CHAPTER 6

INBREEDING, POPULATION SUBDIVISION, AND MIGRATION

In this chapter we consider some of the deep and important implications of the concept of *identity by descent*, which was introduced briefly in Chapter 3 in the context of random genetic drift. Here we show how this concept illuminates the consequences of mating between relatives, not only for the inbred individuals themselves but for genotype frequencies in the population as a whole. Then we show that population subdivision has similarities to inbreeding, because the members of any finite subpopulation are related to one another, though perhaps remotely in time. Finally we consider migration as an evolutionary process that counteracts the tendency for genetic divergence among subpopulations, and discuss applications of coalescence theory to the analysis of real data to make inferences about the history of mutation, migration, random drift, and natural selection among evolving subpopulations.

6.1 INBREEDING

When a mating take place between individuals that are related, the mating constitutes **inbreeding**, and the progeny that result are said to be **inbred**. In human beings, the closest degree of inbreeding usually encountered in most societies is first-cousin mating. But many plants regularly undergo self-fertilization, and some insects regularly practice brother-sister mating. By definition, relatives share one or more common ancestors in their pedigree, and it

is reasonable to suppose that these common ancestors will contribute disproportionately to the genotype of the offspring of a mating between relatives. But how can this effect be measured? The pioneering insight is due to Wright (1922), who formulated a measure of inbreeding called the *inbreeding coefficient* in terms of the correlation between uniting gametes. A later interpretation of the inbreeding coefficient in terms of probability is more transparent (Cotterman 1940; Malécot 1948), and this is the approach we will adopt.

The Inbreeding Coefficient

To make the discussion specific, consider the pedigree in Figure 6.1. It represents the closest degree of inbreeding possible, namely self-fertilization. The curved lines emanating from individual A in generation 0 represent gametes, which join to produce the individual I in generation 1. The black dots on the lines represent the alleles of a gene present in the gametes, which come together to form the genotype of the locus in individual I. The dots are both black to symbolize that they are **identical by descent**, which means that they arose from replication of the same DNA molecule in a previous generation, in this case in generation 0. The **inbreeding coefficient**, typically denoted by the symbol F, is defined as the probability that the two alleles at a locus in an inbred individual are identical by descent. Later in this chapter we will have a need to denote the inbreeding coefficient as F_{IS}, but for now we do not need the subscript and will suppress it.

The concept of identity by descent conceals a subtlety that requires some explanation. As one traces the ancestries of alleles into the past, their ancestral lineages come together, or *coalesce*, reducing the number of ancestral alleles, until ultimately only one common ancestral allele remains. Details of the coalescent process are examined in Chapter 3. Because of coalescence, every allele shares a common ancestral allele with every other allele, and they all are related through DNA replication. Superficially, the coalescent process seems to undermine the concept of the inbreeding coefficient, or at least to render it ambiguous. The ambiguity can be resolved by choosing some arbitrary time in the past and declaring that, at that time, every allele in the population is to be considered as being not identical by descent with any other. This clears the slate because it sets $F = 0$ for all individuals, and thereafter what the inbreeding coefficient actually refers to is the probability of identity by descent subsequent to the time the slate was cleared.

In Figure 6.1, the arbitrary time when all alleles are defined as distinct (not identical by descent) is generation 0. Hence we can write the genotype of individual A in generation 0 as $\alpha_1\alpha_2$, and so by definition α_1 and α_2 are not identical by descent. The probability that the alleles in the inbred offspring I are identical by descent can then be deduced from first principles. Individual I has any one of four possible genotypes with the following probabilities: $\frac{1}{4}\ \alpha_1\alpha_1$, $\frac{1}{4}\ \alpha_1\alpha_2$, $\frac{1}{4}\ \alpha_2\alpha_1$, and $\frac{1}{4}\ \alpha_2\alpha_2$. In the cases $\alpha_1\alpha_1$ and $\alpha_2\alpha_2$, the alleles are identical by descent, and the individual is said to be **autozygous** (the prefix

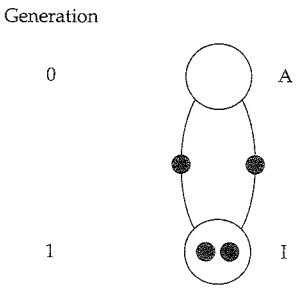


FIGURE 6.1 Pedigree for self-fertilization. Inbred individual I is the result of self-fertilization of the ancestor A. The black dots represent the alleles transmitted by A to I. The inbreeding coefficient of I is defined as the probability that the alleles of a gene in I are identical by descent.

auto means *self*). In the cases $\alpha_1\alpha_2$ and $\alpha_2\alpha_1$, the alleles are not identical by descent, and the individual is said to be **allozygous** (the prefix *allo* means *other*). Note that the concepts of autozygosity and allozygosity have nothing to do with the state of an allele—whether the allele is *A* or *a*, for example. The concepts are concerned only with common ancestry. If the alleles are replicas of a single allele in a common ancestor, they are autozygous; otherwise, they are allozygous.

Because the inbreeding coefficient is equal to the probability of autozygosity, the inbreeding coefficient of individual I is given by $F = \frac{1}{2}$; or to state the result in somewhat different terms, $F = \frac{1}{2}$ is the inbreeding coefficient resulting from one generation of self-fertilization. Two equally valid interpretations of F are:

- F is the probability that any particular gene in an inbred individual has alleles that are identical by descent, or
- F is the overall proportion of genes in an inbred individual that have alleles that are identical by descent.

Because $F = \frac{1}{2}$ in Figure 6.1, this value means that one generation of self-fertilization results in an inbred individual in which 50% of the genes have alleles that are identical by descent (autozygous). Because the span of time in a pedigree is usually short, in this case only one generation, mutation can safely be ignored. Autozygous genotypes must therefore be homozygous, whereas allozygous genotypes can be either homozygous or heterozygous.

Genotype Frequencies with Inbreeding

At the population level, Figure 6.2 illustrates how the concepts of autozygosity and allozygosity are related to those of homozygosity and heterozygosity in a population of inbred organisms. The small circles represent the individuals in a population, with their genotypes indicated for one locus. The population is assumed to be infinite, but to make matters concrete we will focus on these 32 individuals and assume that they are a perfectly rep-

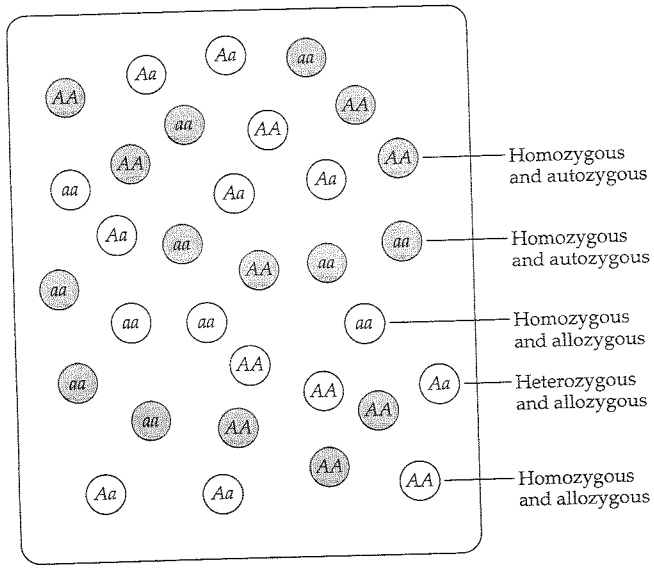


FIGURE 6.2 When there is inbreeding, the alleles in homozygous genotypes may be identical by descent (autozygous, here represented by the shaded circles) or not identical by descent (allozygous). In the absence of mutation, the alleles in heterozygous genotypes must be allozygous.

resentative sample. Some of the circles are shaded, and these represent autozygous genotypes whose alleles are identical by descent. Other circles are unshaded, and these represent allozygous genotypes whose alleles are not identical by descent. Disregarding the possibility of mutation since the time that F was declared to equal 0 (clearing the slate), all autozygous genotypes must be homozygous, but allozygous genotypes may be either homozygous or heterozygous (see Figure 6.2). Since F is the probability of identity by descent, it is also the proportion of individuals whose alleles are autozygous. In this example, $F = \frac{16}{32} = \frac{1}{2}$, which can be determined by counting. Normally one would not be able to distinguish which homozygous genotypes were autozygous and which were allozygous, and here we have chosen them arbitrarily.

The essential point of Figure 6.2 is that two alleles can be *identical by state*, which means that they have the same sequence of nucleotides along the DNA, without being identical by descent. The concept of *identity by descent* pertains to the ancestral origin of an allele and not to its chemical makeup.

Figure 6.2 also illustrates the effect of inbreeding on the genotype frequencies. In this population, the allele frequencies are $p = \frac{16}{32} = \frac{1}{2}$ for A, and $q = \frac{16}{32} = \frac{1}{2}$ for a. These again can be determined by direct counts. With Hardy-Weinberg equilibrium (see Chapter 2), the expected genotype frequencies are $(\frac{1}{2})^2 \times 32 = 8\ AA$, $2(\frac{1}{2})(\frac{1}{2}) \times 32 = 16\ Aa$, and $(\frac{1}{2})^2 \times 32 = 8\ aa$. The genotype counts are actually 12 AA, 8 Aa, and 12 aa. The excess of homozygous genotypes, and deficiency of heterozygous genotypes, are a direct consequence and characteristic of inbreeding.

To understand how inbreeding affects the genotype frequencies, we need only consider the implications of the definition of F for a population of inbred organisms. For this purpose, consider the alleles of a gene present in any one of the inbred organisms. Either of two things must be true: The alleles must either be allozygous (probability $1 - F$) or be autozygous (probability F). If the alleles are allozygous, then the probability that the chosen organism has any particular genotype is simply the probability of that genotype in a random-mating population, because, by chance, the inbreeding has not affected this particular gene. On the other hand, if the alleles are autozygous, then the chosen organism must be homozygous, and the probability of homozygosity for any particular allele is simply the frequency of the allele in the subpopulation as a whole. (Because the alleles in question are autozygous, knowing which allele is present in one chromosome immediately tells you that an identical allele is in the homologous chromosome.) These considerations hold regardless of the number of alleles, but to simplify matters, we consider the case of two alleles A and a at frequencies p and q (with $p + q = 1$). In this case the genotype frequencies are given by

$$AA: \quad p^2(1 - F) + pF = p^2 + pqF \tag{6.1a}$$

$$Aa: \quad 2pq(1 - F) = 2pq - 2pqF \tag{6.1b}$$

$$aa: \quad q^2(1 - F) + pF = q^2 + pqF \tag{6.1c}$$

Equation 6.1a is the probability that an organism has genotype AA; the first term refers to cases in which the alleles are allozygous and the second to cases in which the alleles are autozygous. Similarly, Equation 6.1c is the probability that an organism has genotype aa. Heterozygous Aa genotypes then have the frequency given by Equation 6.1b, since alleles that are heterozygous must be allozygous. The expressions at the far right in Equations 6.1a–c can be obtained by multiplying out those on the left and remembering that $p(1 - p) = q(1 - q) = pq$.

Applying Equation 6.1 to the example in Figure 6.2, we have already shown that $F = \frac{1}{2}$ and also that $p = q = \frac{1}{2}$. From the expressions on the far right in Equation 6, therefore, the expected numbers of each of the three genotypes are $[(\frac{1}{2})^2 + (\frac{1}{2})(\frac{1}{2})(\frac{1}{2})] \times 32 = 12\ AA$, $[2(\frac{1}{2})(\frac{1}{2}) - 2(\frac{1}{2})(\frac{1}{2})(\frac{1}{2})] \times 32 = 8\ Aa$, and $[(\frac{1}{2})^2 + (\frac{1}{2})(\frac{1}{2})(\frac{1}{2})] \times 32 = 12\ aa$. The genotype frequencies shown in Figure 6.2 are therefore in perfect agreement with those expected in the progeny of

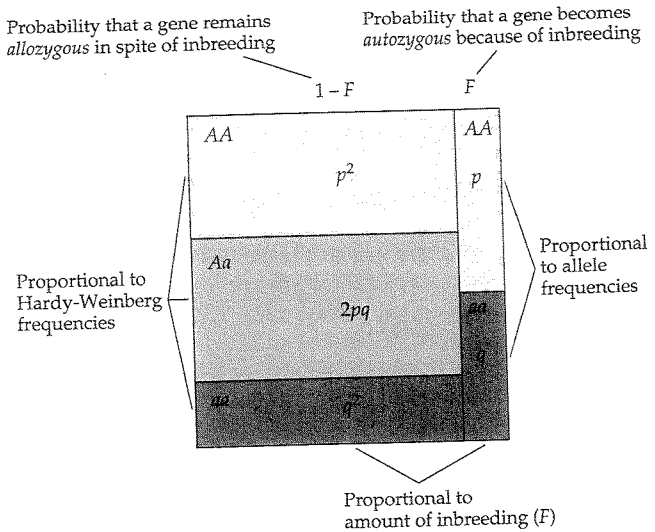


FIGURE 6.3 Graphical representation of the effects of inbreeding on genotype frequencies. Some genes remain allozygous in spite of the inbreeding, and among these the genotype frequencies of *AA*, *Aa*, and *aa* are given by the Hardy-Weinberg principle. Other genes are autozygous because of the inbreeding, and among these the genotype frequencies of *AA* and *aa* are given by the allele frequencies. There are no heterozygotes in the autozygous case because the two alleles present at an autozygous locus are, by definition, identical by descent.

a population of plants in which each individual had undergone one generation of self-fertilization.

The genotype frequencies with inbreeding are summarized graphically in Figure 6.3. The box is divided vertically into two parts, corresponding to genes whose alleles remain allozygous in spite of the inbreeding and those whose alleles are autozygous because of the inbreeding. The division is in the proportion $1 - F : F$. Within the allozygous part of the box, the horizontal panels correspond to the allozygous genotypes *AA*, *Aa*, and *aa*, which are the Hardy-Weinberg frequencies. Within the autozygous part of the box, the horizontal panels correspond to the autozygous genotypes *AA* and *aa*, which are in the proportions $p : q$. Some special cases of Equation 6.1 for the genotype frequencies with inbreeding are given in Table 6.1. When $F = 0$, the genotype frequencies are identical to those of the Hardy-Weinberg principle; and when $F = 1$ (complete inbreeding), all individuals are autozygous, and there is a total absence of heterozygous genotypes.

TABLE 6.1 Genotype Frequencies with Inbreeding

	Frequency in Population			
Genotype	With inbreeding coefficient F		With $F = 0$ (random mating)	With $F = 1$ (complete inbreeding)
AA	$p^2(1-F)$ +	pF	p^2	p
Aa	$2pq(1-F)$		$2pq$	0
aa	$q^2(1-F)$ +	qF	q^2	q
	Allozygous genes	Autozygous genes		

Note also from Equation 6.1 that, while inbreeding does change the genotype frequencies in a population, it does not change the allele frequencies. This is true because, for any value of F, the allele frequency of A is given by $[p^2 + pqF] + (\frac{1}{2})[2pq - 2pqF] = p^2 + pq = p(p + q) = p$. This principle requires the assumption that all genotypes have the same fitness, which is to say that no natural selection takes place. If there is selection, then the allele frequencies can change with inbreeding.

Equation 6.1 generalizes to multiple alleles in a strightforward way. If a gene has multiple alleles $A_1, A_2, \ldots A_n$ at respective frequencies $p_1, p_2, \ldots p_n$ (with $p_1 + p_2 + \cdots + p_n = 1$), then in a population with inbreeding coefficient F, the frequencies of A_iA_i homozygotes and A_iA_j heterozygotes are as follows:

$$A_iA_i: \quad p_i^2(1-F) + p_iF = p_i^2 + p_i(1-p_i)F \qquad [6.2a]$$

$$A_iA_j: \quad 2p_ip_j(1-F) = 2p_ip_j - 2p_ip_jF \qquad [6.2b]$$

CORRELATION BETWEEN UNITING GAMETES Wright's (1922) original conception of the inbreeding coefficient F was as a measure of the correlation between uniting gametes. That this concept is consistent with the probability interpretation is shown for gene with two alleles in Table 6.2. The upper part of the table shows all possible pairs of uniting gametes and their frequencies with inbreeding, with the female gamete on the left and the male gamete on the right. The alleles have been coded with numerical values, *A* with a value 1 and *a* with value 0. Any arbitrary numerical values lead to the same conclusion, but the assignments in Table 6.2 simplify the formulas. The bottom part of the table shows how various expected values are calculated, with the goal of deducing $Cov(xy)$, the covariance between x and y, as well as $V(x)$ and $V(y)$, the variances. By definition, the correlation between uniting gametes r_{UG} (*UG* for uniting gametes) is the ratio of the covariance to the product of the standard deviations, and hence

TABLE 6.2 Correlation between Uniting Gametes

Uniting gametes	Relative frequency	Expected values
$A\ (x=1)\quad A\ (y=1)$	$p^2 + pqF$	$E(x) = p^2 + pqF + pq - pqF = p^2 + pq = p(p+q) = p$
$A\ (x=1)\quad a\ (y=0)$	$pq - pqF$	$E\ (x^2) = p^2 + pqF + pq - pqF = p^2 + pq = p(p+q) = p$
$a\ (x=0)\quad A\ (y=1)$	$pq - pqF$	$E\ (y) = p^2 + pqF + pq - pqF = p^2 + pq = p(p+q) = p$
$a\ (x=0)\quad a\ (y=0)$	$q^2 + pqF$	$E\ (y^2) = p^2 + pqF + pq - pqF = p^2 + pq = p(p+q) = p$
		$E\ (xy) = p^2 + pqF$
		$Cov(xy) = E(xy) - E(x)E(y) = p^2 + pqF - p \times p = pqF$
		$V(x) = E\ (x^2) - [E(x)]^2 = p - p^2 = p(1-p) = pq$
		$V(y) = E\ (y^2) - [E(y)]^2 = p - p^2 = p(1-p) = pq$

$$r_{UG} = \frac{Cov(x,y)}{\sqrt{V(x)V(y)}} = \frac{pqF}{pq} = F \qquad [6.3]$$

Wright, to the end of his long and extraordinarily productive life (he died in 1988 at the age of 98), always preferred his own interpretation of F as a correlation, because under some exceptional circumstances r_{UG} can be negative, and in these cases the probability interpretations fails because a probability cannot be negative.

REDUCTION IN THE FREQUENCY OF HETEROZYGOUS GENOTYPES One of the main effects of inbreeding is that a group of inbred individuals has reduced frequency of heterozygous genotypes, relative to a group of noninbred individuals (see Equation 6.1b). To examine this effect quantitatively, let H_I denote the probability that a gene in an inbred individual is heterozygous, and let H_S denote the proportion of heterozygous genotypes expected with random mating in the subpopulation of which I is a member. With two alleles, Equation 6.1b implies that $H_I = 2pq - 2pqF$, and the Hardy-Weinberg principle implies that $H_S = 2pq$. The proportionate reduction in heterozygosity due to inbreeding, relative to the subpopulation as a whole, is symbolized as F_{IS} and given by the expression

$$F_{IS} = \frac{H_S - H_I}{H_S} = \frac{2pq - (2pq - 2pqF)}{2pq} = F \qquad [6.4]$$

As we shall see Section 6.2, this formulation is particularly useful in thinking about subdivided populations, when both inbreeding and random genetic drift contribute to the overall probability of identity by descent.

PROBLEM 6.1 Plants able to undergo self-fertilization are said to be *self-compatible*. In a population of self-compatible plants, if each plant undergoes self-fertilization a fraction s of the time and otherwise mates randomly, then it can be shown (Crow and Kimura 1970; Hedrick and Cockerham 1986) that F very quickly attains the value $F = s/(2-s)$. *Phlox cuspidata* is self-compatible, and for this species the amount of self-fertilization is estimated at approximately $s = 0.78$ (Levin 1978). From s we can predict the inbreeding coefficient as $F = 0.78/(2 - 0.78) = 0.64$. In a sample of 35 plants from a Texas population of *P. cuspidata*, two alleles of the phosphoglucomutase-2 gene were observed, which we will designate as the *A* and *a* alleles. The sample included were 15 *AA*, 6 *Aa*, and 14 *aa* genotypes (Levin 1978). Are these numbers consistent with the estimate $F = 0.64$? (*Note:* The χ^2 in this case has one degree of freedom because only the allele frequency is estimated from the data; if F also were estimated from the data, rather than being calculated independently from the degree of self-fertilization, then there would be zero degrees of freedom and no goodness-of-fit test would be possible.)

ANSWER The allele frequencies of *A* and *a* are estimated as (30 + 6)/70 = 0.514 and 1 − 0.514 = 0.486, respectively. The hypothesis is that $F = 0.64$, and so $1 - F = 0.36$. The expected numbers of the genotypes *AA*, *Aa*, and *aa* are, respectively, $[(0.514)^2(0.36) + (0.514)(0.64)](35) = 14.8$, $[2(0.514)(0.486)(0.36)](35) = 6.3$, and $[(0.486)^2(0.36) + (0.486)(0.64)](35) = 13.9$. With these expectations, the $\chi^2 = 0.02$ with one degree of freedom, and the associated probability is about 0.96. The fit to the inbreeding model is excellent.

Genetic Effects of Inbreeding

In **outcrossing** species (those that regularly avoid mating between relatives), close inbreeding is generally harmful. The effects are seen most dramatically when inbreeding is complete or nearly complete. In most species of animals, complete autozygosity requires many generations of brother-sister mating. But in *Drosophila melanogaster*, autozygosity of entire chromosomes can be achieved in just a few generations because of the absence of crossing over in the male and the ready availability of genetically marked chromosomes with multiple inversions to prevent crossing over in the female. One widely used inversion chromosome is marked with the dominant mutation *Cy* (for Curly wings), and the critical experimental cross is of the form $Cy/+_i \times Cy/+_i$, where is the ith member of a sample of wildtype chromosomes isolated from a natural population, and the $+_i$ chromosomes are identical by descent. Homozygous *Cy* genotypes do not survive, and so the theoretically expected progeny are $Cy/+_i$ (with curly wings) and $+_i/+_i$ (with straight wings) in a ratio of $\frac{2}{3}:\frac{1}{3}$. If the wildtype chromosome carries one or more mutations that decrease survivorship, then there will be fewer than $\frac{1}{3}$ straight-wing flies,

and if the chromosome carries a recessive lethal mutation, then straight-wing flies will be absent. Control crosses are of the form $Cy/+_i \times Cy/+_j$, where the $+_i$ and $+_j$ chromosomes are not identical by descent. For either type of mating, an estimate $\hat{v}$ of the viability (survivorship) of the +/+ genotype, relative to that of the $Cy/+$ genotype, is given by

$$\hat{v} = \frac{2n_{+/+}}{1+n_{Cy/+}} \qquad (6.5)$$

where $n_{+/+}$ and $n_{Cy/+}$ are the counts of wildtype and *Curly* offspring, respectively (Haldane 1956). The addition of 1 to the denominator makes the estimate of v almost unbiased. When the total number of offspring is large, $\hat{v}$ is essentially equal to two times the number of wildtype offspring divided by the number of *Curly* offspring.

Results of such experiment to estimate the viabilities of autozygous (homozygous) and allozygous (heterozygous) wildtype chromosomes from a natural population are shown in Figure 6.4. It is evident that the homozygous genotypes (shaded histogram) are relatively poor in viability. In fact, about 37% of the homozygous genotypes are lethal. Moreover, among the homozygous genotypes that have viabilities within the normal range of the heterozygous genotypes (open histogram), virtually all can be shown to have reduced fertility (Sved 1975; Simmons and Crow 1977). Inbreeding so close as to make entire chromosomes homozygous is rare in outcrossing species, except in the kind of experiment in Figure 6.4, but the effects are clearly very harmful and provide a new dimension of genetic diversity. In the case of single nucleotide polymorphisms (SNPs), genetic diversity results from common alleles that do not perceptibly impair viability or fertility when homozygous. In the case of inbreeding, the effects are mainly due to rare alleles that are severely detrimental when homozygous. (The fact that the alleles are rare is shown by the small proportion of lethal or near-lethal heterozygous combinations.) Figure 6.4 shows that natural populations of *Drosophila* contain considerable hidden genetic variation in the form of rare deleterious recessive alleles.

Detrimental effects of inbreeding, called **inbreeding depression**, are found in virtually all outcrossing species, and the more intense the inbreeding, the more harmful the effects. Inbreeding in human beings is also generally harmful, but the effect is difficult to measure because the degree of inbreeding is less than that in experimental organisms, and the effects may also vary from population to population. Nevertheless, children of first-cousin matings are, on the average, less capable than noninbred children in any number of ways (for example, higher rate of mortality, lower IQ scores). It should be emphasized, however, that many such children are within the normal range of abilities, and some are quite gifted. Sewall Wright, the celebrated population geneticist, was the child of a first-cousin marriage.

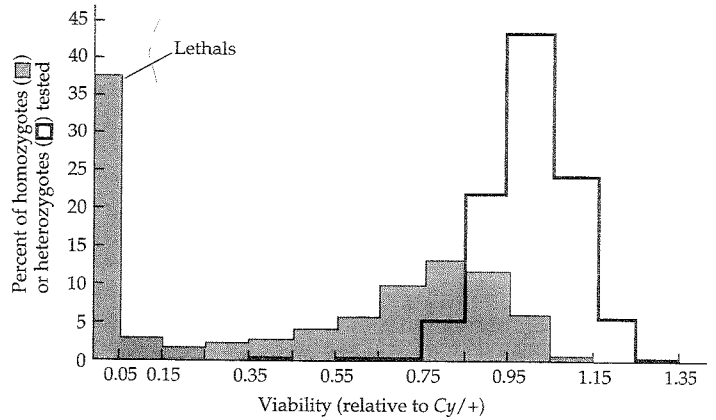


FIGURE 6.4 Viability distributions of wildtype homozygotes (shaded area) and wildtype heterozygotes (black outline) of second chromosomes extracted from *Drosophila melanogaster*. The histograms depict results of testing 691 homozygous combinations and 688 heterozygous combinations. Note that, in this sample, nearly 37% of the wildtype chromosomes are lethal when homozygous, and many more have viabilities substantially below normal. (Data from Mukai et al. 1974.)

In human populations, as in most organisms, the deleterious effects of inbreeding are largely due to the increased homozygosity of rare recessive alleles, and so inbreeding effects in human beings are seen most dramatically in the increased frequency of genetic abnormalities due to harmful recessive alleles among the children of first-cousin matings. The increased frequency of such conditions can be deduced from the genotype frequency given in Equation 6.1c. For the offspring of a first-cousin mating, $F = \frac{1}{16}$, as will be shown in the next section. Suppose that a is a rare deleterious recessive allele with an allele frequency of q. Then, among the children of first-cousin matings, the frequency of aa is expected to be $q^2 + pq(\frac{1}{16})$. On the other hand, among the offspring of matings that take place at random, the frequency of homozygous recessives equals q^2.

Now if c is the proportion of first-cousin matings in a population, then the expected proportion of homozygous aa offspring in the population as a whole that result from the first-cousin matings is given by

$$\frac{c(q^2 + pq/16)}{q^2(1-c) + c(q^2 + pq/16)} = \frac{c(1+15q)}{c+16q-cq} \qquad (6.6)$$

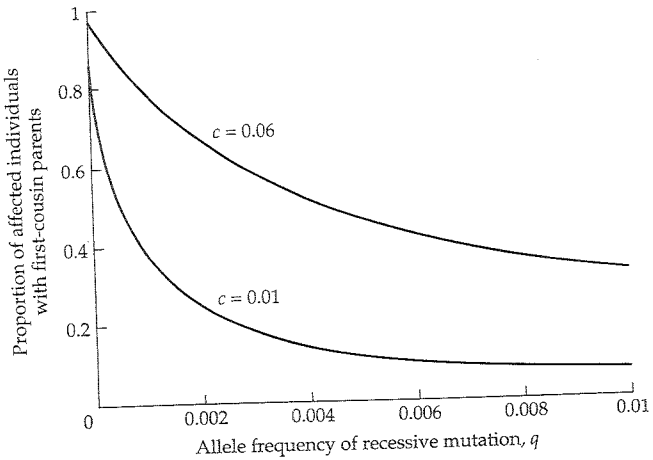


FIGURE 6.5 For rare recessive alleles, even low levels of inbreeding can account for a disproportionate share of homozygous recessive offspring. The curves correspond to overall proportions of first-cousin mating of 1% and 6%, and show that when a recessive allele is rare, a large proportion of homozygous genotypes result from the small proportion of first-cousin matings. The reason is that the offspring of first cousins have a $\frac{1}{16}$ chance of carrying alleles that are identical by descent.

Figure 6.5 shows plots of this proportion for $c = 0.01$ and $c = 0.06$, a range that includes most human populations. Note that, as the recessive allele becomes more rare, the first-cousin matings account for an increasing proportion of all affected children. Consider albinism as an example, which is due to a rare recessive mutation. Although the allele frequency differs among human subpopulations, we will take $q = 0.005$ as typical, which predicts a frequency among the children of nonrelatives of $q^2 = 0.0025\%$, or about one in 40,000. The curves in Figure 6.5 imply that, when the frequency of first cousin matings equals 1% (approximately the value in the Unites States), then the proportion of albino children whose parents are first cousins is 12%. In a population with $c = 0.06$, although first-cousin matings account for 6% of all matings, they account for 46% of matings with albino children.

Calculation of the Inbreeding Coefficient from Pedigrees

Calculation of F from a pedigree is simplified by drawing the pedigree in the form shown in Figure 6.6A, where the lines represent gametes contributed by parents to their offspring. The same pedigree is shown in conventional form in Figure 6.6B. The organisms in gray in part B are not represented in part A

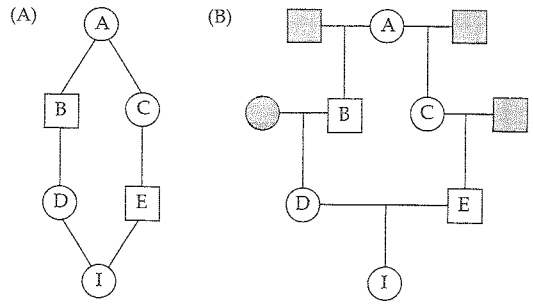


FIGURE 6.6 (A) Convenient way to represent pedigrees for calculation of the inbreeding coefficient. In this case, the pedigree shows a mating between half-first cousins. (B) Conventional representation of the same pedigree as in part A. Squares represent males, circles represent females, and the shaded organisms in part B are not depicted in part A because they do not contribute to the inbreeding of the inbred organism designated I.

because they have no ancestors in common and therefore do not contribute to the inbreeding of the organism denoted I. The inbreeding coefficient F_I of I is the probability that I is autozygous for the alleles of an autosomal gene under consideration. The first step in calculating F_I is to locate all the common ancestors in the pedigree, because an allele could become autozygous in I only if it were inherited through both of I's parents from a common ancestor; in this case, there is only one common ancestor, namely, A. The next step in calculating F_I, which is carried out for each common ancestor in turn, is to trace all the paths of gametes that lead from one of I's parents back to the common ancestor and then down again to the other parent of I. These paths are the paths along which an allele in a common ancestor could become autozygous in I. In Figure 6.6A, there is only one such path: DBACE, in which the common ancestor is underlined for bookkeeping purposes, an especially useful procedure in complex pedigrees.

The third step in calculating F_I is to calculate the probability of autozygosity in I due to each of the paths in turn. For the path DBACE, the reasoning is illustrated in Figure 6.7. Here the black dots represent alleles transmitted along the gametic paths, and the number associated with each step is the probability of identity by descent of the alleles indicated. For all steps except that around the common ancestor, the probability is $\frac{1}{2}$ because, with Mendelian segregation, the probability that a particular allele present in a parent is transmitted to a specified offspring is $\frac{1}{2}$. To understand why $(\frac{1}{2})(1 + F_A)$ is the probability associated with the loop around the common ancestor, denote the alleles in the common ancestor as α_1 and α_2. These symbols are

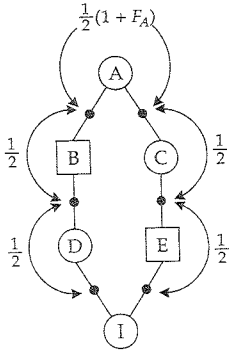


FIGURE 6.7 Loops for the pedigree in Figure 6.6A, showing probabilities that designated alleles (solid dots) are identical by descent. Each loop is independent of the others, so their probabilities multiply. Thus, the inbreeding coefficient of organism I is $F_I = (\frac{1}{2})^5(1 + F_A)$, where F_A represents the inbreeding coefficient of the common ancestor.

used to avoid confusion with conventional allele symbols designating functional types of alleles, such as *A* for dominant and *a* for recessive. The pair of gametes contributed by A could contain $\alpha_1\alpha_1$, $\alpha_2\alpha_2$, $\alpha_1\alpha_2$, or $\alpha_2\alpha_1$, each with a probability of $\frac{1}{4}$ because of Mendelian segregation. In the first two cases, the alleles are clearly identical by descent; in the second two cases, the alleles are identical by descent only if α_1 and α_2 are already identical by descent, which means that A is autozygous. The probability that A is autozygous is, by definition, the inbreeding coefficient of A, F_A. Hence, the probability for the step around the common ancestor A is $(\frac{1}{4}) + (\frac{1}{4}) + (\frac{1}{4})F_A + (\frac{1}{4})F_A = (\frac{1}{2}) + (\frac{1}{2})F_A = (\frac{1}{2})(1 + F_A)$. Because each of the steps in Figure 6.7 is independent of the others, the total probability of autozygosity in I due to the path through A is $(\frac{1}{2}) \times (\frac{1}{2}) \times (\frac{1}{2})(1 + F_A) \times (\frac{1}{2}) \times (\frac{1}{2}) = (\frac{1}{2})^5(1 + F_A)$. Make special note that the exponent on the $(\frac{1}{2})$ is simply the total number of ancestors in the path. In general, if a path through a common ancestor A contains *i* individuals, the probability of autozygosity due to that path is

$$(\tfrac{1}{2})^i(1 + F_A)$$

Thus, the inbreeding coefficient of I in Figure 6.6A is $(\frac{1}{2})^5(1 + F_A)$. Assuming that A itself is not inbred ($F_A = 0$), the inbreeding coefficient of I reduces to $F_I = (\frac{1}{2})^5 = \frac{1}{32}$.

In pedigrees of greater complexity, there is more than one common ancestor and there may be more than one path through any of the common ancestors. The paths are mutually exclusive because autozygosity due to an allele inherited along one path excludes autozygosity due to an allele inherited along a different path. Thus, the total inbreeding coefficient is the sum of the probabilities of autozygosity due to each path considered separately. The whole procedure for calculating *F* is summarized in an example of a first-cousin mating in Figure 6.8. In a first-cousin mating, there are two common ancestors (A and B) and two paths (one each through A and B). The total inbreeding coefficient of I is the sum of the two separate contributions shown in Figure 6.8. If A and B are both noninbred, then $F_A = F_B = 0$, and so $F_I = (\frac{1}{2})^5 + (\frac{1}{2})^5 = \frac{1}{16}$. The result $F_I = \frac{1}{16}$ is the probability that I is autozygous at the specified locus; alternatively, F_I can be interpreted as the average proportion of all genes in I in which the alleles present are autozygous.

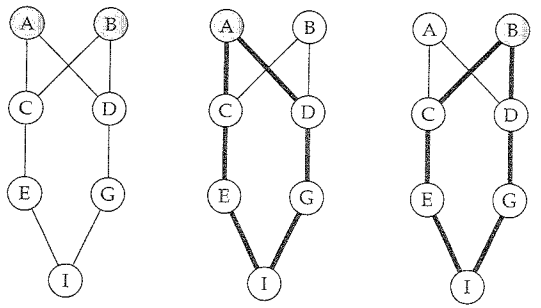


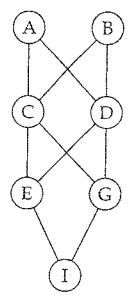
FIGURE 6.8 On the left is a pedigree of individual I, the offspring of a first-cousin mating. On the right are the two paths through common ancestors (heavy lines) used in calculating the inbreeding coefficient of I. Below each path is the contribution to F_I due to that path, calculated as in Figure 6.7. Each path is mutually exclusive of the others, and so their probabilities add. Thus, the total inbreeding coefficient of I is the sum of the two separate contributions. If $F_A = F_B = 0$, then $F_I = \frac{1}{16}$.

In general, for any autosomal gene, the formula for calculating the inbreeding coefficient F_I of an inbred organism I is

$$F_I = \sum_A \left(\frac{1}{2}\right)^i (1 + F_A) \qquad (6.7)$$

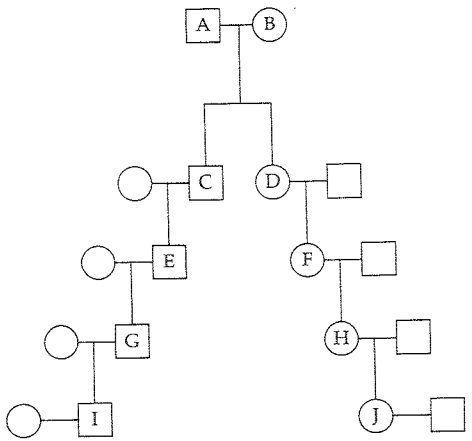
in which the summation over *A* means summation over all possible paths through all common ancestors, *i* is the number of organisms in each path, and A is the common ancestor in each path. Figure 6.9 gives the inbreeding coefficient of an offspring produced by mating between any of several common types of relatives in human pedigrees.

PROBLEM 6.2 The accompanying pedigree depicts two generations of brother-sister mating. Calculate the inbreeding coefficient of I,

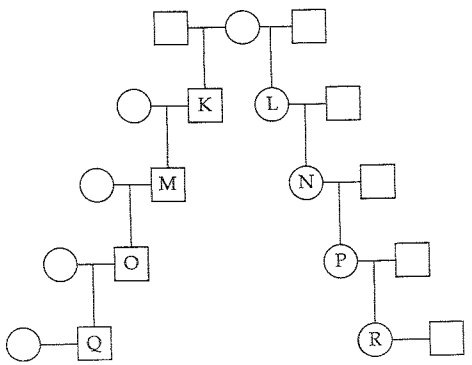


assuming that none of the common ancestors is inbred. (Altogether, there are four common ancestors and six paths.)

ANSWER $F_I = (\frac{1}{2})^3(1 + F_C) + (\frac{1}{2})^3(1 + F_D) + (\frac{1}{2})^5(1 + F_A) + (\frac{1}{2})^5(1 + F_B) + (\frac{1}{2})^5(1 + F_A) + (\frac{1}{2})^5(1 + F_B)$. When the common ancestors are assumed to be noninbred, then $F_A = F_B = F_C = F_D = 0$, and so $F_I = \frac{3}{8}$.



Individuals	Relationship	F of hypothetical offspring
AB	Unrelated	0
AD	Parent-offspring	$\frac{1}{4}$
CD	Full siblings	$\frac{1}{4}$
CF	Uncle-niece	$\frac{1}{8}$
EF	First cousins	$\frac{1}{16}$
EH	First cousins once removed	$\frac{1}{32}$
GH	Second cousins	$\frac{1}{64}$
GJ	Second cousins once removed	$\frac{1}{128}$
IJ	Third cousins	$\frac{1}{256}$



KL	Half siblings	$\frac{1}{8}$
MN	Half first cousins	$\frac{1}{32}$
MP	Half first cousins once removed	$\frac{1}{64}$
OP	Half second cousins	$\frac{1}{128}$
OR	Half second cousins once removed	$\frac{1}{256}$
QR	Half third cousins	$\frac{1}{512}$

FIGURE 6.9 Inbreeding coefficient of the offspring of various types of consanguineous mating.

Regular Systems of Mating

In plant and animal breeding, it is often important to know how rapidly the inbreeding coefficient increases when a strain is propagated by a **regular system of mating**, a systematic and repeated pattern of inbreeding, such as self-fertilization, sib mating, or backcrossing to a standard strain. The reasoning involved in calculating the inbreeding coefficient for any generation is illustrated in Figure 6.10 for repeated self-fertilization. In this figure, the labels $t-1$ and t refer to the inbred organisms after $t-1$ and t generations of self-fertilization. The loop around the ancestor in generation $t-1$ designates the probability that the two indicated alleles are identical by descent. Here the

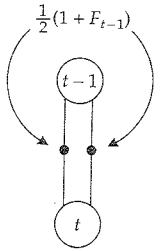


FIGURE 6.10 Increase in F resulting from continued self-fertilization. The organism in generation t is the offspring of self-fertilization of the organism in generation $t-1$. The loop shows that $F_t = (\frac{1}{2})(1+F_{t-1})$.

formula in Equation 6.7 applies with only one path and only one ancestor in the path, and so $F_t = (\frac{1}{2})^1(1+F_{t-1})$, where F_t is the inbreeding coefficient in generation t. This equation is easy to solve in terms of the quantity $1-F_t$, which is often called the *panmictic index* (*panmixia* is an old-fashioned word for *random mating*). Multiplying both sides of the equation for F_t by -1 and then adding $+1$ to each side leads to $1-F_t = 1-(\frac{1}{2})(1+F_{t-1}) = 1-(\frac{1}{2})-(\frac{1}{2})F_{t-1} = (\frac{1}{2})(1-F_{t-1})$, or

$$1-F_t = (\tfrac{1}{2})^t(1-F_0) \qquad [6.8]$$

where F_0 is the inbreeding coefficient in the initial generation when the repeated self-fertilization begins. Self-fertilization therefore leads to an extremely rapid increase in the inbreeding coefficient. When $F_0 = 0$, then $F_1 = \frac{1}{2}$, $F_2 = \frac{3}{4}$, $F_3 = \frac{7}{8}$, $F_4 = \frac{15}{16}$, and so on. The increase in F under self-fertilization and several other regular systems of mating is shown in Figure 6.11. No mat-

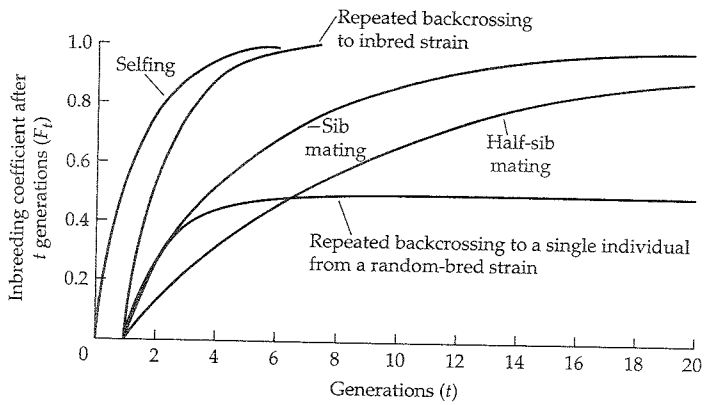


FIGURE 6.11 Theoretical increase in the inbreeding coefficient F for regular systems of mating: selfing, sib mating, half-sib mating, and repeated backcrossing to a single organism from a random-bred strain. In each case, the initial value of F is assumed to be $F_0 = 0$.

ter how much inbreeding has taken place in a population, a single generation of random mating completely erases the effects, and the genotype frequencies return to the Hardy-Weinberg proportions.

Many plants reproduce predominantly by self-fertilization, including crop plants such as soybeans, sorghum, barley, and wheat. As expected of highly self-fertilizing species, each plant is highly homozygous for alleles. Yet in comparing different populations, the proportion of polymorphic genes is comparable to that found in outcrossing species. Polymorphisms are found because self-fertilization does not eliminate genetic variation; it simply reorganizes genetic variation into homozygous genotypes. On the other hand, self-fertilizing species do contain fewer deleterious recessives than do outcrossing species, presumably because the increased frequency of homozygous recessive genotypes permits deleterious mutations to be eliminated from the population by natural selection. The high frequency of homozygous genotypes in naturally self-fertilizing species also impedes recombination producing new gametic types not already present in the parent. Therefore, a predominance of self-fertilization has the effect of slowing the approach to linkage equilibrium because the approach to linkage equilibrium is through recombination in double heterozygotes (*A B/a b* and *A b/a B* in the case of two alleles at each locus); with extreme inbreeding, such doubly heterozygous genotypes are rare. Indeed, the most extreme examples of linkage disequilibrium have been found in predominantly self-fertilizing species such as barley (*Hordeum vulgare*) and wild oats (*Avena barbata*).

Barley, which regularly undergoes more than 99% self-fertilization, provides an extreme example of linkage disequilibrium between two unlinked esterase genes (Clegg et al. 1972). A population that had originated as a complex cross was maintained for 26 generations under normal agricultural conditions without conscious selection. The population was polymorphic for two alleles of an *Esterase-B* gene, which we will designate as alleles *A* and *a*, and also polymorphic for two alleles of an *Esterase-D* gene, which we will designate as alleles *B* and *b*. The gametic types were found in the following proportions. For all practical purposes, these numbers also refer to homozygous genotypes because there is such close inbreeding.

AB	1501	(1642.6)
Ab	754	(613.7)
aB	720	(577.1)
ab	74	(215.6)

(The numbers in parentheses are the expected numbers based on the assumption of linkage equilibrium, calculated as in Chapter 2.) The χ^2 value in this case is 172.7 with one degree of freedom. The associated probability is much less than 0.0001, and so there is undoubtedly linkage disequilibrium. For the above data, the linkage disequilibrium parameter (see Equation 2.13) is $D = -0.046$, which is about 66% of its theoretical minimum. On the other hand, in spite of the small amount of outcrossing in natural populations of barley, the DNA sequences of most genes show evidence for recombination (Morrell et al. 2003).

One of the dramatic successes of plant breeding has come from the crossing of inbred lines to produce high-yielding hybrid corn. Yield of a genetically heterogeneous, outcrossing variety of corn can be improved by selecting the plants with the highest yields in each generation to be the progenitors of the next generation; such artificial selection results in only gradual improvement, however (see Chapter 9). If a large number of self-fertilized lines are established from a heterogeneous population, each line declines in yield as inbreeding proceeds, because of the forced homozygosity of deleterious recessives. Many lines become so inferior that they have to be discontinued. Self-fertilized lines are not likely to become homozygous for exactly the same set of deleterious recessives, however, and when different lines are crossed to produce a hybrid, the hybrid becomes heterozygous for these genes. Alleles favoring high yield in corn are generally dominant, and there may also be genes in which the heterozygous genotypes have a more favorable effect on yield than do the homozygous genotypes; in any case, the hybrid has a much higher yield than either inbred parent. The phenomenon of enhanced hybrid performance is called **hybrid vigor** or **heterosis**. In practice, inbred lines are crossed in many combinations to identify those that produce the best hybrids. Yields of hybrid corn are typically 15–35% greater than yields of outcrossing varieties, and the successful introduction of hybrid corn has been remarkable. Virtually all corn acreage in the United States today is planted with hybrids, as compared to 0.4% of the acreage in 1933 (Sprague 1978).

6.2 POPULATION SUBDIVISION

Most populations are grouped into smaller subpopulations within which mating usually takes place. Such grouping is called **population structure** or **population subdivision**, and it is almost universal among organisms. Many organisms naturally form subpopulations in the form of herds, flocks, schools, colonies, or other types of aggregations. When there is population subdivision, there is almost inevitably some genetic differentiation among the subpopulations. By **genetic differentiation** we mean that the allele frequencies among the subpopulations become different. Genetic differentiation may result from natural selection favoring different genotypes in different subpopulations, but it may also result from random processes in the transmission of alleles from one generation to the next or from chance differences in allele frequency among the initial founders of the subpopulations. The effects of random genetic drift in increasing the variance in allele frequency among subpopulations have already been examined in Chapter 3.

When the subpopulations are completely isolated from migration, then all matings must take place between individuals within each subpopulation. The intra-population mating implies that the individuals within each subpopulation will share some common ancestors, and hence even matings that take place "at random" in the subpopulation are matings that unite individuals who have common ancestors. These common ancestors transmit alleles that are identical by descent that can come together in the progeny of the mating, and a nonzero probability of identity by descent constitutes inbreeding. In other words, population subdivision, in and of itself, results in inbreeding because the individuals in the subpopulation share remote ancestors, even in situations in which the members of each subpopulation choose their mates at random. The relationship between population structure and inbreeding is subtle, but it has profound consequences in population genetics.

Many populations have a **hierarchical population structure**, which means that the subpopulations can be grouped into progressively inclusive levels in which, at each grouping, the next lower levels are included ("nested") within the next higher ones. To consider a concrete example, imagine we were interested in the population structure of a widespread species of freshwater fish. The lowest population level consists of a local interbreeding population of animals within a stream. A stream might contain more than one such local population. The next-higher level in the hierarchy could be the organization of streams into groups feeding the same river. Another higher level could be rivers within watersheds. An even higher level of organization might be watersheds within continents. The aggregation of subpopulations into progressively more inclusive groups can continue for as many levels as is convenient and informative. It is inevitably somewhat arbitrary how the groups at each level are combined to form the next higher level in the hierarchy. The choice of classification is pragmatic: One tries to group the subpopulations in such a way as to highlight the genetic similarities and differences among them. If there were so much migration of fish among subpopulations that all members of the species constituted essentially a single, random-mating population, then there would be no need to define a hierarchical population structure because it would be uninformative. However, most organisms do have significant population structure.

Reduction in Heterozygosity Due to Population Subdivision

One of the important consequences of population structure is a reduction in the average proportion of heterozygous genotypes relative to that expected under random mating. The reason for the reduction in heterozygosity may be understood by considering the somewhat whimsical example in Figure 6.12. The outline is the floor plan of a large barn. The organisms of interest are the mice concentrated primarily into two subpopulations of equal size at the west and east ends of the barn. The movement of mice between the subpopulations is prevented by a large population of hungry and vigilant cats in the central area. The occasional mouse that comes out of its refuge is quickly eaten. (These hypothetical mice have not been endowed with the ingenuity to find alternative routes between the west and east ends of the barn, like sneaking along the rafters.) Because of chance effects in the founding of the subpopulations, the west and east subpopulations are completely homozygous for alternative alleles of a gene. All the mice in the west subpopulation are *AA*, and all those in the east subpopulation are *aa*. In technical terms, the west subpopulation is fixed for the *A* allele (its allele frequency equals 1), and the east subpopulation is fixed for the *a* allele. The genotype frequencies of *AA*, *Aa*, and *aa* in the west subpopulation are 1, 0, and 0, respectively, and those in the east subpopulation are 0, 0, and 1, respectively. Within each subpopulation there is random mating, and the genotype frequencies, though extreme, still satisfy the Hardy-Weinberg principle. In particular, the frequencies of *AA*, *Aa*, and *aa* within each subpopulation are given by p^2, $2pq$, and q^2, where $p = 0$ in the east subpopulation, and $p = 1$ in the west subpopulation. Therefore, within any one of the subpopulations in Figure 6.12, the frequency of heterozygotes equals the frequency expected with HWE.

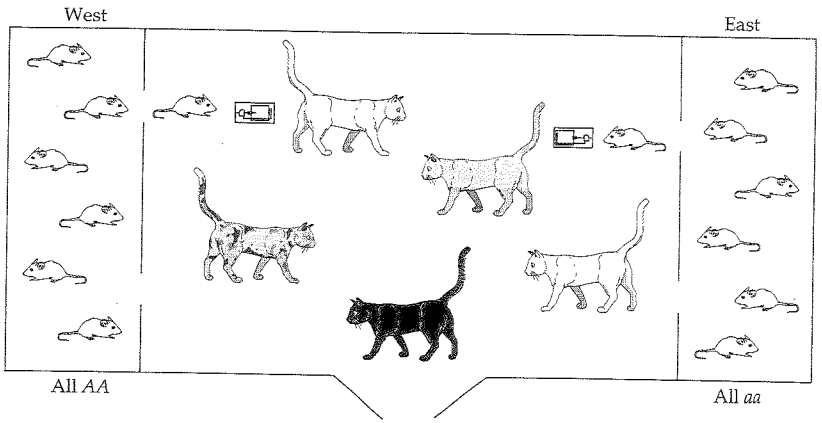


FIGURE 6.12 An extreme example of the general principle that a difference in allele frequency among subpopulations results in a deficiency of heterozygotes. The floor plan is that of a hypothetical barn. The mouse subpopulations in the east and west enclaves are completely isolated because of the cats in the middle. The west subpopulation is fixed for the *A* allele and the east subpopulation for the *a* allele. Trapping mice at random in the area patrolled by the cats would yield an overall allele frequency of $\frac{1}{2}$, but no heterozygous genotypes.

The situation regarding the total mouse population in Figure 6.12 is very different, however, as there is an overall deficiency of heterozygotes. By "total population" in this context, we mean the aggregate of all mice without regard to the population subdivision. Suppose we were unaware of the population structure in the barn. We might then suppose that the barn contained a single randomly mating population. To study the total population of the barn, we trap mice at random in the center area, catching the occasional escapee from the cats. Because the subpopulations are fixed for either A or a, half the time we would trap an AA homozygote and half the time an aa homozygote. Consequently, we estimate the allele frequency of A as $\hat{p} = \frac{1}{2}$. Assuming random mating and Hardy-Weinberg genotype frequencies in the total population, the expected genotype frequencies of AA, Aa, and aa are given by the HWE as $\hat{p}^2$, $2\hat{p}\hat{q}$, and $\hat{q}^2$. Because the overall allele frequency of A among the trapped animals is $\frac{1}{2}$, we would naively expect a fraction $2 \times (\frac{1}{2}) \times (\frac{1}{2}) = \frac{1}{2}$ 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 takes place within each subpopulation—is a consequence of the difference in allele frequency among the subpopulations. Were the allele frequencies in both subpopulations the same, it would not matter whether we sampled from the west subpopulation, the east subpopulation, or from the area in between. We would recover genotypes in Hardy-Weinberg proportions because both subpopulations are genotypically identical and in HWE. In an organism with hierarchically structured subpopulations, there is an analogous deficiency of heterozygotes at each level in the hierarchy. The following section examines the heterozygosities in more detail.

Average Heterozygosity

In the Mojave desert, local populations of the annual plant *Linanthus parryae* are polymorphic for white versus blue flowers. The plant is diminutive, averaging just 1 cm in height, and when the plant is in bloom, the ground cover of white flowers justifies the popular name "desert snow." Blue flowers result from homozygosity for a recessive allele. The geographical distribution of the frequency q of the recessive allele across a region of the Mojave desert is illustrated in Figure 6.13. Each allele frequency is based on an examination of approximately 4000 plants over an area of about 30 square miles (Epling and Dobzhansky 1942).

Judging from the allele-frequency map in Figure 6.13, the highest frequencies of the blue-flower allele are largely concentrated at the west and east ends of the region in question. The unequal allele frequencies across the range imply a decrease in average heterozygosity relative to HWE, analogous to the mouse example in Figure 6.12, though not as extreme. Figure 6.13 shows the estimated allele frequency in each of 30 subpopulations. Suppose

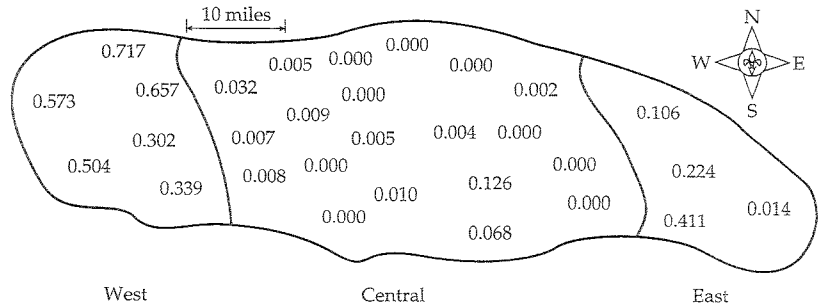


FIGURE 6.13 Estimated frequency of a recessive allele for blue flower color in populations of *Linanthus parryae* in an area of approximately 900 square miles in the Mohave desert. Each allele frequency is based on an examination of approximately 4000 plants over an area of about 30 square miles. (After Wright 1943a.)

each of the subpopulations is regarded as a random-mating unit in HWE for the flower-color alleles. The average heterozygosity among the subpopulations can be denoted as H_S, where the subscript indicates subpopulation. The calculations are shown in the third column in Table 6.3; the heterozygosity in each subpopulation is calculated as $2pq$, where p and q are the estimated frequencies of the alleles for white versus blue flower color, respectively, in each subpopulation. The H_S tabulated at the bottom is the average of all the subpopulation heterozygosities (counting the value 0.000 a total of nine times because of the nine different subpopulations in which $q = 0.000$).

A second hierarchical level of population structure is that of region—west (W), central (C), or east (E). To calculate the heterozygosity expected from HWE in each region, we first estimate the average allele frequency in the region by taking the mean allele frequency across all subpopulations in the region. For example, the average allele frequency q in region E is $(0.106 + 0.224 + 0.411 + 0.014)/4 = 0.1888$. In each region, the heterozygosity expected from HWE is calculated as $2pq$, where p and q are the average allele frequencies in the region. In region E, therefore, the regional heterozygosity equals $2 \times (1 - 0.1888) \times 0.1888 = 0.3062$. The average heterozygosity within regions at the bottom of column 5 is denoted H_R; it is the weighted average of the regional heterozygosities, where each regional heterozygosity is weighted by the number of subpopulations in the region. In this example, $H_R = (6 \times 0.4995 + 20 \times 0.0272 + 4 \times 0.3062)/30 = 0.1589$.

Yet another hierarchical level of population structure in Figure 6.13 is the total population—the aggregate population obtained by conceptually uniting all subpopulations to form a single random mating unit. The average

TABLE 6.3 Hierarchical Structure of *Linanthus parryae*

	Subpopulations		Regions		Total	
Region	Allele frequency	Heterozygosity	Average allele frequency	Heterozygosity	Average allele frequency	Heterozygosity
W	0.573	0.4893				
	0.717	0.4058				
	0.504	0.5000				
	0.657	0.4507				
	0.302	0.4216				
	0.339	0.4482	0.5153	0.4995		
C	9×0.000	0.0000				
	0.032	0.0620				
	0.007	0.0139				
	0.008	0.0159				
	0.005	0.0100				
	0.009	0.0178				
	0.005	0.0100				
	0.010	0.0198				
	0.068	0.1268				
	0.002	0.0040				
	0.004	0.0080				
	0.126	0.2202	0.0138	0.0272		
E	0.106	0.1895				
	0.224	0.3476				
	0.411	0.4842				
	0.014	0.0276	0.1888	0.3062	0.1374	0.2371
Average heterozygosity		$H_S = 0.1424$		$H_R = 0.1589$		$H_T = 0.2371$

Source: Data from Wright 1943a.

allele frequency is the mean allele frequency across all subpopulations, and $q = 0.1374$. Then H_T is calculated as $2pq = 2 \times 0.8626 \times 0.1374 = 0.2371$.

To sum up:

- H_S is the average heterozygosity assuming HWE among organisms within random-mating subpopulations.
- H_R is the average heterozygosity assuming HWE among organisms within regions.
- H_T is the average heterozygosity assuming HWE among organisms within the total area.

The concepts of hierarchical population structure and the various levels of heterozygosity were originally developed by Wright (1943a,b), in his theory of **isolation by distance**, to quantify genetic differences among subgroups at the various levels. The motivation for developing such a method was summarized in the following passage from Wright (1943b):

> Study of statistical differences among local populations is an important line of attack on the evolutionary problem. While such differences can only rarely represent first steps toward speciation in the sense of the splitting of the species, they are important for the evolution of the species as a whole. They provide a possible basis for intergroup selection of genetic systems, a process that provides a more effective mechanism for adaptive advance of the species as a whole than does the mass selection which is all that can occur under panmixia.

Furthermore, the reduction in heterozygosity resulting from population subdivision is intimately related to the reduction in heterozygosity caused by inbreeding due to mating between relatives. As explained earlier, the relation of population structure to inbreeding can be understood by interpreting each subpopulation as a sort of "extended family" or set of interconnected pedigrees. Organisms in the same subpopulation will often share one or more recent or remote common ancestors, and so a mating between organisms in the same subpopulation may result in offspring whose alleles at a locus are identical by descent (autozygous). The larger the subpopulation and the more recently it has been isolated, the smaller the probability of autozygosity, but in any finite subpopulation the probability of autozygosity increases through time.

Wright's F Statistics

To quantify the inbreeding effect of population subdivision, Wright (1921) defined what has come to be called the **fixation index**. This index equals the reduction in heterozygosity expected with random mating at any one level of a population hierarchy relative to another, more inclusive level of the hierarchy. The fixation index is a useful index of genetic differentiation because it allows an objective comparison of the overall effect of population structure among different organisms without getting into details of allele frequencies, observed levels of heterozygosity, and so forth. The genetic symbol for a fixation index is F embellished with subscripts denoting the levels of the hierarchy being compared. For example, F_{SR} is the fixation index of the subpopulations relative to the regional aggregates:

$$F_{SR} = \frac{H_R - H_S}{H_R} \tag{6.9}$$

In words, Equation 6.9 defines F_{SR} as the decrease of heterozygosity among subpopulations within regions ($H_R - H_S$), relative to the heterozygos-

ity among regions (H_R). For the *Linanthus* example in Table 6.3, $F_{SR} = (0.1589 - 0.1424)/0.1589 = 0.1036$.

At the next level of the hierarchy, we may define the fixation index F_{RT} as the proportionate reduction in heterozygosity of the regional aggregates relative to the total combined population:

$$F_{RT} = \frac{H_T - H_R}{H_T} \qquad (6.10)$$

The data in Table 6.3 indicate that $F_{RT} = (0.2371 - 0.1589)/0.2371 = 0.3299$. Comparison of this value with F_{SR} above already makes it clear that there is substantially more variation among regions (as measured by F_{RT}) than there is among subpopulations within regions (as measured by F_{SR}). The comparison of the fixation indices at the two levels gives quantitative expression to the regional differences apparent in Figure 6.13.

The fixation index F_{ST} compares the least inclusive to the most inclusive levels of the population hierarchy and measures all effects of population structure combined:

$$F_{ST} = \frac{H_T - H_S}{H_T} \qquad (6.11)$$

From Table 6.3, $F_{ST} = (0.2371 - 0.1424)/0.2371 = 0.3993$. The overall reduction in average heterozygosity is therefore close to 40% of the total heterozygosity—a very substantial effect.

The **hierarchical *F*-statistics** defined in Equations 6.9 through 6.11 are all types of fixation indices, but they differ in the reference populations: F_{SR} is concerned with subpopulations (S) relative to the regional aggregates (R), F_{RT} is concerned with the regional groupings relative to the total population (T), and F_{ST} is concerned with the subpopulations relative to the total population. The index F_{ST} is the most inclusive measure of population subdivision.

The mathematical relation between the three types of F statistics is demonstrated in the following problem.

PROBLEM 6.3 Show that F_{SR}, F_{RT}, and F_{ST} are related by the equation

$$1 - F_{ST} = (1 - F_{SR})(1 - F_{RT})$$

ANSWER From Equation 6.9, $F_{SR} = 1 - (H_S/H_R)$, or $1 - F_{SR} = H_S/H_R$. Equation 6.10 implies that $F_{RT} = 1 - (H_R/H_T)$, or $1 - F_{RT} = H_R/H_T$. Finally, Equation 6.11 implies that $F_{ST} = 1 - (H_S/H_T)$, or $1 - F_{ST} = H_S/H_T$. Now multiply the expressions for $1 - F_{SR}$ and $1 - F_{RT}$ together to obtain $(1 - F_{SR}) \times (1 - F_{RT}) = (H_S/H_R) \times (H_R/H_T) = H_S/H_T = (1 - F_{ST})$.

For examining the overall level of genetic divergence among subpopulations, F_{ST} is the informative statistic, and the concept has been extended to multiple alleles (Nei 1973). 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 different subpopulations), the observed maximum is usually much less than 1. Wright (1978) has suggested the following qualitative guidelines for the interpretation of F_{ST}:

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

On the other hand, Wright also notes that, among subpopulations, "differentiation is by no means negligible if F_{ST} is as small as 0.05 or even less." Difficulties in interpreting F_{ST} are alleviated somewhat by the use of a standardized version in which F_{ST} is expressed as the proportion of the maximum differentiation possible for the observed level of subpopulation homozygosity (Hedrick 2005).

PROBLEM 6.4 One of the limitations of F_{ST} is that it is does not capture the full range of possibilities that can be found in natural populations. To see this for yourself, consider two subpopulations with two alleles each, A_1 and A_2; in one subpopulation the allele frequencies are $(3 + \sqrt{3})/6 = 0.788675$ and $(3 - \sqrt{3})/6 = 0.211325$, and in the other subpopulation the allele frequencies are reversed. (The choice of these allele frequencies may seem strange, but the rationale for the choice will become clear when you work the problem.) Now consider the same gene in two different subpopulations; one of these subpopulations has alleles A_1 and A_2 at frequencies $\frac{1}{2}$ and $\frac{1}{2}$, and the other has alleles A_3 and A_4 at frequencies $\frac{1}{2}$ and $\frac{1}{2}$. Use Equation 6.11 to calculate F_{ST} for both pairs of subpopulations, and explain why the result seems paradoxical.

ANSWER In the first case, the heterozygosity in each subpopulation is $2 \times (3 + \sqrt{3})/6 \times (3 - \sqrt{3})/6 = \frac{1}{3}$, and hence the average subpopulation heterozygosity is $H_S = \frac{1}{3}$. The average allele frequency for each allele is $\frac{1}{2}$, and hence the total heterozygosity is $H_T = \frac{1}{2}$. In this case, $F_{ST} = [(\frac{1}{2}) - (\frac{1}{3})]/(\frac{1}{2}) = \frac{1}{3}$. In the second case, the heterozygosity in each subpopulation is $2 \times (\frac{1}{2}) \times (\frac{1}{2}) = \frac{1}{2}$, and so $H_S = \frac{1}{2}$. The average allele frequencies are $\frac{1}{4}$ for each of the four alleles, and so $H_T = 1 - (\frac{1}{4})^2 = \frac{3}{4}$. In this case, $F_{ST} = [(\frac{3}{4}) - (\frac{1}{2})]/(\frac{3}{4}) = \frac{1}{3}$, exactly the same as before. The paradox is that the subpopulations have the same value of F_{ST}, when the first two subpopulations differ only in allele frequencies, whereas the second two are so different that they have no alleles in common.