æ

2

NEUTRAL EVOLUTION IN ONE- AND TWO-LOCUS SYSTEMS

19 May 2014

"Variations neither useful nor injurious would not be affected by natural selection, and would be left either a fluctuating element, as perhaps we see in certain polymorphic species, or would ultimately become fixed, owing to the nature of the organism and the nature of the conditions." Darwin (Chapter 4, 1859)

Although the majority of research in evolutionary biology is focused on issues related to natural selection, the rigor of all such analyses depends critically on our understanding of expected patterns of evolution in the absence of selection. The simple reasoning here is that if we are to have much confidence in any adaptive argument, it ought to be possible to firmly reject a simpler neutral hypothesis. Thus, prior to exploring various population-genetic models for adaptive evolution, we first embark on a broad overview of neutral models of evolution. The theory underlying such models brings us into immediate contact with the issue of **genetic drift**, a commonly misunderstood factor in evolutionary biology, but nothing more than the random fluctuations in allele frequencies that necessarily result from the sampling of finite numbers of gametes each generation. The magnitude of such fluctuations increases with decreasing population size, and combined with the input of new alleles by mutation, the incessant stochasticity of the process ensures that populations will evolve even in the absence of selection (Kimura 1983).

Consider, for example, a single heterozygous Bb parent that produces two progeny. There is a 50% probability that one offspring will inherit the B allele and the other the b allele, in which case no net change in allele frequency has been transmitted from parent to offspring. However, there is also a 50% probability that both offspring will inherit the same allele. In extremely large populations, these random changes resulting from gamete sampling tend to average out, leaving the allele frequency in the offspring population very close to that in the parental generation. However, over sufficiently long time scales, the cumulative effects of even small single-generation changes in allele frequencies can become quite pronounced. As will be seen below, if the time scale of interest (t, in generations) is much less than the

average number of reproductive adults in the population (N), random fluctuations in allele frequencies can usually be ignored. This justifies the assumption of an effectively infinite population size as a good first approximation in many applications of population and quantitative genetics. However, for situations in which t is on the order of N or larger, evolution can certainly no longer be viewed as a strictly deterministic process. Rather, any observed evolutionary change must be viewed as one realization of many possible outcomes.

In the following pages, it will be shown that even though finite population size induces stochastic evolutionary change, genetic drift has several *predictable effects.* First, even if mating is completely random, there will still be some long-term trend toward matings among relatives. Because all members of a population must ultimately descend from a narrow ancestral base, the smaller the population size, the greater this tendency will be. Thus, in a tiny **dioecious** (separate-sex) population with a stable adult number of two, all matings must be between full sibs, even though the reproductive pair itself may be a random draw from a larger progeny pool in the preceding generation. It follows that the genetic consequences of finite population size must be similar to those of inbreeding – the average homozygosity at a locus is expected to increase with smaller N, although as noted below, this can be offset in part by replenishment from mutationally derived variation. Second, gamete sampling causes allele frequencies to gradually drift toward zero or one, with the probability of ultimate fixation of any particular allele in the absence of selection being equal to its initial frequency. Third, subdivision of a population into isolated demes results in allele-frequency divergence among demes. The greater the degree of isolation of the subgroups, the more pronounced this differentiation will be.

This and the following two chapters provide a formal basis for these ideas. We first consider matters of one- and two-locus evolution in the context of a population with an idealized mating system and no influence from selective forces. As will be seen in subsequent chapters, such models are often of great utility even when selection is operating, provided the forces of selection are weaker than those associated with random genetic drift. In addition, such theory provides the underlying logic for the development of molecular-marker methods for estimating the power of mutation, recombination, and random genetic drift. These methods, the subject of Chapter 4, are the primary ways we have of describing the past population-genetic environment. Chapter 3 provides a critical link between Chapters 2 and 4 by demonstrating how results derived under the assumption of an ideal random mating population can be extended to a variety of alternative reproductive systems and population structures. In subsequent chapters, the one- and two-locus results introduced here will be used to develop neutral models for the evolution of quantitative traits (Chapter 12) and formal tests of adaptive evolution given a sample of sequences from a genomic region of interest (Chapters 8 and 9).

THE WRIGHT-FISHER MODEL

Because the number of possible types of population structure is literally infinite (involving, for example, various degrees of local inbreeding, geographic subdivision,

and age-specific mortality and fecundity), and temporal variation in population size is also common, it is impossible for us to consider the dynamics of neutral alleles in a fully general sense. Instead, we will focus initially on single finite populations of constant size within which mating is random. Even this simple structure admits to many possible variants, depending for example, on whether there are separate sexes, whether there is variation in family size, and whether generations are overlapping. All of these additional layers of complexity will be taken up in Chapter 3, where it will become clear that most of the results in the current chapter often still hold with an appropriate redefinition of the concept of population size.

Most population geneticists adhere to the **Wright-Fisher model** of drift, whose roots trace to Fisher (1922) and Wright (1931). We assume here a diploid population with a fixed number (N) of **monoecious** (hermaphroditic) adults, random mating (including the possibility of self-fertilization), and discrete generations. The gamete pool produced by the adults is assumed to be effectively infinite, such that the $2N$ gametes that actually contribute to the next generation can be viewed as being sampled with replacement.

Consider a locus with two alleles, B and b, with neither having a selective advantage with respect to the other. If there are i copies of allele B in generation t, the probability that the number in generation $t+1$ is equal to j follows the binomial distribution, with each of the $2N$ sampled gametes having probability $i/(2N)$ of being B,

$$P_{ij} = \binom{2N}{j}(i/2N)^j[1-(i/2N)]^{2N-j} \tag{2.1}$$

This expression holds for all possible values of $i, j = 0, 1, \ldots, 2N$. Letting $\mathbf{P}$ be the $(2N+1)\times(2N+1)$ matrix of these coefficients, the entire probability distribution of the frequency of allele B can then be expressed succinctly as

$$\mathbf{x}(t+1) = \mathbf{P}\mathbf{x}(t) \tag{2.2a}$$

where the elements of the column vector $\mathbf{x}(t)$ are the probabilities that the allele is present in $i = 0, 1, \ldots, 2N$ copies in generation t. If the transition matrix $\mathbf{P}$ remains constant from generation to generation, as it does under the assumptions given above, Equation 2.2a generalizes to

$$\mathbf{x}(t) = \mathbf{P}^t\mathbf{x}(0) \tag{2.2b}$$

which is an example of a **Markov chain**, considered in more detail in Appendix 3.

When considering a single population starting with allele frequency $i/2N$, all of the entries in the initial vector $\mathbf{x}(0)$ are equal to zero, except $x_i(0)$ (corresponding to i copies being present), which equals 1.0. Equation 2.2b then yields the evolution of the probability distribution of the allele frequency over time. That is, the elements of $\mathbf{x}(t)$ denote the frequencies of hypothetical replicate populations at time t that are expected to have allele frequency $i/2N$. The first ($i = 0$) and final ($i = 2N$) elements of $\mathbf{x}(t)$ are of special interest, as they are **absorbing states** – once an allele becomes **lost** (frequency = 0.0) or **fixed** (frequency = 1.0) in a population, it remains at that state indefinitely (barring reintroduction via mutation or migration). As Equation 2.2b is iterated, all of the interior elements of $\mathbf{x}(t)$ eventually converge on zero, and the sum $x_0(t) + x_{2N}(t)$ converges to one. The ultimate probability of fixation of

allele B is given by $x_{2N}(\infty)$, whereas the ultimate probability of loss of allele B (or equivalently, of fixation of allele b) is given by $x_0(\infty)$.

From the elements of $\mathbf{x}(t)$, it is straightforward to compute the expected allele frequency, the variance of allele frequencies among replicate populations, the probabilities of fixation by generation t, etc. However, although this **transition-matrix** approach is exact, it becomes unwieldy with large N, and many useful approximations have been developed (e.g., Gale 1990; Ewens 2004). Some of these approaches will be discussed below, with a most powerful alternative method, the **diffusion approximation**, being covered extensively in Appendix 1.

It should also be noted that the Wright-Fisher model is just one of many possible conceptual frameworks for approximating a randomly mating population. For example, Moran (1962) developed a treatment whereby a single random individual is chosen to reproduce at each point in time, with a single random individual then being chosen to die. Because allele frequencies can change by only a single step during each time interval under this scenario, the **Moran model** turns out to be more analytically tractable than the Wright-Fisher model, although it is also restricted to haploid populations.

Example 2.1. Consider an initially heterozygous individual Bb in a self-fertilizing line maintained by single-progeny descent. With $N = 1$, the only three possible subsequent allele-frequency states in the population are zero, one, or two B alleles. Denoting the initial state of the population by $\mathbf{x}^T(0) = [0, 1, 0]$, the probability that the population is in states 0, 1, or 2 at some future generation t is given by Equation 2.2b with

$$\mathbf{P} = \begin{pmatrix} 1 & 0.25 & 0 \\ 0 & 0.50 & 0 \\ 0 & 0.25 & 1 \end{pmatrix}$$

Although the numerical values for the elements of $\mathbf{P}$ can be obtained directly from Equation 2.1, for this simple example they can also be arrived at intuitively. For example, the elements in the first column of $\mathbf{P}$ denote the probabilities that the population will be in states $j = 0, 1, 2$ in generation $t+1$ given that it is in state 0 in generation t. The only nonzero element in this column is $P_{00} = 1$ because the 0th state is absorbing, i.e., once the population enters this state, it remains there indefinitely.

The probability of being in any particular allele-frequency category in generation t, which follows from Equation 2.2b, is a function of $\mathbf{P}^t$, so for example,

$$\mathbf{P}^2 = \begin{pmatrix} 1 & 0.0375 & 0 \\ 0 & 0.2500 & 0 \\ 0 & 0.0375 & 1 \end{pmatrix}, \quad \mathbf{P}^5 = \begin{pmatrix} 1 & 0.48438 & 0 \\ 0 & 0.03125 & 0 \\ 0 & 0.48438 & 1 \end{pmatrix}, \quad \mathbf{P}^{10} = \begin{pmatrix} 1 & 0.49951 & 0 \\ 0 & 0.00098 & 0 \\ 0 & 0.49951 & 1 \end{pmatrix}$$

and so on. With the initial vector $\mathbf{x}^T(0) = (0, 1, 0)$, only the middle column of $\mathbf{P}^t$ is relevant, giving the following table for the progression of the elements of $\mathbf{x}(t)$ over

time:

t	BB $x_0(t)$	Bb $x_1(t)$	bb $x_2(t)$
0	0.00000	1.00000	0.00000
1	0.25000	0.50000	0.25000
2	0.37500	0.25000	0.37500
3	0.43750	0.12500	0.43750
4	0.46875	0.06250	0.46875
5	0.48438	0.03125	0.48438
...	...	...	...
10	0.49951	0.00098	0.49951
...	...	...	...
15	0.49998	0.00003	0.49998
...	...	...	...
∞	0.50000	0.00000	0.50000

The first and last elements of $\mathbf{x}(t)$ respectively denote the probabilities that the line will have become fixed for the **B** or the **b** allele by time t. Thus, for this particular case, the line eventually becomes completely monomorphic for either the **B** or the **b** allele with equal probability. This meets our intuitive expectations for a neutral locus – in the absence of any directional forces, the two probabilities of fixation are equal to the initial frequencies of the respective alleles.

LOSS OF HETEROZYGOSITY BY RANDOM GENETIC DRIFT

The **sampling variance of an allele frequency** provides one way to succinctly define the stochastic effects of random genetic drift. Consider a large pool of gametes, a fraction p of which carry the B allele, and let $2N$ gametes be randomly drawn to produce a new generation of N individuals. Defining the expected frequencies of genotypes BB, Bb, and bb in the progeny generation by the Hardy-Weinberg proportions p^2, $2p(1-p)$, and $(1-p)^2$, the expected number of B alleles contained in a random offspring is simply $[2 \cdot p^2] + [1 \cdot 2p(1-p)] + [0 \cdot (1-p)^2] = 2p$. However, the expected square of the number of B alleles carried per individual is not $(2p)^2$ but rather $[2^2 \cdot p^2] + [1^2 \cdot 2p(1-p)] + [0^2 \cdot (1-p)^2] = 2p(1+p)$. Thus, the variance of the number of B alleles carried by an individual is $2p(1+p) - (2p)^2 = 2p(1-p)$, while the variance of the total number of B alleles carried in the offspring generation is N times this, $2Np(1-p)$. Because the frequency of allele B is the number of copies divided by $2N$, the sampling variance of the frequency (a second-order moment) is $2Np(1-p)/(2N)^2 = p(1-p)/(2N)$, which is directly proportional to the heterozygosity and inversely proportional to the population size. The expression $p(1-p)/(2N)$ defines the dispersion in allele frequency resulting from a single generation of gamete sampling, conditional on allele frequency p in the parental population.

In the absence of any counteracting evolutionary forces, the dispersive effects of genetic drift will continue each generation, leading to a progressive erosion of

population-level heterozygosity until all loci have eventually become fixed for just a single allele. To evaluate the long-term impact of finite population size on the expected heterozygosity of a locus, we make use of the properties of the **inbreeding coefficient**, f, which denotes the probability that two alleles at a locus in an individual are identical by descent (LW Chapter 7).

There is always a small chance that uniting gametes will derive from related individuals, even in a randomly mating population. For example, in a monoecious population containing only two individuals, there are only four genes residing at each locus, so the probability that one gamete will randomly unite with another containing a direct descendant of the same parental gene is 1/4. With four individuals, there are eight gene copies, and this probability becomes 1/8. Thus, under the idealized Wright-Fisher model, the probability that two direct copies of any parental gene will randomly unite in an offspring is $1/(2N)$. Barring a rare mutation, all such offspring are homozygotes.

Although the quantity $1/(2N)$ may be thought of as the new inbreeding that is incurred each generation, this does not fully describe the build-up of homozygosity in a population. For even if uniting gametes do not carry genes that are direct copies of a parental gene, they may still be identical by descent through inbreeding in a previous generation. Under random mating, the probability of the latter event is simply the inbreeding coefficient of the parental generation. Thus, because the probability of drawing genes that are not direct copies of the same parental gene is $[1 - (1/2N)]$, the expected inbreeding coefficient in generation t is

$$f_t = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right) f_{t-1} \tag{2.3}$$

Subtracting both sides from one yields the recursion formula

$$(1 - f_t) = \left(1 - \frac{1}{2N}\right)(1 - f_{t-1}) \tag{2.4a}$$

which generalizes to

$$(1 - f_t) = \left(1 - \frac{1}{2N}\right)^t (1 - f_0) \tag{2.4b}$$

and finally to

$$(1 - f_t) = \left(1 - \frac{1}{2N}\right)^t \tag{2.4c}$$

if we assume a noninbred base population ($f_0 = 0$). Again, we see the central role that population size plays in the dynamics of genetic variation. As $t \to \infty$, the fraction of the population that is not inbred, $1 - f_t$, approaches zero at a rate that is inversely proportional to N. This **rate of decay of heterozygosity** of $1/(2N)$ was first obtained by Wright (1931). It may be a source of encouragement to the non-mathematically inclined that the brilliant Fisher (1922), using a rather different approach, obtained the wrong answer.

To see the connection between the inbreeding coefficient and the expected heterozygosity in a population, consider a diallelic locus with base-population heterozygosity $2p(1 - p)$. In the descendant population with inbreeding coefficient f, individuals can only be heterozygotes if they carry alleles that are not identical by

descent, the probability of which is $(1-f)$. If two alleles are not identical by descent, they must have been acquired independently, so the probability that a genotype containing a pair of such alleles is a heterozygote is $2p(1-p)$. Thus, the expected heterozygosity of a population with inbreeding coefficient f and initial allele frequency p is $2p(1-p)(1-f)$. This shows that the fractional reduction in heterozygosity relative to the base population is equal to f. Because this argument applies regardless of the initial heterozygosity (and regardless of the number of segregating alleles), Equation 2.4c may be rewritten to describe the expected population heterozygosity at time t,

$$H_t = H_0 \left(1 - \frac{1}{2N}\right)^t \tag{2.5}$$

The time course for the loss of heterozygosity can be clarified by using an exponential approximation to Equation 2.5. Because $(1-x)^t \simeq e^{-xt}$ for $|x| \ll 1$, for N greater than 10 or so,

$$H_t \simeq H_0 e^{-t/(2N)} \tag{2.6a}$$

Rearrangement then leads to the expected time to reach a certain reduction in heterozygosity,

$$t = -2N \ln(H_t/H_0) \tag{2.6b}$$

which shows that the heterozygosity is reduced to half of its initial value in $\sim 1.4N$ generations and to 5% of H_0 in $\sim 6N$ generations. Thus, a population twice the size of another requires twice the number of generations to reach the same expected state.

With a temporally varying population size, Equation 2.6a becomes

$$H_t = H_0 \prod_{i=1}^{t} \left(1 - \frac{1}{2N_i}\right) \simeq H_0 \exp\left[-\sum_{i=1}^{t} 1/(2N_i)\right] \tag{2.7}$$

where the $\prod$ sign denotes a product of terms. This expression illustrates an important point. Because each of the generation-specific terms, $[1-(1/2N_i)]$, is necessarily less than one, *an expansion of population size can reduce the rate of erosion of heterozygosity, but does not eliminate it.*

One significant limitation of the preceding expressions is that they only provide information on the behavior of the *expected* heterozygosity in a population. In reality, fluctuations in allele frequencies resulting from random genetic drift ensure that variation in heterozygosity will arise among loci that start in the same state. In a finite population of size N, the heterozygosity of a diallelic locus can take on $N+1$ discrete values: $0,\ 2(1/2N)[1-(1/2N)],\ \cdots,2(N/2N)[1-(N/2N)]$. Using the transition-matrix approach (Equations 2.2a,b), one can obtain the exact probability distribution of heterozygosity for a locus starting with allele frequency $i/2N$, using the fact that $x_j(t) + x_{2N-j}(t)$ is the probability that the population has heterozygosity $2(j/2N)[1-(j/2N)]$. However, as noted above, this approach becomes computationally difficult as N becomes large.

An alternative approach utilizes a remarkable achievement by Kimura (l955), who used diffusion theory (Appendix 1) to obtain an analytical expression for the

probability density of allele frequency at time t, given the starting value p_0,

$$\varphi(p_t|p_0) = p_0(1-p_0)\sum_{i=1}^{\infty} i(2i+1)(i+1)\cdot \\ F(1-i, i+2, 2, p_0)\cdot F(1-i, i+2, 2, p_t)\cdot e^{-i(i+1)t/(4N)} \tag{2.8}$$

where $F(1-i, i+2, 2, p_0)$ and $F(1-i, i+2, 2, p_t)$ are specific variants of the hypergeometric function (Equation 15.1.1 in Abramowitz and Stegun 1972). Using this expression, $[\varphi(p_t|p_0) + \varphi(1-p_t|p_0)]$ is the probability of heterozygosity $2p_t(1-p_t)$ at time t. We will make more use of Equation 2.8 in the next sections, illustrating in particular its implications for the dispersion of allele frequencies among isolated populations.

PROBABILITIES AND TIMES TO FIXATION OR LOSS

Because Equation 2.8 denotes the probability density of allele frequency p given that the population is still polymorphic,

$$\Omega(p_0, t) = \int_{1/(2N)}^{1-1/(2N)} \varphi(p_t|p_0)dp_t \tag{2.9a}$$

is the probability that both alleles are still present in generation t. The probability that an allele with initial frequency p_0 has been fixed by generation t is

$$p_f(p_0, t) = p_0 + p_0(1-p_0)\sum_{i=1}^{\infty}(2i+1)(-1)^i \\ \cdot F(1-i, i+2, 2, p_0)\cdot e^{-i(i+1)t/4N} \tag{2.9b}$$

whereas the probability of loss of the allele, $p_l(p_0, t)$, is given by Equation 2.9b with $(1-p_0)$ exchanged for p_0 (Kimura 1955). Summing up,

$$\Omega(p_0, t) + p_f(p_0, t) + p_l(p_0, t) = 1 \tag{2.10}$$

As can be seen from the negative exponential terms in the previous expressions, as $t \to \infty$, $\Omega(p_0, t) \to 0$, $p_f(p_0, t) \to p_0$, and $p_l(p_0, t) \to (1-p_0)$. Thus, under neutrality, the probability that a particular allele becomes fixed is simply equal to its initial frequency. It follows that averaging over a very large number of replicate populations, the expected allele frequency remains constant at p_0, with a fraction p_0 of all replicates ultimately reaching allele frequency 1, and the remaining fraction $1-p_0$ having allele frequency 0.

An issue of special interest is the mean time until an allele is absorbed into either state $p = 0$ or $p = 1$. Using diffusion theory (Appendix 1), Kimura and Ohta (1969) obtained expressions for both quantities, and Kimura (1970) presented a description of the entire probability distributions for absorption times. The following example uses a somewhat simpler approach to arrive at results identical to those of Kimura

and Ohta (1969), and provides yet another illustration of how the effects of random genetic drift scale with population size.

Example 2.2. Ewens (2004) used the following line of reasoning to derive the expected time to absorption of a neutral allele under the Wright-Fisher model. Letting δp denote the change in allele frequency in one unit of time, the mean time to absorption for an allele with frequency p may be rewritten as

$$\bar{t}_a(p) = E[\,\bar{t}_a(p + \delta p)] + 1$$

where E denotes an expected value. In words, this expression states that the mean absorption time starting at frequency p is equal to the mean absorption time one time unit later when the allele frequency is $p + \delta p$, plus one. Approximating $\bar{t}_a(p + \delta p)$ by the first three terms in its Taylor series (see LW Equation A1.2) and then taking expectations (only the δp are random terms, the rest are fixed constants), gives

$$\begin{aligned} E[\,\bar{t}_a(p + \delta p)] &\simeq E\left[\,\bar{t}_a(p) + \delta p \frac{\partial\,\bar{t}_a(p)}{\partial p} + \frac{(\delta p)^2}{2}\frac{\partial^2\,\bar{t}_a(p)}{\partial p^2}\right] \\ &= \bar{t}_a(p) + E[\delta p]\frac{\partial\,\bar{t}_a(p)}{\partial p} + \frac{E[(\delta p)^2]}{2}\frac{\partial^2\,\bar{t}_a(p)}{\partial p^2} \end{aligned}$$

Hence, we have

$$\bar{t}_a(p) \simeq \bar{t}_a(p) + E[\delta p]\frac{\partial\,\bar{t}_a(p)}{\partial p} + \frac{E[(\delta p)^2]}{2}\frac{\partial^2\,\bar{t}_a(p)}{\partial p^2} + 1$$

Under neutrality, the expected change in allele frequency is $E(\delta p) = 0$, and as derived above, the expected variance in allele-frequency change is $E[(\delta p)^2] = p(1-p)/(2N)$. Substituting into our approximation and rearranging gives

$$\bar{t}_a(p) - \bar{t}_a(p) - 1 \simeq \frac{p(1-p)}{2 \cdot 2N}\frac{\partial^2\,\bar{t}_a(p)}{\partial p^2}$$

which implies the differential equation

$$\frac{\partial^2\,\bar{t}_a(p)}{\partial p^2} \simeq -\frac{4N}{p(1-p)}$$

Performing the double integration with respect to p leads to the solution

$$\bar{t}_a(p_0) \simeq -4N[p_0 \ln(p_0) + (1-p_0)\ln(1-p_0)] \tag{2.11a}$$

which is the mean time until an allele with initial frequency p_0 is either lost or fixed in a population.

A similar approach can be used to estimate the mean time to fixation for the subset of alleles that specifically become fixed, $\bar{t}_f(p_0)$. The essential modification here is that in estimating $\bar{t}_f(p_0)$, $E(\delta p)$ is no longer equal to zero, because in order for an allele to become fixed, at least one copy must be produced each generation. That is, in the case of conditional fixation, of the $2N$ genes drawn each generation, one is definitely

a B allele, whereas the remaining $2N-1$ genes can be viewed as random, leading to $E(\delta p) = \{(1/2N) + [1-(1/2N)]p\} - p = (1-p)/(2N)$. Similarly, because the states of only $2N-1$ genes are random, $E[(\delta p)^2] = \{(1 \cdot 0) + [(2N-1)p(1-p)]\}/(2N)^2$. Unless N is very small, the approximation $E[(\delta p)^2] = p(1-p)/(2N)$ still holds quite well, and following the procedures utilized above, we then have

$$\left(\frac{2}{p}\right)\frac{\partial \bar{t}_f(p)}{\partial p} + \frac{\partial^2 \bar{t}_f(p)}{\partial p^2} \simeq -\frac{4N}{p(1-p)}$$

The solution of this second-order differential equation requires several steps, which we omit, the final result being

$$\bar{t}_f(p_0) \simeq -\frac{4N(1-p_0)\ln(1-p_0)}{p_0} \tag{2.11b}$$

The mean time to loss of an allele conditional upon loss is identical to the previous expression with $(1-p_0)$ interchanged with p_0,

$$\bar{t}_l(p_0) \simeq -\frac{4Np_0\ln(p_0)}{1-p_0} \tag{2.11c}$$

Finally, because the probability of ultimate fixation of an allele is equal to its initial frequency (p_0) and the probability of ultimate loss is $(1-p_0)$, it follows that

$$\bar{t}_a(p_0) = p_0\bar{t}_f(p_0) + (1-p_0)\bar{t}_l(p_0) \tag{2.11d}$$

Example 8 in Appendix 1 uses diffusion theory to obtain results identical to those just presented.

AGE OF A NEUTRAL ALLELE

A classic result from standard neutral theory is that the age of a neutral allele is proportional to its frequency (Kimura and Ohta 1973; Maruyama 1974; Watterson 1976; Slatkin and Rannala 1997, 2000). In particular, the expected age t (in generations) of an allele with current frequency p is

$$E(t) = -\frac{4Np\ln(p)}{1-p} \tag{2.12}$$

assuming a constant effective population size during the allele's sojourn through the population (Kimura and Ohta 1973). As shown in Figure 2.1, alleles at higher frequency are expected to be older. This result is more than just an esoteric finding, as one class of tests for selection evaluates whether an allele, given its frequency, is too young to be compatible with neutrality (Chapter 9).

-Insert Figure 2.1 Here-

Example 2.3. The mutation *CCR5-δ32* destroys the human *CCR5* receptor, which is used by the HIV virus to enter the cell, leading to significant resistance against HIV infection. This deletion occurs at frequencies up to 14% in Eurasia, but is absent in Africans, Native Americans and East Asians. Assuming a frequency of $p = 0.10$ and an effective population size $N = 5000$ for Caucasians, Stephens et al. (1998) used Equation 2.12 to estimate the age of this allele (under the assumption of neutrality) as

$$\widehat{t} = -\frac{4 \cdot 5000 \cdot 0.1 \log(0.1)}{0.9} = 5116 \text{ generations}$$

However, an independent (and more direct) estimate of the allele's age can be obtained by considering the variation in haplotypes among all sequences carrying this mutation. The δ32 mutation is in strong linkage disequilibrium with allele *215* at the *AFMB* marker (a highly variable tandem-repeat locus), to the extent that 84.8% (39 of 46) of sampled δ32 mutations have the *δ32-215* haplotype. Clearly, the initial δ32 mutation at *CCR5* arose on a chromosome carrying the *215* allele. The recombination fraction between *CCR5* and *AFMB* was estimated by Stephens et al. (1998) to be $c = 0.006$. The probability of the *δ32-215* haplotype remaining intact after τ generations of recombination is just $\pi = (1-c)^{\tau}$, which because $c \ll 1$, implies

$$\tau \simeq -\ln(\pi)/c = -\ln(0.848)/0.006 = 27.5 \text{ generations}$$

Stephens et al. (1998) took these great disparities between age estimates as an indicator that strong selection has promoted the δ32 mutation much more rapidly than would be likely under a pure drift model. Assuming δ32 originated as a single mutation, they estimated the selection coefficient to be between 20% and 40%, depending on assumptions about dominance.

Equation 2.12 is not, however, the whole story, as a very old allele can sometimes be at low frequency, having transiently drifted up to a high frequency before drifting back toward zero frequency by chance. Thus, we expect the confidence interval for estimates of allelic age to be highly asymmetric about the mean. Slatkin and Rannala (2000) provide an approximation for the cumulative probability for the age of an allele, given that its frequency is p in a random sample of n alleles,

$$\Pr(t \leq \tau) = (1-p)^{-1+[n/(1+n\tau/2)]} \tag{2.13}$$

where t is measured in units of $2N$ generations.

Example 2.4. In a random sample of 500 gametes from a population, we find 350 copies of a particular allele **A** yielding an estimated $p = 0.7$. Under the joint assumptions of neutrality and constant population size, what is the 95% confidence interval for the estimated age of this allele? From Equation 2.12, its expected age is

$$\widehat{t} = -\frac{(4N)\,0.7\,\ln(0.7)}{1-0.7} = 1.45N$$

An approximate confidence interval using the approximation given by Equation 2.13 is obtained as follows. Define τ_α as satisfying $\Pr(t \leq \tau_\alpha) = \alpha$. The 95% confidence

interval for allelic age t is given by $(\tau_{0.025}, \tau_{0.975})$. Solving Equation 2.13 for $\tau_{0.025}$, we require

$$(0.3)^{-1+[500/(1+250\tau_{0.025})]} = 0.025$$

the solution of which is $\tau_{0.025} = 0.49N$. The same procedure can be used to find $\tau_{0.975} = 1.96N$, showing that the confidence interval about the mean value is very asymmetric.

Finally, an important caveat is in order with respect to Equation 2.12. Unless the sample size is very large, the estimated frequency of a rare allele can be quite misleading. Consider, for example, a singleton with a sample frequency of $1/n$. Application of Equation 2.12 would imply an estimated age of $4N\ln(n)/[n(1-(1/n))]$ generations. If ten diploid individuals had been sampled, the minimum allele frequency would be 0.05, so that the estimated age of a singleton would be $\sim 0.84N$ generations. As very rare alleles, with true frequencies $\ll 1/n$ will generally either be recorded as singletons or not at all in a sample of size n, it is clear that Equation 2.12 yields upwardly biased estimates of the ages of rare alleles unless the sample size is large enough that the estimate of frequency p is highly accurate.

ALLELE-FREQUENCY DIVERGENCE AMONG POPULATIONS

A natural consequence of allele-frequency drift within populations is the divergence of isolated replicate populations. Suppose a monoecious base population with allele frequency p_0 is suddenly split into several completely isolated subpopulations, each of size N, with random mating within each subpopulation and no selection, migration, or mutation. The variance in allele frequency among subpopulations in generation t is

$$\sigma_p^2(t) = E(p_t^2) - E^2(p_t)$$

Adding and subtracting $E(p_t)$,

$$\begin{aligned}\sigma_p^2(t) &= [E(p_t) - E^2(p_t)] + [E(p_t^2) - E(p_t)] \\ &= E(p_t)[1 - E(p_t)] - E[p_t(1-p_t)]\end{aligned}$$

Because there are no systematic forces causing the allele frequency to increase or decrease, $E(p_t) = p_0$, and the first quantity on the right is $p_0(1-p_0)$. The quantity $E[p_t(1-p_t)]$ is half the expected heterozygosity in a population in generation t, which has already been defined in Equation 2.5. Thus,

$$\sigma_p^2(t) = p_0(1-p_0)\left[1 - \left(1 - \frac{1}{2N}\right)^t\right] \tag{2.14a}$$

which is well approximated by

$$\sigma_p^2(t) \simeq p_0(1-p_0)(1-e^{-t/2N}) \tag{2.14b}$$

for $N > 10$. This shows that the among-population variance asymptotically approaches $p_0(1 - p_0)$, which is half the heterozygosity in the base population. An alternative way to envision this asymptotic result is to note that at fixation the allele frequency has value 1.0 with probability p_0, and otherwise is zero, giving $E(p_0) = 1 \cdot p_0 = E(p_0^2) = 1^2 \cdot p_0 = p_0$. Hence, the among-population variance when all alleles are fixed is just $E(p_0^2) - [E(p_0)]^2 = p_0(1 - p_0)$.

Although Equations 2.14a,b deal with the expected allele-frequency variance, they do not describe the actual form of the distribution of population allele frequencies. However, all of this information is contained in the formulations presented above on the probability distribution of allele frequencies within populations. For example, the transition-matrix approach (Equations 2.2a,b) and the diffusion approximation (Equation 2.8) yield the expected temporal dynamics of the distribution of allele frequencies in different replicate populations, all starting from an identical frequency, p_0 (Figure 2.2).

-Insert Figure 2.2 Here-

Summing up to this point, five significant conclusions can be gleaned with respect to neutral alleles. First, with increasing time, the total probability mass for the allele-frequency distribution at a locus declines because only *segregating* alleles are considered, i.e., the proportions of populations that have experienced gene fixation or loss are ignored. Second, regardless of the starting condition, the distribution becomes flatter with increasing time, such that the frequency of a sufficiently old segregating allele is equally likely anywhere over the (0,1) interval. Third, high-frequency alleles are generally expected to be old. Fourth, the distributions in Figure 2.2 can be interpreted in two different ways: as the probability distribution of allele frequencies over a very large number of replicate populations over time, all starting at an identical state, or as the expected distribution of allele frequencies for the subset of loci with identical starting frequencies within a single population. Finally, the expected allele-frequency distribution is a function of t/N generations, as can be seen from the exponential terms in Equation 2.8. As should be clear by now, this scaling of the temporal dynamics of random genetic drift to the reciprocal of population size is a natural consequence of the fact that the variance of allele-frequency change is inversely proportional to N.

BURI'S EXPERIMENT

Because all populations are finite in size, the theory of random genetic drift is of central significance to all areas of population genetics. It may therefore come as a surprise that highly replicated experiments examining the chance dynamics of allele-frequency change are extremely rare. However, the results of one massive experiment nicely affirm the theoretical expectations outlined above, while making one additional important point. Starting with two homozygous lines of *Drosophila melanogaster,* one of which was fixed for allele bw^{75} and the other for allele bw at the brown locus, Buri (1956) established 212 F_1 hybrid populations. For the following

19 generations, he randomly mated 8 males and 8 females within each population and monitored the changes in allele frequencies in each subline. This could be done in the premolecular era because the genotype at the brown locus determines eye color: $bw^{75}bw^{75}$ = bright red-orange, $bw^{75}bw$ = deep red-brown, and $bwbw$ = white. (Two separate experiments were performed, one with 107 and the other with 105 populations, but the results are so similar that they have been pooled in the following analysis).

In order to evaluate the results in the light of the preceding theory, it is first necessary to demonstrate that the bw^{75} and bw alleles are indeed neutral with respect to each other. This can be done as follows (Figure 2.3, top). In the absence of selection, the expected frequency of the bw^{75} allele averaged over all populations should equal its initial frequency, 0.50, in all generations. Nevertheless, just as the frequency within any population is expected to deviate from 0.5 because of drift, the mean allele frequency in the total aggregate of populations will also vary because the number of populations is finite. The sampling variance of the overall mean frequency is equal to the sum of the expected within- and among-population allele-frequency variances divided by the number of populations, 212. The latter quantity has already been defined in Equation 2.14, while the former is the expected binomial sampling variance divided by the sample size $(2N)$, or $p_0(1-p_0)[1-(1/2N)]^t/(2N)$. The figure shows that although the frequency of the bw^{75} allele averaged over all populations increased to 0.525, it generally remained within two standard errors of the expectation under pure drift. The overall pattern of change in mean allele frequency is therefore compatible with the expectations for a neutral locus subject to random genetic drift.

The dynamics of the among-population divergence (Figure 2.4) are qualitatively very similar to the expected pattern illustrated in the left panel of Figure 2.2. As the population allele frequencies diverge, the initial bell-shaped distribution does indeed become flatter, eventually acquiring a U-shape as populations that are fixed for the bw^{75} or bw alleles accumulate. Had the experiment been extended further in time, the distribution would have eventually consisted of only two classes, populations fixed for bw^{75} and those fixed for bw, with nearly equal frequencies.

Despite the qualitative agreement with theoretical expectations, the rate of divergence illustrated in Figure 2.4 is somewhat greater than that expected for randomly mating populations of 16 individuals. However, this does not necessarily invalidate the theory outlined above, as it is possible that not all 16 potential parents reproduced each generation, and/or that the distribution of family sizes deviated from randomness. Either condition would cause the populations to behave genetically as though they were smaller than the actual size (Chapter 3). With the massive amount of data in Buri's experiment, it is possible to obtain an empirical estimate of this **effective population size** in the following way.

Not including fixed classes, there are 31 possible allele frequencies in Buri's populations (1/32 to 31/32), each of which was observed at various times in one or more of the 212 populations. Focusing on any one allele-frequency class, the sampling variance conditional on the initial allele frequency for this class can then be calculated from the allele frequencies observed in the subsequent generation, and compared to the expected value of $p(1-p)/(2N)$. The 31 points shown in Figure 2.5 provide an empirical description of this function, with an excellent fit being obtained

if it is assumed that the average effective population size was $N \simeq 10.2$ rather than the idealized 16. In other words, the sampling variance of allele frequencies is in very close accord with that expected for an average ideal population of 10.2 randomly mating individuals. Once this change in scale from $N = 16$ to $N = 10.2$ is taken into account, both the erosion of average heterozygosity within populations and the build-up of among-population variance of allele frequencies are quite consistent with the theory outlined above (Figure 2.3, middle and bottom).

-Insert Figures 2.3, 2.4 and 2.5 Here-

HIGHER-ORDER ALLELE-FREQUENCY MOMENTS

In the previous sections, we evaluated the expected values of various population features under neutrality. However, as just noted, in applying such expressions to empirical studies, it is important to keep in mind that the random sampling of allele frequencies across generations will cause the exact behavior of any particular population or group of populations to deviate from the expected pattern. Thus, there is a practical need for expressions for the variance of various population parameters that result from genetic sampling. This in turn requires an understanding of the behavior of higher-order allele-frequency moments. For example, although the *expected* heterozygosity is a function of $2(p - p^2)$, as will be shown below, its *variance* depends on p^3 and p^4.

For a population obeying the features of the idealized Wright-Fisher model, useful expressions can be acquired by noting that the expected value of an allele-frequency moment in generation $t+1$ conditional on allele frequency p_t in the previous generation is

$$\begin{aligned} E[p_{t+1}^k|p_t] &= E[(p_t + \delta p)^k|p_t] \\ &= \sum_{i=0}^{k} \binom{k}{i} p_t^i \, E[(\delta p)^{k-i}|p_t] \end{aligned} \tag{2.15}$$

where δp denotes the change in allele frequency in the previous generation resulting from gamete sampling. For binomial sampling theory, expressions are available for all expected values of powers of δp (e.g., Johnson et al. 2005), so Equation 2.15 can be solved recursively starting with the lower-order moments. For example, in the absence of any directional forces, $E(\delta p) = 0$ and the expected frequency of an allele remains perpetually at its initial value (p_0),

$$E(p_t) = p_0 \tag{2.16a}$$

The second moment is obtained by noting that

$$E(p_{t+1}^2|p_t) = E[(p_t^2 + 2p_t\delta p + \delta p^2)|p_t]$$

Because $E(\delta p) = 0$ and $E(\delta p^2|p_t) = p_t(1-p_t)/(2N)$ under binomial sampling,

$$E(p_{t+1}^2|p_t) = E\left(p_t^2 + \frac{p_t - p_t^2}{2N}\right)$$

Letting $\lambda_1 = 1 - (1/2N)$, this expression can be rearranged to give the recursion equation

$$E(p_{t+1}^2|p_t) - p_0 = [E(p_t^2) - p_0]\lambda_1$$

the general solution of which is

$$E(p_t^2) = p_0 - [p_0(1-p_0)]\lambda_1^t \tag{2.16b}$$

Using expectations for higher-order δp^k terms, expressions for additional moments can be acquired, two of which prove to be particularly useful (Crow and Kimura 1970),

$$E(p_t^3) = p_0 - \frac{3}{2}p_0(1-p_0)\lambda_1^t - \frac{1}{2}p_0(1-p_0)(2p_0-1)(\lambda_1\lambda_2)^t \tag{2.16c}$$

$$E(p_t^4) = p_0 - \frac{18N-11}{10N-6}p_0(1-p_0)\lambda_1^t - p_0(1-p_0)(2p_0-1)(\lambda_1\lambda_2)^t + p_0(1-p_0)\left(p_0(1-p_0) - \frac{2N-1}{10-6}\right)(\lambda_1\lambda_2\lambda_3)^t \tag{2.16d}$$

where $\lambda_i = 1 - (i/2N)$. Modifications for these expressions for populations with separate sexes and 1:1 sex ratios are given by Lynch and Hill (1986).

Example 2.5. The preceding expressions can be used to derive the evolutionary (or drift) variance of heterozygosity at a locus under the assumption of Hardy-Weinberg equilibrium, provided there are only two alleles segregating at the locus. Letting $H_t = 2p_t(1-p_t)$ denote the heterozygosity at generation t, the expected variance of heterozygosity is

$$\begin{aligned}\sigma^2(H_t) &= E\{[2p_t(1-p_t)]^2\} - \{E[2p_t(1-p_t)]\}^2 \\ &= E(4p_t^2) - E(8p_t^3) + E(4p_t^4) - [E(2p_t - p_t^2)]^2\end{aligned}$$

The solution is obtained by substituting Equations 2.16a-d for the expectations of allele-frequency moments, with further simplification made possible by using the approximations $\lambda_1 \simeq e^{-t/2N}$, $\lambda_2 \simeq e^{-2t/2N}$, and $\lambda_3 \simeq e^{-3t/2N}$,

$$\sigma^2(H_t) \simeq H_0\left[\frac{2}{5}e^{-t/2N} + \left(H_0 - \frac{2}{5}\right)e^{-3t/N} - H_0 e^{-t/N}\right] \tag{2.17}$$

This quantity can be viewed as either the variance in heterozygosity that develops at a particular neutral locus among replicate populations starting from the same initial allele frequencies or as the variance in heterozygosity among a pool of loci within the same population with identical initial allele frequencies. As with the expected heterozygosity, the temporal dynamics of the evolutionary variance of heterozygosity

scale inversely with the size of the population. Moreover, the variation in heterozygosity resulting from genetic drift can be quite high, with the standard deviation always exceeding the expected heterozygosity beyond $t = 2N$ generations (Figure 2.6). With more than two alleles per locus, the preceding expressions would need to be modified to account for the negative drift sampling covariance between different alleles at the locus.

-Insert Figure 2.6 Here-

LINKAGE DISEQUILIBRIUM

In the study of multilocus traits, we are naturally interested in combinations of alleles among loci. If the alleles at two loci are independently distributed, the expected frequency of each gamete type can be predicted from the products of the allele frequencies at the two loci. For example, with two alternative alleles (A and a) at one locus having frequencies p and $1-p$, and those (B and b) at another locus having frequencies q and $1-q$, the expected frequencies of gametic types AB, Ab, aB, and ab are pq, $p(1-q)$, $(1-p)q$, and $(1-p)(1-q)$, respectively, under the assumption of independence. A natural measure of the deviation of the frequency of a gametic type from such expectations is the **coefficient of linkage disequilibrium**

$$D_{AB} = p_{AB} - pq \tag{2.18}$$

where p_{AB} denotes the observed frequency of the **AB**th gamete.

This definition has the useful feature of being equivalent to the covariance of the distribution of alleles A and B in the same gametes. To see this, let the random variable x take on a value of one when the allele at the first locus is A and zero otherwise, and likewise let y equal one when the allele at the second locus is B and zero otherwise. Then, $E(xy) = p_{AB} \cdot 1$, $E(x) = \text{freq(A)} \cdot 1 = p$, $E(y) = \text{freq(B)} \cdot 1 = q$, giving the covariance between allele presence at the two loci as $E(xy) - E(x)E(y) = p_{AB} - pq$.

Although in the absence of selection, there will be no tendency for the alleles at different loci to be associated positively versus negatively, historical forces such as migration or nonrandom mating may cause some such correlations. Letting D_0 denote an initial level of disequilibrium, r denote the frequency of recombination between loci, and $\lambda_1 = 1-(1/2N)$, the expected disequilibrium (under random mating) resulting from the joint forces of recombination and gametic sampling is

$$\begin{aligned} E(D_t) &= [(1-r)\lambda_1]^t D_0 \\ &\simeq D_0 e^{-(2Nr+1)t/(2N)} \end{aligned} \tag{2.19}$$

(Hill and Robertson 1966), showing that disequilibrium declines toward zero in the absence of any replenishing forces.

In contrast, the *variance* of D can be quite substantial even when its expected value is zero. The problem can be evaluated by use of the following set of recursion

equations for fourth-order moments of allele frequencies,

$$\begin{pmatrix} E[p(1-p)q(1-q)] \\ E[D(1-2p)(1-2q)] \\ E(D^2) \end{pmatrix}_{t+1} = \lambda_1 \cdot \begin{pmatrix} \lambda_1 & \lambda_1(1-r)/(2N) & 2(1-r)^2/(4N^2) \\ 0 & \lambda_2^2(1-r) & 4\lambda_2(1-r)^2/(2N) \\ 1/(2N) & \lambda_1(1-r)/(2N) & [\lambda_2^2 + (1/4N^2)](1-r)^2 \end{pmatrix} \cdot \begin{pmatrix} E[p(1-p)q(1-q)] \\ E[D(1-2p)(1-2q)] \\ E(D^2) \end{pmatrix}_t \tag{2.20}$$

where, as above, $\lambda_i = 1 - [i/(2N)]$ (Hill and Robertson 1968). The evolutionary variance of D associated with drift among replicated populations or among loci starting from the same allele frequencies is

$$\sigma^2(D) = E(D^2) - E^2(D) \tag{2.21}$$

If $D_0 = 0$, then $E(D_t) = 0$, and $\sigma^2(D_t) = E(D_t^2)$. Ohta and Kimura (1969; their Equations 20-25) obtained a closed-form solution to this expression as well as for the dimensionless squared correlation coefficient

$$r_D^2 = \frac{E(D^2)}{E[p(1-p)q(1-q)]} \tag{2.22}$$

This standardized measure of linkage disequilibrium, introduced by Hill and Robertson (1968), is referred to as the square of the within-gamete correlation of allele frequencies at two loci. However, as Equation 2.22 is defined as a ratio of expectations, it is not a true definition of the squared gametic correlation which is a function of the interdependency of the numerator and denominator. This difference in definition can lead to up to 100-fold differences between values of r_D^2 and the true correlation if allele frequencies are extreme, as is often the case with neutral alleles (Song and Song 2007).

Example 2.6. One common setting generating LD involves a new allele arising as a single copy on a particular background. Let **A** denote the derived allele and assume it arose on a **B** background at a second locus. Initially, all copies of **A** are associated with **B** (there are no **Ab** gametes). Because the sum of the **AB** and **aB** gamete frequencies is just the frequency p_B of **B**, the resulting 2×2 gametic contingency table becomes

	A	a
B	p_A	$p_B - p_A$
b	0	p_b

and the resulting initial values for D and r_D^2 become

$$D_{AB} = p_{AB}p_{ab} - p_{aB}p_{Ab} = p_A p_b - 0 \cdot (p_B - p_A) = p_A p_b$$

$$r_D^2 = \frac{D^2}{p_A p_a p_B p_b} = \frac{(p_A p_b)^2}{p_A p_a p_B p_b} = \frac{p_A p_b}{p_a p_B} = \frac{p_A(1-p_B)}{(1-p_A)p_B}$$

For this example, $r^2 = 1$ only when $p_A = p_B$ (Sved 1971), in other words, when there are no **aB** gametes (**A** and **B** always co-occur), in which case the 2×2 gamete contingency table becomes

	A	a
B	p_A	0
b	0	p_a

, or equivalently

	A	a
B	p_B	0
b	0	p_b

For a newly-arising mutation, $p_A = 1/(2N)$, so initially $D = p_b/(2N)$ and $r_D^2 \simeq (1 - p_B)/(2Np_B)$.

MUTATION-DRIFT EQUILIBRIUM

In the preceding pages, we were largely concerned with the dynamics of gene-frequency change owing to the effects of random genetic drift alone. Under this model, finite population size eventually results in the complete loss of genetic variation (and covariation) within populations, at which point all loci are fixed for ancestral alleles with probabilities equal to their initial frequencies. In reality, mutation will always reintroduce variation at a low rate, which not only offsets some of the loss resulting from drift, but also ensures that neutral loci will continue to diverge among isolated populations. If the time scale of the problem under consideration is short ($t \ll 2N$) and the initial level of within-population variation is high (relative to the mutational rate of production of new heterozygosity per generation), the contribution from mutation will be negligible, and the preceding expressions will be quite adequate. However, for longer-term evolutionary issues, such as the maintenance of variation in natural populations and interspecific divergence, mutation cannot be ignored.

The incorporation of mutation into a neutral model of evolution is relatively straight-forward. Suppose there are k possible alleles at a locus, each with a mutation rate u per generation (the **k-exchangeable alleles model**). The dynamics of heterozygosity can then be obtained by recalling from above that the expected frequency of heterozygotes in generation $t + 1$ in the absence of mutation is $\lambda_1 H_t$, whereas the expected frequency of homozygotes is $1 - \lambda_1 H_t$. Following mutation, the heterozygous state will be retained if: 1) neither allele mutated, the probability of which is $(1 - 2u)$ ignoring the very small probability of double mutations to the same state; or 2) one of the alleles mutated to a different state than the other, the probability of which is $[2u(k - 2)/(k - 1)]$ assuming that all allelic types are equally mutationally exchangeable. On the other hand, homozygotes will be mutationally converted to heterozygotes at rate $2u$. Thus, the expected dynamics of heterozygosity can be expressed as

$$H_{t+1} = H_t \lambda_1 \left((1 - 2u) + \frac{2u(k - 2)}{k - 1} \right) + 2u(1 - \lambda_1 H_t) \tag{2.23}$$

Setting $H_{t+1} = H_t$, the expected value of heterozygosity under drift-mutation balance

is found to be

$$E(H) = \frac{\theta}{1 + [\theta k/(k-1)]} \tag{2.24a}$$

where $\theta = 4Nu$, a result first given by Kimura (1968). Note that θ has the pleasing interpretation of being the ratio of the rates of mutational production of heterozygotes from homozygotes ($2u$) and the rate of loss of heterozygosity by drift ($1/2N$). If a large number of alternative alleles ($k \gg 1$) is assumed, as is reasonable when the unit of analysis is an entire gene, Equation 2.24a reduces to

$$E(H) \simeq \frac{\theta}{1 + \theta} \tag{2.24b}$$

which is equivalent to the **infinite-alleles model** of Kimura and Crow (1964). On the other hand, if the unit of analysis is a nucleotide site, then $k = 4$, and

$$E(H) = \frac{\theta}{1 + (4/3)\theta} \tag{2.24c}$$

where u is now the mutation rate per nucleotide site. Equation 2.24a needs to be modified if alleles mutate at different rates or are not equally mutationally accessible (Kimura 1983; Nei and Kumar 2000), but provided $2Nu \ll 1$, as seems to be generally the case (Chapter 4), then $\widehat{H} \simeq \theta$ regardless of the model assumed. As $2Nu \to \infty$, the infinite-alleles model implies $E(H) \to 1.0$, whereas the $k = 4$ model implies $E(H) \to 0.75$. The latter result is a simple consequence of four segregating alleles with equal frequencies (0.25) when the power of drift is overwhelmed by mutation. A number of other mutational models are possible, and a very general treatment is given by Cockerham (1984), who also considered the transient approach to equilibrium.

As the above drift-mutation models play a central role in the neutral theory of molecular evolution (Kimura 1983), substantial attention has been given to additional details that are obscured by the summary statistic of average heterozygosity. For example, because of the stochastic nature of both mutation and drift, the allele frequencies at any neutral locus are expected to wander stochastically over time, with some loci being transiently fixed for one particular allele, and others being distributed over the remaining spectrum of allele frequencies. Kimura (1968) obtained an expression analogous to Equation 2.8 for the complete probability distribution of the frequency of an allele under the symmetric mutation model described above, starting from an arbitrary allele frequency. Although this expression is quite complicated, a highly useful result is that regardless of the starting point, an equilibrium distribution of allele frequencies p is eventually attained

$$\phi(p) = \frac{\Gamma(\theta + \beta)}{\Gamma(\theta)\Gamma(\beta)}(1-p)^{\theta-1}p^{\beta-1} \tag{2.25a}$$

where $\theta = 4Nu$, $\beta = 4Nu/(k-1)$, and

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx \tag{2.25b}$$

is the **gamma function** (Appendix 2). Equation 2.25a may be viewed as either the expected distribution of allele frequencies over all neutral loci within a single

population in mutation-drift equilibrium or as the distribution of allele frequencies at a particular locus among replicate populations (or species) with identical N and u. Nei and Li (1976) present the theory necessary for predicting the approach to the equilibrium state.

The expected value of any function of population allele frequencies (e.g., homozygosity) can be obtained by simply integrating the function over the density distribution $\phi(p)$. It is useful that Equation 2.25a defines a **beta distribution** (Appendix 2), as many of its properties are already well known. For example, the mean of the distribution, which is the expected allele frequency, is

$$E(p) = \frac{\beta}{\theta + \beta} = \frac{1}{k} \tag{2.26a}$$

and the allele-frequency variance among replicates is

$$\sigma^2(p) = \frac{\theta\beta}{(\theta + \beta)^2(\theta + \beta + 1)} = \frac{k - 1}{k^2[2Nuk(k - 1) + 1]} \tag{2.26b}$$

Expressions for the variance of heterozygosity for a population en route to equilibrium were derived by Li and Nei (1975) and Lessard (1981), and at equilibrium

$$\sigma^2(H) = \frac{2\theta[1 + (\theta/l)]}{[1 + \theta + (\theta/l)]^2[2 + \theta + (\theta/l)][3 + \theta + (\theta/l)]} \tag{2.7a}$$

where $l = k - 1$ (Stewart 1976). Note that as $k \to \infty$ (the infinite-alleles model),

$$\sigma^2(H) = \frac{2\theta(1 + \theta)}{(1 + \theta)^2(2 + \theta)(3 + \theta)} \tag{2.27b}$$

which reduces to $\simeq \theta/3$ for $\theta \ll 1$.

Although the preceding results lead to predicted equilibrium allele frequencies, heterozygosities, etc., viewed in isolation such results obscure the long-term dynamics of neutral mutations. Given a population of size N, $2Nu$ new mutations arise per nucleotide site per generation, each with initial frequency $1/(2N)$. As noted above, the probability of fixation of a neutral allele is simply equal to its initial frequency, so in the long run, at a neutral site, there is an average turnover rate of $2Nu \cdot 1/(2N) = u$ mutations per generation. This simple, but powerful, result tells us that the long-term rate of nucleotide substitution at a neutral site is equal to the mutation rate, *regardless of the size of the population,* a hallmark of the **neutral theory of molecular evolution** (Kimura 1983). This long-term flux occurs in the face of the maintenance of quasi-steady-state within-population heterozygosity, which arises as a consequence of the per-generation balance between the loss of variation by fixation and replenishment by recurrent mutation.

Finally, Ohta and Kimura (1971) and Hill (1975) obtained expressions for the expected values of the two-locus moments described in Equation 2.20, under the infinite-alleles model at stochastic drift-mutation equilibrium. Letting $\rho_L = 4Nr$ and $\theta = 4Nu$, where r is the recombination rate between sites separated by distance L (with c being the recombination rate per site, $r = cL$), and u is the mutation rate per site, Hill's expressions reduce to

$$\begin{aligned} E[p(1 - p)q(1 - q)] &= M(22 + 13\rho_L + 32\theta + \rho_L^2 + 6\rho_L\theta + 8\theta^2) \\ E[D(1 - 2p)(1 - 2q)] &= 8M \\ E[D^2] &= M(10 + \rho_L + 4\theta) \end{aligned} \tag{2.28a}$$

where

$$M = \theta^2/[(\theta+1)(18+13\rho_L+54\theta^2+\rho_L^2+19\rho_L\theta+40\theta^2+6\rho_L\theta^2+8\theta)], \tag{2.28b}$$

and the standardized linkage disequilibrium is given by

$$r_D^2 = \frac{10+\rho_L+4\theta}{22+13\rho_L+32\theta+\rho_L^2+6\rho_L\theta+8\theta^2} \tag{2.29a}$$

As will be seen in Chapter 4, for individual nucleotide sites θ is generally substantially smaller than one, whereas for sites separated by hundreds of base pairs or more, ρ_L is generally greater than one. Under such conditions,

$$r_D^2 \simeq \frac{10+\rho_L}{22+13\rho_L+\rho_L^2} \tag{2.29b}$$

which asymptotically approaches $1/\rho_L$ for large ρ_L. Additional theoretical points of interest are developed in Strobeck and Morgan (1978) and Golding and Strobeck (1980). A more daunting problem is obtaining expressions for the evolutionary variances of these statistics. As the components in Equation 2.28a already involve fourth-order moments of allele frequencies, within and between loci, their variances are functions of moments up to the eighth order. Hill and Weir (1988) have tackled this problem.

Example 2.7. A widely-cited expression for $E(r_D^2)$ originates with Sved (1971; Feldman and Sved 1973),

$$E(r_{IBD}^2) = \frac{1}{1+4N_e r} = \frac{1}{1+\rho_L}$$

Note that this expression is quite different from Equation 2.29a (as no terms for mutation appear), and also different from the mutation-free ($\theta \ll 1$) version given by Equation 2.29b. To understand this discrepancy, it useful to reflect on Sved's simple derivation, which focuses on a different measure of correlation than that outlined above.

The focus here is the conditional probability Q_{AB} that if alleles at site **A** for two random gametes are identical by descent (IBD, owing to common ancestry), those at site **B** on both gametes are also IBD as a consequence of no recombination between the sites during the entire two pathways back to the common ancestor. With this definition, Sved showed that $Q_{AB} = r_{IBD}^2$. Let Q_t denote the value of joint IBD in generation t, and consider how this relates to Q_{t-1} one generation in the past. For a population with effective size N_e, at any particular site a random pair of gametes will be direct copies of a gamete in the preceding generation with probability $1/(2N_e)$, and IBD will exist at both sites provided no recombination has occurred between them, the probability of which is $(1-r)^2$. Alternatively, the two sampled gametes will be sampled from different gametes leading to the preceding generation, the probability of which is $1-1/(2N_e)$. In the latter case, joint IBD will still exist if it happened to have been present for the two chromosomal segments in the preceding generation with no recombination again occuring in either. Summing up, the recursion equation for joint IBD is

$$Q_t = \frac{1}{2N_e}(1-r)^2 + \left(1-\frac{1}{2N_e}\right)(1-r)^2 Q_{t-1}$$

Setting $Q_t = Q_{t-1}$ gives the equilibrium solution as

$$\widetilde{Q} = \frac{1/(2N_e)}{1 - [1 - 1/(2N_e)](1-r)^2} \simeq \frac{1}{1 + 4N_e r}$$

Although this is an equilibrium solution, like Equations 2.29a,b, Sved's Q is not equivalent to the measure r_D^2 noted above, as the latter is concerned with the long-term average disequilibrium associated with identity in state (IIS). Whereas IIS is a directly observable quantity, IBD is not, and hence there are interpretative problems when applying Sved's expression to empirical data. This is because parallel mutation can cause IIS for alleles that are not IBD, and secondary mutations can eliminate IIS for pairs of alleles that are otherwise IBD. Thus, although Sved's expression is widely used, apparently because of its simplicity, Equations 2.29a,b appear to be more appropriate for practical applications to observed molecular variation.

THE DETAILED STRUCTURE OF NEUTRAL VARIATION

While heterozygosity provides a robust measure of genetic variation, as a summary statistic it obscures the details of the underlying structure of this variation, such as the number of alleles and their frequencies. Such features are very important for the estimation of quantities such as θ (Chapter 4) and especially for tests of selection (Chapter 9). Fortunately, at mutation-drift equilibrium, most of these features are just relatively simple functions of θ. The very rich theory on this subject is nicely reviewed by Ewens (2004). There are, however, two alternative ways of thinking about molecular evolution, each appropriate in a different context, and care must be taken in the interpretation of the mutation rate in each model, as in one case, the **the infinite-alleles model**, the unit of observation is the locus (a stretch of DNA), whereas in the **infinite-sites model**, it is the nucleotide site.

The Infinite-alleles Model and the Associated Allele-frequency Spectrum

The infinite-alleles model (briefly introduced above) was developed prior to the DNA-sequencing era, with alleles often being identified by features of their protein products (e.g., electrophoretic mobility on gels) rather than by their underlying DNA sequences. In today's world, different alleles under this model would typically be viewed as different sequences (**haplotypes**) over a region of L nucleotide sites, and the general assumption is that L is large enough that each mutation generates a new haplotype, but small enough that recombination can be ignored, so that mutation is the sole generator of novel sequences. Considering the five short sequences in Figure 2.7, under the infinite-alleles framework, there are three different alleles, although there are only two segregating sites. With allele frequencies 0.4 (AAGACC), 0.4 (AAGGCC), and 0.2 (AAGGCA), the allelic heterozygosity of the sample is 0.64. This is, of course, a rather crude perspective, as it ignores the ways in which alleles

differ from each other (in this example, two pairs of alleles differ at a single site, whereas one pair differs at two sites).

-Insert Figure 2.7 Here-

A key parameter for the infinite-alleles model is the per-*locus* population mutation rate, $\theta_L = 4NuL$, which we distinguish from the more commonly used per-*site* measure, $\theta = 4Nu$, where u is the mutation rate per site. As noted above, under neutrality, provided the population has been at constant size time long enough to be in mutation-drift equilibrium, the expected heterozygosity is given by Equation 2.24b, with $E(H) \approx \theta_L$ for $\theta_L \ll 1$ and $E(H) \approx 1 - 1/\theta_L$ for $\theta_L \gg 1$. If the expected heterozygosity is small, a sample will often be **monomorphic**, consisting of only a single allele, or **dimorphic**, consisting of just two alleles. However, at higher levels of heterozygosity, multiple alleles can be expected, and a natural measure of variation is the number of different alleles in a sample. Insight into this quantity under drift-mutation equilibrium is given by **Ewens' sampling formula** (Ewens 1972), which states the probability of having k different alleles in a sample of size n to be

$$\Pr(k \,|\, \theta_L, n) = \frac{S_n^k \, \theta^k}{S_n(\theta_L)} \tag{2.30a}$$

where

$$S_n(\theta_L) = \theta_L(\theta_L + 1)(\theta_L + 2) \cdots (\theta_L + n - 1) \tag{2.30b}$$

and S_n^k is the coefficient on the θ_L^k term obtained by expanding the polynomial in Equation 2.30b. Ewens' sampling formula opened up the field of formal statistical tests for whether a pattern of allelic variation is consistent with the equilibrium neutral model. For example, using Equation 2.30a, one can ask if an observed estimate of θ_L (obtained in this case as the allelic heterozygosity) is consistent with the observed number k of different alleles (Chapter 7).

Several useful results immediately follow from Equation 2.30a. First, the probability of a monomorphic sample is

$$\Pr(k = 1) = \frac{(n-1)!}{(\theta_L + 1)(\theta_L + 2) \cdots (\theta_L + n - 1)} \tag{2.31a}$$

Second, a bit of algebra gives the mean and variance for the number of alleles in a sample as

$$E(k) = 1 + \theta_L \cdot \sum_{j=2}^{n} \frac{1}{\theta_L + j - 1}, \qquad \sigma^2(k) = \theta_L \cdot \sum_{j=1}^{n-1} \frac{j}{(\theta_L + j)^2} \tag{2.31b}$$

An even more complete description of the segregating allelic variation is given by the **allele-frequency spectrum**, which describes the frequencies of the various different alleles within the sample. Given that the numbering of alleles is arbitrary, the convention is to consider the vector $(n_1, \cdots, n_n)$, where n_i denotes the number of alleles that have exactly i copies in the sample. If the sample is monomorphic, then $n_n = 1$, while if all n alleles are unique (**singletons**), $n_1 = n$. For the example data

set in Figure 2.7, one allele appears as a singleton, while the other two alleles both appear as two copies, giving $n_1 = 1, n_2 = 2$, and $n_3, n_4, n_5 = 0$. The constraint on the n_i is that

$$\sum_{i=1}^{n} i \cdot n_i = n \tag{2.32}$$

A very powerful result, due to Ewens (1972) and Karlin and McGregor (1972), is that conditioned on the observed number of alleles k in the sample, the allele-frequency spectrum is given by

$$\Pr(n_1, n_2, \cdots, n_n \,|\, n,\, k) = \frac{n!}{S_n^k \, 1^{n_1} 2^{n_2} \cdots n^{n_n} \, n_1! n_2! \cdots n_n!} \tag{2.33a}$$

Only nonzero values of n_i are included. Note that Equation 2.33a is *independent* of θ_L. This property arises because k is a sufficient statistic for θ_L under the equilibrium neutral model, so that conditioning on k removes any dependence on θ_L. Similarly, the *joint* probability that a random sample contains k alleles and has an allele-frequency spectrum of $(m_1, \cdots m_k)$, where m_i is the number of copies of allele i in the sample) is

$$\Pr(m_1, \cdots, m_k, k \,|\, n, \theta_L) = \frac{n! \, \theta_L^k}{k! \, m_1 \, m_2 \, \cdots \, m_k \, S_n(\theta_L)} \tag{2.33b}$$

(Ewens 2004). The probability that the sample is monomorphic, Equation 2.31a, directly follows by setting $k = 1, m_1 = n$.

The Infinite-Sites Model and the Associated Site-frequency Spectrum

An alternate framework for summarizing molecular variation is embodied in the **infinite-sites** model, which treats a region as a series of L sites. Each new mutation is again viewed as unique, but now occurring at a novel (not currently segregating) *site*, and as with the infinite-alleles model, recombination within the region is assumed to be insignificant. Because this model allows only a single mutation per site, a particular variant at a polymorphic site is either **ancestral** (original) or **derived** (mutated), and each segregating site is treated as biallelic. With sequence data from one or more outgroups, one can often **polarize** the nucleotides at any particular site, determining with high confidence which is derived. In the absence of such information, the **minor allele frequency**, the frequency of the rarest nucleotide, is reported for each site. In the following expressions, we will make use of both measures of the population mutation rate noted above, with $\theta_L = 4NuL = \theta L$.

The infinite-sites model offers a much richer set of information than can be achieved with the infinite-alleles model. One measure is the **number of segregating sites** S, the number of polymorphic sites in a sample (the loose analog to number of alleles). A second is the **nucleotide diversity** π, the average per-site heterozygosity within the region of interest. Finally, one can consider the **site-frequency spectrum**, the counterpart of the *allele*-frequency spectrum. In this case, instead of counting the number of times each allele appears in the sample, one considers the number of *sites* n_j in the sample with j copies of a nucleotide. The

unfolded frequency spectrum refers to polarized nucleotides, with n_j being the number of sites with j copies of the *derived* nucleotide. For a **folded frequency spectrum**, n_j is the number of sites with j copies of the minor nucleotide, with $j \leq n/2$.

For the example data set in Figure 2.7, four of the six sites are monomorphic, while site 4 has 3 Gs and 2 As, and site 5 has 4 Cs and an A. For a folded frequency spectrum, this gives four sites in the zero class, one site in the one class, and one site in the two class ($n_0 = 4, n_1 = 1, n_2 = 1$). If the nucleotides were polarized, so that (for example) the ancestral states at the six sites were (respectively) A-A-G-G-G-A, the unfolded frequency spectrum would be three in the zero class (sites 1, 2, 3), one in the two class (site 4), one in the four class (site 6), and one in the five class (site 5), or ($n_0 = 3, n_2 = 1, n_4 = 1, n_5 = 1$). Many of the analyses using site-frequency spectrum data condition on only sites that are polymorphic in the sample.

For a very long region ($L \gg 1$) (again assuming neutrality, and mutation-drift equilibrium), the fraction of sites in the entire *population* that have a derived nucleotide at frequency x (for $0 < x < 1$) is given by the (unfolded) **Watterson** (1975) **distribution**,

$$\phi(x) = \frac{\theta}{x} \quad \text{for} \quad \frac{1}{2N} \leq x \leq 1 - \frac{1}{2N} \tag{2.34a}$$

This tells us that under the neutral model most sites are expected to have a low frequency of derived nucleotides in the population. Parameterizing Equation 2.34a for unpolarized alleles, so that $0 < x \leq 0.5$ where x is the minor-allele frequency, gives the **folded Watterson distribution**

$$\phi(x) = \frac{\theta}{x} + \frac{\theta}{1-x} = \frac{\theta}{x(1-x)} \quad \text{for} \quad \frac{1}{2N} \leq x \leq 1/2 \tag{2.34b}$$

The folded and unfolded frequency spectra are very similar over the range where both are defined ($0 < x \leq 0.5$), as high-frequency derived nucleotides are rare under the equilibrium neutral model, due to the fact that most new mutations are lost by drift.

The site-frequency spectrum for a *sample* is not the same as that for the entire population, but follows from the Watterson distribution. For a sample of size n, the number n_i of the L total sites with i derived nucleotides has expected value

$$E(n_i) = \frac{\theta_L}{i}, \quad \text{for} \quad 1 \leq i \leq n-1 \tag{2.35a}$$

(Fu 1995; Ewens 2004). Because $\theta_L = i\,E(n_i)$, Equation 2.35a motivates several infinite-site estimators of θ developed in Chapters 4 and 9. By using different regions of the site-frequency spectrum (i.e., different ranges for i) to estimate θ, various assumptions of the standard neutral model (e.g., constant population size, and absence of selection) can be tested (Chapter 9). Similarly, for a folded frequency spectrum, where n_i now denotes the number of sites with minor-nucleotide frequency i/n,

$$E(n_i) = \frac{\theta_L}{i} + \frac{\theta_L}{n-i} = \frac{\theta_L}{i}\,\frac{n}{n-i}, \quad \text{for} \quad 1 \leq i \leq (n/2) \tag{2.35b}$$

where $(n/2)$ denotes largest integer below or equal to $n/2$.

While the expectations given by Equation 2.35 can be used for method-of-moments estimators of θ (e.g., $\widehat{\theta} = i \cdot n_i$), likelihood estimators (LW Appendix 4) require the full distribution within a sample, not just the expected value. The probability of seeing exactly k derived nucleotides at a site follows from the binomial,

$$\Pr(k \,|\, x, n) = \binom{n}{k} x^k (1-x)^{n-k} \tag{2.36a}$$

In particular, the probability that a sample is polymorphic at a random site is just one minus the probability of a sample monomorphic for either the derived (x^n) or the ancestral $(1-x)^n$ nucleotide,

$$\Pr(1 \le k \le n-1) = 1 - x^n - (1-x)^n \tag{2.36b}$$

The probability of seeing k derived nucleotides at a random site in a sample is the average of Equation 2.36a over the possible x values,

$$\Pr(k \,|\, n) = \binom{n}{k} \int_{1/(2N)}^{1-1/(2N)} x^k (1-x)^{n-k} \phi(x) dx \tag{2.36c}$$

This formula, which forms the null model for several of the likelihood-based tests for a selective sweep (Chapter 8), is the infinite-sites analog to Ewens' sampling formula for the infinite-alleles model (both of which are functions of n and θ). Ewens' formula (Equation 2.30) gives the probability that k alleles are seen in a sample of size n, while Equation 2.36c gives the probability that a randomly-chosen site is segregating k derived nucleotides. The latter formula also defines the probability of k copies of a minor nucleotide under the folded spectrum if Equation 2.34b is used for $\phi(x)$.

THE GENEALOGICAL STRUCTURE OF A POPULATION

The preceding analyses show that a number of summary statistics, such as average levels of heterozygosity and linkage disequilibrium, are defined by the processes of drift, mutation, and recombination in predictable ways, at least for neutral sites. Provided we retain our focus on neutral regions of the genome, it is possible to go quite a bit further, even to the extent of predicting the expected *genealogical* relationships among different sequences sampled within populations. The basic theory, first laid out by Kingman (1982a,b) and now called **coalescent theory**, provides an elegant and powerful approach for solving problems in population genetics and molecular evolution. Kingman (2000) reviews the historical origins of this approach, and detailed overviews can be found in Hudson (1990), Donnelly and Tavaré (1995), Fu and Li (1999), Nordborg (2001), Stephens (2001), Rosenberg and Nordborg (2002), Hein et al. (2005), and Wakely (2006).

Because all of the genes within a population are direct products of past gametic sampling, they are all ultimately related in a genealogical sense. Thus, if one were to sample two alleles in a current population and then follow them back in time, both copies would eventually be traced to a single copy in an ancestral individual, at which point the two alleles are said to have **coalesced.** A key principle is that the

form of the expected gene genealogy for neutral genes, in particular the expected coalescence time, is *completely independent of the mutational process.*

Consider a random sample of n alleles drawn from a current population, assumed to obey all the properties of the idealized Wright-Fisher model, with no recombination within alleles. Focusing initially on just two of the sampled alleles, we first evaluate the probability that both members of the pair are direct copies of a single allele in the preceding generation. Assuming that each individual produces a large number of gametes, because there are $2N$ gene copies in the population each generation, this probability is simply $1/(2N)$, whereas $\lambda_1 = 1 - (1/2N)$ is the probability that coalescence occurs at some earlier generation. Conditional on coalescence not having occurred in generation one, the probability of coalescence two generations in the past is again equal to $1/(2N)$, yielding $\lambda_1(1/2N)$ as the unconditional probability of coalescence two generations back. This simple rule can be generalized to give the probability of coalescence exactly t generations in the past,

$$P_c(t) = \lambda_1^{t-1}(1/2N) \tag{2.37}$$

which defines a geometric distribution, with the sum of $P_c(t)$ over the interval $t = 1$ to ∞ being equal to one. One simple related point is that the probability that the **most recent common ancestor** between two sampled alleles occurred within the last t generations is $1 - \lambda_1^t \simeq 1 - e^{-t/2N}$, namely one minus the probability of no common ancestor over the first t generations into the past.

The average time to coalescence of two randomly sampled genes is simply

$$\bar{t}_c(2) = \sum_{t=1}^{\infty} t \cdot P_c(t) = 2N \tag{2.38}$$

Thus, the expected number of generations required for any two random alleles to trace back to an ancestral copy is simply equal to twice the population size (more precisely, twice the effective population size as defined in Chapter 3).

The logic used to derive this result is easily extended to the entire sample of n gene copies. There are $p_n = n(n-1)/2$ possible pairs of n copies, each of which will or will not coalesce in the preceding generation with respective probabilities $1/(2N)$ and $[1 - (1/2N)]$. If the sample size is much smaller than the population size, the probability of coalescence for any pair in the sample in the preceding generation is simply the product $p_n/(2N)$. Thus, the probability distribution for the coalescence time of one pair within a set of n sequences is

$$P_c(p_n, t) = [1 - (p_n/2N)]^{t-1}[p_n/(2N)] \tag{2.39}$$

The mean time to coalescence of the first pair is then $2N/p_n$ generations (as opposed to $2N$ generations with a single pair). Because at this point two copies have coalesced into one, the sample size has been reduced by one, and the mean time to coalescence of the next pair is found by resetting p_n to $p_{n-1} = (n-1)(n-2)/2$. This procedure can be followed recursively down to the final pair ($p_n = 1$), which again has an expected coalescence time of $2N$ generations (Figure 2.8). The implication of these results is that the expected time for merging n random lineages into $n-1$ lineages,

$$\bar{t}_n = 2N/p_n = \frac{4N}{n(n-1)} \tag{2.40}$$

increases with decreasing sample size.

-Insert Figure 2.8 Here-

The total expected genealogical depth of a sample, obtained by summing the expectations of each coalescence event, is

$$\bar{t}_c(n) = \sum_{i=2}^{n} \frac{4N}{i(i-1)} = 4N\left(1 - \frac{1}{n}\right) \tag{2.41}$$

Thus, under neutrality, the expected time to the most recent common ancestor of all alleles residing at a locus is $4N$ generations. This is equivalent to the mean time to fixation of a neutral mutation, as can be verified by substituting $p_0 = 1/(2N)$ into Equation 2.11b. Notably, the expected distance between the final two nodes in a neutral coalescent tree, $2N$, is at least half the coalescence time for the entire sample. Unfortunately, this fundamental issue is commonly ignored by those who invoke deep splits in a gene genealogy as evidence of an adaptive event.

It is important to note that all of the results in this section were derived without regard to any underlying genetic features of the sampled alleles. However, having determined the expected genealogical features of neutral gene sequences, it is straight-forward to incorporate genetic issues, as mutations will arise randomly along the branches of the genealogy in numbers proportional to time. For example, given the average $2N$ generations separating two randomly sampled alleles, the average number of mutations separating such genes is $2 \cdot 2N \cdot u = 4Nu$, the two arising because each copy is $2N$ generations removed from the common ancestor. Unlike the heterozygosity, which has a maximum value of 1.0, this measure of mutational divergence is unbounded, as it allows for the possibility of multiple mutations per copy.

One of the primary uses of the coalescent derives from its ability to efficiently generate sample distributions of quantities of interest (e.g., the expected pattern of molecular variation among neutral alleles), which provide a formal basis for statistical tests of various evolutionary models (Hudson 2002), including those involving selection (Chapter 9). Although all of the preceding results simply refer to the *expected* coalescent times, each individual coalescence time has considerable evolutionary variance, being equal to the square of its expected value (a feature of geometric distributions). As will be discussed in Chapter 9, a number of useful results have been obtained on the sampling distributions of coalescents, but it is also straight-forward to obtain such information by using computer simulations to construct random genealogies. For each simulated sample, one starts with n distinct lineages, picks two at random, and generates a value for t_n by randomly drawing from an exponential distribution with mean value given by Equation 2.40. After these two samples are joined in the coalescent tree, the remaining sample of $n-1$ distinct lineages is treated in the same way to generate a value of t_{n-1}, and so on, until the last two remaining lineages coalesce to yield t_2. For each branch of the resultant tree, the number of mutations is then drawn using a Poisson distribution with an expectation equal to the product of the mutation rate and the length of the

branch (in generations), e.g., the two branches emanating from the first node have independent numbers of mutations with expectation ut_n. By repeating such **coalescent simulations** several thousands of times, a distribution of interallelic variation under mutation and drift can then be acquired for any mutational model of interest, with the resultant data providing a null model for various tests of selection based on the pattern of variation in an actual sample (Chapter 9).

MUTATION-MIGRATION-DRIFT EQUILIBRIUM

Populations are often structured in space, with distinct **demes** connected by migration to form a **metapopulation**. The joint forces of mutation, migration, and drift structure the genetic variation both within and among demes, and if kept constant eventually lead to equilibrium values. The equilibrium neutral results serve as the framework for tests of abnormally high or low amounts of among-population variation (Chapter 9), corresponding respectively to diversifying and stabilizing selection.

Quantifying Population Structure: F_{ST}

The classic measure of population structure, F_{ST}, was defined by Wright (1943, 1951) as the correlation (identical by descent status) between alleles in different individuals from the same subpopulation. F_{ST} is a measure of the amount of inbreeding introduced by the population structure, being equal to the expected inbreeding coefficient f in a child from two random members from the same subpopulation. More importantly, F_{ST} also measures the fraction of total genetic variance due to differences between subpopulations (indeed, S stands for subpopulation and T for total population). This directly follows from the standard ANOVA identity that the among-group variance equals the within-group covariance (LW Chapter 18), with the later equal to the correlation among group members times the total variance. In particular, consider the distribution of the allele frequencies for a biallelic (B,b) locus over a set of populations. If p_0 denotes the average frequency of B over this set, its total variance is just $p_0(1-p_0)$. F_{ST} is then defined as the fraction of the total variance attributable to the variance in the frequency of allele B among demes ($\sigma^2(p)$),

$$F_{ST} = \frac{\sigma^2(p)}{p_0(1-p_0)}, \tag{2.42}$$

Wright was somewhat ambiguous in his use of F_{ST}, and some confusion has surrounded various interpretations of its true meaning. These are nicely cleaned up by Balding (2003).

Recalling Equation 2.14, which gives the expected variance in the allele frequency between two populations separated from a common ancestor t generations in the past, ignoring mutation and assuming no migration among groups, then

$$F_{ST} = \left[1 - \left(1 - \frac{1}{2N_e}\right)^t\right] \simeq \frac{t}{2N_e} \quad \text{for} \quad t \ll N_e \tag{2.43}$$

Under this model, F_{ST} eventually increases to one, as in the absence of mutation, drift eventually removes all variation within groups, so that only among-group variation remains, with the different demes becoming randomly fixed for alternative alleles. With recurrent mutation and gene flow among demes, however, neither the within- or the among-population variation is ever expected to reach absolute zero, and F_{ST} will be in the (0,1) interval, its magnitude depending on the relative impact of the three contributing forces (mutation, migration, and drift). A variety of methods, including extensions to multiple alleles, have been developed for estimating F_{ST} from samples of alleles from multiple subpopulations under the assumption of the infinite-alleles model (Nei and Chesser 1983; Weir and Cockerham 1984; Weir and Hill 2002; Balding 2003), and others allow for highly mutable alleles with a significant chance of back mutation (Slatkin 1995; Goodman 1997).

A powerful result is that F_{ST} values are closely related to the average coalescence times of alleles within a subpopulation, $\bar{t}_0$, and of alleles drawn randomly from the entire metapopulation, $\bar{t}$,

$$F_{ST} = \frac{\bar{t} - \bar{t}_0}{\bar{t}} = 1 - \frac{\bar{t}_0}{\bar{t}} \tag{2.44}$$

(Slaktin 1991). If $\bar{t}_0$ is very close to $\bar{t}$, there is little among-group differentiation, as pairs of alleles within groups have essentially the same amount of time to diverge mutationally as those among groups. Moreover, Equation 2.44 shows that, because the coalescent structure of a population simply represents a genealogical sampling process and is independent of the mutations that incidentally arise, the equilibrium level of population subdivision is independent of the mutation process.

-Insert Figure 2.8 Here-

Mutation-Migration-Drift Equilibrium Values of F_{ST}

When migration rates are sufficiently high that essentially all demes exchange at least one individual per generation, the physical structure of the metapopulation has no consequences for population subdivision at the genetic level, i.e., $F_{ST} \simeq 0$. On the other hand, in the absence of migration, subpopulations become fixed for different alleles, and F_{ST} approaches 1.0 as most genetic variation is due to among-group differences. When the forces of mutation, migration, and drift operate simultaneously at moderate levels, some intermediate equilibrium level of among-population differentiation is reached such that the loss of new variants within demes by drift is balanced by the spread of novel variants across demes by mutation. As one might expect, the resultant equilibrium F_{ST} value is a function of the population size within each deme and the pattern of migration over demes. There are, however, essentially endless numbers of possible patterns of spatial structure and migration, only two of which we will consider (Figure 2.8).

The simplest structure is Wright's (1951) **island model**, wherein the population consists of d demes, each containing N breeding individuals. Each generation, each deme contributes a fraction m of its genes to a migrant pool, yielding an expected migration rate from any deme to any other of $m/(d-1)$. A remarkable feature of this model is that the equilibrium amount of genetic variation expected within demes

is independent of the level of population subdivision (assuming $m > 0$), a feature known as the **geographic invariance principle** (Maruyama 1971; Nagylaki 1982). Provided there is some potential migratory route between all demes, regardless of the level of migration, the mean coalescence time between random pairs of genes within demes is

$$\bar{t}_0 = 2Nd \tag{2.45a}$$

generations, i.e., twice the sum of the demic population sizes (Li 1976; Strobeck 1987; Hey 1991; Nagylaki 2000). On the other hand, the mean coalescence time for two genes randomly drawn from the entire metapopulation is

$$\bar{t} = 2Nd + \frac{(d-1)^2}{2dm} \tag{2.45b}$$

(Li 1976; Slatkin 1991; Nei and Takahata 1993). While the amount of variation *within* each deme is independent of the magnitude of subdivision, the differentiation *between* demes clearly depends on m. Substituting these values into Equation 2.44 yields

$$F_{ST} = \frac{1}{1 + 4N\dfrac{md^2}{(1-d)^2}} \simeq \frac{1}{1+4Nm} \quad \text{for} \quad d \gg 1 \tag{2.46}$$

This shows that the equilibrium level of population subdivision is completely independent of the mutation rate and largely independent of the number of demes (d), provided the latter is at least moderately large.

The logic behind F_{ST} is readily extended to populations with a **hierarchical** structure (Figure 2.8). In the simplest hierarchical model, the metapopulation is arranged into a series of g groups each consisting of d demes. Within each group, the demes exchange migrants according to the island model with total rate of m_1 (so that $m_1/(d-1)$ is the rate that one deme sends migrants to any other particular deme within its group). There is also much less frequent exchange of migrants between the different groups, i.e., $m_2 \ll m_1$. With this type of structure, it is necessary to consider the degree to which the total genetic variation partitions into components within demes, among demes within a group, and among groups. Each of these components can again be expressed in terms of coalescence times.

Let $\bar{t}_0$ be the mean coalescence time for two individuals from the same deme, $\bar{t}_1$ for two individuals from the same group but different demes, and $\bar{t}_2$ for two individuals from different groups. As with the island model, the geographic invariance principle gives the mean coalescence time for two individuals from the same deme as $\bar{t}_0 = 2Ngd$. In order for two individuals in the same group but different demes to coalesce, they must first trace back to the same deme, which for $m_1 \ll m_2$ requires approximately $t_1' \simeq g(d-1)/(2m_1)$ generations, and then take an additional $\bar{t}_0$ generations to coalesce within that deme, giving $\bar{t}_1 = t_1' + \bar{t}_0$ (Slatkin and Voelm 1991). Finally, for individuals from different groups to coalesce, they first must trace back to the same group, which requires an average of $t_2' = (g-1)/(2m_2)$ generations, then trace back to the same deme within a group (t_1') and finally coalesce within that deme, giving $\bar{t}_2 = t_2' + t_1' + \bar{t}_0$.

Using these results, three F statistics define the partitioning of the total population variation, F_{DG} among demes within groups, F_{GT} among groups within the

total population, and F_{ST} among demes within the total population:

$$\begin{aligned} F_{DG} &= \frac{\bar{t}_1 - \bar{t}_0}{\bar{t}_1} \\ F_{GT} &= \frac{\bar{t}_2 - \bar{t}_1}{\bar{t}_2} \\ F_{ST} &= \frac{\bar{t}_2 - \bar{t}_0}{\bar{t}_2} \end{aligned} \tag{2.47a}$$

(Slatkin and Voelm 1991; Excoffier et al. 2009). Substituting in the values for the various mean coalescence times noted above, and assuming d and g moderately large and $m_2 \ll m_1$, results in further simplification,

$$\begin{aligned} F_{DG} &\simeq \frac{1}{1 + 4Nm_1} \\ F_{GT} \simeq F_{ST} &\simeq \frac{1}{1 + 4Ndm_2} \end{aligned} \tag{2.47b}$$

Note that these expressions are equivalent in form to those for the simple island model, in this case with population size N determining the variation of demes within groups and group size Nd determining that for groups within populations and demes within populations. Such coalescence approaches are readily extended to much more complex situations (e.g., Nordborg 1997; Nagylaki 1998; Pannell and Charlesworth 2000; Pannell 2003).

Literature Cited

Abramowitz, M., and I. A. Stegun. 1972. *Handbook of mathematical functions, with formulas, graphs, and mathematical tables.* Dover Publications, Inc., New York, NY. [2]

Balding, D. G. 2003. Likelihood-based inference for genetic correlation coefficients. *Theor. Pop. Biol.* 63: 221–230. [2]

Buri, P. 1956. Gene frequency in small populations of mutant *Drosophila. Evolution* 10: 367–402. [2]

Cockerham, C. C. 1984. Drift and mutation with a finite number of allelic states. *Proc. Natl. Acad. Sci. USA* 81: 530–534. [2]

Crow, J. F., and M. Kimura. 1970. *An introduction to population genetics theory.* Harper and Row, Publ., New York, NY. [2]

Darwin, C. 1859. *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life.* John Murray, London, UK. [2]

Donnelly, P., and S. Tavaré. 1995. Coalescents and genealogical structure under neutrality. *Ann. Rev. Genetics* 29: 401–421. [2]

Ewens, W. J. 1972. The sampling theory of selectively neutral alleles. *Theor. Pop. Biol.* 3: 87–112. [2]

Ewens, W. J. 2004. *Mathematical population genetics.* 2nd Edition. Springer-Verlag, Berlin, Germany. [2]

Excoffier, L., T. Hofer, and M. Foll. 2009. Detecting loci under selection in a hierarchically structured population. *Heredity* 103: 285–298. [2]

Fisher, R. A. 1922. On the dominance ratio. *Proc. Royal Soc. Edinburgh* 42: 321–341. [2]

Fu, Y.-X. 1995. Statistical properties of segregating sites. *Theor. Pop. Biol.* 48: 172-197. [2]

Fu, Y.-X., and W. H. Li. 1999. Coalescing into the 21st century: an overview and prospects of coalescent theory. *Theor. Pop. Biol.* 56: 1–10. [2]

Gale, J. S. 1990. *Theoretical population genetics.* Unwin Hyman Ltd., London, UK. [2]

Golding, G. B, and C. Strobeck. 1980. Linkage disequilibrium in a finite population that is partially selfing. Genetics 94: 777–789. [2]

Goodman, S. J. 1997. R_{st} Calc: a collection of computer programs for calculating estimates of genetic differentiation from microsatellite data and determining their significance. *Mol. Ecol.* 6: 881–885. [2]

Hein, J., M. H. Schierup, and C. Wiuf. 2005. *Gene genealogies, variation and evolution – a primer in coalescent theory.* Oxford Univ. Press, Oxford, UK. [2]

Hey, J. 1991. A multi-dimensional coalescent process applied to multi-allelic selection models and migration models. *Theor. Pop. Biol.* 39: 30–48. [2]

Hill, W. G. 1975. Linkage disequilibrium among multiple neutral alleles produced by mutation in finite population. *Theor. Pop. Biol.* 8: 117–126. [2]

Hill, W. G., and A. Robertson. 1966. The effect of linkage on limits to artificial selection. *Genet. Res.* 8: 269–294. [2]

Hill, W. G., and A. Robertson. 1968. Linkage disequilibrium in finite populations. *Theor. Appl. Genetics* [2] 38: 226–231. [2]

Hill, W. G., and B. S. Weir. 1988. Variances and covariances of squared linkage disequilibria in finite populations. *Theor. Pop. Biol.* 33: 54–78. [2]

Hudson, R. R. 1990. Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology* 7: 1–43. [2]

Hudson, R. R. 2002. Generating samples under a Wright-Fisher neutral model. *Bioinformatics* 18: 337–338. [2]

Johnson, N. L., A. W. Kemp, and S. Kotz. 2005. *Univariate discrete distributions.* John Wiley & Sons, Inc., Hoboken, NJ. [2]

Karlin, S., and J. L. McGregor. 1972. Addendum to a paper of W. Ewens. *Theor. Pop. Biol.* 3: 113–116. [2]

Kimura, M. 1955. Solution of a process of random genetic drift with a continuous model. *Proc. Natl. Acad. Sci. USA* 41: 144–150. [2]

Kimura, M. 1968. Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. *Genet. Res.* 11: 247–269. [2]

Kimura, M. 1970. The length of time required for a selectively neutral mutant to reach fixation through random frequency drift in a finite population. *Genet. Res.* 15: 131–133. [2]

Kimura, M. 1983. *The neutral theory of molecular evolution.* Cambridge Univ. Press, Cambridge, UK. [2]

Kimura, M. and J. F. Crow. 1964. The number of alleles that can be maintained in a finite population. *Genetics* 49: 725–738. [2]

Kimura, M., and T. Ohta. 1969. The average number of generations until fixation of a mutant gene in a finite population. *Genetics* 61: 763–771. [2]

Kimura, M., and T. Ohta. 1973. The age of a neutral mutation persisting in a finite population. *Genetics* 75: 199–212. [2]

Kingman, J. F. C. 1982a. The coalescent. *Stoch. Proc. Appl.* 13: 235–248. [2]

Kingman, J. F. C. 1982b. On the genealogy of large populations. *J. Appl. Prob.* 19: 27–43. [2]

Kingman, J. F. C. 2000. Origins of the coalescent 1974–1982. *Genetics* 156: 1461–1463. [2]

Lessard, S. 1981. Is the between-population variance negligible in the total variance of heterozygosity? Case of a finite number of loci subject to the infinite-allele model in finite monoecious populations. *Theor. Pop. Biol.* 20: 394–410. [2]

Li, W.-H. 1976. Distribution of nucleotide differences between two randomly chosen cistrons in a subdivided population: the finite island model. *Theor. Pop. Biol.* 10: 303–308. [2]

Li, W.-H., and M. Nei. 1975. Drift variances of heterozygosity and genetic distance in transient states. *Genet. Res.* 25: 229–248. [2]

Lynch, M., and W. G. Hill. 1986. Phenotypic evolution by neutral mutation. *Evolution* 40: 915–935. [2]

Maruyama, T. 1971. An invariant property of subdivided populations. *Genet. Res.* 18: 81–84. [2]

Maruyama, T. 1974. The age of an allele in a finite population. *Genet. Res.* 23: 137–143. [2]

Moran, P. A. P. 1962. *The statistical processes of evolutionary theory.* Clarendon Press, Oxford, UK. [2]

Nagylaki, T. 1982. Geographical invariance in population genetics. *J. Theor. Biol.* 99: 159–172. [2]

Nagylaki, T. 1998. The expected number of heterozygous sites in a subdivided population. *Genetics* 149: 1599–1604. [2]

Nagylaki, T. 2000. Geographical invariance and the strong-migration limit in subdivided populations. *J. Math. Biol.* 41: 123–142. [2]

Nei, M., and R. K. Chesser. 1983. Estimation of fixation indices and gene diversities. Ann. Hum. Genet. 47: 253–259. [2]

Nei, M., and S. Kumar. 2000. *Molecular evolution and phylogenetics.* Oxford Univ. Press, Oxford, UK. [2]

Nei, M., and W.-H. Li. 1976. The transient distribution of allele frequencies under mutation pressure. *Genet. Res.* 28: 205–214. [2]

Nei, M., and N. Takahata. 1993. Effective population size, genetic diversity, and coalescence time in subdivided populations. *J. Mol. Evol.* 37: 240–244. [2]

Nordborg, M. 1997. Structured coalescent processes on different time scales. *Genetics* 146: 1501–1514. [2]

Nordborg, M. 2001. Introduction to coalescent theory *In* D. Balding, M. Bishop, and C. Cannings (eds.), *Handbook of statistical genetics*, pp. 179–212. Wiley, Chichester, UK. [2]

Ohta, T., and M. Kimura. 1969. Linkage disequilibrium due to random genetic drift. *Genet. Res.* 13: 47–55. [2]

Ohta, T., and M. Kimura. 1971. Linkage disequilibrium between two segregating nucleotide sites under the steady flux of mutations in a finite population. *Genetics* 68: 571–580. [2]

Pannell, J. R. 2003. Coalescence in a metapopulation with recurrent local extinction and recolonization. *Evolution* 57: 949–961. [2]

Pannell, J. R., and B. Charlesworth. 2000. Effects of metapopulation processes on measures of genetic diversity. *Phil. Trans. R. Soc. B* 355: 1851–1864. [2]

Rosenberg, N. A., and M. Nordborg. 2002. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nat. Rev. Genet.* 3: 380–390. [2]

Slatkin, M. 1991. Inbreeding coefficients and coalescent times. *Genet. Res.* 58: 167–175. [2]

Slatkin, M. 1995. A measure of population subdivision based on microsatellite allele frequencies. *Genetics* 139: 457–462. [2]

Slatkin, M., and B. Rannala. 1997. Estimating the age of alleles by the use of intraallelic

variability. *Am. J. Hum. Genet.* 60: 447–458. [2]

Slatkin, M., and B. Rannala. 2000. Estimating allele age. *Ann. Rev. Genomics Hum. Genet.* 1: 225–249. [2]

Slatkin, M., and L. Voelm. 1991. F_{ST} is a hieracrchial island model. *Genetics* 127: 627– 629. [2]

Song, Y. S., and J. S. Song. 2007. Analytic computation of the expectation of the linkage disequilibrium coefficient r^2. *Theor. Pop. Biol.* 71: 49–60. [2]

Stephens, J. C., D. E Reich, D. B Goldstein, H. D. Shin, M. W. Smith, M. Carrington, C. Winkler, G. A. Huttley, R. Allikmets, L. Schriml, et al. 1998. Dating the origin of the *CCR5-Delta32* AIDS-resistance allele by the coalescence of haplotypes. *Am. J. Hum. Genet.* 62: 1507-1515. [2]

Stephens, M. 2001. Inference under the coalescent *In* D. Balding, M. Bishop, and C. Cannings (eds.), *Handbook of statistical genetics*, pp. 213–228. Wiley, Chichester, UK. [2]

Stewart, F. M. 1976. Variability in the amount of heterozygosity maintained by neutral mutations. *Theor. Pop. Biol.* 9: 188–201. [2]

Strobeck, C. 1987. Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision. *Genetics* 117: 149–153. [2]

Strobeck, C., and K. Morgan. 1978. The effect of intragenic recombination on the number of alleles in a finite population. *Genetics* 88: 829–844. [2]

Sved, J. A. 1971. Linkage disequilibrium and homozygosity of chromosome segments in finite populations. Theor. Pop. Bio. 2: 125–141. [2]

Sved, J. A., and M. W. Feldman. 1973. Correlation and probability methods for one and two loci. Theor. Pop. Bio. 42: 129–132. [2]

Wakely, J. 2006. *An introduction to coalescent theory.* Roberts & Co. Publ., Greenwood Village, CO. [2]

Watterson, G. A. 1975. On the number of segregation sites. *Theor. Pop. Biol.* 7: 256–276. [2]

Watterson, G. A. 1976. Reversibility and the age of an allele. I. Moran's infinitely many neutral alleles model. *Theor. Pop. Biol.* 10: 239–253. [2]

Weir, B. S., and C. C. Cockerham. 1984. Estimating F-statistics for the analysis of population structure. *Evolution* 38: 1358–1370. [2]

Weir, B. S., and W. G. Hill. 2002. Estimating F-statistics. *Ann. Rev. Genet.* 36: 721–750. [2]

Wright, S. 1931. Evolution in Mendelian populations. *Genetics* 16: 97–159. [2]

Wright, S., 1943. Isolation by distance. *Genetics* 28: 114?-138. [2]

Wright, S. 1951. The genetical structure of populations. *Ann. Eugen.* 15: 323–354. [2]

Figure 2.1. Expected age of a neutral allele, given its frequency p (Equation 2.12). Time is scaled in units of $4N_e$ generations.

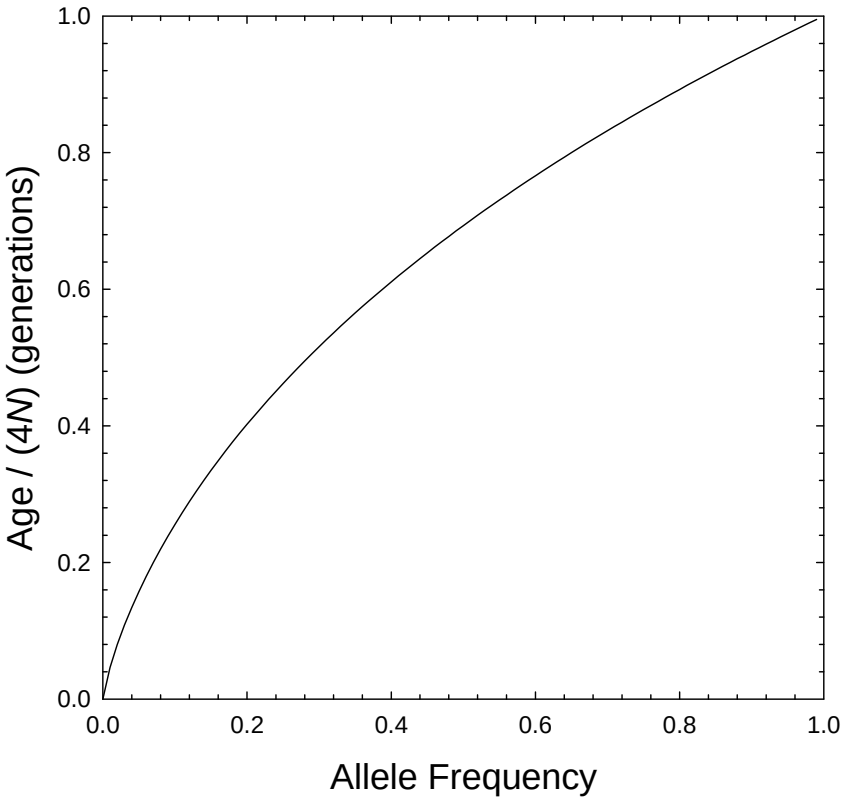

Figure 2.2. Expected probability distributions for the frequencies of *segregating* neutral alleles in replicate, randomly mating populations of size N after t generations of divergence (fixed alleles are ignored). The initial allele frequency in the base population is 0.5 on the left and 0.1 on the right. The abscissa is the population allele frequency, while the ordinate is proportional to the probability of occurrence of that frequency. Note that the time scale is in units of N generations, where N is the population size, so that $t = N$ generations implies 100 generations for a population of size 100 and 10,000 generations for a population of size 10,000. (From Kimura 1955).

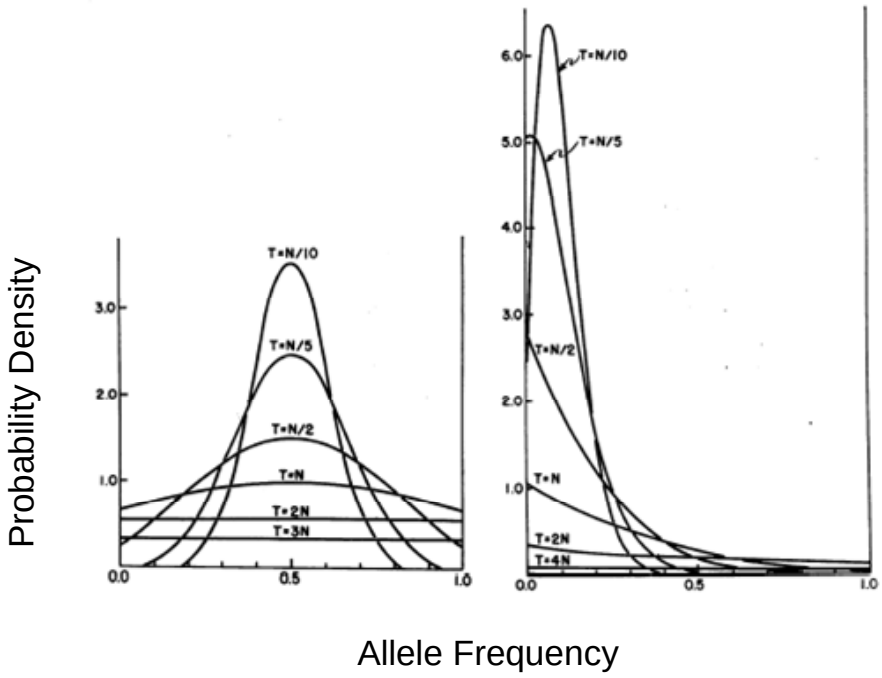

Figure 2.3. Patterns of change in the frequencies of the bw^{75} allele in 212 isolated populations of *Drosophila melanogaster,* each consisting of 8 breeding males and 8 females. **Top**: The average allele frequency over the entire pool of populations. The dotted and solid lines respectively denote upward deviations of two standard errors from the expected value of $p_0 = 1/2$ under the assumption of effective population sizes of 16 and 10.2 individuals. **Middle**: Mean observed heterozygosity compared to the expectations assuming an effective population size of 10.2. The expected heterozygosity is 0.5 in generations 1 and 2 because the base population (generation 0) consisted entirely of heterozygotes, and with separate sexes, an additional generation is required for the unification of alleles that are identical by descent. **Bottom**: Among-line variance of allele frequencies compared with their expectations assuming an effective population size of 10.2. (From Buri 1956).

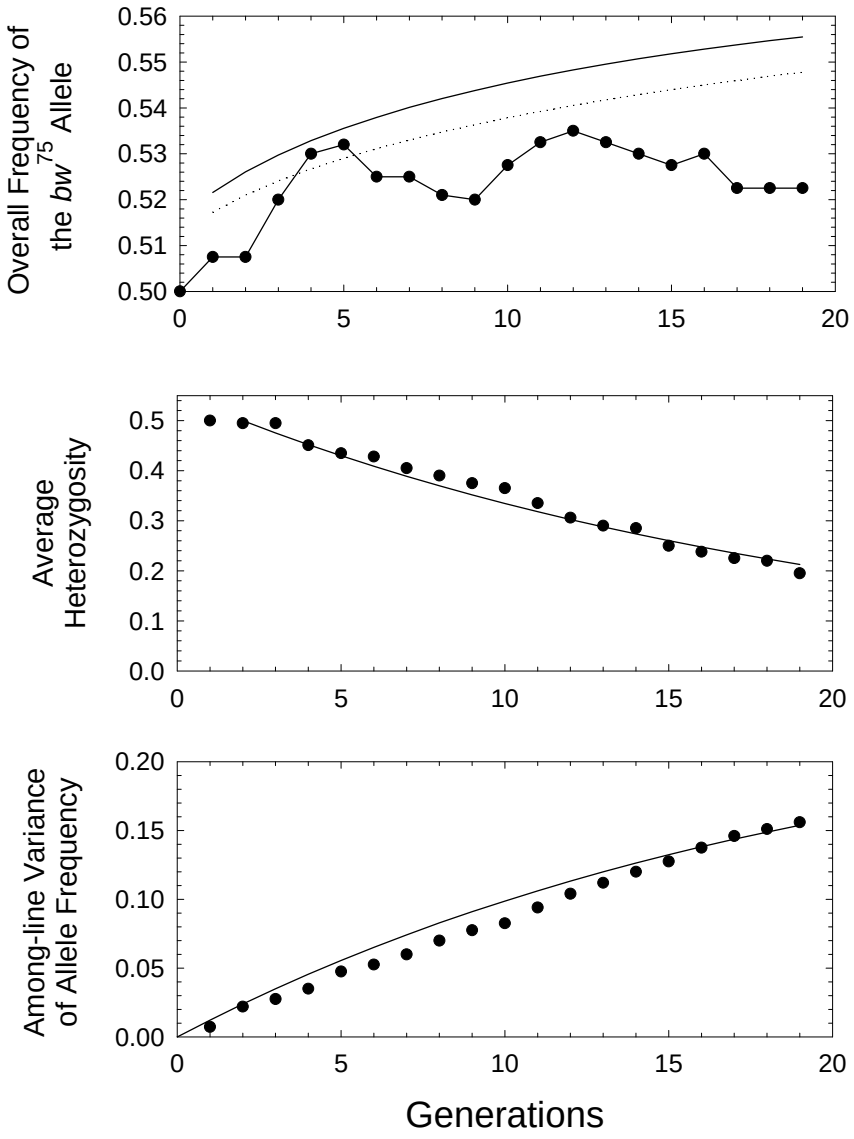

Figure 2.4. Distribution of the number of bw^{75} alleles in 212 populations of *D. melanogaster* each initiated with a frequency of 0.5. Two features are represented in this temporal series: the distribution of frequencies for segregating alleles (1 to 31 copies), the expected form of which is given in Figure 2.2, and the accumulation of fixed alleles (0 or 32 copies). (From Buri 1956).

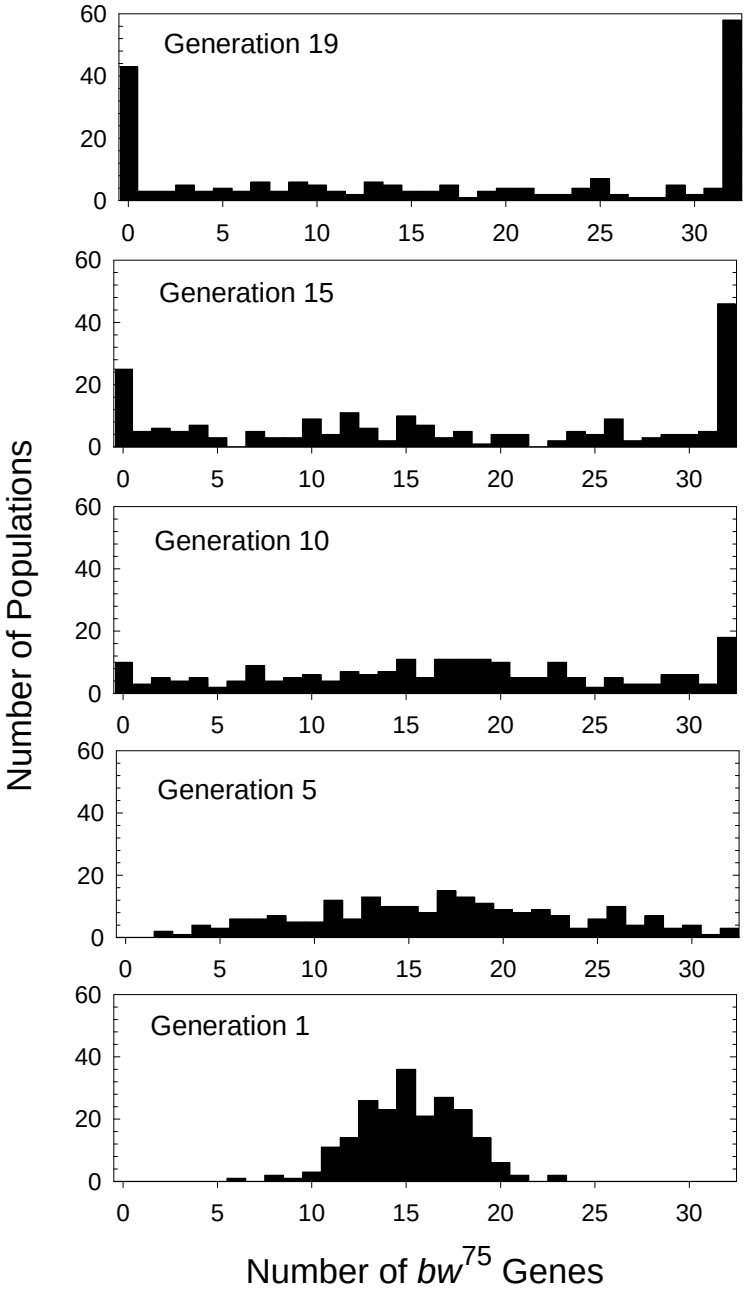

Figure 2.5. Observed sampling variances of allele frequencies for situations in which the donor population contained 1 to 31 bw^{75} genes. The dotted line is the expected pattern, $p(1-p)/2N$, if the actual populations of 8 males and 8 females were randomly mating with equal chances of contributing offspring. The solid line describes the pattern for an average effective population size of 10.2. (From Buri 1956).

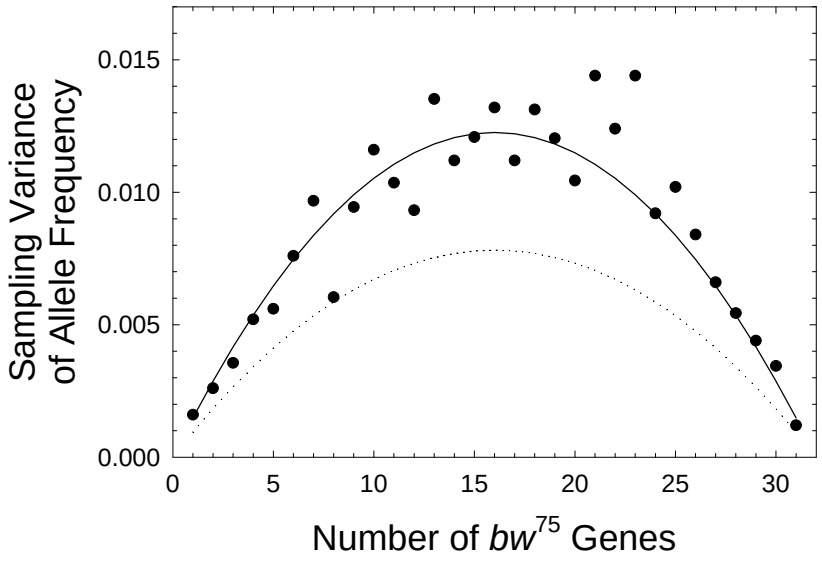

Figure 2.6. Mean heterozygosity and its standard deviation (SD) among replicate populations as a function of time (scaled in units of $2N$ generations). The upper curves assume an initial heterozygosity of $H_0 = 0.5$ and the lower curves of $H_0 = 0.2$. A diallelic locus is assumed, and new variation generated by mutation is ignored. Experimental error resulting from sampling of a finite number of individuals is ignored as well, i.e., we consider only the variance of true population-level heterozygosities resulting from gamete sampling. The expected heterozygosities (solid lines) are obtained with Equation 2.5, whereas the standard deviations (dotted lines) follow from Equation 2.15.

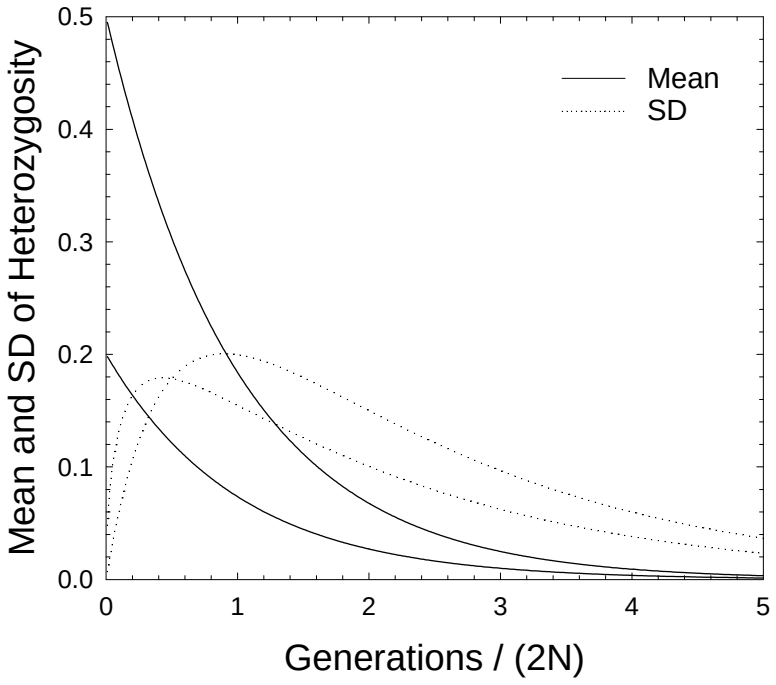

Figure 2.7. An example of the difference between the infinite-alleles and infinite-sites models. Five sequences (horizontal rows) scored at six nucleotide sites are sampled from a population. Variants are denoted in grey. Three of these five sequences are different and are scored as three alleles (or haplotypes) under an infinite-alleles framework (sequences 1 and 3; 2 and 4; and 5). Conversely, only two of the six sites are segregating (4 and 6), giving two polymorphic sites under an infinite-sites framework.

Figure 2.8. Expected coalescence times for a sample of $n = 5$ neutral genes taken from an idealized Wright-Fisher population of size N. The number of gene pairs in each consecutive step of the coalescent process is denoted by p_n, and the expected times to coalescence at each step are equal to $2N/p_n$ generations. The particular lineages that join during each step are arbitrary. Note that over half of the coalescent time for the total lineage of five samples involves the coalescent event between the final branches.

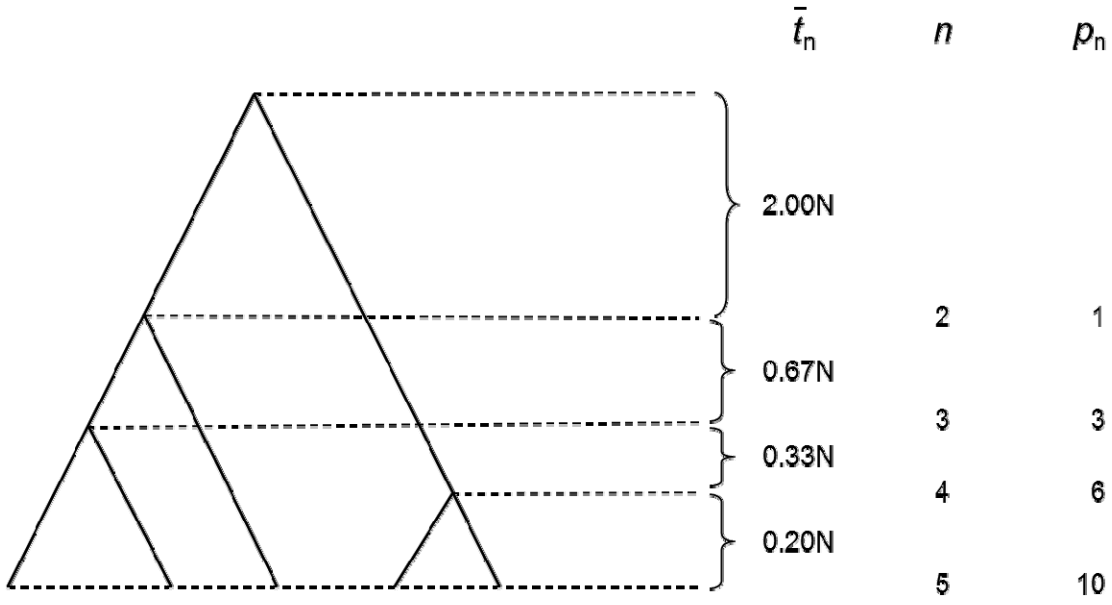

Figure 2.9. Left) The **island model**. Most mating occurs within each subpopulation/deme, but some small amount of equal migration m occurs between them. Hence, a migrant from deme A is equally likely to end up in demes B through D. Right) A **hierarchically-structured population**. Here, migration between A and B and between C and D (m_1) occurs at a much higher level than migration between these two groups (m_2).