyond that, estimates of effective population number $\langle N \rangle$ in natural populations are generally imprecise because the studies are very difficult, and estimates of $\langle \mu \rangle$ (which in this case is the mutation rate to *neutral* alleles) are even more uncertain.

Figure 10(a) shows a second type of test of the adequacy of the neutrality hypothesis in explaining observed levels of genetic variation of allozyme genes. The figure shows the observed distribution of heterozygosity of 74 genes in Caucasians (shaded) along with the computer-generated theoretical distribution expected with the infinite-alleles model (solid lines). The observed average heterozygosity is 0.099, and the theoretical heterozygosity is 0.091. The correspondence between the histograms is fairly good, but the observed distribution seems to include too many genes with heterozygosities in the range of 0.35 to 0.55 (for a possible explanation see Fuerst et al. 1977).

A third type of test of the neutrality hypothesis is shown in Figure 10(b), which presents data on the mean and variance of heterozygosity in 77 ver-

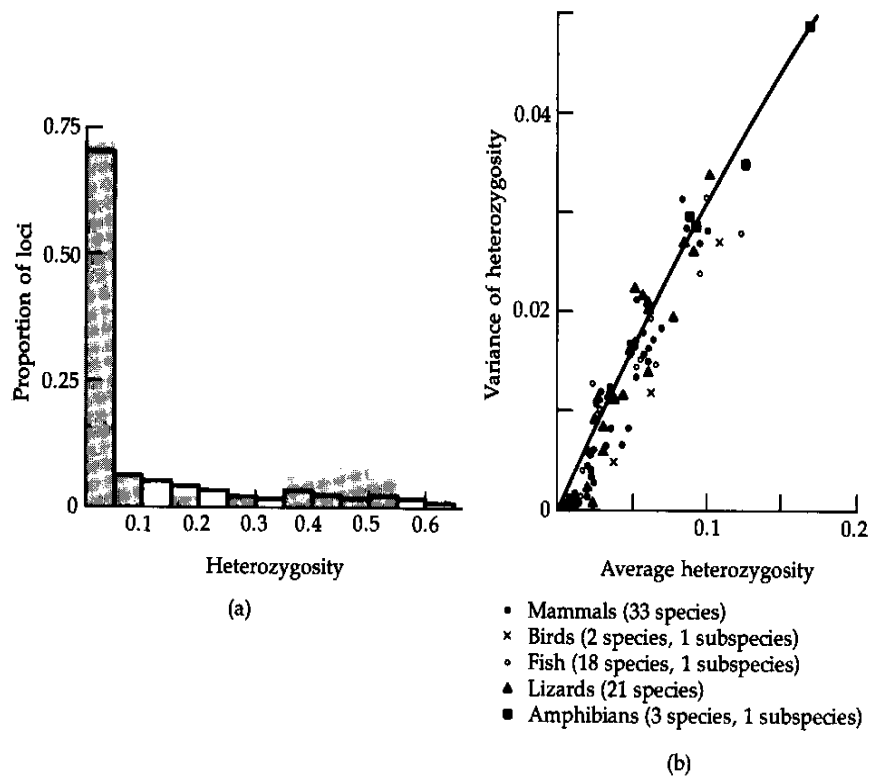


Figure 10. (a) Observed distribution of heterozygosity among genes in Caucasians (shaded) along with theoretical distribution for selectively neutral genes (solid lines). (b) Mean and variance of heterozygosity among genes in vertebrates. The solid line in (b) is the theoretical curve derived by Nei, Fuerst, and Chakraborty for the infinite-alleles model when the mutation rate to neutral alleles varies among genes in such a manner that the variance in mutation rate equals the square of the mean mutation rate. (After Nei et al. 1976.)

tebrate species. The curve is the theoretical expectation from the infinite-alleles model when the rate of selectively neutral mutation varies among genes (Nei et al. 1976). At first glance, the fit in Figure 10b is impressive. On the other hand, the observed points are sufficiently scattered that any number of other curves might fit at least as well. Evidently, statistical comparisons of this sort are too lacking in power to distinguish between the hypotheses.

A brief background on the phrase *lacking in power* may be in order. The neutral theory is useful in being a sort of starting point, or null hypothesis, which provides predictions about the relations among observed quantities that can be confirmed or rejected. Statistical tests of the neutral theory are similar to other types of statistical tests in that two distinct types of possible errors must be balanced. If the tests are too demanding, for example in failing

to allow for the effects of random sampling error, then data may often result in rejection of the hypothesis, even when it is true. This is called *Type I error*. On the other hand, if the statistical test allows too much latitude in the data, then data will seldom result in rejection of the hypothesis, even when it is false. This is called *Type II error*. The trade-off between Type I error and Type II error is that the probability of Type I error cannot be decreased without increasing the probability of Type II error, and vice versa. By convention, statisticians usually adopt a five percent criterion for rejection of the null hypothesis. This is the familiar five-percent level of statistical significance, and it means that there is a five percent chance of rejecting a true hypothesis (Type I error). With this convention, the probability of a Type II error (failing to reject a false hypothesis) falls where it may, and tests which relatively high probabilities of Type II error are said to be *lacking in power*.

Figure 10 is a preliminary introduction to Chapter 3. It is relevant here because of the use of the effective number of alleles and average heterozygosity. Although the comparisons in the figure are lacking in power and hence are inconclusive in their support of the neutrality hypothesis, many other observations and types of data have been brought to bear in assessing the hypothesis. These data often rely on comparisons of nucleotide sequences of DNA in different genes or in different species. These types of comparisons and the conclusions from them are discussed further in Chapter 3.

MIGRATION

In a subdivided population like the one in Figure 2, random genetic drift results in genetic divergence among subpopulations. **Migration**, which refers to the movement of individuals among subpopulations, is a sort of genetic glue that holds subpopulations together and sets a limit to how much genetic divergence can occur. To understand the homogenizing effects of migration, it is useful to study migration in several simple models of population structure.

The Island Model of Migration

In the **island model** of migration, a large population is split into many subpopulations dispersed geographically like islands in an archipelago. Examples of island population structure might include fish in freshwater lakes or slugs in dispersed garden plots. Each subpopulation is assumed to be so large that random genetic drift can be neglected. Consider two alleles A and a with average allele frequencies among the subpopulations equal to \bar{p} and \bar{q} , respectively. Migration is assumed to occur in such a way that, in terms

of allele frequency, the migrants are representative of the subpopulations. That is, among the migrants, the allele frequencies of A and a equal \bar{p} and \bar{q} , respectively. The amount of migration is measured by a number m that equals the probability that a randomly chosen gene in any subpopulation comes from a migrant. Focus attention on a randomly chosen gene in any subpopulation in generation t . This allele could have come from the same subpopulation in generation $t - 1$ (with probability $1 - m$), in which case it is an A allele with probability p_{t-1} , where p_{t-1} represents the allele frequency of A in the subpopulation in question in generation $t - 1$. Alternatively, the allele could have come from a migrant in generation $t - 1$ (with probability m), in which case it is an A allele with probability \bar{p} . (Since random genetic drift, mutation, and natural selection are being ignored, \bar{p} stays the same in all generations.) Altogether,

$$p_t = p_{t-1}(1 - m) + \bar{p}m \quad (2.17)$$

Equation 2.17 is similar to Equation 2.13 for mutation, and its solution in terms of p_0 is

$$p_t = \bar{p} + (p_0 - \bar{p})(1 - m)^t \quad (2.18)$$

where p_0 is the initial frequency of A in the subpopulation in question. As an example of the use of Equation 2.18, suppose there are only two populations with initial allele frequencies of A of 0.2 and 0.8, respectively, with $m = 0.10$ (i.e., 10 percent of the individuals in either subpopulation in any generation are migrants in which the allele frequency of A is $\bar{p} = (0.2 + 0.8)/2 = 0.5$). What is the allele frequency of A in the two populations after 10 generations? For the population with initial allele frequency 0.2, we substitute $p_0 = 0.2$, $\bar{p} = 0.5$, and $m = 0.10$ into Equation 2.18 to obtain $p_{10} = 0.5 + (0.2 - 0.5)(1 - 0.10)^{10}$, or $p_{10} = 0.395$; for the other population, we can substitute $p_0 = 0.8$, $\bar{p} = 0.5$, and $m = 0.10$, so that $p_{10} = 0.5 + (0.8 - 0.5)(1 - 0.10)^{10} = 0.605$.

Another example using Equation 2.18 is shown in Figure 11, where there are five subpopulations (initial frequencies 1, 0.75, 0.50, 0.25, and 0), again with $m = 0.10$. Note how rapidly the allele frequencies converge to the same value (in this case, 0.5). Although Equation 2.18 for migration is mathematically similar to Equation 2.13 for mutation, the biological implications are quite different. Because rates of migration are typically much greater than rates of mutation, changes in allele frequency generally occur much faster with migration.

One-Way Migration

When migration occurs predominantly from one population into another without an equal amount of migration in the reverse direction, then \bar{p} in

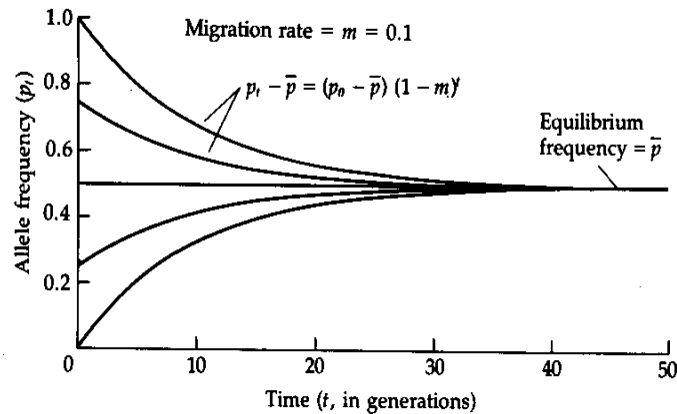


Figure 11. Change of allele frequency with time in five subpopulations exchanging migrants at the rate $m = 0.1$ per generation. Note the rapid convergence to a common equilibrium frequency.

Equation 2.18 should be interpreted as the allele frequency in the population that supplies the migrants. An example is found in racial admixture in the United States. In this case, genetic migration occurs mainly from Whites to Blacks because individuals of mixed racial ancestry are regarded as Blacks. The amount of genetic migration can be estimated using the allele-frequency data in Problem 10 and Equation 2.18. The case of the MN blood groups serves as an example. In West Africa (which may, for the purpose of this problem, be regarded as the ancestral Black population), $\langle p_0 \rangle = 0.474$ for the allele frequency of M. In present-day Claxton Blacks, $\langle p_t \rangle = 0.484$. The Claxton White population may reasonably be regarded as representative of the source of the migrants, and for Claxton Whites $\langle \bar{p} \rangle = 0.507$. Blacks were brought into the United States on a large scale from West Africa about 300 years ago, so $\langle t \rangle$ is about 10 generations. Thus, substituting these estimates into Equation 2.18, we obtain $0.484 = 0.507 + (0.474 - 0.507)(1 - \langle m \rangle)^{10}$, from which $1 - \langle m \rangle = [(0.484 - 0.507)/(0.474 - 0.507)]^{1/10} = 0.965$, or $\langle m \rangle = 0.035$ per generation. This estimate can be interpreted as implying that, in the genetic history of the population of Claxton Blacks, about 3.5 percent of the alleles of the MN gene in any generation were newly introduced by genetic migration from Whites.

Estimate the amount of migration from Whites to Blacks using each of the other genes discussed in Problem 10.

**PROBLEM
15**

ANSWER For Ss, $\langle m \rangle = -0.013$ per generation; for Duffy, $\langle m \rangle = 0.011$; for Kidd, $\langle m \rangle = -0.028$; for Kell, $\langle m \rangle = -0.005$; for G6PD, $\langle m \rangle = 0.039$; for hemoglobin β , $\langle m \rangle = 0.071$.

Problem 15 illustrates some of the difficulties in estimating racial admixture from allele frequencies. The positive values of $\langle m \rangle$ vary widely, and the negative values are not consistent with the proposed model of migration. Cavalli-Sforza and Bodmer (1971) remark that "The weakness of the analysis is mostly due to the uncertainty of the origin of black Americans . . . and the variability of gene frequencies in the probable area of the slave markets in West Africa. In addition, it is unavoidable that gene frequencies have changed somewhat from their original values, due to drift or, in some cases, selection. The opportunities for admixture, and the time available for it, must also have varied widely." The most reliable gene among those in Problem 15 is probably that for the Duffy blood groups, because the Fy^a allele is virtually nonexistent in all of West Africa. For this gene, the estimate of $\langle m \rangle$ is about one percent per generation, which is consistent with the average value for a large number of other genes (Cavalli-Sforza and Bodmer 1971).

Isolate Breaking and Wahlund's Principle

Isolate breaking refers to the fusion of formerly isolated subpopulations by migration. Fusion of populations reduces the frequency of homozygous genotypes, a phenomenon called **Wahlund's principle**. To reinforce this idea, imagine two isolated subpopulations with allele frequencies of a of q_1 and q_2 . The average frequency of homozygous aa genotypes in the subpopulations is, from Table 1, $\bar{q}^2(1 - F) + \bar{q}F$, where \bar{q} represents the average allele frequency of a over all subpopulations (denoted q_0 in Table 1) and F represents the fixation index F_{ST} . Were the subpopulations to fuse into one, the allele frequency of a in the fused population would be \bar{q} . After one generation of random mating, the frequency of homozygous aa genotypes is \bar{q}^2 . Therefore, the average frequency of aa genotypes before population fusion is always greater than the frequency after fusion in the amount

$$\begin{aligned} \Delta R &= \bar{q}^2(1 - F) + \bar{q}F - \bar{q}^2 \\ &= -\bar{q}^2F + \bar{q}F \\ &= \bar{q}(1 - \bar{q})F \end{aligned} \tag{2.19}$$

In human populations, the main effect of isolate breaking (fusion of populations) is to decrease the overall frequency of children born with genetic defects resulting from homozygous recessive genes, particularly harmful recessive genes with frequencies that are, for whatever reasons, relatively high in one of the populations. Examples of harmful recessive genes at high frequency in certain human populations include the alleles for α_1 -antitrypsin deficiency ($\langle q \rangle = 0.024$) and cystic fibrosis ($\langle q \rangle = 0.022$) in Caucasians, sickle-cell anemia in many Negro populations ($\langle q \rangle = 0.05$ in American Blacks, up to $\langle q \rangle = 0.1$ in some African populations), albinism in the Hopi and some

other Southwest American Indians ($\langle q \rangle = 0.07$), and Tay-Sachs disease among Ashkenazi Jews ($\langle q \rangle = 0.013$).

To illustrate the effect of isolate breaking, imagine a population of gray squirrels that, because of a founder effect, has a high frequency of albinism, equal to 16 percent. (Albinism is inherited absence of pigment resulting from a homozygous recessive gene.) In a nearby forest is another population in which the albino mutation is absent, so the allele frequency in this population is 0. Overall, the average frequency of albinos in the two populations is $(0.16 + 0)/2 = 8$ percent. Were the two populations to fuse and undergo random mating, the allele frequency of the albino mutation in the fused population would be $[(0.16)^{1/2} + (0)^{1/2}]/2 = 0.2$, and the frequency of the homozygous genotype would equal $(0.2)^2 = 4$ percent. That is, the frequency of albinos in the fused population is smaller than the average frequency among the individual isolates.

Tay-Sachs disease is an autosomal-recessive degenerative disorder of the brain that usually leads to death in infancy or early childhood. Among Ashkenazi Jews the incidence of the condition is about 1 in 6000 births, but among non-Jews the incidence is about 1 in 500,000 births (Myriantopoulos and Aronson 1966). What incidence of the disease would be expected among the offspring of matings between Ashkenazi Jews and non-Jews? If these offspring were to mate randomly among themselves, what incidence of the disease would be expected among their offspring?

PROBLEM

16

ANSWER If the allele frequencies of the Tay-Sachs mutation among Ashkenazi Jews and non-Jews are denoted q_1 and q_2 , respectively, then, in the initial populations, $\langle q_1 \rangle = (1/6000)^{1/2} = 0.0129$ and $\langle q_2 \rangle = (1/500,000)^{1/2} = 0.0014$. In offspring of matings between Jews and non-Jews, the expected frequency of homozygous recessives is $\langle q_1 \rangle \langle q_2 \rangle = (0.0129)(0.0014) = 1.806 \times 10^{-5}$, or about 1 in 55,000 births. Within the hybrid population, the allele frequency of the harmful recessive is $\bar{q} = (1/2)[\langle q_1 \rangle(1 - \langle q_2 \rangle) + \langle q_2 \rangle(1 - \langle q_1 \rangle)] + \langle q_1 \rangle \langle q_2 \rangle = 0.0072$; the first term in this expression is for the heterozygous genotypes and the second term for the homozygous recessives. (Theoretically, the estimate should be adjusted to take into account the nonreproduction of homozygous recessives, but the correction is very small and can be neglected.) With random mating within the hybrid population, the frequency of homozygous recessives is $\bar{q}^2 = (0.0072)^2 = 5.184 \times 10^{-5}$, or about 1 in 19,000 births. Subsequent generations have approximately the same gene and genotype frequencies as the second generation.

As indicated in Equation 2.19, the reduction in frequency of recessive homozygotes due to population fusion can be expressed in terms of the fixation index as $\bar{p}\bar{q}F$. The reduction in frequency can also be expressed in terms of the variance in allele frequency among subpopulations. The **variance** measures the degree of dispersion of a set of numbers from the mean, or how

closely the individual numbers cluster around the mean. Among any set of numbers, the variance can be calculated as the mean of the squares (often called the **mean square**) minus the square of the mean. To take an extreme example, imagine two subpopulations with allele frequencies 1.0 and 0, respectively. The mean square is $(1/2)[(1)^2 + (0)^2] = 0.5$ and the mean is $(1/2)[1 + 0] = 0.5$, so the variance σ^2 equals 0.5 [the mean square] $- (0.5)^2$ [the squared mean] $= 0.25$. The allele frequencies in the two populations in this example are as different as they can possibly be, so the maximum possible variance in allele frequency for two populations is 0.25. As another example, suppose the two allele frequencies are 0.75 and 0.25; then $\sigma^2 = (1/2)[(0.75)^2 + (0.25)^2] - [(1/2)(0.75 + 0.25)]^2 = 0.0625$. Note in this example that the deviations of the allele frequencies from the mean are half as large as in the previous example—0.25 instead of 0.50—but the variance is only one-quarter as large. This disproportionate reduction in the variance occurs because the variance depends on the square of the deviation of each allele frequency from the mean. As a final example, suppose the allele frequencies in the two populations are both 0.5. Then $\sigma^2 = (1/2)[(0.5)^2 + (0.5)^2] - [(1/2)(0.5 + 0.5)]^2 = 0$, which is simply the mathematical manner of stating that the populations have the same allele frequency (i.e., there is no variation).

We are now in a position to express Wahlund's principle in terms of the variance in allele frequency. Consider the two populations in Figure 12 before they become fused. The average allele frequency of *a*, denoted \bar{q} , equals $(1/2)(q_1 + q_2)$. The mean square, denoted \bar{q}^2 , is the average of the squares, which equals $(1/2)(q_1^2 + q_2^2)$. The variance in allele frequency σ^2 is numerically equal to the mean square minus the squared mean. In symbols, $\sigma^2 = \bar{q}^2 - \bar{q}^2$. As shown in Figure 12, the average frequency of homozygous *aa* genotypes prior to population fusion is greater than the frequency after fusion by the amount σ^2 . Because the decrease in homozygosity with population fusion is the same whether stated in terms of *F* or σ^2 , it follows that $\bar{p}\bar{q}F = \sigma^2$, or

$$F_{ST} = \frac{\sigma^2}{\bar{p}\bar{q}} \quad (2.20)$$

Equation 2.20 provides a convenient method of estimating *F* from allele-frequency data. For example, in the three pairs of hypothetical populations discussed earlier, $\bar{p} = \bar{q} = 0.5$ in all cases. For the populations with allele frequencies of 1 and 0, $F = \sigma^2/\bar{p}\bar{q} = 0.25/(0.5)(0.5) = 1$, which is simply a statement in terms of the fixation index F_{ST} that the populations are as different as they can be. Similarly, for the populations with allele frequencies of 0.75 and 0.25, $F = 0.0625/(0.5)(0.5) = 0.25$; and for the populations with allele frequencies of 0.5 and 0.5, $F = 0/(0.5)(0.5) = 0$.

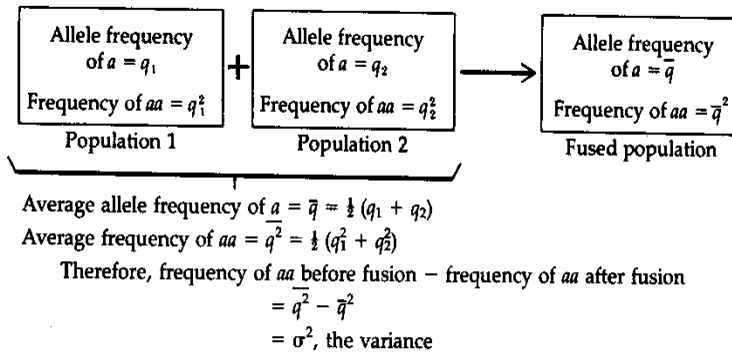


Figure 12. Illustration of Wahlund's principle (isolate breaking). The frequency of homozygous recessives upon population fusion and random mating is less than the average frequency before fusion by an amount equal to the variance in allele frequency. The variance is a measure of the dispersion or spread among a group of numbers, and it is numerically equal to the mean of the squares of the numbers minus the square of the mean of the numbers.

For the MN and Duffy blood groups in Problem 10, verify that $F_{ST} = \sigma^2/\bar{p}\bar{q}$ in the comparisons "West African Blacks versus Claxton Blacks" and "West African Blacks versus Claxton Whites."

**PROBLEM
17**

ANSWER West African Blacks versus Claxton Blacks: $\langle p_1 \rangle =$ estimated frequency M allele in West African Blacks = 0.474, $\langle p_2 \rangle =$ estimated frequency of M allele in Claxton Blacks = 0.484; $\langle \sigma^2 \rangle = [(\rho_1^2 + \rho_2^2)/2] - [(\rho_1 + \rho_2)/2]^2 = 2.5 \times 10^{-5}$; $\langle \bar{p} \rangle = (\rho_1 + \rho_2)/2 = 0.479$, so $\langle \bar{q} \rangle = 0.521$. $\langle F_{ST} \rangle = \langle \sigma^2 \rangle / \langle \bar{p} \rangle \langle \bar{q} \rangle = 0.0001$, which agrees with the answer given in Problem 10. For the Duffy blood group, $\langle p_1 \rangle = 0$, $\langle p_2 \rangle = 0.045$, $\langle \sigma^2 \rangle = 5.063 \times 10^{-4}$, $\langle \bar{p} \rangle = 0.0225$, $\langle \bar{q} \rangle = 0.9775$, $\langle F_{ST} \rangle = 0.0230$. West African Blacks versus Claxton Whites: for MN, $\langle p_1 \rangle = 0.474$, $\langle p_2 \rangle = 0.507$, $\langle \sigma^2 \rangle = 2.722 \times 10^{-4}$, $\langle \bar{p} \rangle = 0.4905$, $\langle \bar{q} \rangle = 0.5095$, $\langle F_{ST} \rangle = 0.0011$. For Duffy, $\langle p_1 \rangle = 0$, $\langle p_2 \rangle = 0.422$, $\langle \sigma^2 \rangle = 0.04452$, $\langle \bar{p} \rangle = 0.211$, $\langle \bar{q} \rangle = 0.789$, $\langle F_{ST} \rangle = 0.2674$. As in Problem 10, parameters have been estimated by equating them with observed values, which ignores certain small corrections for sampling variation.

How Migration Limits Genetic Divergence

It is remarkable how little migration is required to prevent significant genetic divergence among subpopulations resulting from random genetic drift. The effect can be seen quantitatively by considering the model of random drift depicted in Figure 2, but permitting migration at a rate m according to the island model. The expression for F_t derived in Figure 5 is still valid, provided that neither allele α_i or allele α_j has been replaced by a migrant allele.

Therefore $F_t = [1/2N + (1 - 1/2N)F_{t-1}](1 - m)^2$. At equilibrium, $F_t = F_{t-1} = \hat{F}$, so $\hat{F} = [1/2N + (1 - 1/2N)\hat{F}](1 - m)^2$, which for reasonably small values of m is approximately

$$\hat{F} = \frac{1}{4Nm + 1} \quad (2.21)$$

Equation 2.21 for migration has the same form as Equation 2.16 for mutation, except that Nm replaces $N\mu$. This emphasizes the theoretical similarity between the effects of mutation and migration. However, it must be emphasized that rates of migration are typically much greater than rates of mutation, so the practical implications of Equations 2.21 and 2.16 for homozygosity are very different.

Since the actual number of migrants per generation equals Nm , Equation 2.21 implies that \hat{F} decreases as the number of migrants increases. Indeed, the decrease in \hat{F} with increasing Nm is extremely rapid. When $Nm = 0$, $\hat{F} = 1$. However, when

$Nm = 0.25$ (one migrant every fourth generation), $\hat{F} = 0.50$

$Nm = 0.5$ (one migrant every second generation), $\hat{F} = 0.33$

$Nm = 1$ (one migrant every generation), $\hat{F} = 0.20$

$Nm = 2$ (two migrants every generation), $\hat{F} = 0.11$

The implication is that migration is a potent force acting against genetic divergence resulting from genetic drift among subpopulations. For example, the mouse barns in Problem 5 would need only 1.8 migrant mice per barn per generation in order to maintain a value of $\hat{F} = 0.12$. [That is, $4Nm + 1 = (0.12)^{-1}$, so $Nm = 1.8$.]

Estimates of Migration

One method of estimating genetic migration in natural populations relies on the finding that, in theoretical models, the logarithm of Nm decreases approximately as a linear function of the average frequency of private alleles in samples from the subpopulations (Slatkin 1985a). *Private alleles* are those found in only one subpopulation. Data on the average frequency of private alleles in samples has been compiled and analyzed by Slatkin (1985a), and the resulting estimates of Nm and equilibrium values of F_{ST} are summarized in Table 3. There is obviously considerable variation in Nm among organisms. However, many of the values of Nm are approximately two or smaller, which gives considerable scope for genetic divergence resulting from random genetic drift.

A second kind of approach to estimating Nm in natural populations is illustrated in Figure 13, which gives the distribution of estimated values of

Table 3. Estimates of Nm and \hat{F}_{ST} .

Species	Type of organism	Estimated Nm	Estimated \hat{F}_{ST}
<i>Stephanomeria exigua</i>	Annual plant	1.4	0.152
<i>Mytilus edulis</i>	Mollusc	42.0	0.006
<i>Drosophila willistoni</i>	Insect	9.9	0.025
<i>Drosophila pseudoobscura</i>	Insect	1.0	0.200
<i>Chanos chanos</i>	Fish	4.2	0.056
<i>Hyla regilla</i>	Frog	1.4	0.152
<i>Plethodon ouachitae</i>	Salamander	2.1	0.106
<i>Plethodon cinereus</i>	Salamander	0.22	0.532
<i>Plethodon dorsalis</i>	Salamander	0.10	0.714
<i>Batrachoseps pacifica</i> ssp. 1	Salamander	0.64	0.281
<i>Batrachoseps pacifica</i> ssp. 2	Salamander	0.20	0.556
<i>Batrachoseps campi</i>	Salamander	0.16	0.610
<i>Lacerta melisellensis</i>	Lizard	1.9	0.116
<i>Peromyscus californicus</i>	Mouse	2.2	0.102
<i>Peromyscus polionotus</i>	Mouse	0.31	0.446
<i>Thomomys bottae</i>	Gopher	0.86	0.225

(Data from Slatkin 1985a.)

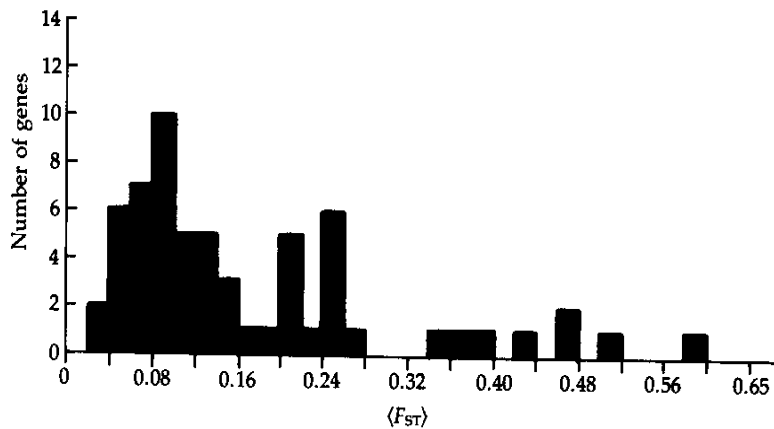


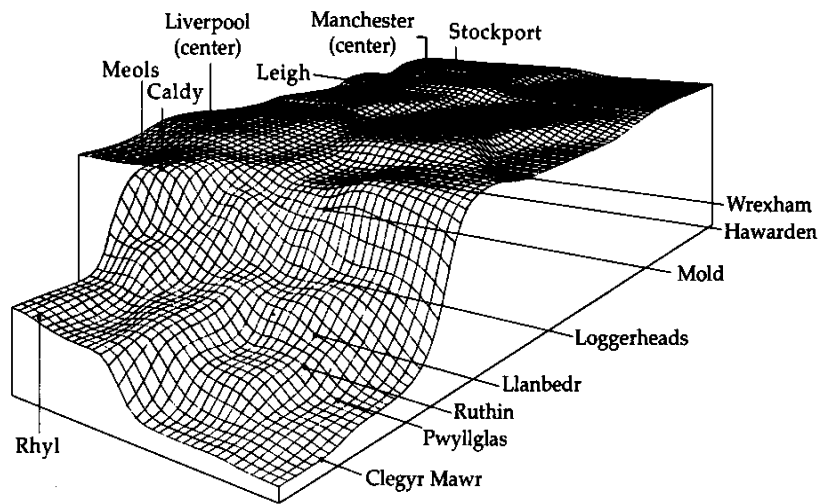
Figure 13. Distribution of estimated values of F_{ST} for 61 genes among natural populations of *Drosophila melanogaster*. While the average value of $\langle F_{ST} \rangle$ suggests migration at a level of Nm between 1 and 2, about one-third of the genes demonstrate $\langle F_{ST} \rangle$ values greater than 0.20. (From Singh and Rhomberg, personal communication.)

$\langle F_{ST} \rangle$ among 61 genes in natural populations of *Drosophila melanogaster* (Singh and Rhomberg, personal communication). The average of the estimated values is $F_{ST} = 0.16$, which, assuming equilibrium, estimates $(4Nm + 1)^{-1}$ (Equation 2.21). Therefore, $\langle Nm \rangle = [(1/0.16) - 1]/4 = 1.3$. This estimate is within the range for other *Drosophila* species in Table 3. However, there are many genes in Figure 13 which have $\langle F_{ST} \rangle$ values greater than 0.30.

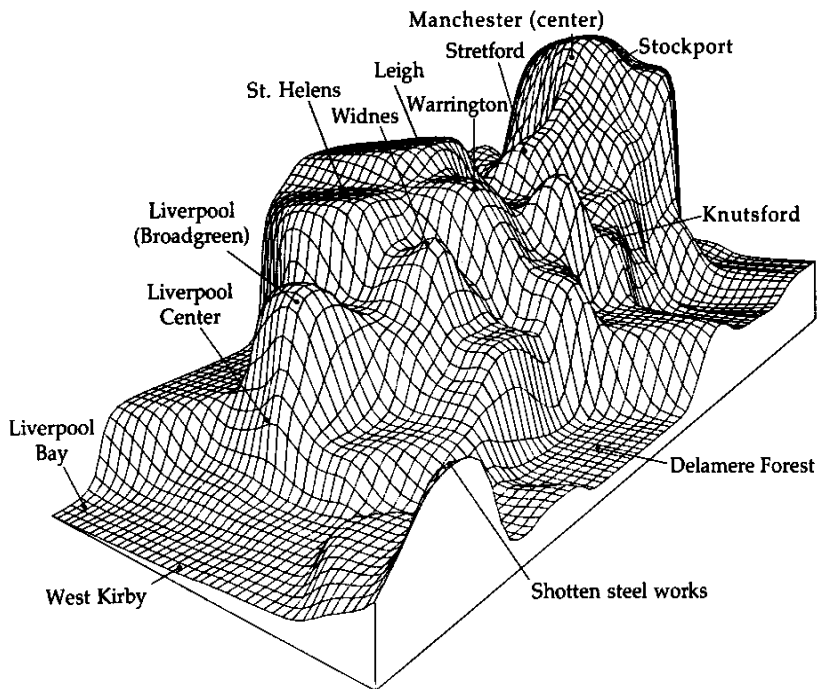
Patterns of Migration

Migration in actual populations is more complex than is assumed in the island model of migration. In nature, migrants come primarily from nearby populations. To the extent that nearby populations have similar allele frequencies, the effects of migration are smaller, and sometimes much smaller, than predicted by the island model. Populations in nature may be strung out along one dimension, such as a river bank. Populations may also be distributed regularly in two dimensions, or there may be one large population with an internal genetic structure caused by the tendency for mating to occur between individuals born in the same region. Analysis of the effects of migration in such complex population structures is usually difficult (Slatkin 1985b). Among humans, migration rates depend on age, sex, marital status, socioeconomic status, population density, and many other factors. Migration rates can change rapidly, moreover, so a full-blown theory of migration has to be extremely complex.

The effects of migration on genetic differentiation of populations are seen dramatically in Figure 14. Panel (a) pertains to the moth *Biston betularia*, panel (b) to the moth *Gonodontis bidentata*. Both species have evolved melanic (black) forms in response to heavy air pollution (see discussion relating to Problem 9 in Chapter 1), and the graphs give the frequency of the melanic forms in the two species. The geographical area in (a) includes Liverpool and Manchester, as viewed from rural Wales. Note the fall-off in frequency of melanics in the nonindustrial areas toward the front of the graph. *B. betularia* occurs in low population densities and must fly relatively long distances to find a mate. The resulting high rate of migration hinders differentiation of populations, hence the smooth surface. The geographical area in (b) is smaller than that in (a), but the view is from the same perspective. *G. bidentata* occurs in high population densities and the migration rate is low, so there is substantial genetic differentiation among populations, as evidenced by the bumpy surface of the graph.



(a)



(b)

Figure 14. (a) Distribution of melanic individuals of *Biston betularia* over an area including Liverpool and Manchester, as viewed from rural Wales. (b) Distribution of melanic individuals of *Gonodontis bidentata* over a smaller area than in (a), but viewed from the same perspective. (From Bishop and Cook 1975.)