

## CHAPTER 2

# Genotype frequencies

### 2.1 Mendel's model of particulate genetics

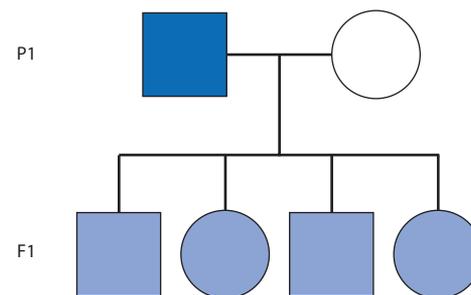
- Mendel's breeding experiments.
- Independent assortment of alleles.
- Independent segregation of loci.
- Some common genetic terminology.

In the nineteenth century there were several theories of heredity, including inheritance of acquired characteristics and blending inheritance. Jean-Baptiste Lamarck is most commonly associated with the discredited hypothesis of inheritance of acquired characteristics (although it is important to recognize his efforts in seeking general causal explanations of evolutionary change). He argued that individuals contain "nervous fluid" and that organs or features (phenotypes) employed or exercised more frequently attract more nervous fluid, causing the trait to become more developed in offspring. His widely known example is the long neck of the giraffe, which he said developed because individuals continually stretched to reach leaves at the tops of trees. Later, Charles Darwin and many of his contemporaries subscribed to the idea of blending inheritance. Under blending inheritance, offspring display phenotypes that are an intermediate combination of parental phenotypes (Fig. 2.1).

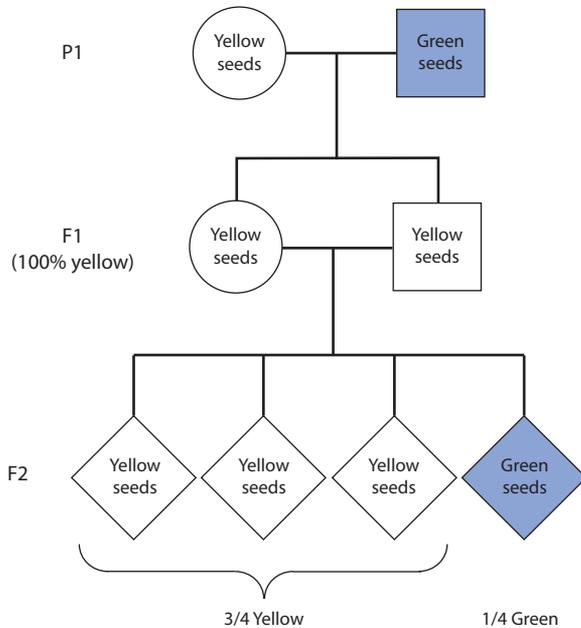
From 1856 to 1863, the Augustinian monk Gregor Mendel carried out experiments with pea plants that demonstrated the concept of particulate inheritance. Mendel showed that phenotypes are determined by discrete units that are inherited intact and unchanged through generations. His hypothesis was sufficient to explain three common observations: (i) phenotype is sometimes identical between parents and offspring; (ii) offspring phenotype can differ from that of the parents; and (iii) "pure" phenotypes of earlier generations could skip generations and reappear in later generations. Neither blending inheritance nor inheritance of acquired characteristics are satisfactory explanations for all of these observations. It

is hard for us to fully appreciate now, but Mendel's results were truly revolutionary and served as the very foundation of population genetics. The lack of an accurate mechanistic model of heredity severely constrained biological explanations of cause and effect up to the point that Mendel's results were "rediscovered" in the year 1900.

It is worthwhile to briefly review the experiments with pea plants that Mendel used to demonstrate independent assortment of both alleles within a locus and of multiple loci, sometimes dubbed Mendel's first and second laws. We need to remember that this was well before the Punnett square (named after Reginald C. Punnett), which originated in about 1905. Therefore, the conceptual tool we would use now to predict progeny genotypes from parental genotypes was a thing of the future. So in revisiting Mendel's experiments we will not use the Punnett square in an attempt to follow his logic. Mendel only observed the phenotypes of generations of pea plants that he had hand-pollinated. From these phenotypes and their patterns of inheritance he inferred the



**Figure 2.1** The model of blending inheritance predicts that progeny have phenotypes that are the intermediate of their parents. Here "pure" blue and white parents yield light blue progeny, but these intermediate progeny could never themselves be parents of progeny with pure blue or white phenotypes identical to those in the P1 generation. Crossing any shade of blue with a pure white or blue phenotype would always lead to some intermediate shade of blue. By convention, in pedigrees females are indicated by circles and males by squares, whereas P refers to parental and F to filial.



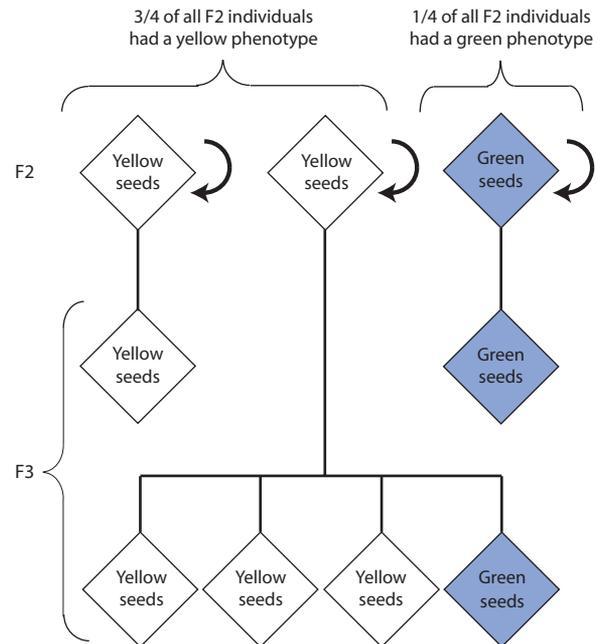
**Figure 2.2** Mendel's crosses to examine the segregation ratio in the seed coat color of pea plants. The parental plants (P1 generation) were pure breeding, meaning that if self-fertilized all resulting progeny had a phenotype identical to the parent. Some individuals are represented by diamonds since pea plants are hermaphrodites and can act as a mother, a father, or can self-fertilize.

existence of heritable factors. His experiments were actually both logical and clever, but are now taken for granted since the basic mechanism of particulate inheritance has long since ceased to be an open question. It was Mendel who established the first and most fundamental prediction of population genetics: expected genotype frequencies.

Mendel used pea seed coat color as a phenotype he could track across generations. His goal was to determine, if possible, the general rules governing inheritance of pea phenotypes. He established "pure"-breeding lines (meaning plants that always produced progeny with phenotypes like themselves) of peas with both yellow and green seeds. Using these pure-breeding lines as parents, he crossed a yellow- and a green-seeded plant. The parental cross and the next two generations of progeny are shown in Fig. 2.2. Mendel recognized that the F1 plants had an "impure" phenotype because of the F2 generation plants, of which three-quarters had yellow and one-quarter had green seed coats.

His insightful next step was to self-pollinate a sample of the plants from the F2 generation (Fig. 2.3). He considered the F2 individuals with yellow and green seed coats separately. All green-seeded F2 plants produced green progeny and thus were "pure"

green. However, the yellow-seeded F2 plants were of two kinds. Considering just the yellow F2 seeds, one-third were pure and produced only yellow-seeded progeny whereas two-thirds were "impure" yellow since they produced both yellow- and green-seeded progeny. Mendel combined the frequencies of the F2 yellow and green phenotypes along with the frequencies of the F3 progeny. He reasoned that three-quarters of all F2 plants had yellow seeds but these could be divided into plants that produced pure yellow F3 progeny (one-third) and plants that produced both yellow and green F3 progeny (two-thirds). So the ratio of pure yellow to impure yellow in the F2 was  $(\frac{1}{3} \times \frac{3}{4}) = \frac{1}{4}$  pure yellow to  $(\frac{2}{3} \times \frac{3}{4}) = \frac{1}{2}$  "impure" yellow. The green-seeded progeny comprised one-quarter of the F2 generation and all produced green-seeded progeny when self-fertilized, so that  $(1 \times \frac{1}{4} \text{ green}) = \frac{1}{4}$  pure green. In total, the ratios of phenotypes in the F2 generation were 1 pure yellow : 2 impure yellow : 1 pure green or 1 : 2 : 1. Mendel reasoned that "the ratio of 3 : 1 in which the distribution of the dominating and recessive traits take place in the first generation therefore resolves itself into the ratio of 1 : 2 : 1 if one differentiates the



**Figure 2.3** Mendel self-pollinated (indicated by curved arrows) the F2 progeny produced by the cross shown in Figure 2.2. Of the F2 progeny that had a yellow phenotype (three-quarters of the total), one-third produced all progeny with a yellow phenotype and two-thirds produced progeny with a 3 : 1 ratio of yellow and green progeny (or three-quarters yellow progeny). Individuals are represented by diamonds since pea plants are hermaphrodites.

meaning of the dominating trait as a hybrid and as a parental trait" (quoted in Orel 1996). During his work, Mendel employed the terms "dominating" (which became dominant) and "recessive" to describe the manifestation of traits in impure or heterozygous individuals.

With the benefit of modern symbols of particulate heredity, we could diagram Mendel's monohybrid cross with pea color in the following way.

P1	Phenotype	Yellow × green
	Genotype	GG      gg
	Gametes produced	G      g
F1	Phenotype	All "impure" yellow
	Genotype	Gg
	Gametes produced	G, g

A Punnett square could be used to predict the phenotypic ratios of the F2 plants:

	G	g
G	GG	Gg
g	Gg	gg

F2	Phenotype	3 Yellow : 1 green
	Genotype	GG      Gg      gg
	Gametes produced	G      G, g      g

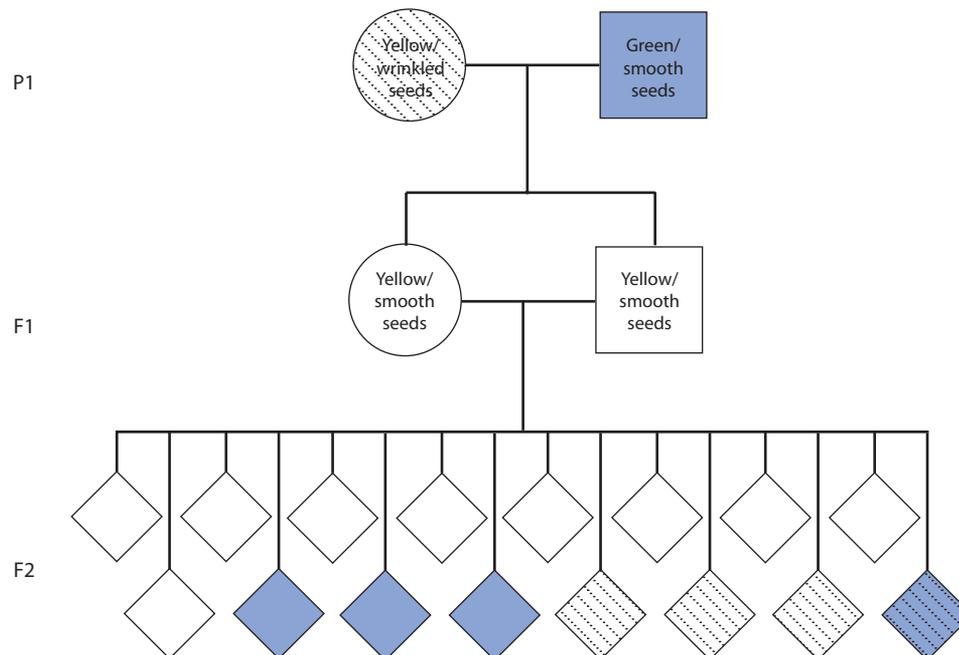
and another Punnett square could be used to predict the genotypic ratios of the two-thirds of the yellow F2 plants:

	G	g
G	GG	Gg
g	Gg	gg

**Mendel's first "law"** Predicts independent segregation of alleles at a single locus: two members of a gene pair (alleles) segregate separately into gametes so that half of the gametes carry one allele and the other half carry the other allele.

Individual pea plants obviously have more than a single phenotype and Mendel followed the inheritance of other characters in addition to seed coat color. In one example of his crossing experiments, Mendel tracked the simultaneous inheritance of both seed coat color and seed surface condition (either wrinkled ("angular") or smooth). He constructed an initial cross among pure-breeding lines identical to what he had done when tracking seed color inheritance, except now there were two phenotypes (Fig. 2.4).

**Figure 2.4** Mendel's crosses to examine the segregation ratios of two phenotypes, seed coat color (yellow or green) and seed coat surface (smooth or wrinkled), in pea plants. The hatched pattern indicates wrinkled seeds while white indicates smooth seeds. The F2 individuals exhibited a phenotypic ratio of 9 round/yellow : 3 round/green : 3 wrinkled/yellow : 1 wrinkled/green.



The F<sub>2</sub> progeny appeared in the phenotypic ratio of 9 round/yellow : 3 round/green : 3 wrinkled/yellow : 1 wrinkled/green.

How did Mendel go from this F<sub>2</sub> phenotypic ratio to the second law? He ignored the wrinkled/smooth phenotype and just considered the yellow/green seed color phenotype in self-pollination crosses of F<sub>2</sub> plants just like those for the first law. In the F<sub>2</sub> progeny, 12/16 or three-quarters had a yellow seed coat and 4/16 or one-quarter had a green seed coat, or a 3 yellow : 1 green phenotypic ratio. Again using self-pollination of F<sub>2</sub> plants like those in Fig. 2.3, he showed that the yellow phenotypes were one-quarter pure and one-half impure yellow. Thus, the segregation ratio for seed color was 1 : 2 : 1 and the wrinkled/smooth phenotype did not alter this result. Mendel obtained an identical result when focusing instead on the wrinkled/smooth phenotype and ignoring the seed color phenotype.

Mendel concluded that a phenotypic segregation ratio of 9 : 3 : 3 : 1 is the same as combining two independent 3 : 1 segregation ratios of two phenotypes since  $(3 : 1) \times (3 : 1) = 9 : 3 : 3 : 1$ . Similarly, multiplication of two (1 : 2 : 1) phenotypic ratios will predict the two phenotype ratio  $(1 : 2 : 1) \times (1 : 2 : 1) = 1 : 2 : 1 : 2 : 4 : 2 : 1 : 2 : 1$ . We now recognize that dominance in the first two phenotype ratios masks the ability to distinguish some of the homozygous and heterozygous genotypes, whereas the ratio in the second case would result if there was no dominance. You can confirm these conclusions by working out a Punnett square for the F<sub>2</sub> progeny in the two-locus case.

**Mendel's second "law"** Predicts independent assortment of multiple loci: during gamete formation, the segregation of alleles of one gene is independent of the segregation of alleles of another gene.

Mendel performed similar breeding experiments with numerous other pea phenotypes and obtained similar results. Mendel described his work with peas and other plants in lectures and published it in 1866 in the *Proceedings of the Natural Science Society of Brnn* in German. It went unnoticed for nearly 35 years. However, Mendel's results were eventually recognized and his paper was translated into several languages. Mendel's rediscovered hypothesis of particulate

inheritance was also bolstered by evidence from microscopic observations of cell division that led Walter Sutton and Theodor Boveri to propose the chromosome theory of heredity in 1902.

Much of the currently used terminology was coined as the field of particulate genetics initially developed. Therefore, many of the critical terms in genetics have remained in use for long periods of time. However, the meanings and connotations of these terms have often changed as our understanding of genetics has also changed.

Unfortunately, this has led to a situation where words can sometimes mislead. A common example is equating *gene* and *allele*. For example, it is commonplace for news media to report scientific breakthroughs where a "gene" has been identified as causing a particular phenotype, often a debilitating disease. Very often what is meant in these cases is that an *allele* with the phenotypic effect has been identified. Both unaffected and affected individuals all possess the gene, but they differ in their alleles and therefore in their genotype. If individuals of the same species really differed in their gene content (or loci they possessed), that would provide evidence of additions or deletions to genomes. For an interesting discussion of how terminology in genetics has changed – and some of the misunderstandings this can cause – see Judson (2001).

**Gene** Unit of particulate inheritance; in contemporary usage usually means an exon or series of exons, or a DNA sequence that codes for an RNA or protein.

**Locus** (plural **loci**, pronounced "low-sigh") Literally "place" or location in the genome; in contemporary usage is the most general reference to *any* sequence or genomic region, including non-coding regions.

**Allele** Variant or alternative form of the DNA sequence at a given locus.

**Genotype** The set of alleles possessed by an individual at one locus; the genetic composition of an individual at one or many loci.

**Phenotype** The morphological, biochemical, physiological, and behavioral attributes of an individual; synonymous with character, trait.

**Dominant** Where the expressed phenotype of one allele takes precedence over the expressed phenotype of another allele. The allele associated with the expressed phenotype is said to be dominant. Dominance is seen on a continuous scale that ranges between “complete” dominance (one allele completely masks the phenotype of another allele so that the phenotype of a heterozygote is identical to a homozygote for the dominant allele), “partial,” or “incomplete” dominance (masking effect is incomplete so that the phenotype of a heterozygote is intermediate to both homozygotes) and includes over- and under-dominance (phenotype is outside the range of phenotypes seen in the homozygous genotypes). The lack of dominance (heterozygote is exactly intermediate to phenotypes of both homozygotes) is sometimes termed “codominance” or “semi-dominance.”

**Recessive** The expressed phenotype of one allele is masked by the expressed phenotype of another allele. The allele associated with the concealed phenotype is said to be recessive.

## 2.2 Hardy–Weinberg expected genotype frequencies

- Hardy–Weinberg and its assumptions.
- Each assumption is a population genetic process.
- Hardy–Weinberg is a null model.
- Hardy–Weinberg in haplo-diploid systems.

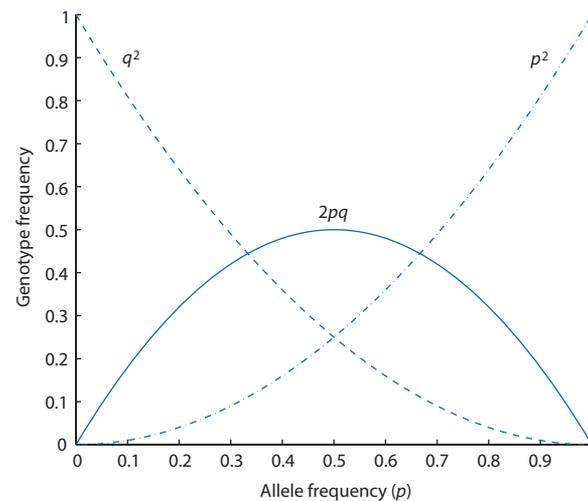
Mendel’s “laws” could be called the original expectations in population genetics. With the concept of particulate genetics established, it was possible to make a wide array of predictions about genotype and allele frequencies as well as the frequency of phenotypes with a one-locus basis. Still, progress and insight into particulate genetics was gradual. Until 1914 it was generally believed that rare

(infrequent) alleles would disappear from populations over time. Godfrey H. Hardy (1908) and Wilhelm Weinberg (1908) worked independently to show that the laws of Mendelian heredity did not predict such a phenomenon (see Crow 1988). In 1908 they both formulated the relationship that can be used to predict allele frequencies given genotype frequencies or predict genotype frequencies given allele frequencies. This relationship is the well-known Hardy–Weinberg equation

$$p^2 + 2pq + q^2 = 1 \quad (2.1)$$

where  $p$  and  $q$  are allele frequencies for a genetic locus with two alleles.

Genotype frequencies predicted by the Hardy–Weinberg equation can be summarized graphically. Figure 2.5 shows Hardy–Weinberg expected genotype frequencies on the  $y$  axis for each genotype for any given value of the allele frequency on the  $x$  axis. Another graphical tool to depict genotype and allele frequencies simultaneously for a single locus with two alleles is the De Finetti diagram (Fig. 2.6). As we will see, De Finetti diagrams are helpful when examining how population genetic processes dictate allele and genotype frequencies. In both graphs it is apparent that heterozygotes are most frequent when the frequency of the two alleles is equal to 0.5. You can also see that when an allele is rare, the corresponding homozygote genotype is even rarer since the genotype frequency is the square of the allele frequency.



**Figure 2.5** Hardy–Weinberg expected genotype frequencies for AA, Aa, and aa genotypes ( $y$  axis) for any given value of the allele frequency ( $x$  axis). Note that the value of the allele frequency not graphed can be determined by  $q = 1 - p$ .

### Interact box 2.1 Genotype frequencies

PopGene.S<sup>2</sup> (short for Population genetics simulation software) is a population genetics simulation program that will be featured in several Interact boxes. Here we will use PopGene.S<sup>2</sup> to explore interactive versions of Figs 2.5 and 2.6. Using the program will require that you download it from a website and install it on a computer running Windows. Simulations that can be explored with PopGene.S<sup>2</sup> will be featured in Interact boxes throughout this book.

Find Interact box 2.1 on the text web page and click on the link for **PopGene.S<sup>2</sup>**. The PopGene.S<sup>2</sup> website has download and installation instructions (and lists computer operating system requirements). But don't worry: the program is small, runs on most Windows computers, and is simple to install. After you have PopGene.S<sup>2</sup> installed according to the instructions provided in the PopGene.S<sup>2</sup> website, move on to Step 1 to begin the simulation.

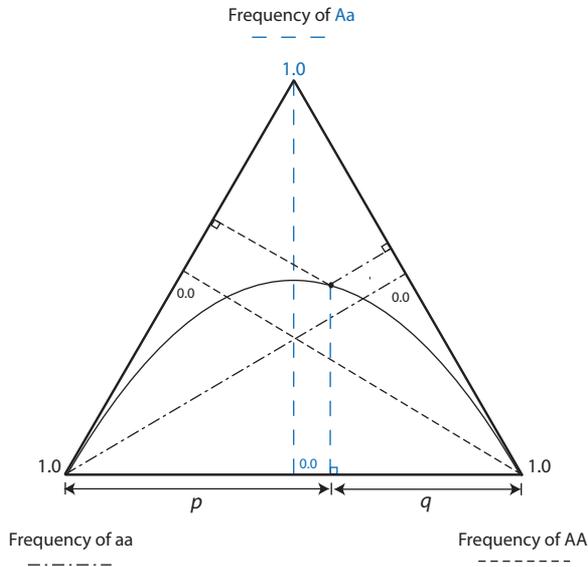
**Step 1** Open PopGene.S<sup>2</sup> and click once on the information box to make it disappear. Click on the **Allele and Genotype Frequencies** menu and select **Genotype frequencies**. A new window will open that contains a picture like Fig. 2.5 and some fields where you can enter genotype frequency values. Enter 0.25 into  $P(AA)$  to specify the frequency of the AA genotype. After entering each value the program will update the  $p$  and  $q$  values (the allele frequencies). Now enter 0.5 into  $P(Aa)$ . Once two genotype frequencies are entered the third genotype frequency is determined and the program will display the value. Click the **OK** button. A dot will appear on the graph corresponding to the value of  $P(Aa)$  and the frequency of the allele. Changing the  $P(AA)$  and  $P(Aa)$  frequencies to different values and then clicking OK again will add a new dot. Try several values for the genotype frequencies at different allele frequencies, both in and out of Hardy–Weinberg expected genotype frequencies.

**Step 2** Leave the **Genotype frequencies** window open but move it to one side to make room for another window. Now click on the **Mating Models** menu and select **Autosomal locus**. A new window will open that contains a triangular graph like that in Fig. 2.6. To display a set of genotype frequencies, enter the frequencies for  $P(AA)$  and  $P(Aa)$  in the text boxes and click **OK**. (Frequency of the aa genotype or  $P(aa)$  is calculated since the three genotype frequencies must sum to one.) The point on the De Finetti diagram representing the user-entered genotype frequencies will be plotted as a red square and displayed under the heading of "Initial frequencies". Hardy–Weinberg expected genotype frequencies are plotted as a blue square based on the allele frequencies that correspond to the user-entered genotype frequencies. These Hardy–Weinberg expected genotype frequencies are displayed under the heading of "Panmixia frequencies". Try several sets of different genotype frequencies to see values for genotype frequencies at different allele frequencies, both in and out of Hardy–Weinberg expected genotype frequencies.

**Step 3** Compare the two ways of visualizing genotype frequencies by plotting identical genotype frequencies in each window.

A single generation of reproduction where a set of conditions, or assumptions, are met will result in a population that meets Hardy–Weinberg expected genotype frequencies, often called Hardy–Weinberg equilibrium. The list of assumptions associated with this prediction for genotype frequencies is long. The set of assumptions includes:

- the organism is diploid,
- reproduction is sexual (as opposed to clonal),
- generations are discrete and non-overlapping,
- the locus under consideration has two alleles,
- allele frequencies are identical among all mating types (i.e. sexes),
- mating is random (as opposed to assortative),
- there is random union of gametes,
- population size is very large, effectively infinite,
- migration is negligible (no population structure, no gene flow),
- mutation does not occur or its rate is very low,
- natural selection does not act (all individuals and gametes have equal fitness).



**Figure 2.6** A De Finetti diagram for one locus with two alleles. The triangular coordinate system results from the requirement that the frequencies of all three genotypes must sum to one. Any point inside or on the edge of the triangle represents all three genotype frequencies of a population. The parabola describes Hardy–Weinberg expected genotype frequencies. The dashed lines represent the frequencies of each of the three genotypes between zero and one. Genotype frequencies at any point can be determined by the length of lines that are perpendicular to each of the sides of the triangle. A practical way to estimate genotype frequencies on the diagram is to hold a ruler parallel to one of the sides of the triangle and mark off the distance on one of the frequency axes. The point on the parabola is a population in Hardy–Weinberg equilibrium where the frequency of AA is 0.36, the frequency of aa is 0.16, and the frequency of Aa is 0.48. The perpendicular line to the base of the triangle also divides the bases into regions corresponding in length to the allele frequencies. Any population with genotype frequencies not on the parabola has an excess (above the parabola) or deficit (below the parabola) of heterozygotes compared to Hardy–Weinberg expected genotype frequencies.

These assumptions make intuitive sense when each is examined in detail (although this will probably be more apparent after more reading and simulation). As we will see later, Hardy–Weinberg holds for any number of alleles, although equation 2.1 is valid for only two alleles. Many of the assumptions can be thought of as assuring random mating and production of all possible progeny genotypes. Hardy–Weinberg genotype frequencies in progeny would not be realized if the two sexes have different allele frequencies even if matings take place between random pairs of parents. It is also possible that just by chance not all genotypes would be produced if only a small number of parents mated, just like flipping

a fair coin only a few times may not produce an equal number of heads and tails. Natural selection is a process that causes some genotypes in either the parental or progeny generations to be more frequent than others. So it is logical that Hardy–Weinberg expectations would not be met if natural selection were acting. In a sense, these assumptions define the biological processes that make up the field of population genetics. Each assumption represents one of the conceptual areas where population genetics can make testable predictions via expectation in order to distinguish the biological processes operating in populations. This is quite a set of accomplishments for an equation with just three terms!

Despite all of this praise, you might ask: what good is a model with so many restrictive assumptions? Are all these assumptions likely to be met in actual populations? The Hardy–Weinberg model is not necessarily meant to be an exact description of any actual population, although actual populations often exhibit genotype frequencies predicted by Hardy–Weinberg. Hardy–Weinberg provides a **null model**, a prediction based on a simplified or idealized situation where no biological processes are acting and genotype frequencies are the result of random combination. Actual populations can be compared with this null model to test hypotheses about the evolutionary forces acting on allele and genotype frequencies. The important point and the original motivation for Hardy and Weinberg was to show that the process of particulate inheritance itself does not cause any changes in allele frequencies across generations. Thus, changes in allele frequency or departures from Hardy–Weinberg expected genotype frequencies must be caused by processes that alter the outcome of basic inheritance.

**Null model** A testable model of no effect. A prediction or expectation based on the simplest assumptions to predict outcomes. Often, null models make predictions based on purely random processes, random samples, or variables having no effect on an outcome.

In the final part of this section we will explore genotype frequency expectations adjusted to account for ploidy (number of homologous chromosomes)

differences between males and females as seen in chromosomal sex determination and haplo-diploid organisms. In chromosomal sex determination as seen in mammals, birds, and Lepidoptera (butterflies and moths), one sex is determined by possession of two identical chromosomes (the homogametic sex) and the other sex determined by possession of two different chromosomes (the heterogametic sex). In mammals females are homogametic (XX) and males heterogametic (XY), whereas in birds and Lepidoptera the opposite is true, with heterogametic females (ZW) and homogametic males (ZZ). In haplo-diploid species such as bees and wasps (Hymenoptera), males are haploid (hemizygous) for all chromosomes whereas females are diploid for all chromosomes.

Predicting genotype frequencies at one locus in these cases under random mating and the other assumptions of Hardy–Weinberg requires keeping track of allele or genotype frequencies in both sexes and loci on specific chromosomes. An effective method is to draw a Punnett square that distinguishes the sex of an individual as well as the gamete types

that can be generated at mating (Table 2.1). The Punnett square shows that genotype frequencies in the diploid sex are identical to Hardy–Weinberg expectations for autosomes, whereas genotype frequencies are equivalent to allele frequencies in the haploid sex. One consequence of different chromosome types between the sexes is that fully recessive phenotypes are more common in the heterogametic sex, where a single chromosome determines the phenotype and recessive phenotypes appear at the allele frequency. However, in the homogametic sex, fully recessive phenotypes appear at the frequency of the recessive genotype (e.g.  $q^2$ ) since they are masked in heterozygotes. Some types of color blindness in humans are examples of traits due to genes on the X chromosome (called “X-linked” traits) that are more common in men than in women due to haplo-diploid inheritance.

Later, in section 2.4, we will examine two categories of applications of Hardy–Weinberg expected genotype frequencies. The first set of applications arises when we assume (often with supporting evidence)

**Table 2.1** Punnett square to predict genotype frequencies for loci on sex chromosomes and for all loci in males and females of haplo-diploid species. Notation in this table is based on birds where the sex chromosomes are Z and W (ZZ males and ZW females) with a diallelic locus on the Z chromosome possessing alleles A and a at frequencies  $p$  and  $q$ , respectively. In general, genotype frequencies in the homogametic or diploid sex are identical to Hardy–Weinberg expectations for autosomes, whereas genotype frequencies are equal to allele frequencies in the heterogametic or haploid sex.

Hemizygous or haploid sex			Diploid sex		
Genotype	Gamete	Frequency	Genotype	Gamete	Frequency
ZW	Z-A Z-a W	$p$ $q$	ZZ	Z-A Z-a	$p$ $q$
Expected genotype frequencies under random mating					
Homogametic sex					
Z-A Z-A			$p^2$		
Z-A Z-a			$2pq$		
Z-a Z-a			$q^2$		
Heterogametic sex					
Z-A W			$p$		
Z-a W			$q$		

that the assumptions of Hardy–Weinberg are true. We can then compare several expectations for genotype frequencies with actual genotype frequencies to distinguish between several alternative hypotheses. The second type of application is where we examine what results when assumptions of Hardy–Weinberg are not met. There are many cases where population genotype frequencies can be used to reveal the action of various population genetic processes. Before that, the next section builds a proof of the Hardy–Weinberg prediction that inheritance per se will not alter allele frequencies.

### 2.3 Why does Hardy–Weinberg work?

- A simple proof of Hardy–Weinberg.
- Hardy–Weinberg with more than two alleles.

The Hardy–Weinberg equation is one of the most basic expectations we have in population genetics. It is very likely that you were already familiar with the Hardy–Weinberg equation before you picked up this book. But where does Hardy–Weinberg actually come from? What is the logic behind it? Let's develop a simple proof that Hardy–Weinberg is actually true. This will also be our first real foray into the type of algebraic argument that much of population genetics is built on. Given that you start out knowing the conclusion of the Hardy–Weinberg tale, this gives you the opportunity to focus on the style in which it is told. Algebraic or quantitative arguments are a central part of the language and vocabulary of population genetics, so part of the task of learning population genetics is becoming accustomed to this mode of discourse.

We would like to prove that  $p^2 + 2pq + q^2 = 1$  accurately predicts genotype frequencies given the values of allele frequencies. Let's start off by making some explicit assumptions to bound the problem. The assumptions, in no particular order, are:

- 1 mating is random (parents meet and mate according to their frequencies);
- 2 all parents have the same number of offspring (equivalent to no natural selection on fecundity);
- 3 all progeny are equally fit (equivalent to no natural selection on viability);
- 4 there is no mutation that could act to change an A to a or an a to A;
- 5 it is a single population that is very large;
- 6 there are two and only two mating types.

Now let's define the variables we will need for a case with one locus that has two alleles (A and a).

$N$  = Population size of individuals ( $N$  diploid individuals have  $2N$  alleles)

Allele frequencies:

$$p = \text{frequency(A allele)} \\ = (\text{total number of A alleles})/2N$$

$$q = \text{frequency(a allele)} \\ = (\text{total number of a alleles})/2N$$

$$p + q = 1$$

Genotype frequencies:

$$X = \text{frequency(AA genotype)} \\ = (\text{total number of AA genotypes})/N$$

$$Y = \text{frequency(Aa genotype)} \\ = (\text{total number of Aa genotypes})/N$$

$$Z = \text{frequency(aa genotype)} \\ = (\text{total number of aa genotypes})/N$$

$$X + Y + Z = 1$$

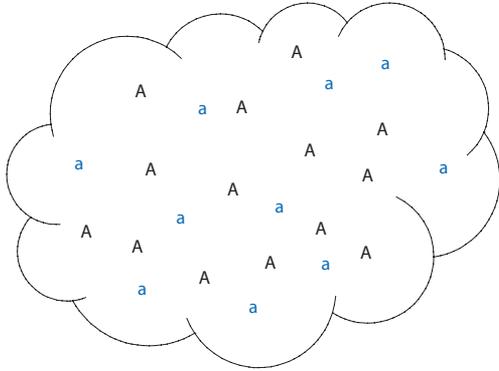
We do not distinguish between the heterozygotes Aa and aA and treat them as being equivalent genotypes. Therefore, we can express allele frequencies in terms of genotype frequencies by adding together the frequencies of A-containing and a-containing genotypes:

$$p = X + \frac{1}{2}Y \quad (2.2)$$

$$q = Z + \frac{1}{2}Y \quad (2.3)$$

Each homozygote contains two alleles of the same type while each heterozygote contains one allele of each type so the heterozygote genotypes are each weighted by half.

With the variables defined, we can then follow allele frequencies across one generation of reproduction. The first step is to calculate the probability that parents of any two particular genotypes will mate. Since mating is assumed to be random, the chance that two genotypes will mate is just the product of their individual frequencies. As shown in Fig. 2.7, random mating can be thought of as being like gas atoms in a balloon. As with gas atoms, each genotype or gamete bumps into others at random, with the probability of a collision (or mating or union) being the product of the frequencies of the two objects



**Figure 2.7** A schematic representation of random mating as a cloud of gas where the frequency of A alleles is  $14/24$  and the frequency of a alleles is  $10/24$ . Any given A has a frequency of  $14/24$  and will encounter another A with probability of  $14/24$  or an a with the probability of  $10/24$ . This makes the frequency of an A–A collision  $(14/24)^2$  and an A–a collision  $(14/24)(10/24)$ , just as the probability of two independent events is the product of their individual probabilities. The population of A and a alleles is assumed to be large enough so that taking one out of the cloud will make almost no change in the overall frequency of its type.

colliding. To calculate the probabilities of mating among the three different genotypes we can make a table to organize the resulting mating frequencies. This table will predict the mating frequencies among genotypes in the initial generation, which we will call generation  $t$ .

A parental mating frequency table (generation  $t$ ) is shown below.

Moms		Dads		
		AA	Aa	aa
	Frequency . . .	$X$	$Y$	$Z$
AA	$X$	$X^2$	$XY$	$XZ$
Aa	$Y$	$XY$	$Y^2$	$YZ$
aa	$Z$	$ZX$	$ZY$	$Z^2$

The table expresses parental mating frequencies in the currency of genotype frequencies. For example, we expect matings between AA moms and Aa dads to occur with a frequency of  $XY$ .

Next we need to determine the frequency of each genotype in the offspring of any given parental mating pair. This will require that we predict the offspring genotypes resulting from each possible parental mating. We can do this easily with a Punnett square. We will use the frequencies of each parental

mating (above) together with the frequencies of the offspring genotypes. Summed for all possible parental matings, this gives the frequency of offspring genotypes one generation later, or in generation  $t + 1$ . A table will help organize all the frequencies, like the offspring frequency table (generation  $t + 1$ ) shown below.

Parental mating	Total frequency	Offspring genotype frequencies		
		AA	Aa	aa
AA $\times$ AA	$X^2$	$X^2$	0	0
AA $\times$ Aa	$2XY$	$XY$	$XY$	0
AA $\times$ aa	$2XZ$	0	$2XZ$	0
Aa $\times$ Aa	$Y^2$	$Y^2/4$	$(2Y^2)/4$	$Y^2/4$
Aa $\times$ aa	$2YZ$	0	$YZ$	$YZ$
aa $\times$ aa	$Z^2$	0	0	$Z^2$

In this table, the total frequency is just the frequency of each parental mating pair taken from the parental mating frequency table. We now need to partition this total frequency of each parental mating into the frequencies of the three progeny genotypes produced. Let's look at an example. Parents with AA and Aa genotypes will produce progeny with two genotypes: half AA and half Aa (you can use a Punnett square to show this is true). Therefore, the AA  $\times$  Aa parental matings, which have a total frequency of  $2XY$  under random mating, are expected to produce  $(1/2)2XY = XY$  of each of AA and Aa progeny. The same logic applies to all of the other parental matings. Notice that each row in the offspring genotype frequency table sums to the total frequency of each parental mating.

The columns in the offspring genotype frequency table are the basis of the final step. The sum of each column gives the total frequencies of each progeny genotype expected in generation  $t + 1$ . Let's take the sum of each column, again expressed in the currency of genotype frequencies, and then simplify the algebra to see whether Hardy and Weinberg were correct.

$$\begin{aligned}
 \text{AA} &= X^2 + XY + Y^2/4 \\
 &= (X + 1/2Y)^2 \text{ (recall that } p = X + 1/2Y) \\
 &= p^2
 \end{aligned}$$

$$\begin{aligned}
 aa &= Y^2/4 + YZ + Z^2 \\
 &= (Z + 1/2Y)^2 \text{ (recall that } q = Z + 1/2Y\text{)} \\
 &= q^2 \\
 \\
 Aa &= XY + 2XZ + 2Y^2/4 + YZ \\
 &= 2(XY/2 + XZ + Y^2/4 + YZ/2) \\
 &= 2(X + Y/2)(Z + Y/2) \\
 &= 2pq \qquad (2.4)
 \end{aligned}$$

So we have proved that progeny genotype and allele frequencies are identical to parental genotype and allele frequencies over one generation or that  $f(A)_t = f(A)_{t+1}$ . The major conclusion here is that *genotype frequencies remain constant over generations as long as the assumptions of Hardy–Weinberg are met*. In fact, we have just proved that under Mendelian heredity genotype and allele frequencies should not change over time unless one or more of our assumptions is not met. This simple model of expected genotype frequencies has profound conclusions! In fact, Hardy–Weinberg expected genotype frequencies serve as one of the most basic tools to test for the action of biological processes that alter genotype and allele frequencies.

You might wonder whether Hardy–Weinberg applies to loci with more than two alleles. For the last point in this section let's explore that question. With three alleles at one locus (allele frequencies symbolized by  $p$ ,  $q$ , and  $r$ ), Hardy–Weinberg expected genotype frequencies are  $p^2 + q^2 + r^2 + 2pq + 2pr + 2qr = 1$ . These genotype frequencies are obtained by expanding  $(p + q + r)^2$ , a method that can be applied to any number of alleles at one locus. In general, expanding the squared sum of the allele frequencies will show:

- the frequency of any homozygous genotype is the squared frequency of the single allele that composes the genotype ( $[\text{allele frequency}]^2$ );
- the frequency of any heterozygous genotype is twice the product of the two allele frequencies that comprise the genotype ( $2[\text{allele 1 frequency}] \times [\text{allele 2 frequency}]$ ); and
- there are as many homozygous genotypes as there are alleles and  $\frac{N(N-1)}{2}$  heterozygous genotypes where  $N$  is the number of alleles.

Do you think it would be possible to prove Hardy–Weinberg for more than two alleles at one locus? The answer is absolutely yes. This would just require constructing larger versions of the parental genotype

mating table and expected offspring frequency table as we did for two alleles at one locus.

## 2.4 Applications of Hardy–Weinberg

- Apply Hardy–Weinberg to estimate the frequency of an observed genotype in a forensic DNA typing case.
- The  $\chi^2$  test gauges whether observed and expected differ more than expected by chance.
- Assume Hardy–Weinberg to compare two genetic models.

In the previous two sections we established the Hardy–Weinberg expectations for genotype frequencies. In this section we will examine three ways that expected genotype frequencies are employed in practice. The goal of this section is to become familiar with realistic applications as well as hypothesis tests that compare observed and Hardy–Weinberg expected genotype frequencies. In this process we will also look at a specific method to account for sampling error (see Appendix).

### Forensic DNA profiling

Our first application of Hardy–Weinberg can be found in newspapers on a regular basis and commonly dramatized on television. A terrible crime has been committed. Left at the crime scene was a biological sample that law-enforcement authorities use to obtain a multilocus genotype or DNA profile. A suspect in the crime has been identified and subpoenaed to provide a tissue sample for DNA profiling. The DNA profiles from the suspect and from the crime scene are identical. The DNA profile is shown in Table 2.2. Should we conclude that the suspect left the biological sample found at the crime scene?

**Table 2.2** Example DNA profile for three simple tandem repeat (STR) loci commonly used in human forensic cases. Locus names refer to the human chromosome (e.g. D3 means third chromosome) and chromosome region where the SRT locus is found.

Locus	D3S1358	D21S11	D18S51
Genotype	17, 18	29, 30	18, 18



To answer this critical question we will employ Hardy–Weinberg to predict the expected frequency of the DNA profile or genotype. Just because two DNA profiles match, there is not necessarily strong evidence that the individual who left the evidence DNA and the suspect are the same person. It is possible that there are actually two or more people with identical DNA profiles. Hardy–Weinberg and Mendel’s second law will serve as the bases for us to estimate just how frequently a given DNA profile should be observed. Then we can determine whether two unrelated individuals sharing an identical DNA profile is a likely occurrence.

To determine the expected frequency of a one-locus genotype, we employ the Hardy–Weinberg equation (2.1). In doing so, *we are implicitly accepting that all of the assumptions of Hardy–Weinberg are approximately met.* If these assumptions were not met, then the Hardy–Weinberg equation would not provide an accurate expectation for the genotype frequencies! To determine the frequency of the three-locus genotype in Table 2.2 we need allele frequencies for those loci, which are found in Table 2.3. Starting with the locus D3S1358, we see in Table 2.3 that the 17-repeat allele has a frequency of 0.2118 and the 18-repeat allele a frequency of 0.1626. Then using Hardy–Weinberg, the 17, 18 genotype has an expected frequency of  $2(0.2118)(0.1626) = 0.0689$  or 6.89%. For the two other loci in the DNA profile of Table 2.2 we carry out the same steps.

D21S11 29-Repeat allele frequency = 0.1811  
30-Repeat allele frequency = 0.2321  
Genotype frequency  
=  $2(0.1811)(0.2321) = 0.0841$   
or 8.41%

D18S51 18-Repeat allele frequency = 0.0918  
Genotype frequency =  $(0.0918)^2$   
= 0.0084 or 0.84%

The genotype for each locus has a relatively large chance of being observed in a population. For example, a little less than 1% of white US citizens (or about 1 in 119) are expected to be homozygous for the 18-repeat allele at locus D18S51. Therefore, a match between evidence and suspect DNA profiles homozygous for the 18 repeat at that locus would not be strong evidence that the samples came from the same individual.

Fortunately, we can combine the information from all three loci. To do this we use the **product**

**rule**, which states that the probability of observing multiple independent events is just the product of each individual event. We already used the product rule in the last section to calculate the expected frequency of each genotype under Hardy–Weinberg by treating each allele as an independent probability. Now we just extend the product rule to cover multiple genotypes, *under the assumption that each of the loci is independent by Mendel’s second law* (the assumption is justified here since each of the loci is on a separate chromosome). The expected frequency of the three locus genotype (sometimes called the probability of identity) is then  $0.0689 \times 0.0841 \times 0.0084 = 0.000049$  or 0.0049%. Another way to express this probability is as an **odds ratio**, or the reciprocal of the probability (an approximation that holds when the probability is very small). Here the odds ratio is  $1/0.000049 = 20,408$ , meaning that we would expect to observe the three-locus DNA profile once in 20,408 white US citizens.

**Product rule** The probability of two (or more) independent events occurring simultaneously is the product of their individual probabilities.

**Odds ratio** The number of events divided by the number of non-events; one over the expected sample size required to observe a single instance or event.

Now we can return to the question of whether two unrelated individuals are likely to share an identical three-locus DNA profile by chance. One out of every 20,408 white US citizens is expected to have the genotype in Table 2.2. Although the three-locus DNA profile is considerably less frequent than a genotype for a single locus, it is still does not approach a unique, individual identifier. Therefore, there is a finite chance that a suspect will match an evidence DNA profile by chance alone. Such DNA profile matches, or “inclusions,” require additional evidence to ascertain guilt or innocence. In fact, the term prosecutor’s fallacy was coined to describe failure to recognize the difference between a DNA match and guilt (for example, a person can be present at a location and not involved in a crime). Only when DNA profiles do not match, called an “exclusion,” can a suspect be unambiguously and absolutely excluded as the source of a biological sample at a crime scene.

### Problem box 2.1 The expected genotype frequency for a DNA profile

Calculate the expected genotype frequency and odds ratio for the 10-locus DNA profile below. Allele frequencies are given in Table 2.3.

D3S1358	17, 18
vWA	17, 17
FGA	24, 25
Amelogenin	X, Y
D8S1179	13, 14
D21S11	29, 30
D18S51	18, 18
D5S818	12, 13
D13S317	9, 12
D7S820	11, 12

What does the amelogenin locus tell us and how did you assign an expected frequency to the observed genotype? Is it likely that two unrelated individuals would share this 10-locus genotype by chance? For this genotype, would a match between a crime scene sample and a suspect be convincing evidence that the person was present at the crime scene?

Current forensic DNA profiles use 10–13 loci to estimate expected genotype frequencies. Problem 2.1 gives a 10-locus genotype for the same individual in Table 2.2, allowing you to calculate the odds ratio for a realistic example. In Chapter 4 we will reconsider the expected frequency of a DNA profile with the added complication of allele-frequency differentiation among human racial groups.

#### Testing for Hardy–Weinberg

A common use of Hardy–Weinberg expectations is to test for deviations from its null model. Populations with genotype frequencies that do not fit Hardy–Weinberg expectations are evidence that one or more of the evolutionary processes embodied in the assumptions of Hardy–Weinberg are acting to determine genotype frequencies. Our null hypothesis is that genotype frequencies meet Hardy–Weinberg expectations within some degree of estimation error. Genotype frequencies that are not close to Hardy–Weinberg expectations allow us to reject this null hypothesis. The processes in the list of assumptions then become possible alternative hypotheses to explain observed genotype frequencies. In this section we will work through a hypothesis test for Hardy–Weinberg equilibrium.

The first example uses observed genotypes for the MN blood group, a single locus in humans that has

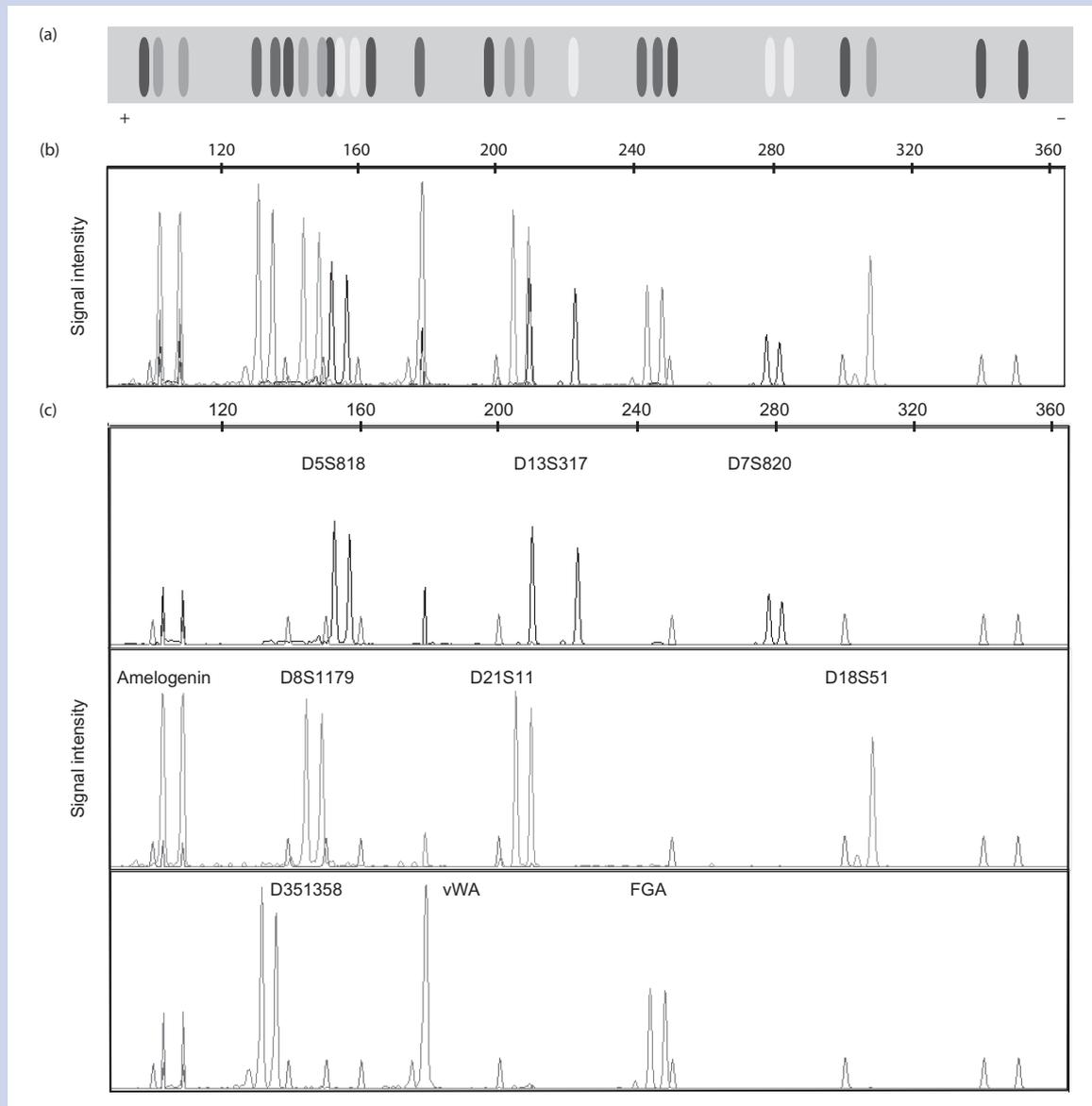
### Box 2.1 DNA profiling

The loci used for human DNA profiling are a general class of DNA sequence marker known as simple tandem repeat (STR), simple sequence repeat (SSR), or microsatellite loci. These loci feature tandemly repeated DNA sequences of one to six base pairs (bp) and often exhibit many alleles per locus and high levels of heterozygosity. Allelic states are simply the number of repeats present at the locus, which can be determined by electrophoresis of PCR amplified DNA fragments. STR loci used in human DNA profiling generally exhibit Hardy–Weinberg expected genotype frequencies, there is evidence that the genotypes are selectively “neutral” (i.e. not affected by natural selection), and the loci meet the other assumptions of Hardy–Weinberg. STR loci

are employed widely in population genetic studies and in genetic mapping (see reviews by Goldstein & Pollock 1997; McDonald & Potts 1997).

This is an example of the DNA sequence found at a microsatellite locus. This sequence is the 24.1 allele from the FGA locus (Genbank accession no. AY749636; see Fig. 2.8). The integral repeat is the 4 bp sequence CTTT and most alleles have sequences that differ by some number of full CTTT repeats. However, there are exceptions where alleles have sequences with partial repeats or stutters in the repeat pattern, for example the TTTCT and CTC sequences imbedded in the perfect CTTT repeats. In this case, the 24.1 allele is 1 bp longer than the 24 allele sequence. *(continued)*

## Box 2.1 (continued)



**Figure 2.8** The original data for the DNA profile given in Table 2.2 and Problem box 2.1 obtained by capillary electrophoresis. The PCR oligonucleotide primers used to amplify each locus are labeled with a molecule that emits blue, green, or yellow light when exposed to laser light. Thus, the DNA fragments for each locus are identified by their label color as well as their size range in base pairs. (a) A simulation of the DNA profile as it would appear on an electrophoretic gel (+ indicates the anode side). Blue, green, and yellow label the 10 DNA-profiling loci, shown here in grayscale. Other DNA fragments are size standards (originally in red) with a known molecular weight used to estimate the size in base pairs of the other DNA fragments in the profile. (b) The DNA profile for all loci and the size-standard DNA fragments as a graph of color signal intensity by size of DNA fragment in base pairs. (c) A simpler view of trace data for each label color independently with the individual loci labeled above the trace peaks. A few shorter peaks are visible in the yellow, green, and blue traces of (c) that are not labeled as loci. These artifacts, called pull-up peaks, are caused by intense signal from a locus labeled with another color (e.g. the yellow and blue peaks in the location of the green-labeled amelogenin locus). A full-color version of this figure is available on the textbook website.

```

GCCCCATAGGTTTTGAACTCACAGATTAACCTGTAACCAAATAAAATTAGGCATTATTTACAAGCTA
GTTT CTTT CTTT CTTT TTTCT CTTT CTTT CTTT CTTT CTTT CTTT CTTT CTTT CTTT CTTT
CTTT CTTT CTTT CTTT CTTT CTC CTC CTC CTTT CTC CTTT CTTT TTTGCTGGCA
ATTACAGACAAATCAA

```

**Table 2.4** Expected numbers of each of the three MN blood group genotypes under the null hypotheses of Hardy–Weinberg. Genotype frequencies are based on a sample of 1066 Chukchi individuals, a native people of eastern Siberia (Roychoudhury and Nei 1988).

Frequency of M = $\hat{p}$ = 0.4184 Frequency of N = $\hat{q}$ = 0.5816			
Genotype	Observed	Expected number of genotypes	Observed – expected
MM	165	$N \times \hat{p}^2 = 1066 \times (0.4184)^2 = 186.61$	-21.6
MN	562	$N \times 2\hat{p}\hat{q} = 1066 \times 2(0.4184)(0.5816) = 518.80$	43.2
NN	339	$N \times \hat{q}^2 = 1066 \times (0.5816)^2 = 360.58$	-21.6

two alleles (Table 2.4). First we need to estimate the frequency of the M allele, using the notation that the estimated frequency of M is  $\hat{p}$  and the frequency of N is  $\hat{q}$ . Note that the “hat” superscripts indicate that these are allele-frequency *estimates* (see Chapter 1). The total number of alleles is  $2N$  given a sample of  $N$  diploid individuals. We can then count up all of the alleles of one type to estimate the frequency of that allele.

$$\hat{p} = \frac{2 \times \text{Frequency}(\text{MM}) + \text{frequency}(\text{MN})}{2N} \quad (2.5)$$

$$\hat{p} = \frac{2 \times 165 + 562}{2 \times 1066} = \frac{892}{2132} = 0.4184 \quad (2.6)$$

Since  $\hat{p} + \hat{q} \approx 1$ , we can estimate the frequency of the N allele by subtraction as  $\hat{q} = 1 - \hat{p} = 1 - 0.4184 = 0.5816$ .

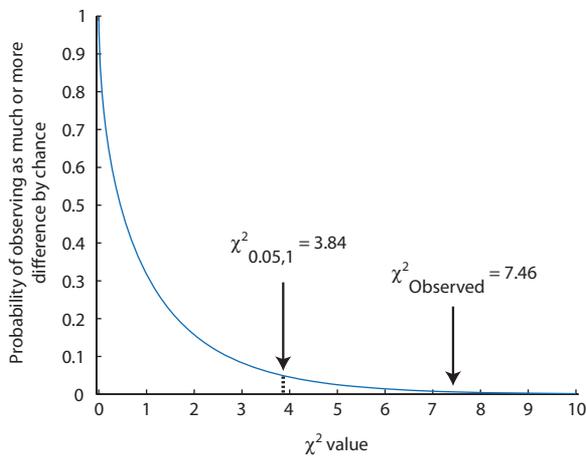
Using these allele frequencies allows calculation of the Hardy–Weinberg expected genotype frequency and number of individuals with each genotype, as shown in Table 2.4. In Table 2.4 we can see that the match between the observed and expected is not perfect, but we need some method to ask whether the difference is actually large enough to conclude that Hardy–Weinberg equilibrium does not hold in the sample of 1066 genotypes. Remember that any allele-frequency estimate ( $\hat{p}$ ) could differ slightly from the true parameter ( $p$ ) due to chance events as well as due to random sampling in the group of genotypes used to estimate the allele frequencies. Asking whether genotypes are in Hardy–Weinberg proportions is actually the same as asking whether a coin is “fair.” With a fair coin we expect one-half heads and one-half tails if we flip it a large number of times. But even with a fair coin we can get something other than exactly

50 : 50 even if the sample size is large. We would consider a coin fair if in 1000 flips it produced 510 heads and 490 tails. However, the hypothesis that a coin is fair would be in doubt if we observed 250 heads and 750 tails given that we expect 500 of each.

In more general terms, the expected frequency of an event,  $p$ , times the number of trials or samples,  $n$ , gives the expected number events or  $np$ . To test the hypothesis that  $p$  is the frequency of an event in an actual population, we compare  $np$  with  $n\hat{p}$ . Close agreement suggests that the parameter and the estimate are the same quantity. But a large disagreement instead suggests that  $p$  and  $\hat{p}$  are likely to be different probabilities. The Chi-squared ( $\chi^2$ ) distribution is a statistical test commonly used to compare  $np$  and  $n\hat{p}$ . The  $\chi^2$  test *provides the probability of obtaining the difference (or more) between the observed ( $n\hat{p}$ ) and expected ( $np$ ) number of outcomes by chance alone if the null hypothesis is true*. As the difference between the observed and expected grows larger it becomes less probable that the parameter and the parameter estimate are actually the same but differ in a given sample due to chance. The  $\chi^2$  statistic is:

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \quad (2.7)$$

The  $\chi^2$  formula makes intuitive sense. In the numerator there is the difference between the observed and Hardy–Weinberg expected number of individuals. This difference is squared, like a variance, since we do not care about the direction of the difference but only the magnitude of the difference. Then in the denominator we divide by the expected number of individuals to make the squared difference relative. For example, a squared difference of 4 is small if the expected number is 100 (it is 4%) but relatively



**Figure 2.9** A  $\chi^2$  distribution with one degree of freedom. The  $\chi^2$  value for the Hardy–Weinberg test with MN blood group genotypes as well as the critical value to reject the null hypothesis are shown (see text for details). The area under the curve to the right of the arrow indicates the probability of observing that much or more difference between the observed and expected outcomes.

larger if the expected number is 8 (it is 50%). Adding all of these relative squared differences gives the total relative squared deviation observed over all genotypes.

$$\chi^2 = \frac{(-21.6)^2}{181.61} + \frac{(43.2)^2}{518.80} + \frac{(-21.6)^2}{360.58} = 7.46 \quad (2.8)$$

We need to compare our statistic to values from the  $\chi^2$  distribution. But first we need to know how much information, or the degrees of freedom (commonly abbreviated as df), was used to estimate the  $\chi^2$  statistic. In general, degrees of freedom are based on the number of categories of data:  $df = \text{no. of classes compared} - \text{no. of parameters estimated} - 1$

for the  $\chi^2$  test itself. In this case  $df = 3 - 1 - 1 = 1$  for three genotypes and one estimated allele frequency (with two alleles: the other allele frequency is fixed once the first has been estimated).

Figure 2.9 shows a  $\chi^2$  distribution for one degree of freedom. Small deviations of the observed from the expected are more probable since they leave more area of the distribution to the right of the  $\chi^2$  value. As the  $\chi^2$  value gets larger, the probability that the difference between the observed and expected is just due to chance sampling decreases (the area under the curve to the right gets smaller). Another way of saying this is that as the observed and expected get increasingly different, it becomes more improbable that our null hypothesis of Hardy–Weinberg is actually the process that is determining genotype frequencies. Using Table 2.5 we see that a  $\chi^2$  value of 7.46 with 1 df has a probability between 0.01 and 0.001. The conclusion is that the observed genotype frequencies would be observed less than 1% of the time in a population that actually had Hardy–Weinberg expected genotype frequencies. Under the null hypothesis we do not expect this much difference or more from Hardy–Weinberg expectations to occur often. By convention, we would reject chance as the explanation for the differences if the  $\chi^2$  value had a probability of 0.05 or less. In other words, if chance explains the difference in five trials out of 100 or less then we reject the hypothesis that the observed and expected patterns are the same. The critical value above which we reject the null hypothesis for a  $\chi^2$  test is 3.84 with 1 df, or in notation  $\chi^2_{0.05,1} = 3.84$ . In this case, we can clearly see an excess of heterozygotes and deficits of homozygotes and employing the  $\chi^2$  test allows us to conclude that Hardy–Weinberg expected genotype frequencies are not present in the population.

**Table 2.5**  $\chi^2$  values and associated cumulative probabilities in the right-hand tail of the distribution for 1–5 df.

df	Probability					
	0.5	0.25	0.1	0.05	0.01	0.001
1	0.4549	1.3233	2.7055	3.8415	6.6349	10.8276
2	1.3863	2.7726	4.6052	5.9915	9.2103	13.8155
3	2.3660	4.1083	6.2514	7.8147	11.3449	16.2662
4	3.3567	5.3853	7.7794	9.4877	13.2767	18.4668
5	4.3515	6.6257	9.2364	11.0705	15.0863	20.5150

### Interact box 2.2 $\chi^2$ test

The program PopGene.S<sup>2</sup> (refer to the link for **Interact box 2.1** to obtain PopGene.S<sup>2</sup> if necessary) can be used to carry out a  $\chi^2$  test for one locus with two alleles under the null hypothesis of Hardy–Weinberg genotype frequencies. Launch PopGene.S<sup>2</sup> and select **Chi-square test** under the **Allele and Genotype Frequencies** menu. Try these examples.

- Input the observed genotype frequencies in Table 2.4 to confirm the calculations and  $\chi^2$  value.
- Test the null hypothesis of Hardy–Weinberg genotype frequencies for these data from a sample of 459 Yugoslavians: MM, 144; MN, 201; NN, 114.

### Assuming Hardy–Weinberg to test alternative models of inheritance

Biologists are all probably familiar with the ABO blood groups and are aware that mixing blood of different types can cause blood cell lysis and possibly result in death. Although we take this for granted now, there was a time when blood types and their patterns of inheritance defined an active area of clinical research. It was in 1900 that Karl Landsteiner of the University of Vienna mixed the blood of the people in his laboratory to study the patterns of blood cell agglutination (clumping). Landsteiner was awarded the Nobel Prize

for Medicine in 1930 for his discovery of human ABO blood groups. Not until 1925, due to the research of Felix Bernstein, was the genetic basis of the ABO blood groups resolved (see Crow 1993a).

Landsteiner observed the presence of four blood phenotypes A, B, AB, and O. A logical question was, then, “what is the genetic basis of these four blood group phenotypes?” We will test two hypotheses (or models) to explain the inheritance of ABO blood groups that coexisted for 25 years. The approach will use the frequency of genotypes in a sample population to test the two hypotheses rather than an approach such as examining pedigrees. The hypotheses are that the four blood group phenotypes are explained by either two independent loci with two alleles each with one allele completely dominant at each locus (hypothesis 1) or a single locus with three alleles where two of the alleles show no dominance with each other but both are completely dominant over a third allele (hypothesis 2). Throughout, we will assume that Hardy–Weinberg expected genotype frequencies are met in order to determine which hypothesis best fits the available data.

Our first task is a straightforward application of Hardy–Weinberg in order to determine the expected frequencies of the blood group genotypes. The genotypes and the expected genotype frequencies are shown in Table 2.6. Look at the table but cover up the expected frequencies with a sheet of paper. The genotypes given for the two hypotheses would both explain the observed pattern of four blood groups. Hypothesis 1 requires complete dominance of the A and B alleles at their respective loci. Hypothesis 2 requires A and B to have no dominance with each other but complete dominance when paired with the O allele.

**Table 2.6** Hardy–Weinberg expected genotype frequencies for the ABO blood groups under the hypotheses of (1) two loci with two alleles each and (2) one locus with three alleles. Both hypotheses have the potential to explain the observation of four blood group phenotypes. The notation  $f_x$  is used to refer to the frequency of allele  $x$ . The underscore indicates any allele; for example,  $A_$  means both AA and Aa genotypes. The observed blood type frequencies were determined for Japanese people living in Korea (from Bernstein (1925) as reported in Crow (1993a)).

Blood type	Genotype		Expected genotype frequency		Observed (total = 502)
	Hypothesis 1	Hypothesis 2	Hypothesis 1	Hypothesis 2	
O	aa bb	OO	$f_a^2 f_b^2$	$f_O^2$	148
A	$A_ bb$	AA, AO	$(1 - f_a^2) f_b^2$	$f_A^2 + 2f_A f_O$	212
B	aa $B_$	BB, BO	$f_a^2 (1 - f_b^2)$	$f_B^2 + 2f_B f_O$	103
AB	$A_ B_$	AB	$(1 - f_a^2)(1 - f_b^2)$	$2f_A f_B$	39

Now let's construct several of the expected genotype frequencies (before you lift that sheet of paper). The O blood group under hypothesis 1 is the frequency of a homozygous genotype at two loci (aa bb). The frequency of one homozygote is the square of the allele frequency:  $f_a^2$  and  $f_b^2$  if we use  $f_x$  to indicate the frequency of allele  $x$ . Using the product rule or Mendel's second law, the expected frequency of the two-locus genotype is the product of frequencies of the one-locus genotypes,  $f_a^2$  and  $f_b^2$ . For the next genotype under hypothesis 1 (A\_ bb), we use a little trick to simplify the amount of notation. The genotype A\_ means AA or Aa; in other words, any genotype but aa. Since the frequencies of the three genotypes at one locus must sum to one, we can write  $f_{A\_}$  as  $1 - f_{aa}$  or  $1 - f_a^2$ . Then the frequency of the A\_ bb genotype is  $(1 - f_a^2)f_b^2$ . You should now work out and write down the other six expected genotype frequency expressions: then lift the paper and compare your work to Table 2.6.

The next step is to compare the expected genotype frequencies for the two hypotheses with observed genotype frequencies. To do this we will need to estimate allele frequencies under each hypothesis and use these to compute the expected genotype frequencies. (Although these allele frequencies are parameter estimates, the "hat" notation is not used for the sake of readability.) For the hypothesis of two loci (hypothesis 1)  $f_b^2 = (148 + 212)/502 = 0.717$  so we can estimate the allele frequency as  $f_b = \sqrt{f_b^2} = \sqrt{0.717} = 0.847$ . The other allele frequency at that locus is then determined by subtraction:  $f_B = 1 - 0.847 = 0.153$ . Similarly for the second locus  $f_a^2 = (148 + 103)/502 = 0.50$  and  $f_a = \sqrt{f_a^2} = \sqrt{0.50} = 0.707$ , giving  $f_A = 1 - 0.707 = 0.293$  by subtraction.

### Problem box 2.2 Proving allele frequencies are obtained from expected genotype frequencies

Can you use algebra to prove that adding together expected genotype frequencies under hypotheses 1 and 2 in Table 2.7 gives the allele frequencies shown in the text? For the genotypes of hypothesis 1, show that  $f(aa\ bb) + f(A\_bb) =$  the frequency of the bb genotype. For hypothesis 2 show the observed genotype frequencies that can be used to estimate the frequency of the B allele starting off with the relationship  $f_A + f_B + f_O = 1$  and then solving for  $f_B$  in terms of  $f_A$  and  $f_O$ .

For the hypothesis of one locus with three alleles (hypothesis 2) we estimate the frequency of any of the alleles by using the relationship that the three allele frequencies sum to one. This basic relationship can be reworked to obtain the expected genotype frequency expressions into expressions that allow us to estimate the allele frequencies (see Problem box 2.2). It turns out that adding together all expected genotype frequency terms for two of the alleles estimates the square of one minus the other allele. For example,  $(1 - f_B)^2 = f_O^2 + f_A^2 + 2f_Af_O$ , and checking in Table 2.7 this corresponds to  $(148 + 212)/502 = 0.717$ . Therefore,  $1 - f_B = 0.847$  and  $f_B = 0.153$ . Using similar steps,  $(1 - f_A)^2 = f_O^2 + f_B^2 + 2f_Bf_O =$

**Table 2.7** Expected numbers of each of the four blood group genotypes under the hypotheses of (1) two loci with two alleles each and (2) one locus with three alleles. Estimated allele frequencies are based on a sample of 502 individuals.

Blood	Observed	Expected number of genotypes	Observed - expected	(Observed - expected) <sup>2</sup> /expected
Hypothesis 1 ( $f_A = 0.293$ , $f_a = 0.707$ , $f_B = 0.153$ , $f_b = 0.847$ )				
O	148	$502(0.707)^2(0.847)^2 = 180.02$	-40.02	8.90
A	212	$502(0.500)(0.847)^2 = 180.07$	31.93	5.66
B	103	$502(0.707)^2(0.282) = 70.76$	32.24	14.69
AB	39	$502(0.500)(0.282) = 70.78$	-31.78	14.27
Hypothesis 2 ( $f_A = 0.293$ , $f_B = 0.153$ , $f_O = 0.554$ )				
O	148	$502(0.554)^2 = 154.07$	-6.07	0.24
A	212	$502((0.293)^2 + 2(0.293)(0.554)) = 206.07$	5.93	0.17
B	103	$502((0.153)^2 + 2(0.153)(0.554)) = 96.85$	6.15	0.39
AB	39	$502(2(0.293)(0.153)) = 45.01$	-6.01	0.80

$(148 + 103)/502 = 0.50$ . Therefore,  $1 - f_A = 0.707$  and  $f_A = 0.293$ . Finally, by subtraction  $f_O = 1 - f_B - f_A = 1 - 0.153 - 0.293 = 0.554$ .

The number of genotypes under each hypothesis can then be found by using the expected genotype frequencies in Table 2.6 and the estimated allele frequencies. Table 2.7 gives the calculation for the expected numbers of each genotype under both hypotheses. We can also calculate a Chi-squared value associated with each hypothesis based on the difference between the observed and expected genotype frequencies. For hypothesis 1  $\chi^2 = 43.52$ , whereas  $\chi^2 = 1.60$  for hypothesis 2. Both of these tests have one degree of freedom (4 genotypes – 2 for estimated allele frequencies – 1 for the test), giving a critical

value of  $\chi_{0.05,1}^2 = 3.84$ . Clearly the hypothesis of three alleles at one locus is the better fit to the observed data. Thus, we have just used genotype frequency data sampled from a population with the assumptions of Hardy–Weinberg equilibrium as a means to distinguish between two hypotheses for the genetic basis of blood groups.

## 2.5 The fixation index and heterozygosity

- The fixation index ( $F$ ) measures deviation from Hardy–Weinberg expected heterozygote frequencies.
- Examples of mating systems and  $F$  in wild populations.
- Observed and expected heterozygosity.

The mating patterns of actual organisms frequently do not exhibit the random mating assumed by Hardy–Weinberg. In fact, many species exhibit mating systems that create predictable deviations from Hardy–Weinberg expected genotype frequencies. The term **assortative mating** is used to describe non-random mating. **Positive assortative mating** describes the case when individuals with like genotypes or phenotypes tend to mate. **Negative assortative mating** occurs when individuals with unlike genotypes or phenotypes tend to mate (also called disassortative mating). Both of these general types of non-random mating will impact expected genotype frequencies in a population. This section describes the impacts of non-random mating on genotype frequencies and introduces a commonly used measure of non-random mating that can be utilized to estimate mating patterns in natural populations.

Mating among related individuals, termed **consanguineous mating** or **biparental inbreeding**, increases the probability that the resulting progeny are homozygous compared to random mating. This occurs since relatives, by definition, are more likely than two random individuals to share one or two alleles that were inherited from ancestors they share in common (this makes mating among relatives a form of assortative mating). Therefore, when related individuals mate their progeny have a higher chance of receiving the same allele from both parents, giving them a greater chance of having a homozygous genotype. **Sexual autogamy** or **self-fertilization** is an extreme example of consanguineous mating where an individual can mate with itself by virtue of possessing reproductive organs of both sexes. Many plants and some animals, such as the nematode

### Problem box 2.3 Inheritance for corn kernel phenotypes

Corn kernels are individual seeds that display a wide diversity of phenotypes (see Fig. 2.10 and Plate 2.10). In a total of 3816 corn seeds, the following phenotypes were observed:

Purple, smooth	2058
Purple, wrinkled	728
Yellow, smooth	769
Yellow, wrinkled	261

Are these genotype frequencies consistent with inheritance due to one locus with three alleles or two loci each with two alleles?



**Figure 2.10** Corn cobs demonstrating yellow and purple seeds that are either wrinkled or smooth. For a color version of this image see Plate 2.10.

*Caenorhabditis elegans*, are hermaphrodites that can mate with themselves.

There are also cases of disassortative mating, where individuals with unlike genotypes have a higher probability of mating. A classic example in mammals is mating based on genotypes at major histocompatibility complex (MHC) loci, which produce proteins involved in self/non-self recognition in immune response. Mice are able to recognize individuals with similar MHC genotypes via odor, and based on these odors avoid mating with individuals possessing a similar MHC genotype. Experiments where young mice were raised in nests of either their true parents or foster parents (called cross-fostering) showed that mice learn to avoid mating with individuals possessing odor cues similar to their nest-mates' rather than avoiding MHC-similar individuals per se (Penn & Potts 1998). This suggests mice learn the odor of family members in the nest and avoid mating with individuals with similar odors, indirectly leading to disassortative mating at MHC loci as well as the avoidance of consanguineous mating. One hypothesis to explain the evolution of disassortative mating at MHC loci is that the behavior is adaptive since progeny with

higher heterozygosity at MHC loci may have more effective immune response. There is also evidence that humans prefer individuals with dissimilar MHC genotypes (Wedekind & Furi 1997).

The effects of non-random mating on genotype frequencies can be measured by comparing Hardy–Weinberg expected frequency of heterozygotes, which assumes random mating, with observed heterozygote frequencies in a population. A quantity called the **fixation index**, symbolized by  $F$  (or sometimes  $f$ , although it never will be in this book, since  $f$  is reserved for the inbreeding coefficient as introduced later in section 2.6), is commonly used to compare how much heterozygosity is present in an actual population relative to expected levels of heterozygosity under random mating:

$$F = \frac{H_e - H_o}{H_e} \quad (2.9)$$

where  $H_e$  is the Hardy–Weinberg expected frequency of heterozygotes based on population allele frequencies and  $H_o$  is the observed frequency of heterozygotes. Dividing the difference between the expected and

### Interact box 2.3 Assortative mating and genotype frequencies

The impact of assortative mating on genotype frequencies can be simulated in PopGene.S<sup>2</sup>. The program models several non-random mating scenarios that can be selected under the **Mating Models** menu. The results in each case are presented on a De Finetti diagram, where genotype and allele frequencies can be followed over multiple generations.

Start with the **Positive w/o dominance** model of mating. In this case only like genotypes are able to mate (e.g. AA mates only with AA, Aa mates only with Aa, and aa mates only with aa). Take the time to write out Punnett squares to predict progeny genotype frequencies for each of the matings that takes place. Enter initial genotype frequencies of  $P(AA) = 0.25$  and  $P(Aa) = 0.5$  ( $P(aa)$  is determined by subtraction) as a logical place to start. At first, run the simulation for 30 generations. With these values and mating patterns, what happens to the frequency of heterozygotes? What happens to the allele frequencies? Next try other initial genotype frequencies that vary the allele frequencies and that are both in and out of Hardy–Weinberg proportions.

Next run both the **Positive with dominance** and **Negative** (Dissassortative matings) models. In the **Positive with dominance** model, the AA and Aa genotypes have identical phenotypes. Mating can therefore take place among any pairing of genotypes with the dominant phenotype or between aa individuals with the recessive phenotype. In the **Negative** mating model, only unlike genotypes can mate. Take the time to write out Punnett squares to predict progeny genotype frequencies for each of the matings that takes place. In each case, use a set of the same genotype frequencies that you employed in the **Positive w/o dominance** mating model. How do all types of non-random mating affect genotype frequencies? How do they affect allele frequencies?

observed heterozygosity by the expected heterozygosity expresses the difference in the numerator as a percentage of the expected heterozygosity. Even if the difference in the numerator may seem small, it may be large relative to the expected heterozygosity. Dividing by the expected heterozygosity also puts  $F$  on a convenient scale of  $-1$  and  $+1$ . Negative values indicate heterozygote excess and positive values indicate homozygote excess relative to Hardy–Weinberg expectations. In fact, the fixation index can be interpreted as the correlation between the two alleles sampled to make a diploid genotype (see the Appendix for an introduction to correlation if necessary). Given that one allele has been sampled from the population, if the second allele tends to be identical there is a positive correlation (e.g. A and then A or a and then a), if the second allele tends to be different there is a negative correlation (e.g. A and then a or a and then A), and if the second allele is independent there is no correlation (e.g. equally likely to be A or a). With random mating, no correlation is expected between the first and second allele sampled to make a diploid genotype.

**Assortative mating** Mating patterns where individuals do not mate in proportion to their genotype frequencies in the population; mating that is more (positive assortative mating) or less (negative assortative mating) frequent with respect to genotype or genetically based phenotype frequency than expected by random combination.

**Consanguineous mating** Mating between related individuals that can take the form

of biparental inbreeding (mating between two related individuals) or sexual autogamy (self-fertilization).

**Fixation index ( $F$ )** The proportion by which heterozygosity is reduced or increased relative to the heterozygosity in a randomly mating population with the same allele frequencies.

Let's work through an example of genotype data for one locus with two alleles that can be used to estimate the fixation index. Table 2.8 gives observed counts and frequencies of the three genotypes in a sample of 200 individuals. To estimate the fixation index from these data requires an estimate of allele frequencies first. The allele frequencies can then be used to determine expected heterozygosity under the assumptions of Hardy–Weinberg. If  $p$  represents the frequency of the B allele,

$$\hat{p} = \frac{142 + \frac{1}{2}(28)}{200} = 0.78 \quad (2.10)$$

using the genotype counting method to estimate allele frequency (Table 2.8 uses the allele counting method). The frequency of the b allele,  $q$ , can be estimated directly in a similar fashion or by subtraction ( $\hat{q} = 1 - \hat{p} = 1 - 0.78 = 0.22$ ) since there are only two alleles in this case. The Hardy–Weinberg expected frequency of heterozygotes is  $H_e = 2\hat{p}\hat{q} = 2(0.78)(0.22) = 0.343$ . It is then simple to estimate the fixation index using the observed and expected heterozygosities:

$$\hat{F} = \frac{0.343 - 0.14}{0.343} = 0.59 \quad (2.11)$$

**Table 2.8** Observed genotype counts and frequencies in a sample of  $N = 200$  individuals for a single locus with two alleles. Allele frequencies in the population can be estimated from the genotype frequencies by summing the total count of each allele and dividing it by the total number of alleles in the sample ( $2N$ ).

Genotype	Observed	Observed frequency	Allele count	Allele frequency
BB	142	$\frac{142}{200} = 0.71$	284 B	$\hat{p} = \frac{284 + 28}{400} = 0.78$
Bb	28	$\frac{28}{200} = 0.14$	28 B, 28 b	
bb	30	$\frac{30}{200} = 0.15$	60 b	$\hat{q} = \frac{60 + 28}{400} = 0.22$

**Table 2.9** Estimates of the fixation index ( $\hat{F}$ ) for various species and breeds based on pedigree or molecular genetic marker data.

Species	Mating system	$\hat{F}$	Method	Reference
Human				
<i>Homo sapiens</i>	Outcrossed	0.0001–0.046	Pedigree	Jorde 1997
Snail				
<i>Bulinus truncates</i>	Selfed and outcrossed	0.6–1.0	Microsatellites	Viard et al. 1997
Domestic dogs				
Breeds combined	Outcrossed	0.33	Allozyme	Christensen et al. 1985
German Shepherd	Outcrossed	0.10		
Mongrels	Outcrossed	0.06		
Plants				
<i>Arabidopsis thaliana</i>	Selfed	0.99	Allozyme	Abbott and Gomes 1989
<i>Pinus ponderosa</i>	Outcrossed	–0.37	Allozyme	Brown 1979

In this example there is a clear deficit of heterozygotes relative to Hardy–Weinberg expectations. The population contains 59% fewer heterozygotes than would be expected in a population with the same allele frequencies that was experiencing random mating and the other conditions set out in the assumptions of Hardy–Weinberg. Interpreted as a correlation between the allelic states of the two alleles in a genotype, this value of the fixation index tells us that the two alleles in a genotype are much more frequently of the same state than expected by chance.

In biological populations, a wide range of values has been observed for the fixation index (Table 2.9). Fixation indices have frequently been estimated with allozyme data (see Box 2.2). Estimates of  $\hat{F}$  are generally correlated with mating system. Even in species where individuals possess reproductive organs of one sex only (termed **dioecious** individuals), mating among relatives can be common and ranges from infrequent to almost invariant. In other cases, mating is essentially random or complex mating and social systems have evolved to prevent consanguineous mating. Pure-breed dogs are an example where mating among relatives has been enforced by humans to develop lineages with specific phenotypes and behaviors, resulting in high fixation indices in some breeds. Many plant species possess both male and female sexual functions (hermaphrodites) and exhibit an extreme form of consanguineous mating, self-fertilization, that causes rapid loss of heterozygosity. In the case of Ponderosa pines in Table 2.9, the excess of heterozygotes may be due to natural selection against homozygotes

at some loci (inbreeding depression). This makes the important point that departures from Hardy–Weinberg expected genotype frequencies estimated by the fixation index are potentially influenced by processes in addition to the mating system. Genetic loci free of the influence of other processes such as natural selection are often sought to estimate  $\hat{F}$ . In addition,  $\hat{F}$  can be estimated using the average of multiple loci, which will tend to reduce bias since loci will differ in the degree they are influenced by other processes and outliers will be apparent.

Extending the fixation index to loci with more than two alleles just requires a means to calculate the expected heterozygosity ( $H_e$ ) for an arbitrary number of alleles at one locus. This can be accomplished by adding up all of the expected frequencies of each possible homozygous genotype and subtracting this total from one or summing the expected frequencies of all heterozygous genotypes:

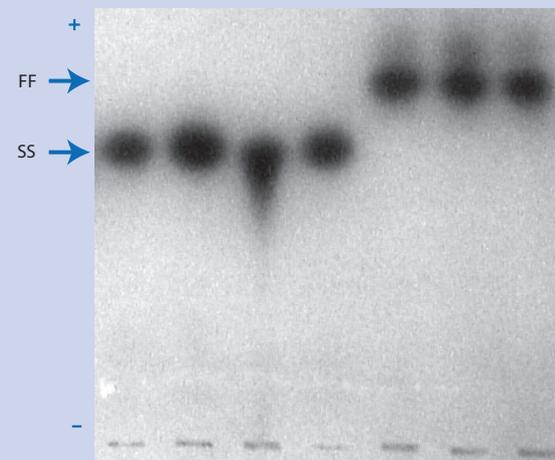
$$H_e = 1 - \sum_{i=1}^k p_i^2 = \sum_{i=1}^{k-1} \sum_{j=i+1}^k 2p_i p_j \quad (2.12)$$

where  $k$  is the number of alleles at the locus, the  $p_i^2$  and  $2p_i p_j$  terms represent the expected genotype frequencies based on allele frequencies, and the  $\sum_{i=1}^k$  (pronounced “sigma”) indicates summation of the frequencies of the  $k$  homozygous genotypes. This quantity is also called the **gene diversity** (Nei 1987). The expected heterozygosity can be adjusted

### Box 2.2 Protein locus or allozyme genotyping

Determining the genotypes of individuals at enzymatic protein loci is a rapid technique to estimate genotype frequencies in populations. Protein analysis was the primary molecular genotyping technique for several decades before DNA-based techniques became widely available. Alleles at loci that code for proteins with enzymatic function can be ascertained in a multi-step process. First, fresh tissue samples are ground up under conditions that preserve the function of proteins. Next, these protein extracts are loaded onto starch gels and exposed to an electric field. The electrical current results in electrophoresis where proteins are separated based on their ratio of molecular charge to molecular weight. Once electrophoresis is complete, the gel is then “stained” to visualize specific enzymes. The primary biochemical products of protein enzymes are not themselves visible. However, a series of biochemical reactions in a process called enzymatic coupling can be used to eventually produce a visible product (often nitro blue tetrazolium or NBT) at the site where the enzyme is active (see Fig. 2.11). If different DNA sequences at a protein enzyme locus result in different amino acid sequences that differ in net charge, then multiple alleles will appear in the gel after staining. The term allozyme (also known as isozyme) is used to describe the multiple allelic staining variants at a single protein locus. Allozyme electrophoresis and staining detects

only a subset of genetic variation at protein coding loci. Amino acid changes that are charge-neutral and nucleotide changes that are synonymous (do not alter the amino acid sequence) cannot be detected by allozyme electrophoresis methods. Refer to Manchenko (2003) for a technical introduction and detailed methods of allozyme detection.



**Figure 2.11** An allozyme gel stained to show alleles at the phosphoglucosmutase or PGM locus in striped bass and white bass. The right-most three individuals are homozygous for the faster-migrating allele (FF genotype) while the left-most four individuals are homozygous for the slower-migrating allele (SS genotype). No double-banded heterozygotes (FS genotype) are visible on this gel. The + and – indicate the cathode and anode ends of the gel, respectively. Wells where the individuals samples were loaded into the gel can be seen at the bottom of the picture. Gel picture kindly provided by J. Epifanio.

for small samples by multiplying  $H_e$  by  $2N/(2N - 1)$  where  $N$  is the total number of genotypes (Nei & Roychoudhury 1974), a correction that makes little difference unless  $N$  is about 50 or fewer individuals. In a similar manner, the observed heterozygosity ( $H_o$ ) is the sum of the frequencies of all heterozygotes observed in a sample of genotypes:

$$\hat{H}_o = \sum_{i=1}^h H_i \quad (2.13)$$

where the observed frequency of each heterozygous genotype  $H_i$  is summed over the  $h = k(k - 1)/2$

heterozygous genotypes possible with  $k$  alleles. Finally, both  $H_e$  and  $H_o$  can be averaged over multiple loci to obtain mean heterozygosity estimates for two or more loci. Heterozygosity provides one of the basic measures of genetic variation in population genetics.

The fixation index as a measure of deviation from expected levels of heterozygosity is a critical concept that will appear in several places later in this text. The fixation index plays a conceptual role in understanding the effects of population size on heterozygosity (Chapter 3) and also serves as an estimator of the impact of population structure on the distribution of genetic variation (Chapter 4).

## 2.6 Mating among relatives

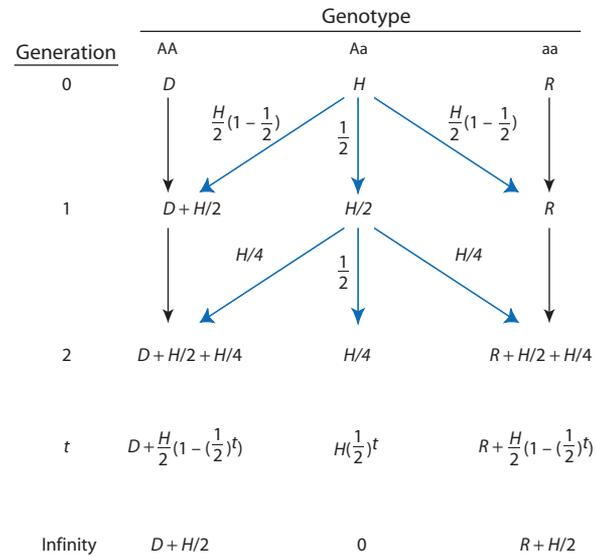
- Consanguineous mating alters genotype frequencies but not allele frequencies.
- Mating among relatives and the probability that two alleles are identical by descent.
- Inbreeding depression and its possible causes.
- The many meanings of inbreeding.

The previous section of this chapter showed how non-random mating can increase or decrease the frequency of heterozygote genotypes compared to the frequency that is expected with random mating. The last section also introduced the fixation index as well as ways to quantify heterozygosity in a population. This section will build on that foundation to show two concepts: (i) the consequences of non-random mating on allele and genotype frequencies in a population and (ii) the probability that two alleles are identical by descent. The focus will be on positive genotypic assortative mating (like genotypes mate) or inbreeding since this will eventually be helpful to understand genotype frequencies in small populations. The end of this section will consider some of the consequences of inbreeding and the evolution of autogamy.

### Impacts of inbreeding on genotype and allele frequencies

Let's develop an example to understand the impact of inbreeding on genotype and allele frequencies in a population. Under complete positive assortative mating or selfing, individuals mate with another individual possessing an identical genotype. Figure 2.12 diagrams the process of positive genotypic assortative mating for a diallelic locus, following the frequencies of each genotype through time. Initially, the frequency of the heterozygote is  $H$  but this frequency will be halved each generation. A Punnett square for two heterozygotes shows that half of the progeny are heterozygotes ( $H/2$ ). The other half of the progeny are homozygotes ( $H/2$ ), composed of one-quarter of the original heterozygote frequency of each homozygote genotype ( $H/2(1 - 1/2)$ ). It is obvious that matings among like homozygotes will produce only identical homozygotes, so the homozygote genotypes each yield a constant frequency of homozygous progeny each generation. In total, however, the frequency of the homozygous genotypes increases by a factor

of  $\frac{H}{2} \left(1 - \frac{1}{2}\right)$  each generation due to homozygous



**Figure 2.12** The impact of complete positive genotypic assortative mating (like genotypes mate) or self-fertilization on genotype frequencies. The initial genotype frequencies are represented by  $D$ ,  $H$ , and  $R$ . When either of the homozygotes mates with an individual with the same genotype, all progeny bear their parent's homozygous genotype. When two heterozygote individuals mate, the expected genotype frequencies among the progeny are one half heterozygous genotypes and one quarter of each homozygous genotype. Every generation the frequency of the heterozygotes declines by one-half while one-quarter of the heterozygote frequency is added to the frequencies of each homozygote (diagonal arrows). Eventually, the population will lose all heterozygosity although allele frequencies will remain constant. Therefore, assortative mating or self-fertilization changes the packing of alleles in genotypes but not the allele frequencies themselves.

progeny of the heterozygous genotypes. If the process of complete assortative mating continues, the population rapidly loses heterozygosity and approaches a state where the frequency of heterozygotes is 0.

As an example, imagine a population where  $p = q = 0.5$  that has Hardy–Weinberg genotype frequencies  $D = 0.25$ ,  $H = 0.5$ , and  $R = 0.25$ . Under complete positive assortative mating, what would be the frequency of heterozygotes after five generations? Using Fig. 2.12, at time  $t = 5$  heterozygosity would be  $H(1/2)^5 = H(1/32) = 1/64$  or 0.016. This is a drastic reduction in only five generations.

Genotype frequencies change quite rapidly under complete assortative mating, but what about allele frequencies? Let's employ the same example population with  $p = q = 0.5$  and Hardy–Weinberg genotype frequencies  $D = 0.25$ ,  $H = 0.5$ , and  $R = 0.25$  to answer the question. For both of the homozygous genotypes, the initial frequencies would be  $D = R = (0.5)^2 = 0.25$ . In Fig. 2.12, the contribution of each

homozygote genotype frequency from mating among heterozygotes after five generations is  $H/2(1 - (1/2)^5) = H/2(1 - 1/32) = H/2(31/32)$ . With the initial frequency of  $H = 0.5$ ,  $H/2(31/32) = 0.242$ . Therefore, the frequencies of both homozygous genotypes are  $0.25 + 0.242 = 0.492$  after five generations. It is also apparent that the total increase in homozygotes ( $31/32$ ) is exactly the same as the total decrease in heterozygotes ( $1/32$ ), so that the allele frequencies in the population have remained constant. After five generations of assortative mating in this example, genotypes are much more likely to contain two identical alleles than they are to contain two unlike alleles. This conclusion is also reflected in the value of the fixation index for this example,  $\hat{F} = (0.5 - 0.016)/0.5 = 0.968$ . In general, *positive assortative mating or inbreeding changes the way in which alleles are "packaged" into genotypes*, increasing the frequencies of all homozygous genotypes by the same total amount that heterozygosity is decreased, but allele frequencies in a population do not change.

The fact that allele frequencies do not change over time can also be shown elegantly with some simple algebra. Using the notation in Fig. 2.12 and defining the frequency of the A allele as  $p$  and the a allele as  $q$  with subscripts to indicate generation, allele frequencies can be determined by the genotype counting method as  $p_0 = D_0 + 1/2H_0$  and  $q_0 = R_0 + 1/2H_0$ . Figure 2.12 also provides the expressions for genotype frequencies from one generation to the next:  $D_1 = D_0 + 1/4H_0$ ,  $H_1 = 1/2H_0$ , and  $R_1 = R_0 + 1/4H_0$ . We can then use these expressions to predict allele frequency in one generation:

$$p_1 = D_1 + 1/2H_1 \quad (2.14)$$

as a function of genotype frequencies in the previous generation using substitution for  $D_1$  and  $H_1$ :

$$p_1 = D_0 + 1/4H_0 + 1/2(1/2H_0) \quad (2.15)$$

which simplifies to:

$$p_1 = D_0 + 1/2H_0 \quad (2.16)$$

and then recognizing that the right hand side is equal to the frequency of A in generation 0:

$$p_1 = p_0 \quad (2.17)$$

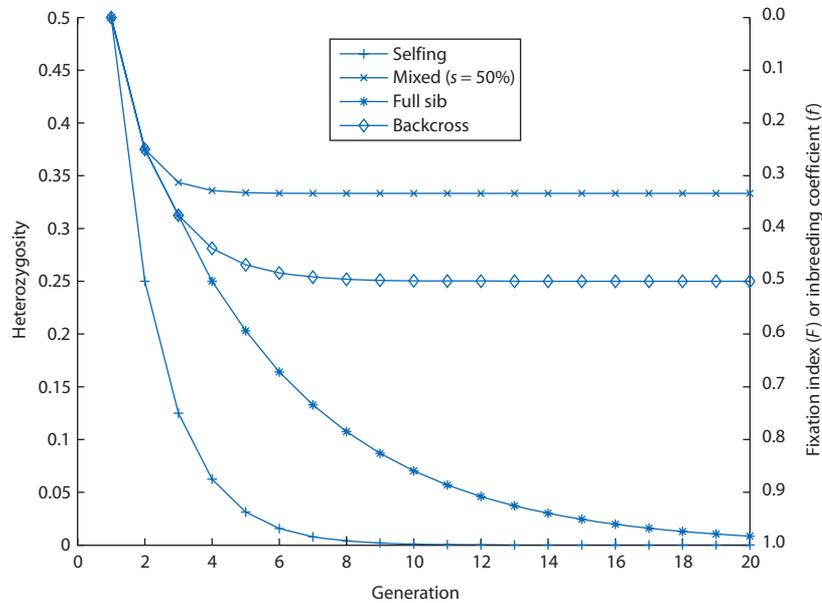
Thus, allele frequencies remain constant under complete assortative mating. As practice, you should carry out the algebra for the frequency of the a allele.

Under complete self-fertilization heterozygosity declines very rapidly. There can also be partial self-fertilization in a population (termed **mixed mating**), where some matings are self-fertilization and others are between two individuals (called **outcrossing**). In addition, many organisms are not capable of self-fertilization but instead engage in biparental inbreeding to some degree. In general, these forms of inbreeding will reduce heterozygosity compared to random mating, although they will not drive heterozygosity toward zero as in the case of complete selfing. The rate of decline in heterozygosity can be determined for many possible types of mating systems and a few examples are shown in Fig. 2.13. Regardless of the specifics of the form of consanguineous mating that occurs, it remains true that inbreeding causes alleles to be packaged more frequently as homozygotes (heterozygosity declines) and inbreeding does not alter allele frequencies in a population.

#### ***Inbreeding coefficient and autozygosity in a pedigree***

The effects of consanguineous mating can also be thought of as increasing the probability that two alleles at one locus in an individual are inherited from the same ancestor. Such a genotype would be homozygous and considered **autozygous** since the alleles were inherited from a common ancestor. If the two alleles are not inherited from the same ancestor in the recent past, we would call the genotype **allozygous** (*allo-* means other). You are probably already familiar with autozygosity, although you may not recognize it as such. Two times the probability of autozygosity (since diploid individuals have two alleles) is commonly expressed as the **degree of relatedness** among relatives. For example, full siblings (full brothers and sisters) are one-half related and first cousins are one-eighth related. Using a pedigree and tracing the probabilities of inheritance of an allele, the autozygosity and the basis of average relatedness can be seen.

Inbreeding in the autozygosity sense, often called the **coefficient of inbreeding** ( $f$ ), can best be seen in a pedigree such as that shown in Fig. 2.14. Fig. 2.14a gives a hypothetical pedigree for four generations. The pedigree can be used to determine the probability that the fourth-generation progeny, labeled G, have autozygous genotypes due to individual A being a common ancestor of both their maternal and paternal parents. To make the process simpler, Fig. 2.14b strips away all of the external

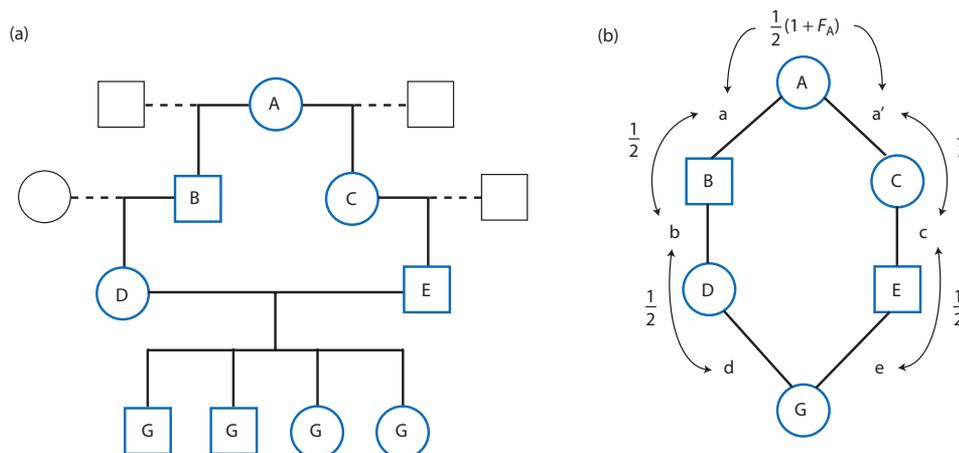


**Figure 2.13** The impact of various systems of consanguineous mating or inbreeding on heterozygosity, the fixation index ( $F$ ), and the inbreeding coefficient ( $f$ ) over time. Initially, the population has allele frequencies of  $p = q = 0.5$  and all individuals are assumed randomly mated. Since inbreeding does not change allele frequencies, expected heterozygosity ( $H_e$ ) remains 0.5 for all 20 generations. As inbreeding progresses, observed heterozygosity declines and the fixation index and inbreeding coefficient increase. Selfing is 100% self-fertilization whereas mixed mating is 50% of the population selfing and 50% randomly mating. Full sib is brother–sister or parent–offspring mating. Backcross is one individual mated to its progeny, then to its grand progeny, then to its great-grand progeny and so on, a mating scheme that is difficult to carry on for many generations. Change in the coefficient of inbreeding over time is based on the following recursion equations: selfing  $f_{t+1} = 1/2(1 + F_t)$ ; mixed  $f_{t+1} = 1/2(1 + F_t)(s)$  where  $s$  is the selfing rate; full sib  $f_{t+2} = 1/4(1 + 2f_{t+1} + f_t)$ ; backcross  $f_{t+1} = 1/4(1 + 2f_t)$  (see Falconer & MacKay 1996 for detailed derivations).

ancestors and shows only the paths where alleles could be inherited in the progeny from individual A.

To begin the process of determining the autozygosity for G, it is necessary to determine the probability that A transmitted the same allele to individuals B and C

and C, or in notation  $P(a = a')$ . With two alleles designated 1 and 2, there are only four possible patterns of allelic transmission from A to B and C, shown in Fig. 2.15. In only half of these cases do B and C inherit an identical allele from A, so  $P(a = a') = 1/2$ .



**Figure 2.14** Average relatedness and autozygosity as the probability that two alleles at one locus are identical by descent. (a) A pedigree where individual A has progeny that are half-siblings (B and C). B and C then produce progeny D and E, which in turn produce offspring G. (b) Only the paths of relatedness where alleles could be inherited from A, with curved arrows to indicate the probability that gametes carry alleles identical by descent. Upper-case letters for individuals represent diploid genotypes and lower-case letters indicate allele copies within the gametes produced by the genotypes. The probability that A transmits a copy of the same allele to B and C depends on the degree of inbreeding for individual A, or  $F_A$ .

Path of inheritance	a	a'	a	a'	a	a'	a	a'
Allele	1	1	2	2	1	2	2	1
	⏟				⏟			
	Alleles identical by descent (IBD)				Alleles not identical by descent			

**Figure 2.15** The possible patterns of transmission from one parent to two progeny for a locus with two alleles. Half of the outcomes result in the two progeny inheriting an allele that is identical by descent. The a and a' refer to paths of inheritance in the pedigree in Fig. 2.14b.

This probability would still be  $1/2$  no matter how many alleles were present in the population, since the probability arises from the fact that diploid genotypes have only two alleles.

To have a complete account of the probability that B and C inherit an identical allele from A, we also need to take into account the past history of A's genotype since it is possible that A was itself the product of mating among relatives. If A was the product of some level of biparental inbreeding, then the chance that it transmits alleles identical by descent to B and C is greater than if A was from a randomly mating population. Another way to think of it is, with A being the product of some level of inbreeding instead of random mating, the chances that the alleles transmitted to B and C are not identical (see Fig. 2.14b) will be less than  $1/2$  by the amount that A is inbred. If the degree to which A is inbred (or the probability that A is autozygous) is  $F_A$ , then the total probability that B and C inherit the same allele is:

$$P(a = a') = 1/2 + 1/2 F_A = 1/2(1 + F_A) \quad (2.18)$$

If  $F_A$  is 0 in equation 2.18, then the chance of transmitting the same allele to B and C reduces to the  $1/2$  expected in a randomly mating population.

For the other paths of inheritance in Fig. 2.14, the logic is similar to determine the probability that an allele is identical by descent. For example, what is the probability that the allele in gamete d is identical by descent to the allele in gamete b, or  $P(b = d)$ ? When D mated it passed on one of two alleles, with a probability of  $1/2$  for each allele. One allele was inherited from each parent, so there is a  $1/2$  chance of transmitting a maternal or paternal allele. This makes  $P(b = d) = 1/2$ . (Just like with individual A,  $P(b = d)$  could also be increased to the extent that B was inbred, although random mating for all genotypes but A is assumed here for simplicity.) This same logic applies to all other paths in the pedigree that connect A and the

progeny G. The probability of a given allele being transmitted along a path is independent of the probability along any other path, so the probability of autozygosity (symbolized as  $f$  to distinguish it from the pre-existing autozygosity of individual A) over the entire pedigree for any of the G progeny is:

$$\begin{aligned} f_G = f_{DE} &= \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2}(1 + f_A) \times \frac{1}{2} \times \frac{1}{2} \\ &\quad P(b = d) \quad P(a = b) \quad P(a = a') \quad P(a' = c) \quad P(c = e) \\ &= (1/2)^5(1 + f_A) = 1/32(1 + f_A) \end{aligned} \quad (2.19)$$

since independent probabilities can be multiplied to find the total probability of an event. This is equivalent to the average relatedness among half-cousins. In general for pedigrees,  $f = (1/2)^i(1 + F_A)$  where A is the common ancestor and  $i$  is the number of paths or individuals over which alleles are transmitted. A trick is to write down the chain of individuals starting with the common ancestor and ending with the individuals of interest and count the individuals along paths of inheritance (not including the individuals of interest). That gives  $\text{GDBACEG}$  or five ancestors, yielding a result identical to equation 2.19.

Although it is useful to determine the inbreeding coefficient (autozygosity) for a specific pedigree, the more general point is to see mating among relatives as a process that increases autozygosity in a population. When individuals have common relatives, the chance that their genotype contains loci with alleles identical by descent is increased. Further, the inbreeding coefficient or autozygosity measured for a specific pedigree is identical in concept and interpretation to the departure from Hardy–Weinberg measured by the fixation index (assuming most of the deviation from Hardy–Weinberg expectations is caused by non-random mating and not other population genetic processes). Both express the probability that two alleles in a genotype are identical due to common ancestry.

The departure from Hardy–Weinberg expected genotype frequencies, the autozygosity or inbreeding coefficient, and the fixation index are all interrelated. Another way of stating the results that were developed in Fig. 2.12 is that  $f$  measures the degree to which Hardy–Weinberg genotype proportions are not met, due to inbreeding:

$$\begin{aligned} D &= p^2 + fpq \\ H &= 2pq - f2pq \\ R &= q^2 + fpq \end{aligned} \quad (2.20)$$

With consanguineous mating, the decline in heterozygosity is proportional to the increase in the inbreeding coefficient, shown by substituting  $H_e$  for  $2pq$  in equation 2.20 and rearranging to give

$$H = H_e(1 - f) \quad (2.21)$$

where  $H_e$  is the Hardy–Weinberg expected heterozygosity based on population allele frequencies. Rearranging equation 2.21 in terms of the inbreeding coefficient gives:

$$f = 1 - \frac{H}{H_e} \quad (2.22)$$

This is really exactly the same quantity as the fixation index (equation 2.9)

$$f = 1 - \frac{H}{H_e} = \frac{H_e}{H_e} - \frac{H}{H_e} = \frac{H_e - H}{H_e} \quad (2.23)$$

The inbreeding coefficient and the fixation index are measures of excess homozygosity and therefore are just different ways of expressing the heterozygosity. Returning to Fig. 2.13 helps show the equivalence of the inbreeding coefficient, the fixation index, and the decline in heterozygosity in several specific cases of regular consanguineous mating. Remember that in all cases in Fig. 2.13, the Hardy–Weinberg expected heterozygosity is  $1/2$ .

**Allozygous genotype** A homozygous or heterozygous genotype composed of two alleles not inherited from a recent common ancestor.

**Autozygosity or inbreeding coefficient ( $f$ )** The probability that two alleles in a homozygous genotype are identical by descent.

**Autozygous genotype** A homozygous genotype composed of two identical alleles that are inherited from a common ancestor.

**Identity by descent** Transmission from a common ancestor.

**Relatedness** The expected proportion of alleles between two individuals that are identical by descent; twice the autozygosity.

### Phenotypic consequences of inbreeding

The process of consanguineous mating or inbreeding is associated with changes in the mean phenotype within a population. These changes arise from two general causes: changes in genotype frequencies in a population per se and fitness effects associated with changes in genotype frequencies.

The mean phenotype of a population will be impacted by the changes in genotype frequency caused by inbreeding. To show this it is necessary to introduce terminology to express the phenotype associated with a given genotype, a topic covered in much greater detail and explained more fully in Chapters 9 and 10 in this volume. We will assign AA genotypes the phenotype  $+a$ , heterozygotes the phenotype  $d$ , and aa homozygotes the phenotype  $-a$ . Each genotype contributes to the overall phenotype based on how frequent it is in the population. The mean phenotype in a population is then the sum of each genotype-frequency-weighted phenotype (Table 2.10). When there is no dominance, the phenotype of the heterozygotes is exactly intermediate between the phenotypes of the two homozygotes and  $d = 0$ . In that case, it is easy to see that inbreeding will not change the mean phenotype in the population since both homozygous genotypes increase by the same amount and their effects on the mean phenotype cancel out (mean =  $ap^2 + afpq + d2pq - df2pq - aq^2 - afpq$ , where the heterozygote terms are crossed out since  $d = 0$ ). When there is some degree of dominance (positive  $d$  indicates the phenotype of Aa is like that of AA while negative  $d$  indicates the phenotype of Aa is like that of aa), then the mean phenotype of the population will change with consanguineous mating since heterozygotes will become less frequent. If dominance is in the direction of the  $+a$  phenotype ( $d > 0$ ), then inbreeding will reduce the population mean because the heterozygote frequency will drop. Similarly, if dominance is in the direction of  $-a$  ( $d < 0$ ) then inbreeding will increase the population mean again because the heterozygote frequency decreases. It is also true in the case of dominance that a return to random mating will restore the frequencies of heterozygotes and return the population mean to its original value before inbreeding. These changes in the population mean phenotype are simply a consequence of changing the genotype frequencies when there is no change in the allele frequencies.

There is a wealth of evidence that inbreeding has **deleterious** (harmful or damaging) consequences

**Table 2.10** The mean phenotype in a population that is experiencing consanguineous mating. The inbreeding coefficient is  $f$  and  $d = 0$  when there is no dominance.

Genotype	Phenotype	Frequency	Contribution to population mean
AA	$+a$	$p^2 + fpq$	$ap^2 + afpq$
Aa	$d$	$2pq - f2pq$	$d2pq - df2pq$
aa	$-a$	$q^2 + fpq$	$-aq^2 - afpq$

Population mean:  $ap^2 + d2pq - df2pq - aq^2 = a(p - q) + d2pq(1 - f)$

and is associated with a decline in the average phenotype in a population, a phenomenon referred to as **inbreeding depression**. Since the early twentieth century, studies in animals and plants that have been intentionally inbred provide ample evidence that decreased performance, growth, reproduction, viability (all measures of fitness), and abnormal phenotypes are associated with consanguineous mating. A related phenomenon is **heterosis** or hybrid vigor, characterized by beneficial consequences of increased heterozygosity such as increased viability and reproduction, or the reverse of inbreeding depression. One example is the heterosis exhibited in corn, which has led to the nearly universal use of F1 hybrid seed for agriculture in developed countries.

**Heterosis** The increase in performance, survival, and ability to reproduce of individuals possessing heterozygous loci (hybrid vigor); increase in the population average phenotype associated with increased heterozygosity.  
**Inbreeding depression** The reduction in performance, survival, and ability to reproduce of individuals possessing homozygous loci; decrease in population average phenotype associated with consanguineous mating that increases homozygosity.

There is evidence that humans experience inbreeding depression, based on observed phenotypes in the offspring of couples with known consanguinity. For example, mortality among children of first-cousin marriages was 4.5% greater than for marriages between unrelated individuals measured in a range of human populations (see review by Jorde 1997). Human studies have utilized existing parental pairs with relatively low levels of inbreeding, such as uncle/niece, first cousins, or second cousins, in contrast

to animal and plant studies where both very high levels and a broad range of inbreeding coefficients are achieved intentionally. Drawing conclusions about the causes of variation in phenotypes from such observational studies requires extreme caution, since the prevalence of consanguineous mating in humans is also correlated with social and economic variables such as illiteracy, age at marriage, duration of marriage, and income. These latter variables are therefore not independent of consanguinity and can themselves contribute to variation in phenotypes such as fertility and infant mortality (see Bittles et al. 1991, 2002).

The Mendelian genetic causes of inbreeding depression have been a topic of population genetics research for more than a century. There are two classical hypotheses to explain inbreeding depression and changes in fitness as the inbreeding coefficient increases (Charlesworth & Charlesworth 1999; Carr & Dudash 2003). Both hypotheses predict that levels of inbreeding depression will increase along with consanguineous mating that increases homozygosity, although for different reasons (Table 2.11). The first hypothesis, often called the **dominance hypothesis**, is that increasing homozygosity increases the phenotypic expression of fully and partly recessive alleles with deleterious effects. The second hypothesis is that inbreeding depression is the result of the decrease in the frequency of heterozygotes that occurs with consanguineous mating. This explanation supposes that heterozygotes have higher fitness than homozygotes (heterosis) and is called the **overdominance hypothesis**. In addition, the fitness interactions of alleles at different loci (epistasis; see Chapter 9) may also cause inbreeding depression, a hypothesis that is particularly difficult to test (see Carr & Dudash 2003). These causes of inbreeding depression may all operate simultaneously.

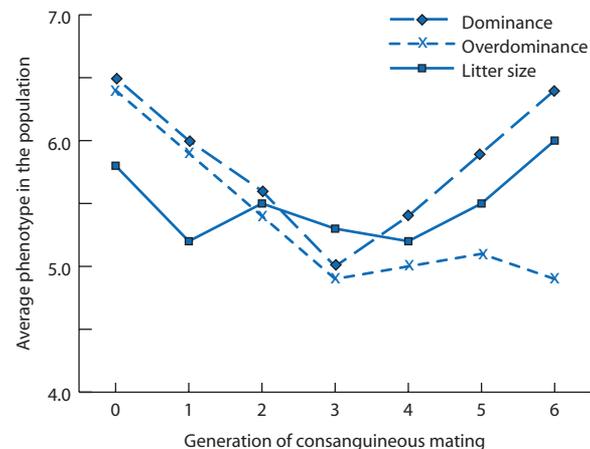
These dominance and overdominance hypotheses make different testable predictions about

**Table 2.11** A summary of the Mendelian basis of inbreeding depression under the dominance and overdominance hypotheses along with predicted patterns of inbreeding depression with continued consanguineous mating.

Hypothesis	Mendelian basis	Low-fitness genotypes	Changes in inbreeding depression with continued consanguineous mating
Dominance	Recessive and partly recessive deleterious alleles	Only homozygotes for deleterious recessive alleles	Purging of deleterious alleles that is increasingly effective as degree of recessiveness increases
Overdominance	Heterozygote advantage or heterosis	All homozygotes	No changes as long as consanguineous mating keeps heterozygosity low

how inbreeding depression (measured as the average phenotype of a population) will change over time with continued consanguineous mating. Under the dominance hypothesis, recessive alleles that cause lowered fitness are more frequently found in homozygous genotypes under consanguineous mating. This exposes the deleterious phenotype and the genotype will decrease in frequency in a population by natural selection (individuals homozygous for such alleles have lower survivorship and reproduction). This reduction in the frequency of deleterious alleles by natural selection is referred to as **purging of genetic load**. Purging increases the frequency of alleles that do not have deleterious effects when homozygous, so that the average phenotype in a population then returns to the initial average it had before the onset of consanguineous mating. In contrast, the overdominance hypothesis does not predict a purging effect with consanguineous mating. With consanguineous mating, the frequency of heterozygotes will decrease and not recover until mating patterns change (see Fig. 2.12). Even if heterozygotes are frequent and have a fitness advantage, each generation of mating and Mendelian segregation will reconstitute the two homozygous genotypes so purging cannot occur. These predictions highlight the major difference between the hypotheses. Inbreeding depression with overdominance arises from genotype frequencies in a population while inbreeding depression with dominance is caused by the frequency of deleterious recessive alleles in a population. Models of natural selection that are relevant to inbreeding depression on population genotype and allele frequencies receive detailed coverage in Chapter 6 in this volume.

Inbreeding depression in many animals and plants appears to be caused, at least in part, by deleterious recessive alleles consistent with the dominance hypothesis (Byers & Waller 1999; Charlesworth & Charlesworth 1999; Crnokrak & Barrett 2002). A classic example of inbreeding depression and recovery of the population mean for litter size in mice is shown in Fig. 2.16. Model research organisms



**Figure 2.16** A graphical depiction of the predictions of the dominance and overdominance hypotheses for the genetic basis of inbreeding depression. The line for dominance shows purging and recovery of the population mean under continued consanguineous mating expected if deleterious recessive alleles cause inbreeding depression. However, the line for overdominance as the basis of inbreeding depression shows no purging effect since heterozygotes continue to decrease in frequency. The results of an inbreeding depression experiment with mice show that litter size recovers under continued brother–sister mating as expected under the dominance hypothesis (Lynch 1977). Only two of the original 14 pairs of wild-caught mice were left at the sixth generation. Not all of the mouse phenotypes showed patterns consistent with the dominance hypothesis.

such as mice, rats, and *Drosophila*, intentionally inbred by schemes such as full-sib mating for 10s or 100s of generations to create highly homozygous, so-called pure-breeding lines, are also not immune to inbreeding depression. Such inbred lines are often founded from multiple families and many of these family lines go extinct from low viability or reproductive failure with habitual inbreeding. This is another type of purging effect due to natural selection that leaves only those lines that exhibit less inbreeding depression, which could be due to dominance, overdominance, or epistasis. Purging is not universally observed in all species and it is likely that inbreeding depression has several genetic causes within species as well as different predominant causes among different species.

The social and economic correlates of inbreeding depression in humans mentioned above are a specific example of environmental effects on phenotypes. Inbreeding depression can be more pronounced when environmental conditions are more severe or limiting. For example, in the plant rose pink (*Sabatia angularis*), progeny from self-fertilizations showed decreasing relative performance when grown in the greenhouse, a garden, and their native habitat, consistent with environmental contributions to the expression of inbreeding depression (Dudash 1990). In another study, the number of surviving progeny for inbred and random-bred male wild mice (*Mus domesticus*) were similar under laboratory conditions, but inbred males sired only 20% of the surviving progeny that random bred males did when under semi-natural conditions due to male–male competition (Meagher et al. 2000). However, not all studies show environmental differences in the expression of inbreeding depression. As an example, uniform levels of inbreeding depression were shown by mosquitoes grown in the laboratory and in natural tree holes where they develop as larvae and pupae in the wild (Armbruster et al. 2000).

The degree of inbreeding depression also depends on the phenotype being considered. In plants, traits early in the life cycle such as germination less often show inbreeding depression than traits later in the life cycle such as growth and reproduction (Husband & Schemske 1996). A similar pattern is apparent in animals, with inbreeding depression most often observed for traits related to survival and reproduction.

Inbreeding depression is a critical concept when thinking about the evolution of mating patterns in plants and animals. Suppose that a single locus

determines whether an individual will self or out-cross and the only allele present in a population is the outcrossing allele. Then imagine that mutation produces an allele at that locus, which, when homozygous, causes an individual to self-fertilize. Such a selfing allele would have a transmission advantage over outcrossing alleles in the population. To see this, consider the number of allele copies at the mating locus transmitted from parents to progeny. Parents with outcrossing alleles mate with another individual and transmit one allele to their progeny. Self-fertilizing parents, however, are both mom and dad to their offspring and transmit two alleles to their progeny. In a population of constant size where each individual contributes an average of one progeny to the next generation, the selfing allele is reproduced twice as fast as an outcrossing allele and would rapidly become fixed in the population (Fisher 1999; see Lande & Schemske 1985). Based on this two-fold higher rate of increase of the selfing allele, complete self-fertilization would eventually evolve unless some disadvantage counteracted the increase of selfed progeny in the population. Inbreeding depression where the average fitness of outcrossed progeny exceeds the average fitness of selfed progeny by a factor of two could play this role. If outcrossed progeny are at a two-fold advantage due to inbreeding depression, then complete outcrossing would evolve. Explaining the existence of populations that engage in intermediate levels of selfing and outcrossing, a mating system common in plants, remains a challenge under these predictions (Byers & Waller 1999).

### *The many meanings of inbreeding*

Unfortunately, the word inbreeding is used as a generic term to describe multiple distinct, although inter-related, concepts in population genetics (Jacquard 1975; Templeton & Read 1994). Inbreeding can apply to:

- consanguinity or kinship of two different individuals;
- autozygosity of two alleles either within an individual or sampled at random;
- the fixation index and Hardy–Weinberg expected and observed genotype frequencies, especially when there is an excess of homozygotes;
- inbreeding depression;
- the description of the mating system of a population or species (as in inbred);

- genetic subdivision of a species into populations that exchange limited levels of gene flow such that individual populations increase in autozygosity;
- the increase in homozygosity in a population due to its finite size.

These different concepts all relate in some way to either the autozygosity or to genotype frequencies in a population, so the connection to inbreeding is clear. Awareness of the different ways the word inbreeding is used as well as an understanding of these different uses will prevent confusion, which can often be avoided simply by using more specific terminology. Remembering that the concepts are interrelated under the general umbrella of inbreeding can also help in realizing the equivalence of the population genetic processes in operation. The next chapter will show how finite population size is equivalent in its effects to inbreeding. Chapter 4 will take up the topic of population subdivision.

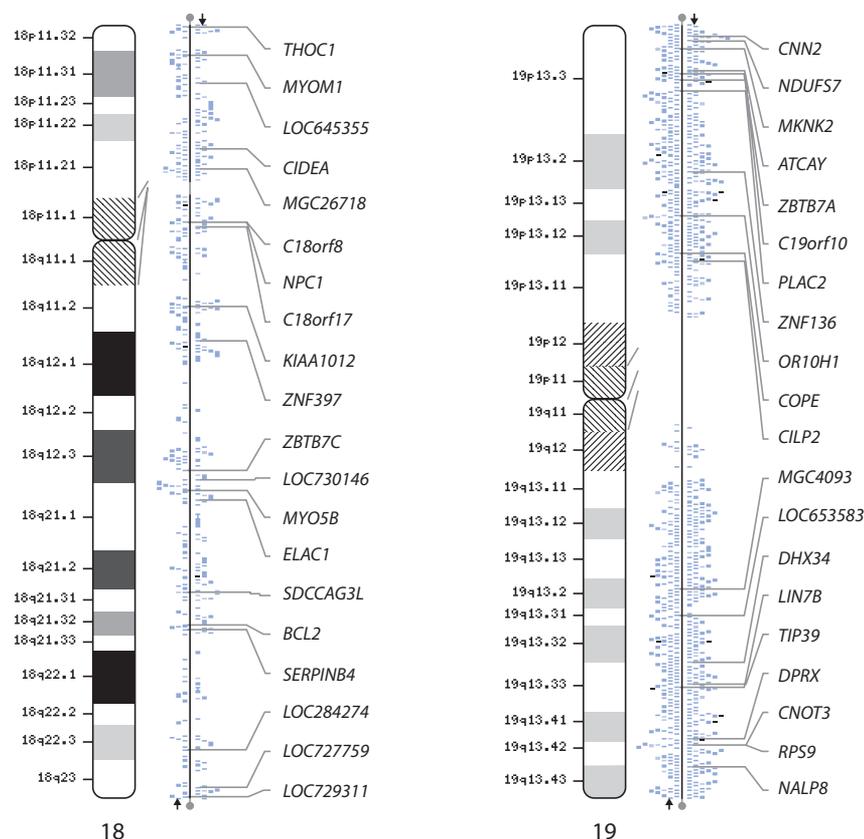
## 2.7 Gametic disequilibrium

- Estimating gametic disequilibrium with  $D$ .
- Approach to gametic equilibrium over time.
- Causes of gametic disequilibrium.

In 1902 Walter Sutton and Theodor Boveri advanced the chromosome theory of heredity. They observed cell division and hypothesized that the discrete bodies seen separating into sets at meiosis and mitosis contained hereditary material that was transmitted from parents to offspring. At the time the concept of chromosomal inheritance presented a paradox. Mendel's second law says that gamete **haplotypes (haploid genotype)** should appear in frequencies proportional to the product of allele frequencies. This prediction conflicted with the chromosome theory of heredity since there are not enough chromosomes to represent each hereditary trait.

To see the problem, take the example of *Homo sapiens* with a current estimate of around 30,000 genes in the nuclear genome. However, humans have only 23 pairs of chromosomes. There are a large number of loci but a small number of chromosomes. So if chromosomes are indeed hereditary molecules, many genes must be on the same chromosome (on average about 1300 genes per chromosome for humans if there are 30,000 genes). This means that some genes are physically **linked** by being located on the same chromosome (see Fig. 2.17). The solution to the paradox is the process of **recombination**. Sister chromatids touch at random points during

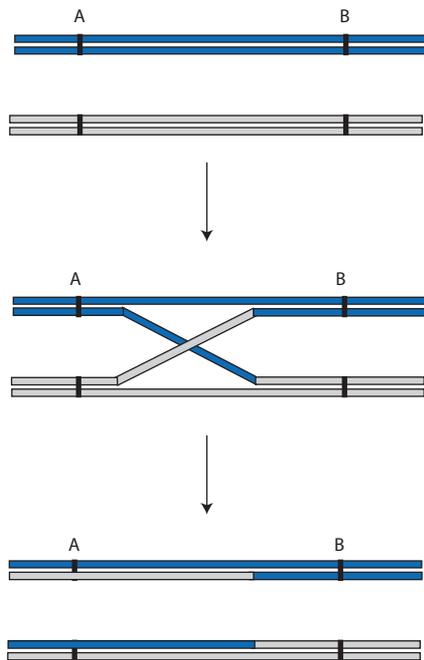
**Figure 2.17** Maps for human chromosomes 18 (left) and 19 (right) showing chromosome regions, the physical locations of identified genes and open reading frames (labeled orf) along the chromosomes, and the names and locations of a subset of genes. Chromosome 18 is about 85 million bp and chromosome 19 is about 67 million bp. Maps from NCBI Map Viewer based on data as of January 2008.



meiosis and exchange short segments, a process known as crossing-over (Fig. 2.18).

Linkage of loci has the potential to impact multi-locus genotype frequencies and violate Mendel's law of independent segregation, which assumes the absence of linkage. To generalize expectations for genotype frequencies for two (or more) loci requires a model that accounts explicitly for linkage by including the rate of recombination between loci. The effects of linkage and recombination are important determinants of whether or not expected genotype frequencies under independent segregation of two loci (Mendel's second law) are met. Autosomal linkage is the general case that will be used to develop expectations for genotype frequencies under linkage.

The frequency of a two-locus gamete haplotype will depend on two factors: (i) allele frequencies and (ii) the amount of recombination between the two loci. We can begin to construct a model based on the recombination rate by asking what gametes are generated by the genotype  $A_1A_2B_1B_2$ . Throughout



**Figure 2.18** A schematic diagram of the process of recombination between two loci, A and B. Two double-stranded chromosomes (drawn in different colors) exchange strands and form a Holliday structure. The crossover event can resolve into either of two recombinant chromosomes that generate new combinations of alleles at the two loci. The chance of a crossover event occurring generally increases as the distance between loci increases. Two loci are independent when the probability of recombination and non-recombination are both equal to  $1/2$ . Gene conversion, a double crossover event without exchange of flanking strands, is not shown.

this section loci are indicated by the letters, alleles at the loci by the numerical subscripts and allele frequencies indicated by  $p_1$  and  $p_2$  for locus A and  $q_1$  and  $q_2$  for locus B. The problem is easier to conceptualize if we draw the two locus genotype as being on two lines akin to chromosomal strands

$$\begin{array}{cc} A_1 & B_1 \\ \hline A_2 & B_2 \end{array}$$

Given this physical arrangement of the two loci, what are the gametes produced during meiosis with and without recombination events?

$A_1B_1$  and  $A_2B_2$  "Coupling" gametes: alleles on the same chromosome remain together (term coined by Bateson and Punnett).

$A_1B_2$  and  $A_2B_1$  "Repulsion" gametes: alleles on the same chromosome seem repulsed by each other and pair with alleles on the opposite strand (term coined by Thomas Hunt Morgan).

The **recombination fraction**, symbolized as  $r$  (or sometimes  $c$ ), refers to the total frequency of gametes resulting from recombination events between two loci. Using  $r$  to express an arbitrary recombination fraction, let's build an expectation for the frequency of coupling and repulsion gametes. If  $r$  is the rate of recombination, then  $1 - r$  is the rate of non-recombination since the frequency of all gametes is one, or 100%. Within each of these two categories of gametes (coupling and repulsion), two types of gametes are produced so the frequency of each gamete type is half that of the total frequency for the gamete category. We can also determine the expected frequencies of each gamete under random association of the alleles at the two loci based on Mendel's law of independent segregation.

Gamete	Frequency		
	Expected	Observed	
$A_1B_1$	$p_1q_1$	$g_{11} = (1 - r)/2$	$1 - r$ is the frequency of all coupling gametes.
$A_2B_2$	$p_2q_2$	$g_{22} = (1 - r)/2$	
$A_1B_2$	$p_1q_2$	$g_{12} = r/2$	$r$ is the frequency of all recombinant gametes.
$A_2B_1$	$p_2q_1$	$g_{21} = r/2$	

The recombination fraction,  $r$ , can be thought of as the probability that a recombination event will occur between two loci. With independent assortment, the coupling and repulsion gametes are in equal frequencies and  $r$  equals  $1/2$  (like the chances of getting heads when flipping a coin). Values of  $r$  less than  $1/2$  indicate that recombination is less likely than non-recombination, so coupling gametes are more frequent. Values of  $r$  greater than  $1/2$  are possible and would indicate that recombinant gametes are more frequent than non-recombinant gametes (although such a pattern would likely be due to a process such as natural selection eliminating coupling gametes from the population rather than recombination exclusively).

The expected frequencies of gametes produced by all possible genotypes for two diallelic loci including the contributions of recombination are derived in Table 2.12. Summing the frequencies of each gamete produced by all genotypes gives the gamete frequencies that will found the next generation. This gives expected gamete frequencies for the more general case of a randomly mating population rather than for a single genotype. The table shows how only two of 10 possible genotypes contribute to the production of recombinant gametes. Most genotypes produce recombinant gametes that are identical to non-recombinant gametes (e.g. the  $A_1B_1/A_1B_2$  genotype produces  $A_1B_1$  and  $A_1B_2$  coupling gametes and  $A_1B_1$  and  $A_1B_2$  repulsion gametes).

We can utilize observed gamete frequencies to develop a measure of the degree to which alleles are

associated within gamete haplotypes. This quantity is called the **gametic disequilibrium** (or sometimes **linkage disequilibrium**) **parameter** and can be expressed by:

$$D = \begin{array}{cc} g_{11}g_{22} & - \\ \text{(coupling term)} & \text{(repulsion term)} \end{array} \quad (2.24)$$

where  $g_{xy}$  stands for a gamete frequency.  $D$  is the difference between the product of the coupling gamete frequencies and the product of the repulsion gamete frequencies. This makes intuitive sense: with independent assortment the frequencies of the coupling and repulsion gamete types are identical and cancel out to give  $D = 0$ , or **gametic equilibrium**. Another way to think of the gametic disequilibrium parameter is as a measure of the difference between observed and expected gamete frequencies:  $g_{11} = p_1q_1 + D$ ,  $g_{22} = p_2q_2 + D$ ,  $g_{12} = p_1q_2 - D$ , and  $g_{21} = p_2q_1 - D$  (note that observed and expected gamete frequencies cannot be negative). In this sense,  $D$  measures the deviation of gamete frequencies from what is expected under independent assortment. Since  $D$  can be both positive as well as negative, both coupling and repulsion gametes can be in excess or deficit relative to the expectations of independent assortment.

Gametic disequilibrium can be measured using several estimators, including the squared correlation coefficient ( $\rho^2$ , where  $\rho$  is pronounced "rho"), where  $\rho^2 = D^2/(p_1p_2q_1q_2)$ , which has a range of  $-1$  to  $+1$ . Different estimators of gametic disequilibrium

**Table 2.12** Expected frequencies of gametes for two diallelic loci in a randomly mating population with a recombination rate between the loci of  $r$ . The first eight genotypes have non-recombinant and recombinant gametes that are identical. The last two genotypes produce novel recombinant gametes, requiring inclusion of the recombination rate to predict gamete frequencies. Summing down each column of the table gives the total frequency of each gamete in the next generation.

Genotype	Expected frequency	Frequency of gametes in next generation			
		$A_1B_1$	$A_2B_2$	$A_1B_2$	$A_2B_1$
$A_1B_1/A_1B_1$	$(p_1q_1)^2$	$(p_1q_1)^2$			
$A_2B_2/A_2B_2$	$(p_2q_2)^2$		$(p_2q_2)^2$		
$A_1B_1/A_1B_2$	$2(p_1q_1)(p_1q_2)$	$(p_1q_1)(p_1q_2)$		$(p_1q_1)(p_1q_2)$	
$A_1B_1/A_2B_1$	$2(p_1q_1)(p_2q_1)$	$(p_1q_1)(p_2q_1)$			$(p_1q_1)(p_2q_1)$
$A_2B_2/A_1B_2$	$2(p_2q_2)(p_1q_2)$		$(p_2q_2)(p_1q_2)$	$(p_2q_2)(p_1q_2)$	
$A_2B_2/A_2B_1$	$2(p_2q_2)(p_2q_1)$		$(p_2q_2)(p_2q_1)$		$(p_2q_2)(p_2q_1)$
$A_1B_2/A_1B_2$	$(p_1q_2)^2$			$(p_1q_2)^2$	
$A_2B_1/A_2B_1$	$(p_2q_1)^2$				$(p_2q_1)^2$
$A_2B_2/A_1B_1$	$2(p_2q_2)(p_1q_1)$	$(1-r)(p_2q_2)(p_1q_1)$	$(1-r)(p_2q_2)(p_1q_1)$	$r(p_2q_2)(p_1q_1)$	$r(p_2q_2)(p_1q_1)$
$A_1B_2/A_2B_1$	$2(p_1q_2)(p_2q_1)$	$r(p_1q_2)(p_2q_1)$	$r(p_1q_2)(p_2q_1)$	$(1-r)(p_1q_2)(p_2q_1)$	$(1-r)(p_1q_2)(p_2q_1)$

have slightly different strengths and weaknesses (see Hedrick 1987; Flint-Garcia et al. 2003). The discussion here will focus on the classical estimator  $D$  to develop the conceptual basis of measuring gametic disequilibrium and to understand the genetic processes that cause it.

**Gametic disequilibrium** The non-random association or combination of alleles at multiple loci in a sample of gametes or haplotypes.

**Linkage** Co-inheritance of loci caused by physical location on the same chromosome.

**Recombination fraction** The proportion of “repulsion” or recombinant gametes produced by a double heterozygote genotype each generation.

Now that we have developed an estimator of gametic disequilibrium, it can be used to understand how allelic association at two loci changes over time or its dynamic behavior. If a population starts out with some level of gametic disequilibrium, what happens to  $D$  over time with recombination? Imagine a population with a given level of gametic disequilibrium at the present time ( $D_{t=n}$ ). How much gametic disequilibrium was there a single generation before the present at generation  $n - 1$ ? Recombination will produce  $r$  recombinant gametes each generation so that:

$$D_{tn} = (1 - r)D_{tn-1} \quad (2.25)$$

Since gametic disequilibrium decays by a factor of  $1 - r$  each generation,

$$\begin{aligned} D_{tn} &= (1 - r)D_{tn-1} = (1 - r)^2 D_{tn-2} \\ &= (1 - r)^3 D_{tn-3} \dots \end{aligned} \quad (2.26)$$

We can predict the amount of gametic disequilibrium over time by using the amount of disequilibrium initially present ( $D_{t0}$ ) and multiplying it by  $(1 - r)$  raised to the power of the number of generations that have elapsed:

$$D_{tn} = D_{t0}(1 - r)^n \quad (2.27)$$

Figure 2.19 shows the decay of gametic disequilibrium over time using equation 2.27. Initially there are only coupling gametes in the population and no repulsion gametes, giving a maximum amount of

gametic disequilibrium. As  $r$  increases, the approach to gametic equilibrium ( $D = 0$ ) is more rapid. Equation 2.27 and Fig. 2.19 both assume that there are no other processes acting to counter the mixing effect of recombination. Therefore, the steady-state will always be equal frequencies of all gametes ( $D = 0$ ), with the recombination rate determining how rapidly gametic equilibrium is attained.

### Interact box 2.4 Decay of gametic disequilibrium and a $\chi^2$ test

The program PopGene.S<sup>2</sup> has two simulation modules under the **Gametic Disequilibrium** menu that demonstrate two features of gametic disequilibrium covered so far.

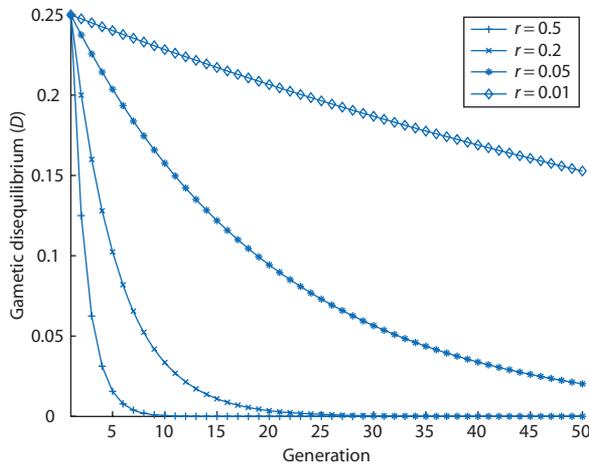
The **Magnitude of  $D$**  simulation can be used to model and graph the decay in  $D$  over time given an initial value of  $D$  and a value for the recombination rate. This module can be used to produce graphs like that shown in Fig. 2.19.

The **Gamete frequencies** simulation can be used to conduct a hypothesis test using a  $\chi^2$  test that the observed level of gametic disequilibrium is significantly different than expected under random segregation. Another way to carry out this test without the aid of PopGene.S<sup>2</sup> is to use

$$\chi^2 = \frac{D^2 N}{p_1 p_2 q_1 q_2} \quad (2.28)$$

where  $N$  is the total sample size of gametes,  $D$  is the gametic disequilibrium parameter, and  $p$  and  $q$  are the allele frequencies at two diallelic loci. The  $\chi^2$  value has 1 df and can be compared with the critical value found in Table 2.5.

One potential drawback of  $D$  in equation 2.24 is that its maximum value depends on the allele frequencies in the population. This can make interpreting an estimate of  $D$  or comparing estimates of  $D$  from different populations problematic. For example, it is possible that two populations have very strong association among alleles within gametes (e.g. no repulsion gametes), but the two populations differ



**Figure 2.19** The decay of gametic disequilibrium ( $D$ ) over time for four recombination rates. Initially, there are only coupling ( $P_{11} = P_{22} = 1/2$ ) and no repulsion gametes ( $P_{12} = P_{21} = 0$ ). Gametic disequilibrium decays as a function of time and the recombination rate ( $D_{t_n} = D_{t_0}(1 - r)^n$ ) assuming a single large population, random mating, and no counteracting genetic processes. If all gametes were initially repulsion, gametic disequilibrium would initially equal  $-0.25$  and decay to 0 in an identical fashion.

in allele frequency so that the maximum value of  $D$  in each population is also different. If all alleles are not at equal frequencies in a population, then the frequencies of the two coupling or the two repulsion gametes are also not equal. When  $D < 0$ ,  $D_{max}$  is the larger of  $-p_1q_1$  or  $-p_2q_2$ , whereas when  $D > 0$ ,  $D_{max}$  is the smaller of  $p_1q_2$  or  $p_2q_1$ .

A way to avoid these problems is to express  $D$  as the percentage of its largest value:

$$D' = D/D_{max} \quad (2.29)$$

This gives a measure of gametic disequilibrium that is normalized by the maximum or minimum value  $D$  can assume given population allele frequencies. Even though a given value of  $D$  may seem small in the absolute, it may be large relative to  $D_{max}$  given the population allele frequencies.

Measures of gametic disequilibrium can be used to test fundamental hypotheses regarding the processes that shape genotype frequencies in natural populations. In epidemiology, pathogens are often considered to reproduce predominantly clonally even if capable of sex and recombination. This assumption has implications for clinical treatment of infections, the emergence of new virulent strains, and vaccine development strategies. Clonal reproduction would accompany high levels of gametic disequilibrium since recombination would not occur, a hypothesis

that can be tested with genetic marker data. The protozoan parasite *Toxoplasma gondii* is one such example. It infects all mammals and birds and causes toxoplasmosis, an illness in humans, and was considered clonal despite sexual reproduction that occurs in cats. To test this hypothesis, Lehmann et al. (2004) sampled *T. gondii* from pigs, chickens, and cats, and then genotyped the protozoa at seven loci. The results revealed normalized gametic disequilibrium ( $|D'|$ ) between pairs of loci that ranged from 0.35 to 0.96. The two loci known to be physically linked showed the highest values of  $D'$  while all others have less gametic disequilibrium. This pattern is inconsistent with clonal reproduction, which would maintain gametic disequilibrium at all loci regardless of the physical distance between loci.

$D$  is frequently called the linkage disequilibrium parameter rather than the gametic disequilibrium parameter. This is a misnomer, since physical linkage only dictates the rate at which allelic combinations approach independent assortment. Recombination, determined by the degree of linkage, only causes a reduction in gametic disequilibrium over time, but it cannot cause an increase in gametic disequilibrium. Processes other than linkage are responsible for the production of deviations from independent assortment of alleles at multiple loci in gametes. Using the term gametic disequilibrium reminds us that the deviation from random association of alleles at two loci is a pattern seen in gametes or haplotypes. Although linkage can certainly contribute to this pattern, so can a number of other population genetic processes. It is even possible that several processes operating simultaneously produce a given pattern of gametic disequilibrium. Processes that maintain or increase gametic disequilibrium include those discussed in the following sections.

### Physical linkage

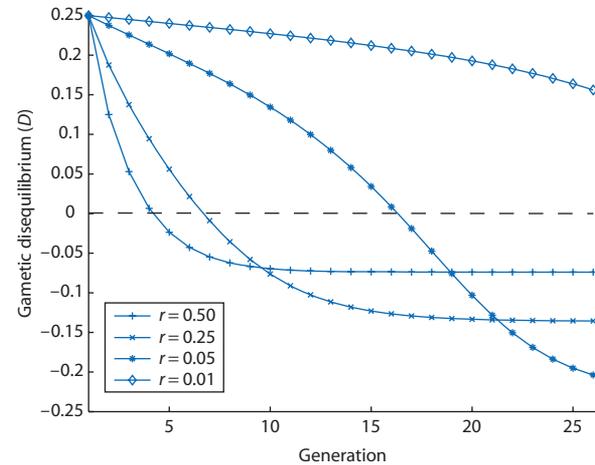
Linkage is the physical association of loci on a chromosome that causes alleles at the loci to be inherited in their original combinations. This association of alleles at loci on the same chromosome is broken down by crossing over and recombination. The probability that a recombination event occurs between two loci is a function of the distance along the chromosome between two loci. Loci that are very far apart (or on separate chromosomes) have recombination rates approaching 50% and are said to be unlinked. Loci located very near each other on the same chromosome might have recombination

rates of 5 or 1% and would be described as tightly linked. Therefore, the degree of physical linkage of loci dictates the recombination rate and thereby the decay of gametic disequilibrium.

Linkage-like effects can be seen in some chromosomes and genomes where gametic disequilibrium is expected to persist over longer time scales due to exceptional inheritance or recombination patterns. Organisms such as birds and mammals have chromosomal sex determination, as with the well-known X and Y sex chromosome system in humans. Loci located on X chromosomes experience recombination normally whereas those on Y chromosomes experience no recombination. This is caused by the Y chromosome lacking a homologous chromosome to pair with at meiosis since YY genotypes do not exist. In addition, we would expect that the rate of decay of gametic disequilibrium for X chromosomes is about half that of autosomes with comparable recombination rates, since X recombination takes place only in females (XX) at meiosis, and not at all in males (XY). Organelle genomes found in mitochondria and chloroplasts are a case where gametic disequilibrium persists indefinitely since these genomes are uniparentally inherited and do not experience observable levels of recombination.

### Natural selection

Natural selection is a process that can continuously counteract the randomizing effects of recombination. Imagine a case where genotypes have different rates of survival or different fitnesses. In such a case natural selection will reduce the frequency of lower



**Figure 2.20** The decay of gametic disequilibrium ( $D$ ) over time when both strong natural selection and recombination are acting. Initially, there are only coupling ( $P_{11} = P_{22} = 1/2$ ) and no repulsion gametes ( $P_{12} = P_{21} = 0$ ). The relative fitness values of the AAbb and aaBB genotypes are 1 while all other genotypes have a fitness of  $1/2$ . Unlike in Fig. 2.19, gametic disequilibrium does not decay to zero over time due to the action of natural selection.

fitness genotypes, which will also reduce the number of gametes these genotypes contribute to forming the next generation. At the same time that natural selection is acting, recombination is also working to randomize the associations of alleles at the two loci. Figure 2.20 shows an example of this type of natural selection acting in concert with recombination to maintain gametic disequilibrium.

The action of natural selection acting on differences in gamete fitness can produce steady states other than  $D = 0$ , expected eventually under even free

## Interact box 2.5

### Gametic disequilibrium under both recombination and natural selection

To simulate the combined action of recombination and natural selection on gametic disequilibrium, try the program **Populus**, which can be obtained by following the link on the text website.

In the Java version of Populus, use the **Natural Selection** menu to select the **Two-Locus Selection** simulation. Set  $p_{AB} = p_{ab} = 0.5$  and  $p_{Ab} = p_{aB} = 0.0$  as a case where there is maximum gametic disequilibrium initially. Use fitness values of  $w_{AaBb} = 1$ , all others = 0.5 and  $w_{AAbb} = w_{aaBB} = 1$ , all others = 0.5 to generate strong natural selection (relative fitness values are explained in Chapter 6). Finally, try recombination values of  $r = 0.5$  and 0.05. Focus your attention on the **D vs. t** plot. What do the two different fitness cases do to levels of gametic disequilibrium and how effective is recombination in opposing or accelerating this effect?

For those who would like to see the details behind the recombination and natural selection model in Populus, a spreadsheet version of this model is available in Microsoft Excel format. The spreadsheet model will allow you to see all the calculations represented by formulas along with a graph of gametic disequilibrium over time.

recombination. In such cases, the population reaches a balance where the action of natural selection to increase  $D$  and the action of recombination to decrease  $D$  cancel each other out. The point where the two processes are exactly equal in magnitude but opposite in their effects is where gametic disequilibrium will be maintained in a population. It is important to recognize that the amount of steady-state gametic disequilibrium depends on which genotypes have high fitness values, so there are also plenty of cases where natural selection and recombination act in concert to accelerate the decay of gametic disequilibrium more rapidly than just recombination alone.

### Mutation

Alleles change from one form to another by the random process of mutation, which can either increase or decrease gametic disequilibrium. First consider the case of mutation producing a novel allele not found previously in the population. Since a new allele is present in the population as only a single copy, it is found only in association with the other alleles on the chromosome strand where it originated. Thus, a novel allele produced by mutation would initially increase gametic disequilibrium. Should the novel allele persist in the population and increase in frequency, then recombination will work to randomize the other alleles found with the novel allele and eventually dissipate the gametic disequilibrium. Mutation can also produce alleles identical to those currently present in a population. In that case, mutation can contribute to randomizing the combinations of alleles

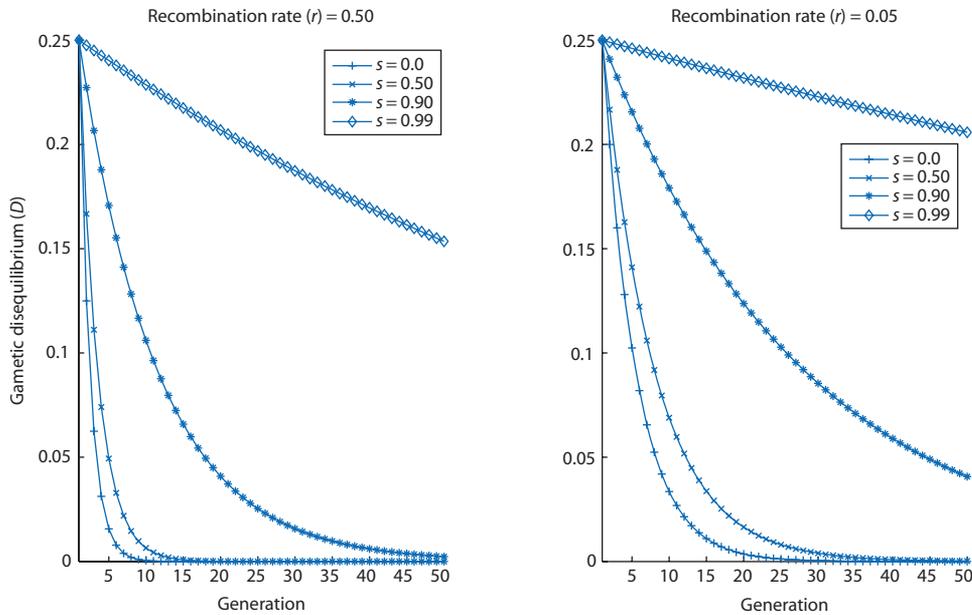
at different loci and thereby decrease levels of gametic disequilibrium. On the other hand, if the population is at gametic equilibrium mutation can create gametic disequilibrium by changing the frequencies of gamete haplotypes. However, it is important to recognize that mutation rates are often very low and the gamete frequency changes caused by mutation are inversely proportional to population size, so that mutation usually makes a modest contribution to overall levels of gametic disequilibrium.

### Mixing of diverged populations

The mixing of two genetically diverged populations, often termed admixture, can produce substantial levels of gametic disequilibrium. This is caused by different allele frequencies in the two source populations that result in different gamete frequencies at gametic equilibrium. Recombination acts to produce independent segregation but it does so only based on the allele frequencies within a group of mating individuals. Table 2.13 gives an example of gametic disequilibrium produced when two populations with diverged allele frequencies are mixed equally to form a third population. In the example, the allele-frequency divergence is large and admixture produces a new population where gametic disequilibrium is 64% of its maximum value. In general, gametic disequilibrium due to the admixture of two diverged populations increases as allele frequencies become more diverged between the source populations, and the initial composition of the mixture population approaches equal proportions of the source populations.

**Table 2.13** Example of the effect of population admixture on gametic disequilibrium. In this case the two populations are each at gametic equilibrium given their respective allele frequencies. When an equal number of gametes from each of these two genetically diverged populations are combined to form a new population, gametic disequilibrium results from the diverged gamete frequencies in the founding populations. The allele frequencies are: population 1  $p_1 = 0.1, p_2 = 0.9, q_1 = 0.1, q_2 = 0.9$ ; population 2  $p_1 = 0.9, p_2 = 0.1, q_1 = 0.9, q_2 = 0.1$ . In population 1 and population 2 gamete frequencies are the product of their respective allele frequencies as expected under independent segregation. In the mixture population, all allele frequencies become the average of the two source populations (0.5) with  $D_{max} = 0.25$ .

Gamete/ $D$	Gamete frequency	Population 1	Population 2	Mixture population
$A_1B_1$	$g_{11}$	0.01	0.81	0.41
$A_2B_2$	$g_{22}$	0.81	0.01	0.41
$A_1B_2$	$g_{12}$	0.09	0.09	0.09
$A_2B_1$	$g_{21}$	0.09	0.09	0.09
$D$		0.0	0.0	0.16
$D'$		0.0	0.0	0.16/0.25 = 0.64



**Figure 2.21**  
The decay of gametic disequilibrium ( $D$ ) over time with random mating ( $s = 0$ ) and three levels of self-fertilization. Initially, there are only coupling ( $P_{11} = P_{22} = 1/2$ ) and no repulsion gametes ( $P_{12} = P_{21} = 0$ ). Self-fertilization slows the decay of linkage disequilibrium appreciably even when there is free recombination ( $r = 0.5$ ).

### Mating system

As covered earlier in this chapter, self-fertilization and mating between relatives increases homozygosity at the expense of heterozygosity. An increase in homozygosity causes a reduction in the effective rate of recombination because crossing over between two homozygous loci does not alter the gamete haplotypes produced by that genotype. The effective recombination fraction under self-fertilization is:

$$r_{\text{effective}} = r \left( 1 - \frac{s}{2-s} \right) \quad (2.30)$$

where  $s$  is the proportion of progeny produced by self-fertilization each generation. This is based on the expected inbreeding coefficient at equilibrium  $F_{eq} = \frac{s}{2-s}$  (Haldane 1924). Figure 2.21 shows the decay in gametic disequilibrium predicted by equation 2.30 for four self-fertilization rates in the cases of free recombination ( $r = 0.5$ ) and tight linkage ( $r = 0.05$ ). Self-fertilization clearly increases the persistence of gametic disequilibrium, with marked effects at high selfing rates. In fact, the predominantly self-fertilizing plant *Arabidopsis thaliana* exhibits gametic disequilibrium over much longer regions of chromosome compared to outcrossing plants and animals (see review by Flint-Garcia et al. 2003).

### Chance

It is possible to observe gametic disequilibrium just by chance in small populations or small samples of gametes. Recombination itself is a random process in terms of where crossing over events occur in the genome. As shown in the Appendix, estimates are more likely to approach their true values as larger samples are taken. This applies to mating patterns and the number of gametes that contribute to surviving progeny in biological populations. If only a few individuals mate (even at random) or only a few gametes found the next generation, then this is a small “sample” of possible gametes that could deviate from independent segregation just by chance. When the chance effects due to population size and recombination are in equilibrium, the effects of population size can be summarized approximately by:

$$\rho^2 = \frac{1}{1 + 4N_e r} \quad (2.31)$$

where  $N_e$  is the genetic effective population size and  $r$  is the recombination fraction per generation (Hill & Robertson 1968; Ohta & Kimura 1969; the basis of this type of equation is derived in Chapter 4). As shown in Fig. 2.22, when the product of  $N_e$  and  $r$  is small, chance sampling contributes to maintaining some gametic disequilibrium since only a

few gametes contribute to the next generation when  $N_e$  is small or only a few recombinant gametes exist when  $r$  is small. The lesson is that  $D$  as we have used it in this section assumes a large population size

(similar to Hardy–Weinberg) so that actual gamete frequencies approach those expected based on allele frequencies, an assumption that is not always met in actual populations.

### Interact box 2.6 Estimating genotypic disequilibrium

In practice, the recombination fraction for two loci can be measured by crossing a double heterozygote with a double homozygote and then counting the recombinant gametes. However, this basic experiment cannot be carried out with species that cannot easily be mated in controlled crosses. An alternative but approximate means to test for gametic equilibrium is to examine the joint frequencies of genotypes at pairs of loci. If there is independent segregation at the two loci then the genotypes observed at one locus should be independent of the genotypes at the other locus. Such contingency table tests are commonly employed to determine whether genotypes at one locus are independent of genotypes at another locus.

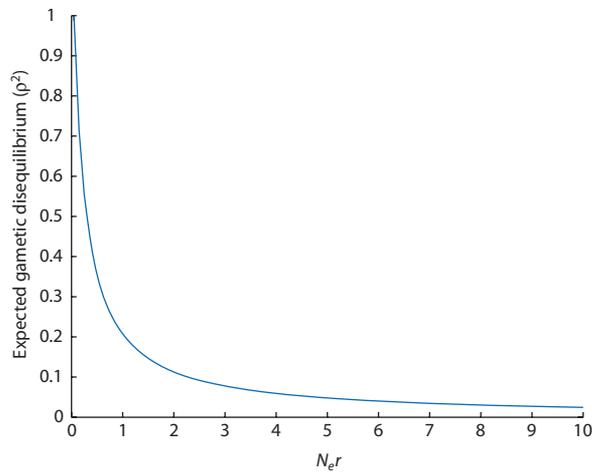
Contingency table tests involve tabulating counts of all genotypes for pairs of loci. In Table 2.14, genotypes observed at two microsatellite loci (AC25-6#10 and AT150-2#4) within a single population (the Choptank river) of the striped bass *Morone saxatilis* are given. The genotypes of 50 individuals are tabulated with alleles at each locus are represented with numbers. For example, there were 15 fish that had a 22 homozygous genotype for locus AC25-6#10 and also had a 44 homozygous genotype for locus AT150-2#4. This joint frequency of homozygous genotypes is unlikely if genotypes at the two loci are independent, in which case the counts should be distributed randomly with respect to genotypes.

In the striped bass case shown here, null alleles (microsatellite alleles that are present in the genome but not reliably amplified by PCR) are probably the cause of fewer than expected heterozygotes that lead to a non-random joint distribution of genotypes (Brown et al. 2005). Thus, the perception of gametic disequilibrium can be due to technical limitations of genotyping techniques in addition to population genetic processes such as physical linkage, self-fertilization, consanguineous mating, and structured populations that cause actual gametic disequilibrium.

**Genepop on the Web** can be used to construct genotype count tables for pairs of loci and carry out statistical tests that compare observed to those expected by chance. Instructions on how to use Genepop and an example of striped bass microsatellite genotype data set in Genepop format are available on the text website along with a link to the Genepop site.

**Table 2.14** Joint counts of genotype frequencies observed at two microsatellite loci in the fish *Morone saxatilis*. Alleles at each locus are indicated by numbers (e.g. 12 is a heterozygote and 22 is a homozygote).

Genotype at locus AT150-2#4	Genotype at locus AC25-6#10					Row totals
	12	22	33	24	44	
22	0	0	1	0	0	1
24	1	4	0	4	1	10
44	2	15	0	0	0	17
25	0	3	0	0	0	3
45	0	8	0	1	0	9
55	1	1	0	0	0	2
26	0	1	0	2	0	3
46	1	3	0	0	0	4
56	0	0	0	1	0	1
Column totals	5	35	1	8	1	50



**Figure 2.22** Expected levels of gametic disequilibrium ( $\rho^2$ ) due to the combination of finite effective population size ( $N_e$ ) and the recombination rate ( $r$ ). Chance sampling will maintain gametic disequilibrium if very few recombinant gametes are produced (small  $r$ ), the population is small so that gamete frequencies fluctuate by chance (small  $N_e$ ), or if both factors in combination cause chance fluctuations in gamete frequency (small  $N_e r$ ).

### Chapter 2 review

- Mendel's experiments with peas lead him to hypothesize particulate inheritance with independent segregation of alleles within loci and independent assortment of multiple loci.
- Expected genotype frequencies predicted by the Hardy–Weinberg equation (for any number of alleles) show that Mendelian inheritance should lead to constant allele frequencies across generations. This prediction has a large set of assumptions about the absence of many population genetic processes. Hardy–Weinberg expected genotype frequencies therefore serve as a null model used as a standard of reference.
- The null model of Hardy–Weinberg expected genotype frequencies can be tested directly or assumed to be approximately true in order to test other hypotheses about Mendelian inheritance.
- The fixation index ( $F$ ) measures departures from Hardy–Weinberg expected genotype frequencies (excess or deficit of heterozygotes) that can be caused by patterns of mating.
- Mating among relatives or consanguineous mating causes changes in genotype frequencies (specifically a decrease in heterozygosity) but no changes in allele frequencies.
- Consanguineous mating can also be viewed as a process that increases the chances that alleles

descended from a common ancestor are found together in a diploid genotype (autozygosity).

- The fixation index, the autozygosity, and the coefficient of inbreeding are all interrelated measures of changes in genotype frequencies with consanguineous mating.
- Consanguineous mating may result in inbreeding depression, which ultimately is caused by overdominance (heterozygote advantage) or dominance (deleterious recessive alleles).
- The gametic disequilibrium parameter ( $D$ ) measures the degree of non-random association of alleles at two loci. Gametic disequilibrium is broken down by recombination.
- A wide variety of population genetic processes – natural selection, chance, admixture of populations, mating system, and mutation – can maintain and increase gametic disequilibrium even between loci without physical linkage to reduce recombination.

### Further reading

For a detailed history of Gregor Mendel's research in the context of early theories of heredity as well as the analysis of Mendel's results by subsequent generations of scientists see:

Orel V. 1996. *Gregor Mendel: the First Geneticist*. Oxford University Press, Oxford.

For recent perspectives on whether or not Gregor Mendel may have fudged his data, see a set of articles published together:

Myers JR. 2004. An alternative possibility for seed coat color determination in Mendel's experiment. *Genetics* 166: 1137.

Novitiski E. 2004. Revision of Fisher's analysis of Mendel's garden pea experiments. *Genetics* 166: 1139–40.

Novitiski E. 2004. On Fisher's criticism of Mendel's results with the garden pea. *Genetics* 166: 1133–6.

To learn more about the population genetics of DNA typing in criminal investigation consult:

Commission on DNA Forensic Science. 1997. *The evaluation of forensic DNA evidence. Proceedings of the National Academy of Sciences USA* 94: 5498–500 (an excerpt from the Executive Summary of the 1996 National Research Council Report).

Gill P. 2002. Role of short tandem repeat DNA in forensic casework in the UK – past, present, and future perspectives. *BioTechniques* 32: 366–85.

National Research Council, Commission on DNA Forensic Science. 1996. *The Evaluation of Forensic DNA Evidence: an Update*. National Academy of Sciences Press, Washington, DC.

To learn more about the Mendelian genetics of the ABO blood group, see the brief history:

Crow JF. 1993. Felix Bernstein and the first human marker locus. *Genetics* 133: 4–7.

Inbreeding depression has a large literature and probing its causes employs a wide array of methods. A good single source from which to learn more is:

Lynch M and Walsh B. 1998. *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Sunderland, MA.

For more detail on ways to estimate gametic disequilibrium, consult:

Gaut BS and Long AD. 2003. The lowdown on linkage disequilibrium. *Plant Cell* 15: 1502–6.

A review of estimators of gametic disequilibrium, the genetic processes that influence its levels, and extensive references to past papers can be found in:

Flint-Garcia SA, Thornsberry JM, and Buckler ES. 2003. Structure of linkage disequilibrium in plants. *Annual Review of Plant Biology* 54: 357–74.

### Problem box answers

#### Problem box 2.1 answer

Using the allele frequencies in Table 2.3 we can calculate the expected genotype frequencies for each locus:

D3S1358:  $2(0.2118)(0.1626) = 0.0689$ ;  
 D21S11:  $2(0.1811)(0.2321) = 0.0841$ ;  
 D18S51:  $(0.0918)^2 = 0.0084$ ;  
 vWA:  $(0.2628)^2 = 0.0691$ ;  
 FGA:  $2(0.1378)(0.0689) = 0.0190$ ;  
 D8S1179:  $2(0.3393)(0.2015) = 0.1367$ ;  
 D5S818:  $2(0.3538)(0.1462) = 0.0992$ ;  
 D13S317:  $2(0.0765)(0.3087) = 0.0472$ ;  
 D7S820:  $2(0.2020)(0.1404) = 0.0567$ .

As is evident from the allele designations, the amelogenin locus resides on the sex chromosomes and can be used to distinguish chromosomal males and females. It is a reasonable approximation to say that half of the population is male and assign a frequency of 0.5 to the amelogenin genotype. The expected frequency of the ten-locus genotype is therefore  $0.0689 \times 0.0841 \times 0.0084 \times 0.0691 \times 0.0190 \times 0.1367 \times 0.0992 \times 0.0472 \times 0.0567 \times 0.5 = 1.160 \times 10^{-12}$ . The odds ratio is one in 862,379,847,814. This 10-locus DNA profile is effectively a unique identifier since the current human population is approximately 6.5 billion and we would expect to observe this exact 10-locus genotype only once in a population 132 times larger than the current human population. In fact, it is likely that this 10-locus genotype has only occurred once in all of the humans who have ever lived.

#### Problem box 2.2 answer

For hypothesis 1 the observed frequency of the bb genotype is given by:

$$f(aa \text{ bb}) + f(A\_ \text{ bb}) = fa^2fb^2 + (1 - fa^2)fb^2$$

Expanding the second term gives:

$$= fa^2fb^2 + fb^2 - fa^2fb^2 = fb^2$$

The other allele frequencies can be obtained by similar steps. This must be true under Mendel's second law if the two loci are truly independent.

For hypothesis 2 the frequency of the B allele can be obtained from the fact that all the allele frequencies must sum to 1:

$$fA + fB + fO = 1$$

Then subtracting  $fB$  from each side gives:

$$fA + fO = 1 - fB$$

Squaring both sides gives:

$$(fA + fO)^2 = (1 - fB)^2$$

The left side of which can be expanded to:

$$(fA + fO)^2 = (fA + fO)(fA + fO) \\ = fO^2 + 2fAfO + fA^2$$

And then:

$$(1 - fB)^2 = fO^2 + 2fAfO + fA^2$$

The last expression on the right is identical to the sum of the first and second expected genotype frequencies for hypothesis 2 in Table 2.3. Expressions for the frequency of the A and O alleles can also be obtained in this fashion.

### Problem box 2.3 answer

The first step is to hypothesize genotypes under the two models of inheritance, as shown in Tables 2.6 and 2.7 for blood groups. Then these genotypes can be used to estimate allele frequencies (symbolized here with a  $P$  to indicate probability). For the hypothesis of two loci with two alleles each:

$$P_{a^2} = \frac{769 + 261}{3816} = 0.270, P_a = 0.52$$

$$P_{b^2} = \frac{728 + 261}{3816} = 0.260, P_b = 0.51$$

$$P_A = 1 - P_a = 0.48, P_B = 1 - P_b = 0.49$$

For the hypothesis of one locus with three alleles:

$$(1 - P_B)^2 = (P_A + P_C)^2 = P_A^2 + 2P_A P_C + P_C^2 \\ = \frac{728 + 261}{3816} = 0.260$$

$$1 - P_B = \sqrt{0.260} = 0.51$$

$$P_B = 0.49$$

$$(1 - P_A)^2 = (P_B + P_C)^2 = P_B^2 + 2P_B P_C + P_C^2 \\ = \frac{769 + 261}{3816} = 0.270$$

$$1 - P_A = \sqrt{0.270} = 0.52$$

$$P_A = 0.48$$

$$P_C = 1 - P_A - P_B = 0.03$$

The expected numbers of each genotype as well as the differences between the observed and expected genotype frequencies are worked out in the tables. For the hypothesis of two loci with two alleles each,  $\chi^2 = 0.266$ , whereas  $\chi^2 = 19,688$  for the hypothesis of one locus with three alleles. Both of these tests have one degree of freedom (4 genotypes – 2 for estimated allele frequencies – 1 for the test), giving a critical value of  $\chi_{0.05,1}^2 = 3.84$  from Table 2.5. The deviations between observed and expected genotype frequencies could easily be due to chance under the hypothesis of two loci with two alleles each. However, the observed genotype frequencies are extremely unlikely under the hypothesis of three alleles at one locus since the deviations between observed and expected genotype frequencies are very large.

Phenotype	Genotype	Observed	Expected number of genotypes	Observed– expected	(Observed–expected) <sup>2</sup> / expected
Hypothesis 1: two loci with two alleles each					
Purple/smooth	A_B_	2058	3816 (1 – 0.52 <sup>2</sup> )(1 – 0.51 <sup>2</sup> ) = 2060.0	2.0	0.002
Purple/wrinkled	A_bb	728	3816 (1 – 0.52 <sup>2</sup> )(0.51) <sup>2</sup> = 724.2	3.8	0.020
Yellow/smooth	aaB_	769	3816 (0.52) <sup>2</sup> (1 – 0.51 <sup>2</sup> ) = 763.5	5.5	0.040
Yellow/wrinkled	aabb	261	3816 (0.52) <sup>2</sup> (0.51) <sup>2</sup> = 268.4	–7.4	0.204
Hypothesis 2: one locus with three alleles					
Purple/smooth	AB	2058	3816 (2(0.48)(0.49)) = 1795	63.0	38.5
Purple/wrinkled	AA, AC	728	3816 ((0.48) <sup>2</sup> + 2(0.48)(0.03)) = 989.1	–261.1	68.9
Yellow/smooth	BB, BC	769	3816 ((0.49) <sup>2</sup> + 2(0.49)(0.03)) = 1028.4	–259.4	63.7
Yellow/wrinkled	CC	261	3816 (0.03) <sup>2</sup> = 3.4	257.6	19,517.0