1.1 Relevance of Population Genetics 4

1.2 Phenotypic Variation in Natural Populations 5

Continuous Variation: The Normal Distribution 5

Mean and Variance 6

The Central Limit Theorem 8

Discrete Mendelian Variation 10

1.3 Multiple-Factor Inheritance 12

1.4 Maintenance of Genetic Variation 14

1.5 Molecular Population Genetics 15

Electrophoresis 16

Allele Frequencies and Genotype Frequencies 18

Polymorphism and Heterozygosity 20

Allozyme Polymorphisms 21

Inferences from Allozyme Polymorphisms 23

1.6 Polymorphisms in DNA Sequences 25

Restriction Enzymes 25

The Polymerase Chain Reaction 28

Single Nucleotide Polymorphisms 32

Synonymous and Nonsynonymous Polymorphisms 32

Segregating Sites and Nucleotide Mismatches 34

1.7 Utility of Genetic Polymorphisms 38

CHAPTER 1

Genetic and Phenotypic Variation

The science of **population genetics** deals with Mendel's laws and other genetic principles as they apply to entire populations of organisms. The organisms may be human beings, animals, plants, or microbes. The populations may be natural, agricultural, or experimental. The environment may be city, farm, field, or forest. The habitat may be soil, water, or air. Because of its wide-ranging purview, population genetics cuts across many fields of modern biology. A working knowledge of population genetics has become essential in genetics, genomics, evolutionary biology, computational biology, systematics, plant breeding, animal breeding, ecology, natural history, forestry, horticulture, conservation, and wildlife management. A basic understanding of population genetics is also useful in medicine, law, biotechnology, molecular biology, cell biology, sociology, and anthropology.

Population genetics also includes the study of the various forces that result in evolutionary changes in species through time. Individual organisms are characterized by their **genotype**, or their genetic constitution, and by their **phenotype**, or the traits that they manifest. There is often a complex relationship between genotype and phenotype, because phenotype may depend on the interactions of different genes as well as effects of the environment. By defining the genetic framework within which evolution takes place, the principles of population genetics are basic to a broad evolutionary perspective on biology. From an experimental point of view, evolution provides a wealth of testable hypotheses for all other branches of biology. Many oddities in biology become comprehensible in the light of evolution: They result from shared ancestry among organisms, and they attest to the unity of life on earth.

1.1 RELEVANCE OF POPULATION GENETICS

Practical applications of population genetics are extensive. Many applications, particularly those relevant to human beings, also have important implications in ethics and social policy. Among the applications of population genetics in medicine, agriculture, conservation, and research are:

- Genetic counseling of parents and other relatives of patients with hereditary diseases.
- Genetic mapping and identification of genes for disease susceptibility in human beings, including breast cancer, colon cancer, diabetes, and schizophrenia.
- Implications of population screening for carriers of disease genes, confidentiality of results, and maintenance of health insurability.
- Statistical interpretation of the significance of matching DNA types found between a suspect and a blood or semen sample from the scene of a crime.
- Design of studies to sample and preserve a record of genetic variation among human populations throughout the world.
- Improvement in the performance of domesticated animals and crop plants.
- Organization of mating programs for the preservation of endangered species in zoos and wildlife refuges.
- Sampling and preservation of germ plasms of potentially beneficial plants and animals that may soon vanish from the wild.
- Interpretation of differences in the nucleotide sequences of genes or amino acid sequences of proteins among members of the same or closely related species.
- Analysis of genes and genomes among diverse species to determine their evolutionary relationships and to test hypotheses about the evolutionary process.

Genetic variation in populations became a subject of scientific inquiry in the late nineteenth century prior even to the rediscovery of the Mendel (1866) paper on heredity. The leading exponent of the study of hereditary differences in human populations was Francis Galton (1822–1911). Galton was a pioneer in the application of statistics to biology. He used statistical methods to study physical traits such as eye color and fingerprint ridges as well as behavioral traits such as temperament and musical ability. Galton was among the first to examine the statistical relations between the distributions of phenotypic traits in successive generations. He is regarded as the founder of biometry, the application of statistics to biological problems. Prior to 1900, Galton's work was carried out with no knowledge of the theory of inheritance proposed by Gregor Mendel (1822–1884).

1.2 PHENOTYPIC VARIATION IN NATURAL POPULATIONS

Galton and Mendel exemplify opposite approaches to the study of inherited traits. Mendel's point of departure in the study of genetics was **discrete variation**, in which phenotypic differences among organisms can be assigned to a small number of clearly distinct classes, such as round versus wrinkled peas. Galton's point of departure was **continuous variation**, in which the phenotypes of organisms are measured on a quantitative scale, like height or weight, and in which the phenotypes grade imperceptibly from one category into the next. As material for the study of phenotypic variation, Galton's choice was good: Most of the differences among normal people that are visible to the unaided eye are differences in continuous traits—height, weight, skin color, hair color, facial features, running speed, shoe size, and so forth. The same is true of phenotypic variation in other organisms. On the other hand, as material for the study of genetic variation, Mendel's choice was good (Hartl and Orel 1992; Orel 1996): The result of segregation is revealed most clearly in pedigrees of discrete, simple Mendelian traits. By **segregation** we mean that the two forms of a gene present in an individual, say *A* and *a*, separate in the formation of reproductive cells, so that each reproductive cell receives exactly one copy of either *A* or *a*.

Continuous Variation: The Normal Distribution

With continuous traits, not only do the phenotypes grade into one another, but the traits also usually present difficulties for genetic analysis. The problems are of two principal types:

- Most continuous traits are influenced by slightly different DNA sequences in two or more genes, hence the segregation of the differences in one gene in pedigrees is obscured by the segregation of differences in other genes that affect the trait.
- Most continuous traits are influenced by environmental factors as well as by genes, and so genetic segregation is obscured by environmental effects.

These problems are not insurmountable in organisms with a sufficiently high density of genetic markers scattered throughout the genome (the complement of chromosomes) because the genetic markers can be tracked in pedigrees along with the continuous trait of interest. Organisms with sufficiently dense genetic maps include human beings, laboratory animals, and many domesticated animals and crop plants.

In Galton's time, however, studies of continuous traits based on linkage of genetic markers along the chromosome were unknown. Why, then, did

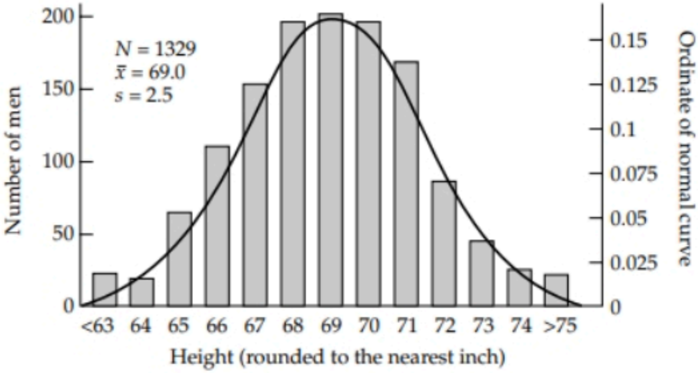


FIGURE 1.1 Distribution of heights among 1329 British men. (Data from Galton 1889.)

Galton focus on continuous traits? The reasons is that they have a sort of regularity—a statistical predictability—of their own. For many continuous traits, when the phenotypes are grouped into suitable intervals and plotted as a bar graph, the distribution of phenotypes conforms closely to the familiar bell-shaped, symmetrical curve known as the **normal distribution**. For example, a bar graph of Galton's data on the heights of 1329 men, rounded to the nearest inch, is plotted in Figure 1.1. The smooth curve is the normal distribution that best fits the data. The equation of the normal curve, more properly called the *normal probability density function*, is:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{1.1}$$

where x ranges from $-\propto$ to $+\propto$, and where $\pi = 3.14159$ and $e = 2.71828$ are constants. The location of the peak of the distribution along the x axis is determined by the parameter μ, which is the **mean**, or average, of the phenotypic values. The degree to which the phenotypes are clustered around the mean is determined by the parameter σ^2, which is the **variance** of the distribution. Mathematically, the variance is the average of the squared difference of each phenotypic value from the mean; that is, it is the average of the values of $(x - \mu)^2$. How μ and σ^2 are estimated from data is considered next.

Mean and Variance

The values of μ and σ^2 are called **parameters**, meaning that they are fixed numerical constants representing some feature or property of a population, in this case the mean and variance, respectively. Although they are constants, their values are unknown, and so they must be estimated from a **sample** chosen to represent the entire population. For the height data, the sample is tabulated in Table 1.1, in which f_i is the number of men whose height is x_i, rounded to the nearest inch. (The fact that the shortest and tallest men are

TABLE 1.1 Heights of 1329 Men

Height interval (i)	Height range (in.)	Nearest inch (x_i)	Number of men (f_i)	$f_i \times x_i$	$f_i \times x_i^2$
1	<63.5	63	23	1,449	91,287
2	63.5–64.5	64	20	1,280	81,920
3	64.5–65.5	65	64	4,160	270,400
4	65.5–66.5	66	110	7,260	479,160
5	66.5–67.5	67	155	10,385	695,795
6	67.5–68.5	68	199	13,532	920,176
7	68.5–69.5	69	203	14,007	966,483
8	69.5–70.5	70	198	13,860	970,200
9	70.5–71.5	71	171	12,141	862,011
10	71.5–72.5	72	88	6,336	456,192
11	72.5–73.5	73	47	3,431	250,463
12	73.5–74.5	74	27	1,998	147,852
13	>74.5	75	24	1,800	135,000
Totals			1,329 (Σf_i)	91,639 ($\Sigma f_i x_i$)	6,326,939 ($\Sigma f_i x_i^2$)

Source: Data from Galton 1889.

grouped in opposite tails of the distribution makes no difference, because these men account for only a small proportion of the total sample.) Also tabulated are the products $f_i \times x_i$ and $f_i \times x_i^2$ as well as their sums.

The mean μ of the distribution is estimated as the mean of the sample, which is conventionally denoted $\bar{x}$ (also sometimes as $\hat{\mu}$):

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} \tag{1.2}$$

In this example, $\bar{x} = 91{,}639/1329 = 68.95$ inches.

Likewise, the variance σ^2 of the distribution is estimated as the variance of the sample, which is conventionally denoted s^2 (also sometimes as $\hat{\sigma}^2$):

$$s^2 = \frac{\sum f_i (x_i - \bar{x})^2}{\sum f_i} = \frac{\sum f_i x_i^2}{\sum f_i} - (\bar{x})^2 \tag{1.3}$$

The expression in the middle follows directly from the definition of the variance: It is the average of the squared deviations from the mean, and for any value of x_i, $(x_i - \bar{x})$ is its deviation from the mean. The expression on the right is identical arithmetically but easier to apply in practice. In the example in Table 1.1, $s^2 = 6{,}326{,}939/1329 - (68.96)^2 = 6.11$. (This value may differ slightly from your own calculation according to the number of significant digits you carried along before rounding off.) If the sample size is small (say, less than 50), then a slightly better estimate of the variance is obtained by multiplying the expression in Equation 1.3 by $n/(n-1)$, where n is the total size of the sample (in this case, 1329).

Closely related to the variance is the **standard deviation** of the distribution, which is the square root of the variance. The standard deviation is a natural quantity to consider in view of the units of measurement. In Table 1.1, for example, each measurement is in inches. The mean is also in inches. However, because the variance is the average of the squared deviations, the variance has the unit of inches-squared, which seems more appropriate for an area than for a height. Taking the square root of the variance restores the correct unit of measure: in this example, inches. The estimate of the standard deviation is conventionally denoted s (also sometimes as $\hat{\sigma}$) and it is calculated as the square root of the quantity in Equation 1.3. In the height example, $s = 2.47$ (which may again differ slightly from your own calculation because of round-off error). The estimate s of the standard deviation is often called the *standard error*.

For a normal distribution, the proportions 68%, 95%, and 99.7% are the proportions of observations expected to fall within 1, 2, or 3 standard errors of the mean, respectively. These emerge directly from Equation 1.1 because the proportion of observations falling with any specified range of x equals the integral of Equation 1.1 across the specified range. For the normal distribution, the integral between the limits $\mu \pm \sigma$ equals 0.6827, that between $\mu \pm 2\sigma$ equals 0.9545, and that between $\mu \pm 3\sigma$ equals 0.9973. In data analysis, $\overline{x}$ and s are used in place of μ and σ. Why do we use two symbols for the mean and two for the standard deviation? Because there is an important difference between $\overline{x}$ and μ and between s and σ. The symbols μ and σ represent the values of the mean and standard deviation in the entire population. The true values of these parameters are unknown and can only be estimated from samples taken from the population. The symbols $\overline{x}$ and s represent the estimates of μ and σ based on a sample, and the different symbols are used to emphasize that the estimates will differ from one sample to the next, and so $\overline{x}$ and s only approximate μ and σ.

Incidentally, the integral of the normal distribution between the limits $\mu \pm 4\sigma$ equals 0.9999; this result says that fewer than one in 10,000 observations falls more than four standard deviations from the mean.

The Central Limit Theorem

Galton was immensely impressed with the observation that many natural phenomena follow the normal distribution. He writes:

> I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the "law of frequency of error" [the normal distribution]. Whenever a large sample of chaotic elements is taken in hand and marshaled in the order of their magnitude, this unexpected and most beautiful form of regularity proves to have been latent all along. The law would have been personified by the Greeks if they had known of it. It reigns with serenity and complete self-effacement amidst the wildest confusion. The larger the mob and the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of unreason.

It is, indeed, remarkable to consider that pure, blind chance is the reason for this "unexpected and most beautiful form of regularity." This principle is useful in practice, too. Modern computers can generate uniformly distributed random numbers in a variety of ways. (In a set of uniformly distributed random numbers, one number is as likely to be sampled as any other.) A common way to generate a single sample from a normal distribution is to generate 12 uniformly distributed random numbers on a computer and simply add them up!

The theoretical basis of the normal distribution is known in probability theory as the **central limit theorem**. Roughly speaking, the central limit theorem states that the sum of a large number of independent random quantities always converges to the normal distribution. For our purposes, "independent" in this context means that information about any one of the observations gives no improvement in the ability to predict any other of the observations. A large number of independent random quantities is apparently what Galton meant by "a large sample of chaotic elements." The central limit theorem explains in part why so many continuously distributed traits conform to the normal distribution. Most continuous traits are *multifactorial*, meaning that they are influenced by "many factors," typically several or many genes acting together with environmental factors. Among human beings, for example, the obvious differences between normal people in hair color, eye color, skin color, stature, weight, and other such traits are not usually traceable to single genes. They result from the combined effects of several or many genes as well as numerous environmental effects acting together as "a large sample of chaotic elements," which often produce, in the aggregate, a normal distribution of phenotypes.

It should be emphasized that the "large number" of random elements specified in the central limit theorem need not be excessive. As an example, Figure 1.2 is a bar graph of 100 observations in which each "observation" consists of the sum of nine consecutive random numbers chosen with equal probability from anywhere in the range (–1, +1). For the sum of nine random numbers in this range, the theoretical mean equals 0 and the theoretical standard deviation equals 1.73; the sample values were $\overline{x} = -0.12$ and $s = 1.70$. Expressed as a difference from the mean in multiples of the standard devia-

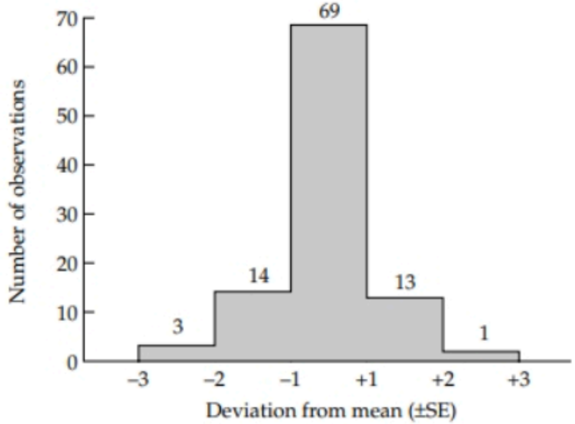


FIGURE 1.2 Distribution of 100 values of the sum of nine random numbers from the interval (–1, +1).

tion, the number of observations in each category is shown at the top of the bar in Figure 1.2. Because the expected numbers are 2.5, 13.5, 68, 13.5, and 2.5, the fit to a normal distribution is obviously very good. In this example, therefore, fewer than 10 "chaotic elements," when added together, yields "this unexpected and most beautiful form of regularity."

PROBLEM 1.1 At an International Health Exhibition in London in 1884, Galton set up an "anthropometric laboratory" that carried out tens of thousands of measurements covering a wide range of human traits. Among the traits was "strength of pull," expressed as the number of pounds that a person could pull with one arm against a resisting force in a sort of arm-wrestling contraption (Galton 1889). The data for 519 males aged 23–26 years fell into the following categories (the number in parentheses is the number of males in each category): 40–50 lbs (10), 50–60 (42), 60–70 (140), 70–80 (168), 80–90 (113), 90–100 (22), 100–110 (24). Using the midpoint of each category as the strength of pull for all males in that category, estimate the mean and standard deviation of strength of pull. Assuming that strength of pull has a normal distribution with parameters equal to these estimates, what is the expected proportion of males whose strength of pull exceeds 112 pounds?

ANSWER The values of x_i are 45, 55, 65, and so forth. Then $\sum f_i = 519$, $\sum f_i x_i = 38{,}675$, and $\sum f_i x_i^2 = 2{,}963{,}375$. Hence, $\overline{x} = 74.5$ lbs, $s^2 = 156.8$ lbs^2, and so $s = 12.5$ lbs. (Answers may differ slightly because of round-off error.) A strength of pull of 112 lbs is three standard errors above the mean; hence a proportion of only (1 – 0.997)/2 = 0.0015 (about one in 667) males is expected to have a phenotype exceeding this value.

Discrete Mendelian Variation

Discrete Mendelian variation (also called simple Mendelian variation) refers to phenotypic differences resulting from segregation of the alleles of a single gene. Environmental effects on the trait are small enough, relative to hereditary differences, that the transmission of alleles determining the trait can be traced through pedigrees. An example of discrete Mendelian variation is the inheritance of red, pink, or white flower color in snapdragon, *Antirrhinum majus* (Figure 1.3). Snapdragons, like human beings, are *diploid* organisms that have two copies of each chromosome, one inherited from each parent. Any gene therefore has a partner on its matching chromosome. Each of the

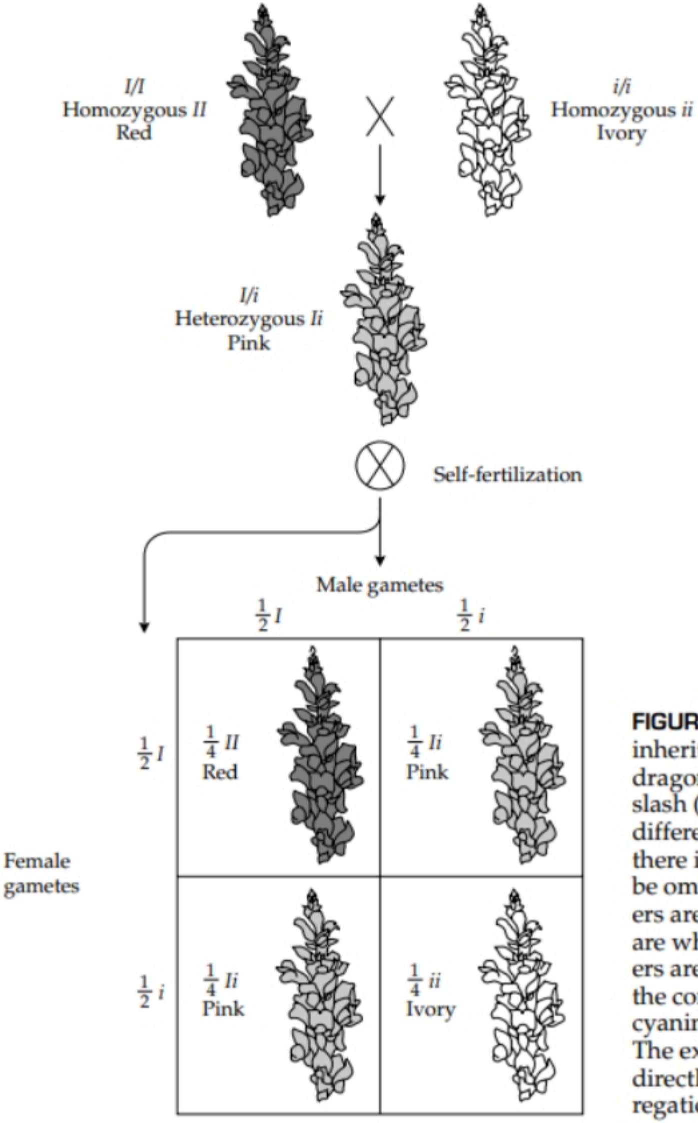


FIGURE 1.3 Simple Mendelian inheritance of flower color in snapdragon (*Antirrhinum majus*). The slash (e.g., *I/I*) separates alleles in different chromosomes, and when there is no ambiguity the slash may be omitted. Homozygygous *II* flowers are red, homozygous *ii* flowers are white, and heterozygous *Ii* flowers are pink. The color results from the concentration of a red anthocyanin pigment in cells of the petals. The example is a classic in showing directly the result of Mendelian segregation in the cross *Ii* × *Ii*.

possible forms of a gene encoded in its DNA sequence is called an **allele** of the gene. When the two alleles in an individual are indistinguishable, the genotype of the individual is said to be **homozygous** (*II* or *ii* in Figure 1.3), and when they are different the individual is said to be **heterozygous** (in this example, *Ii*). This example is exceptionally convenient for genetic studies because of the intermediate phenotype of the heterozygote. The result of segregation of the *I* and *i* alleles is outwardly manifested in the 1 : 2 : 1 ratio of plants with red, pink, or ivory-colored flowers.

Natural populations rarely show discrete complex phenotypes that segregate in simple Mendelian fashion of the sort exemplified by flower color in snapdragons. In human populations, for example, although simple Mendelian inheritance does account for many inherited disorders, each of these disorders is individually quite rare. Examples include cystic fibrosis, phenylketonuria, sickle-cell anemia, and hemophilia.

Because most of the variation in phenotype among normal individuals in natural populations is multifactorial, the pattern of inheritance of these traits shows no clear-cut evidence for Mendelian segregation and nothing resembling any of the simple numerical ratios that Mendel originally discovered in his pea-breeding experiments. The absence of such ratios caused a great controversy in the early 1900s immediately after the rediscovery of Mendel's paper. On the one side were the disciples of Galton, called "biometricians," who dismissed the significance of Mendel's discovery, claiming that Mendel's postulated segregating factors were not only irrelevant for continuous traits but also inadequate to explain the observed correlations in traits between relatives. On the other side were the so-called "Mendelians," who argued that segregation of multiple interacting genes could explain continuous traits just as well as discrete traits. The acrimonious dispute between the biometricians and the Mendelians continued for nearly 20 years. The implications of multifactorial inheritance of discrete traits was the focus of a 1918 paper by the statistician Ronald Aylmer Fisher (1890–1962) entitled "The correlation between relatives on the supposition of Mendelian inheritance." The type of model underlying Fisher's analysis is discussed next.

1.3 MULTIPLE-FACTOR INHERITANCE

Fisher examined a mathematical model of multifactorial inheritance and deduced the expected correlations between relatives. He showed that the kinds of data available for continuous traits were not only compatible with Mendelian inheritance but were also predicted by it.

The spirit of Fisher's model is shown in Figure 1.4, which illustrates the genetic variation expected among the progeny of a cross between genotypes that are heterozygous for each of three unlinked genes. (Genes are said to be *unlinked* when they undergo segregation independently of one another, as if they were in different chromosomes.) The alleles of the genes are represented

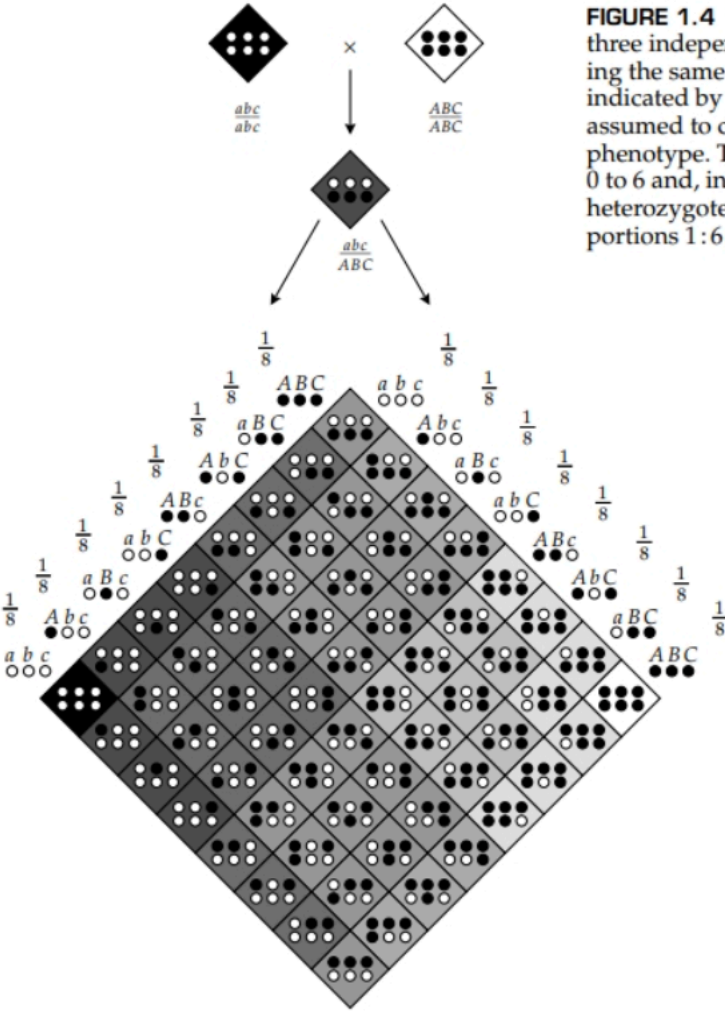


FIGURE 1.4 Result of segregation of three independent pairs of alleles affecting the same trait. Each allele that is indicated by an uppercase letter is assumed to contribute one unit to the phenotype. The phenotypes range from 0 to 6 and, in the cross between triple heterozygotes, are formed in the proportions 1 : 6 : 15 : 20 : 15 : 6 : 1.

A/*a*, *B*/*b*, and *C*/*c*, and the genetic variation resulting from segregation and independent assortment is evident in the various degrees of shading. If we assume a trait in which each uppercase allele adds one unit to the phenotype and in which each lowercase allele is without effect, then the *aa bb cc* genotype has a phenotype of 0 and the *AA BB CC* genotype has a phenotype of 6. Thus there are seven possible phenotypes (0–6) among the progeny. The distribution of phenotypes is shown in the bar graph in Figure 1.5. The smooth curve is the normal distribution approximating the data, which has a mean of 3 and a variance of 1.5. In Figure 1.4, we have assumed that all of the variation in phenotype results from differences in genotype. If there were

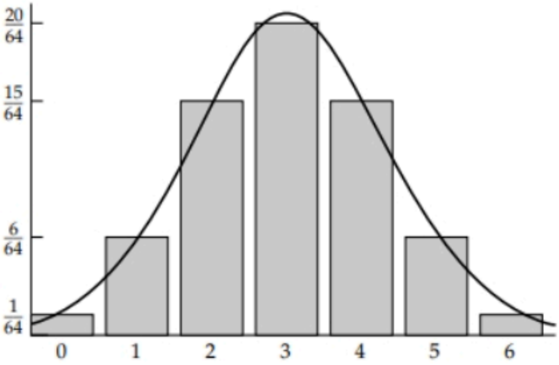


FIGURE 1.5 Distribution of phenotypes from the cross in Figure 1.4 and the approximating normal distribution. The normal curve has mean 3 and variance 1.5.

also random environmental factors affecting the trait, as well as a greater number of genes, then the bars in Figure 1.5 would become less distinct and a normal distribution approximated even better. The result is the central limit theorem at work producing Galton's "supreme law of unreason."

Fisher's model was a good deal more complex than that in Figure 1.4, allowing for differences in the effects of alleles, differences in allele frequency, various types of dominance relations, and the effects of random environmental factors. The work was pathbreaking in demonstrating that continuous variation could be explained by multiple interacting Mendelian factors. Fisher's model was complex for its time and the paper a difficult one. It is not clear even now what practical role Fisher's paper may have played in ending the controversy between the biometricians and the Mendelians. Not many people seem to have read it. (One wag called it a paper you should not read unless you have read it before.) On the other hand, it is the seminal paper that marked the reconciliation of the theories of Galton and Mendel.

1.4 MAINTENANCE OF GENETIC VARIATION

Because Darwin's theory of evolution by means of natural selection requires the presence of genetic variation among individuals, population geneticists have been interested in this issue since the field came into existence in the early 1900s. The main issues were the extent of the differences in genotype from one individual to another, and the processes by which the genetic variation was maintained from one generation to the next. Because the genes underlying multifactorial traits are not revealed by segregation in pedigrees, early population studies were constrained to examine special cases of discrete variation. Classical examples include color or pattern variation within populations of flowers, insects, or snails; variation in blood groups in humans due to differences in carbohydrate antigens present on the surface of red blood cells recognized by protein antibodies of the immune system; and variation in chromosomes in *Drosophila* due to inversions that could be detected by studying the giant chromosomes present in the larval salivary glands. Each of these examples yielded important insights into evolutionary processes, but all were so different that none could be generalized. Each system also held a possible bias due to the effects of differences in genotype on the relative fitness of the organisms.

Limited as they were, the results were variously interpreted to give support to either of two models for the abundance and maintenance of genetic variation. One view, called the *classical hypothesis*, asserted that genetic variation was uncommon and was composed largely of harmful mutant alleles maintained in the population by a balance between recurrent harmful mutations and negative selection. The other model, called the *balance hypothesis*, supposed that genetic variation was abundant and maintained by selection either favoring heterozygous genotypes or genotypes that were rare. In the classical hypothesis, genetic variation was mostly bad; in the balance hypothesis, mostly good. Each side gave some ground to the other, the classical view conceding some cases of balancing selection, and the balance view conceding the existence of harmful mutations. In the meantime, both hypotheses overlooked another important alternative—that much of the genetic variation in natural populations might have little or no significant effect on the organism's fitness, a model that later became known as the **neutral theory**.

1.5 MOLECULAR POPULATION GENETICS

The classical hypothesis and the balance hypothesis sat across the table glowering at each other through most of the 1950s and 1960s. The issues could not be resolved without an unbiased method for studying genetic variation that could be widely applied to a large number of genes in a variety of organisms. This method finally became possible with the direct study of genes and their products using techniques described in this section, but it came at the price of disconnecting genotype from phenotype. Because the mechanisms of transcription, RNA processing, and translation are relatively free of the gene interactions and environmental effects, the correspondence between DNA sequences and alleles is one-to-one: different alleles have different DNA sequences irrespective of whether the alleles affect phenotype. Likewise, alleles that differ in a protein-coding region may result in different amino acid sequences, irrespective of what the protein does in metabolism or how the difference in sequence affects the organism.

The study of molecules is therefore an efficient way to detect simple Mendelian variation—and therein lies a paradox. As evolutionary biologists, population geneticists are interested in observable phenotypes that are likely to be subject to natural selection: morphology, rate of development, mating behavior, age of reproduction, longevity, and so forth (in short, the types of traits that attracted Galton). On the other hand, genetic studies are most readily carried out by detecting differences between molecules resulting from simple Mendelian inheritance. The paradox is that differences in molecules among healthy organisms are not usually related in any obvious way to differences in phenotype. Thus, there is a gap in being unable to specify exactly which types of molecular differences underlie the evolutionary process. The irony of the situation is similar to that described by the physiologist Albert Szent-Gyorgyi:

> My own scientific life was a descent from higher to lower dimensions, led by the desire to understand life. I went from animals to cells, from cells to bacteria, from bacteria to molecules, from molecules to electrons. The story had its irony, for molecules and electrons have no life at all. On my way, life ran out between my fingers.

The gap between genotype and phenotype results from the complex interactions between genes and environment in the determination of physiology, development, and behavior. In evolutionary biology, the complexity is even greater because the key issue is the relative ability of organisms to survive and reproduce in their environments. Nevertheless, the disconnect between differences in molecules and evolutionary adaptations is by no means inevitable, permanent, or insurmountable. It is already clear that the study of the relation between genetic variation and evolutionary adaptation must be high on the agenda of evolutionary biology, and already there are many examples in which the relation is quite well established.

Electrophoresis

New and improved methods for studying macromolecules are continually being created, especially for DNA and proteins. Almost as quickly as they appear, alert population geneticists have applied them to studies of genetic variation in natural populations. Although there are many such experimental procedures that differ in a myriad of details, most of the methods are based on novel combinations of a few simple principles.

One of the most widely applied principles for the study of macromolecules is **electrophoresis,** in which macromolecules in solution move in response to an electric field (Smithies 1954, 1995; Shaw 1965; Lewontin and Hubby 1966). Electrophoresis can be used to separate either protein molecules or nucleic acids. The supporting material that holds the macromolecules is usually some sort of gel, which can be in the form of a horizontal slab, a vertical sandwich between two glass plates, or a cylinder contained within the wall of a glass or plastic tube. Opposite sides of the gel are arranged to make contact with a buffered solution and electrodes. Each sam-

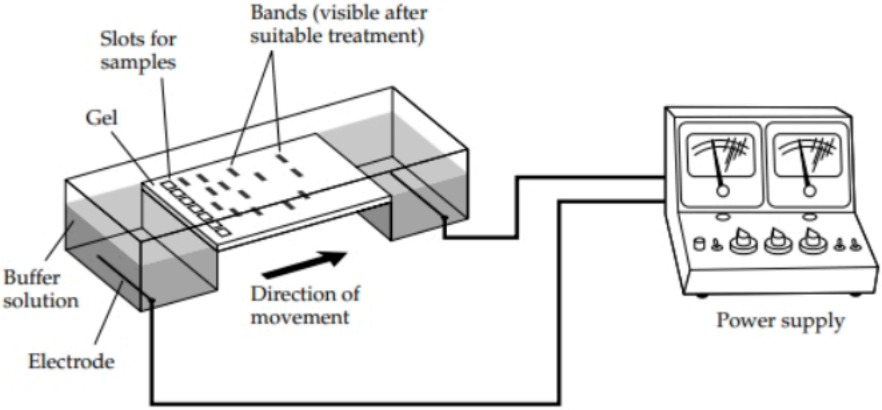


FIGURE 1.6 One type of laboratory apparatus for electrophoresis. The procedure is widely used to separate protein or DNA molecules. In conventional gels, DNA fragments smaller than about 20 kb (1 kb = 1000 nucleotide pairs) migrate approximately in proportion to the logarithm of their molecular weights.

ple of material containing the macromolecules to be separated is placed at or near one end of the slab or tube, and an electric current is applied across the gel for several hours. Molecules in the samples—usually proteins or nucleic acids are of greatest interest—move through the gel in response to the electric field. Molecules of different size and charge move at different rates. Double-stranded DNA molecules move in primarily in relation to their size, whereas protein molecules move primarily in relation to their ionic charge as well as their size. After the electrophoresis is finished, the positions of the molecule or molecules of interest are revealed by any of several procedures.

A typical laboratory setup for protein electrophoresis is diagrammed in Figure 1.6. Protein electrophoresis is used primarily to study enzyme molecules, and the position to which a particular enzyme migrates is revealed by soaking the gel in a solution containing a substrate for the enzyme along with a dye that precipitates where the enzyme-catalyzed reaction takes place. A dark band thus appears in the gel at the position of the enzyme. If the enzyme present in a sample has an amino acid replacement that results in a difference in the overall ionic charge of the molecule, then the enzyme will have a somewhat altered electrophoretic mobility and move at a different rate. The electrophoretic mobility changes because enzymes of the same size and shape move at a rate determined largely by the ratio of the number of positively charged amino acids (primarily lysine, arginine, and histidine) to the number of negatively charged ones (principally aspartic acid and glutam-

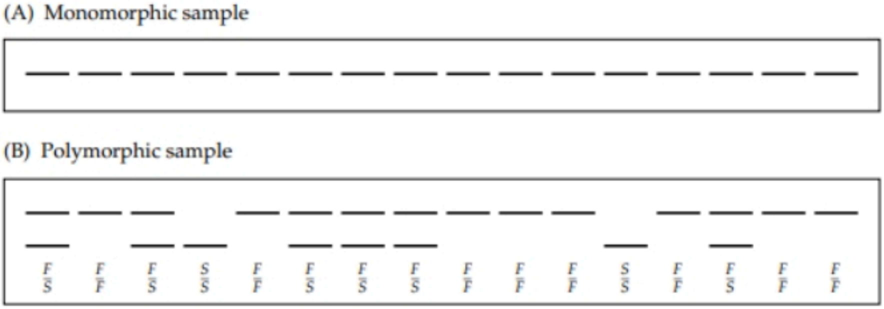


FIGURE 1.7 Monomorphism and polymorphism. (A) Hypothetical gel showing protein monomorphism. All samples have an enzyme with the same electrophoretic mobility. (B) Hypothetical gel showing allozyme polymorphism. Eight samples are homozygous for an allele (*F*) that codes for a rapidly migrating enzyme; two samples are homozygous for a different allele (*S*) that codes for a slowly migrating enzyme; and six samples are heterozygous (*F*/*S*) and therefore exhibit enzyme bands corresponding to both alleles.

ic acid). Electrophoresis can therefore be used to detect a mutation that results in a difference in electrophoretic mobility of the enzyme it encodes.

One possible result of an electrophoresis experiment is shown in the hypothetical gel in Figure 1.7A, in which all samples manifest an enzyme with the same electrophoretic mobility. The result indicates a *monomorphic sample* because there is only one electrophoretic pattern observed. Another kind of result is shown in Figure 1.7B, in which *polymorphism* is observed in the types of electrophoretic patterns. When polymorphic enzyme bands are observed, genetic tests typically indicate that organisms with only a fast-migrating enzyme are homozygous for a *fast* allele (*F*/*F*) and those with only a slow-migrating enzyme are homozygous for a *slow* allele (*S*/*S*). Organisms with both enzyme bands are heterozygous for the alleles (*F*/*S*). Simple Mendelian inheritance of the polymorphism is indicated by, for example, the finding that matings of two heterozygotes produce, on the average, $\frac{1}{4}$ F/F, $\frac{1}{2}$ F/S, and $\frac{1}{4}$ S/S progeny. Two enzyme bands appear in heterozygotes whenever the active enzyme consists of a single polypeptide chain (rather than two or more polypeptide chains aggregated together) because heterozygotes produce a different polypeptide chain from each allele.

Allele Frequencies and Genotype Frequencies

Enzymes that differ in electrophoretic mobility as a result of allelic differences in a single gene are called **allozymes**. Hence, allozyme variation in a population is usually an indication of simple Mendelian genetic variation. As we shall see later in this chapter, allozyme variation is widespread in almost all natural populations studied by electrophoresis, including organisms such as bacteria, plants, *Drosophila*, mice, and human beings.

To convey some sense of how population genetic data are analyzed, consider a population containing an allozyme polymorphism with *F* and *S* alleles at different frequencies. By the **allele frequency** of a specified allele, we mean the proportion of all alleles of the gene that are of the specified type. Suppose we carried out electrophoresis of the enzyme on a sample of 400 members of a population and found 165 F/F, 190 F/S, and 45 S/S. (Here we use the slash to separate the symbol for each allele; if there is no ambiguity, the slash is optional.) In this sample, the observed numbers of *F* and *S* alleles are therefore:

$$F\text{: } 2 \times 165 + 190 = 520$$
$$S\text{: } 190 + 2 \times 45 = 280$$

The factors of 2 are included for the homozygous genotypes because each *FF* genotype contains two *F* alleles and each *SS* genotype contains two *S* alleles. The total number of alleles in the sample equals $2 \times 400 = 800$. Therefore, if we let p represent the frequency of the *F* allele and q represent the frequency of the *S* allele (with $p + q = 1$ because these are the only alleles of the gene in question), then we can estimate p and q from the observations as:

$$\hat{p} = 520/800 = 0.650$$
$$\hat{q} = 280/800 = 0.350$$

For proportions such as these, the estimated standard errors of the estimated allele frequencies are given by $\sqrt{(\hat{p}\,\hat{q}/n)}$ where n is the number of alleles in the sample. In this case the estimated standard error of $\hat{p}$ (and also of $\hat{q}$) equals $\sqrt{(0.650 \times 0.350\ /\ 800)} = 0.0169$

Note that, if the *F* and *S* alleles were combined into genotypes at random (with independence), the expected frequencies of three genotypes can be calculated by multiplication as p^2 *FF*, $2pq$ *FS*, and q^2 *SS*. Therefore, assuming random combination into genotypes, the expected numbers of the three genotypes are:

$$FF\text{: } (0.65)^2 \times 400 = 169$$
$$FS\text{: } 2 \times 0.65 \times 0.35 \times 400 = 182$$
$$SS\text{: } (0.35)^2 \times 400 = 49$$

Hence, the observed numbers in this hypothetical population are very close to those expected with random combinations of alleles. The proportions p^2, $2pq$, and q^2 for the three genotypes when two alleles are combined at random constitute the **Hardy-Weinberg principle**, which is one of the basic principles in population genetics. The Hardy-Weinberg principle is discussed in detail in Chapter 2.

PROBLEM 1.2 Suppose that a random sample of 400 individuals from a different population includes 185 *F/F*, 150 *F/S*, and 65 *S/S* genotypes. Estimate the allele frequency *p* of *F* and *q* of *S*. Assuming random combinations of alleles in the genotypes, what are the expected numbers of the three genotypes? Do the observed data seem to fit the expectations?

ANSWER Among the total of 800 alleles, the observed number of *F* alleles is $2 \times 185 + 150 = 520$ and that of *S* alleles is $150 + 2 \times 65 = 280$. Therefore, $\hat{p} = 520/800 = 0.65$ and $\hat{q} = 280/800 = 0.35$. Note that the estimated allele frequencies are the same as in the earlier example, even though the observed numbers of the genotypes are different. With random combinations of alleles in the genotypes, the expected numbers are again 169 *F/F*, 182 *F/S*, and 49 *S/S*. Compared with the observed numbers, there appear to be too many homozygous genotypes and too few heterozygous genotypes. A statistical method for deciding whether or not the fit is satisfactory is discussed in Chapter 2.

Polymorphism and Heterozygosity

Polymorphism of a gene in a sample is usually of interest only insofar as it indicates polymorphism of the gene in the population as a whole. In a population, a **polymorphic gene** is one for which the most common allele has a frequency of less than 0.95. In practical terms, for a gene with two alleles, this definition implies that a random sample as small as 11 individuals from a population with Hardy-Weinberg genotype frequencies (p^2, $2pq$, and q^2) would be expected to include at least one heterozygous genotype, because $2 \times 0.05 \times 0.95 \times 11 = 1$. Observed frequencies of heterozygous genotypes for genes encoding enzymes vary widely but are typically around 15% in invertebrates and 7% in vertebrates (Figure 1.8). The cutoff at 0.95 is arbitrary, but it serves to focus attention on those genes in which allelic variation is common. In any large population, rare alleles are observed for virtually every gene. An allele is considered a *rare allele* if its frequency is less than 0.005; in human beings, between one and two people per thousand are heterozygous for rare alleles of any gene. Many rare alleles are deleterious and are presumably maintained in the population by recurrent mutation. The definition of polymorphism is an attempt to focus on genes that have alleles with frequencies too high to be explained solely by recurrent mutation to harmful alleles. With the 0.95 definition of polymorphism given above, and if alleles are combined at random into genotypes, then roughly 10% of a population with Hardy-Weinberg frequencies is heterozygous for the most common allele, because $2 \times 0.95 \times 0.05 = 0.095 \approx 10\%$.

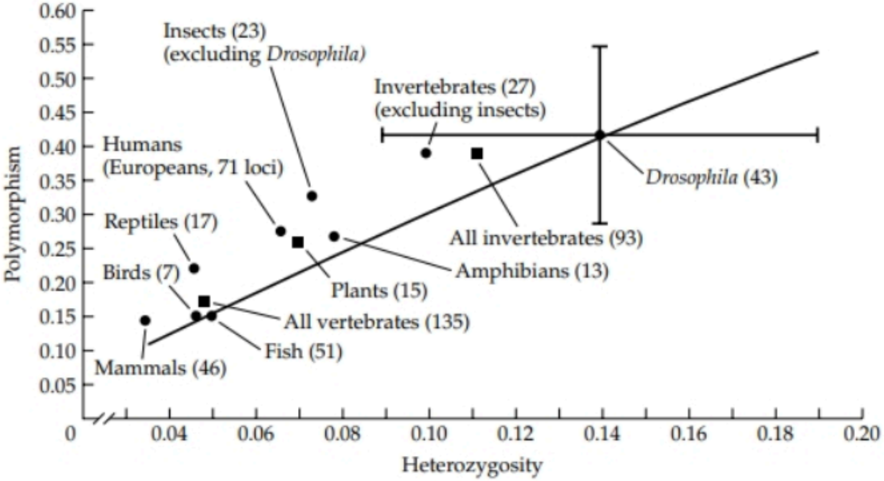


FIGURE 1.8 Estimated levels of heterozygosity and proportion of polymorphic genes derived from allozyme studies of various groups of plants and animals. The curve denotes the theoretically expected relation under the dubious assumption that all allozyme variation is selectively neutral. The number of species studied is shown in parenthesis beside each point. Squares denote averages for plants, invertebrates, and vertebrates. The bars across the *Drosophila* point indicate the standard error within which about 68% of the species are expected to fall. Other groups have similarly large standard errors. (Data from Nevo 1978.)

Allozyme Polymorphisms

Figure 1.8 summarizes the results of electrophoretic surveys of 14 to 71 (mostly around 20) genes in populations of 243 species. Each point in the figure gives the type of organism studied and the number of species examined. The axis labeled Polymorphism refers to the estimated proportion of genes that are polymorphic by the 0.95 criterion. The axis labeled Heterozygosity refers to the average heterozygosity in each group. The average heterozygosity is the estimated proportion of genes expected to be heterozygous in an average organism; it is estimated as the proportion of heterozygous genotypes for each gene averaged over all genes. For example, the data for Europeans include an English population in which 10 enzyme genes were examined (Harris 1966). Of the 10 genes, three were found to be polymorphic, from which the estimated proportion of polymorphic genes in the genome is 3/l0 = 0.30. The observed proportion of heterozygous genotypes for each of the three polymorphic genes was 0.509 (for red-cell acid phosphatase), 0.385 (for phosphoglucomutase), and 0.095 (for adenylate kinase); the average heterozygosity in this sample—taking into account the additional seven genes for which the observed heterozygosity was 0—is therefore $(0.509 + 0.385 + 0.095 + 7 \times 0)/10 = 0.099$. A more extensive electrophoretic survey of 104 genes in a sample including all major human races gave estimates of polymorphism of 0.32 and heterozygosity of 0.06 (Harris et al. 1977).

The vertical and horizontal bars on the point corresponding to *Drosophila* indicate the size of the standard error of the estimate. The bars indicate the limits of polymorphism and heterozygosity within which about 68% of the species are expected to fall. Among *Drosophila* species, approximately 68% have a proportion of polymorphic genes in the range 0.30–0.56 and an average heterozygosity in the range 0.09–0.19. If such bars were attached to each point, their lengths would be comparable to those for *Drosophila*, indicating substantial variability in polymorphism and heterozygosity among species within groups.

Figure 1.8 indicates a positive relationship between amount of polymorphism and degree of heterozygosity. This relationship is as expected because the greater the fraction of polymorphic genes in a population, the more genes that are expected to be heterozygous on the average. Consider an idealized population in which each new mutation encodes a protein whose electrophoretic mobility is distinct from all others present in the population and in which each new mutant allele is selectively neutral (that is, it has a negligible effects on survival and reproduction). Because of recurrent mutation, the alleles in a population gradually turn over in time as some are lost while others become polymorphic. Under these conditions, and restricting our attention to autosomal genes in a diploid species, the expected proportion of polymorphic loci P is given by

$$\ln[1-P] = \theta \ln(0.05) \approx -3\theta \tag{1.4}$$

(Kimura and Ohta 1971) where $\theta = 4N\mu$ is the product of the population size (N) and the mutation rate (μ) per gene per generation. (The symbol ln refers to the logarithm in base e.) The value 0.05 emerges from the definition of polymorphism in which the frequency of most common allele is less than 0.95, because $0.05 = 1 - 0.95$.

Under the same assumptions, the expected magnitude of the heterozygosity (H) can be shown to equal

$$H = \frac{\theta}{1+\theta} \tag{1.5}$$

(Kimura and Crow 1964). Consequently, for genes in an ideal population undergoing successive neutral mutations, the expected relationship between heterozygosity and polymorphisms can be obtained by eliminating θ between Equations 1.4 and 1.5, with the result that

$$\ln[1-P] = \frac{-3H}{1-H} \tag{1.6}$$

This is the relationship shown by the curve in Figure 1.8.

The overall mean polymorphism in Figure 1.8 is 0.26 ± 0.15, and the mean heterozygosity is 0.07 ± 0.05. Vertebrates have the lowest average amount of genetic variation among the groups in Figure 1.8, plants come next, and invertebrates have the highest. *Drosophila* is the most genetically variable group of higher organisms so far studied, and mammals the least variable. Human beings are fairly typical of large mammals. The one obvious conclusion that can be reached from Figure 1.8 is that allozyme polymorphisms are widespread among higher organisms. Genetic variation is even more prevalent among some prokaryotes. For example, natural isolates of the mammalian intestinal bacteria *Escherichia coli* exhibit levels of genetic polymorphism two or three times greater than vertebrates (Selander et al. 1987).

Although genetic polymorphisms are widespread, they are not universal. For example, both major subspecies of the cheetah *Acinonynx jubatus* are virtually monomorphic (O'Brien et al. 1987). A survey of 49 enzymes among 30 animals from the East African subspecies (*A. j. raineyi*) yielded only two polymorphic genes and estimates of polymorphism of 0.04 and heterozygosity of 0.01; among 98 animals from the South African species (*A. j. jubatus*), the estimate of polymorphism was 0.02 and that of heterozygosity 0.0004. Most unusual was the finding of skin-graft acceptance between unrelated cheetahs from the South African subspecies. Graft acceptance means that the cheetah population is monomorphic for the major histocompatibility locus that triggers graft rejection, which is abundantly polymorphic in other mammals. Apparently, the cheetah, which was worldwide in its range at one time but presently numbers less than 20,000 animals, underwent at least two severe constrictions in population number resulting in the loss of most of its genetic variability.

Inferences from Allozyme Polymorphisms

The generality of estimates of polymorphism based on electrophoresis is somewhat uncertain (Lewontin 1974b, 1991). The amount of polymorphism may be underestimated because conventional electrophoresis fails to detect many amino acid replacements. For example, in a study of 14 myoglobin proteins from various species including cetaceans (whales, dolphins and porpoises), no more than eight could be distinguished by conventional electrophoresis; however, 13 could be distinguished by varying the pH value of the electrophoresis buffer (McLellan and Inouye 1986). Some amino acid replacements can be detected because they render the enzyme sensitive to high temperatures; a test for temperature sensitivity increased the number of identified alleles of the gene coding for xanthine dehydrogenase in *Drosophila pseudoobscura* from 6 to 37 and increased the estimate of average heterozygosity from 0.44 to 0.73 (Singh et al. 1976). On the other hand, although more elaborate techniques reveal additional alleles of genes known to be polymorphic, thus increasing estimates of heterozygosity, genes classified as monomorphic by means of routine electrophoresis tend to remain monomorphic, and so estimates of polymorphism remain much the same as before.

Electrophoretic surveys might also overestimate the amount of polymorphism because the enzymes typically surveyed are those found in relatively high concentration in tissues or body fluids ("Group I enzymes") and often lack the high substrate specificity of enzymes implicated in central metabolic processes ("Group II enzymes"). For example, among 10 Group I and 11 Group II enzymes in *Drosophila*, estimates of polymorphism and heterozygosity were 0.70 and 0.24 in the former and 0.27 and 0.04 in the latter (Gillespie and Langley 1974). In summary, protein electrophoresis is a convenient method for detecting polymorphisms, but it is difficult to extrapolate from electrophoretic surveys of enzymes to the entire genome because the enzymes may not be representative.

The high levels of polymorphism observed for allozymes immediately cast the classical hypothesis into doubt. This hypothesis asserted that genetic variation consists largely of highly deleterious alleles maintained by recurrent mutation. The classical hypothesis predicts that allozyme polymorphisms should be rare, whereas Figure 1.8 indicates that they are common. But beyond mere estimates of the magnitude of polymorphism and heterozygosity, other data also throw the alternative balance hypothesis into doubt. The balance hypothesis predicts that genetic variation should be common because it is maintained either by selection favoring heterozygous genotypes or selection favoring rare genotypes. This type of selection predicts very strong harmful effects of inbreeding (mating between close relatives), but the inbreeding effects actually observed are relatively mild.

The seemingly good fit between the data in Figure 1.8 and the theoretical curve for neutrality from Equation 1.6 might be taken as support for the neutral theory, but the data conceal a multitude of complications. Some individual genes show too much heterozygosity for their level of polymorphism, and other genes show too little heterozygosity. More refined analyses of the DNA sequences of alleles of individual genes using statistical methods discussed in Chapters 4 and 7 show that Figure 1.8 presents a picture painted with too broad a brush. Among the many loci represented are some for which most polymorphic alleles appear to be slightly deleterious, others for which the polymorphisms appear to be maintained by some form of selection, and still others showing no marked departure from the patterns expected were the polymorphic alleles selectively neutral or nearly neutral.

1.6 POLYMORPHISMS IN DNA SEQUENCES

Electrophoresis is also one of the mainstays for the study of genetic variation in DNA sequences, because DNA molecules are negatively charged and therefore will move in an electric field. Although standard DNA purification procedures typically shear double-stranded DNA molecules at random positions into fragments of about 50 kb (1 kb = 1000 nucleotide pairs), a number of methods can be used to produce fragments of a specific length.

Restriction Enzymes

DNA fragments of a specific size can be produced by means of any of a class of enzymes called **restriction enzymes**, which cleave double-stranded DNA wherever there is a particular, short nucleotide sequence called the enzyme's *restriction site*. Because the cleavage sites are highly specific, the size of any DNA fragment produced is determined by the distance between adjacent restriction sites. Examples of restriction enzymes and their restriction sites are shown in Figure 1.9, where the cuts are made at the positions of the

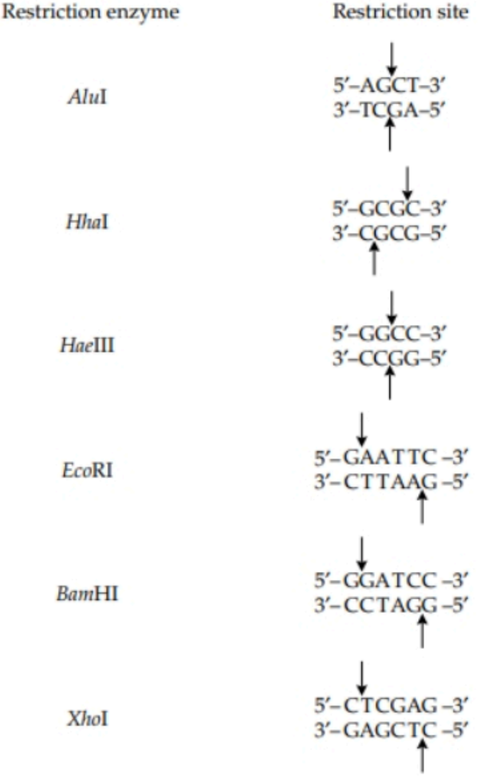


FIGURE 1.9 Restriction enzymes cleave DNA molecules at sites of specific, short nucleotide sequences. More than 500 different restriction enzymes are commercially available. They are essential tools in DNA analysis and gene cloning. The cleavage site in each DNA strand is indicated by the arrow.

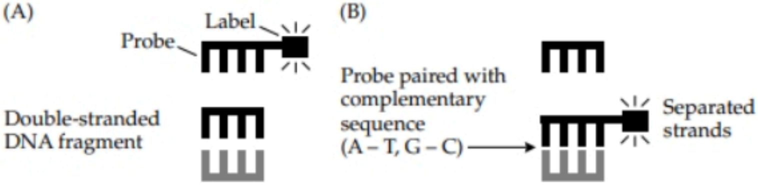

(A)
Label
Probe
Double-stranded DNA fragment
(B)
Probe paired with complementary sequence (A – T, G – C)
Separated strands

FIGURE 1.10 Nucleic acid probes are based on the principle that individual strands with complementary nucleotide sequences of sufficient length can form stable double-stranded molecules. (A) A probe that has exactly the same nucleotide sequence (region in black) as one strand of a double-stranded DNA molecule. (B) If the DNA strands are separated and then allowed to come together again, then in the presence of an excess of probe molecules, the complementary strand will undergo hybridization with the probe rather than with its original partner.

arrows. For example, the enzyme *Alu*I cuts at sites of the four-nucleotide sequence 5′–AGCT–3′, and *Eco*RI cuts at the six-nucleotide sequence 5′–GAATTC–3′. The nucleotide sequence of only one DNA strand need be specified, because in double-stranded DNA, the nucleotide A pairs with T and the nucleotide G pairs with C. The symbols 5′ and 3′ are used to denote the polarity (left-to-right directionality) of the strands. In double-stranded DNA, each strand has a polarity opposite the other, hence the sequence 5′–GAATTC–3′ is paired with the sequence 3′–CTTAAG–5′. As illustrated in Figure 1.9, most restriction enzymes used in population studies have restriction sites consisting of either four nucleotides or six nucleotides.

Because of the specific cleavage sites, digestion of genomic DNA with a restriction enzyme yields a set of fragments of different sizes according to the distances between adjacent restriction sites. These fragments are separated by size by means of electrophoresis, and then any fragment of interest is identified as illustrated in Figure 1.10. Because complementary nucleotide strands can pair with one another, a stretch of single-stranded DNA is able to pair with a complementary region of a strand in a double-stranded molecule, provided that the strands of the double-stranded molecule are first separated either chemically or by heat. The small stretch of single stranded DNA is usually called a **probe**. A probe can range in size from 24 nucleotides to many thousands of nucleotides, and it is usually labeled in some fashion to be able to emit fluorescent or visible light or to undergo radioactive decay. The label can be attached to one end of the molecule as shown in Figure 1.10A, or it can be incorporated into individual nucleotides along the probe. A probe as short as that shown in Figure 1.10 would not work in practice, because pairing of such short regions is easily disrupted by thermal motion. The diagram will nevertheless suffice to make the point that a DNA (or RNA) probe of suffi-

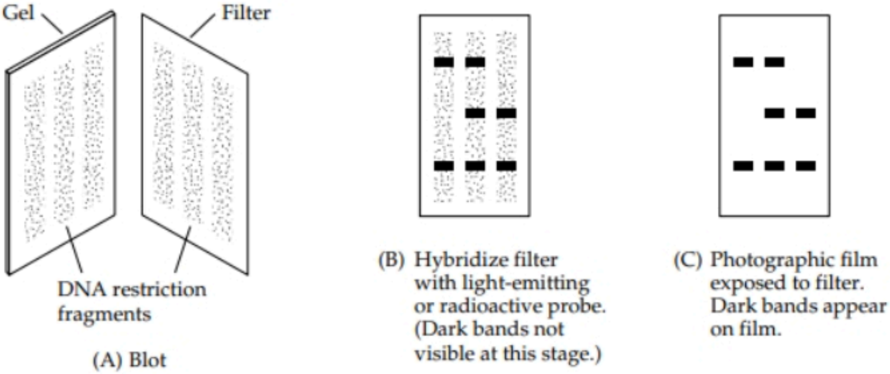

Gel
Filter
DNA restriction fragments
(A) Blot
(B) Hybridize filter with light-emitting or radioactive probe. (Dark bands not visible at this stage.)
(C) Photographic film exposed to filter. Dark bands appear on film.

FIGURE 1.11 Southern blot procedure. (A) DNA fragments separated by electrophoresis are transferred and chemically attached to a filter. (B) The filter is mixed with labeled probe DNA, which hybridizes with and sticks to homologous DNA molecules in the filter. (C) After washing, the filter is exposed to photographic film, which develops dark bands caused by radioactive or light emission from the probe.

cient length will hybridize with a complementary (or nearly complementary) sequence in any strand of DNA (or RNA). As noted by the shading in Figure 1.10B, a probe will generally pair only with one strand along a stretch of double-stranded DNA, because the base sequence in the same region of the other strand is identical and therefore not able to pair with the sequence in the probe.

Hybridization of a restriction fragment with a probe is the principle underlying the **Southern blot** procedure illustrated in Figure 1.11. DNA restriction fragments that have been separated by electrophoresis are rendered single-stranded by soaking in a solution of sodium hydroxide, then blotted onto a nitrocellulose or nylon filter where subsequent chemical treatment attaches them (Figure 1.11A). The filter is then bathed in a solution containing labeled probe DNA (part B). As the solution cools, the probe DNA strands form double-stranded molecules with their complementary counterparts on the filter, and careful washing removes all of the probe DNA that has remained unpaired. The filter is sandwiched with photographic film, where light emission or radioactive disintegrations from the bound probe result in visible bands (part C).

Genetic differences resulting in the presence or absence of restriction sites can be identified because they change the length of characteristic restriction fragments. An example is illustrated in Figure 1.12. The upper part of each panel shows the location of restriction sites in the DNA molecules in a diploid genotype. The *a*-type molecule contains one additional restriction site

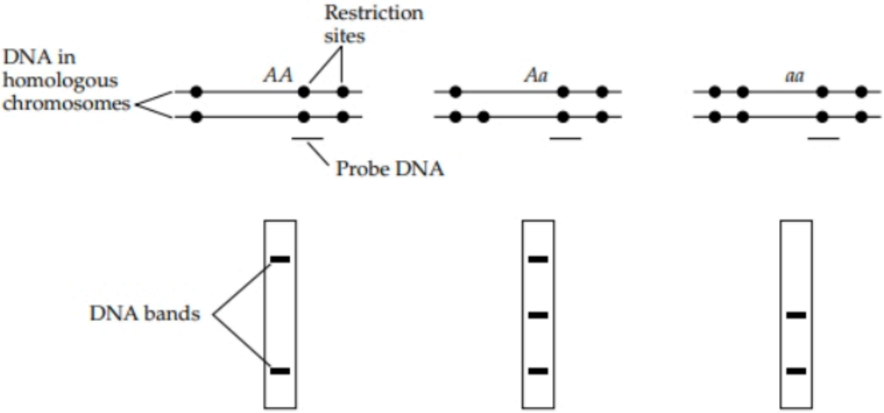


FIGURE 1.12 Restriction fragment length polymorphisms (RFLPs) result from the presence or absence of particular restriction sites in DNA. In this example, the DNA molecule designated A contains three restriction sites, and the one designated a contains four. Genotypes *AA*, *Aa*, and *aa* each yield a different pattern of bands in a Southern blot using the indicated probe DNA.

not present in the *A*-type molecule. The lower part of the figure demonstrates that, with suitable probe DNA, all three genotypes can be distinguished by their pattern of restriction fragments. A difference in the length of a restriction fragment found segregating in natural populations is called a **restriction fragment length polymorphism** or **RFLP**. Because RFLPs are abundant in the genomes of most organisms, they have been widely used to study genetic variation in DNA sequences in natural populations.

The Polymerase Chain Reaction

The **polymerase chain reaction (PCR)** results in the amplification of specific DNA fragments. PCR is of great utility in population genetics, either for the production of probe DNA or for the direct determination of the amount of nucleotide sequence variation present in natural populations. The method is outlined in Figure 1.13. The original DNA sequence to be amplified is shown in black and the newly synthesized DNA strands in gray. The small circles represent short, chemically synthesized sequences of single-stranded DNA (*oligonucleotides*) that are complementary in sequence to the ends of the region to be amplified. The oligonucleotides are called *primer sequences* because they pair with complementary strands at the ends of the sequence to be amplified and are used as primers for chain elongation by DNA polymerase. Primer oligonucleotides are typically 20–30 nucleotides in length. DNA to be used as the template in a PCR reaction is first mixed with both

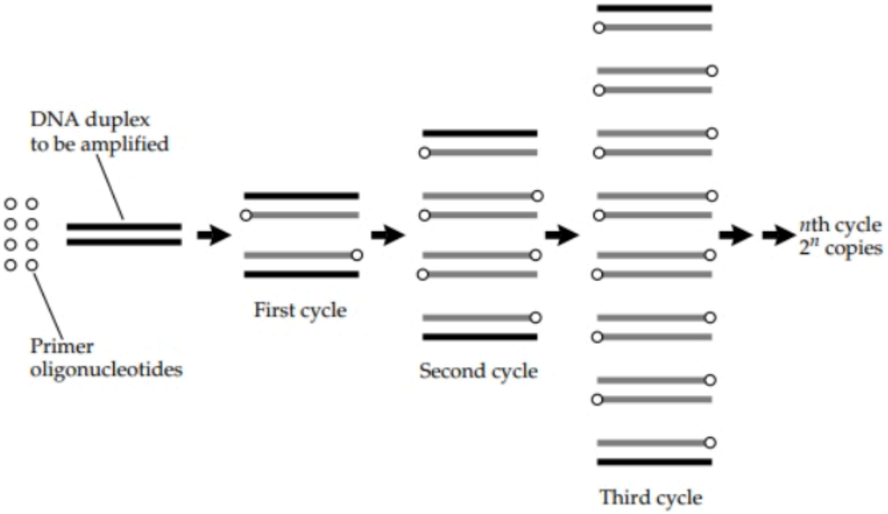


FIGURE 1.13 The polymerase chain reaction (PCR). Short primer oligonucleotides are used as primers to initiate DNA replication from opposite ends of a DNA duplex to be amplified. After each round of replication, the DNA is heated to separate the strands and then cooled to allow new primers to anneal. Repeated rounds of replication result in an exponential increase in the number of target molecules.

primers along with a thermostable DNA polymerase in a buffer solution. The PCR amplification takes place in cycles. In the first cycle, the DNA is heated to separate the strands and then cooled in the presence of a vast excess of the primer oligonucleotides. Then elongation of the primers produces double-stranded molecules. The second cycle of PCR is similar to the first but, after the second cycle, there are four copies of each original molecule. The cycle is repeated from 20 to 30 times, each resulting in a doubling of the number of molecules. The theoretical result of n rounds of amplification is 2^n copies of each template molecule originally present. In practice the reaction does not proceed with perfect efficiency, and the efficiency varies because of the hybridization kinetics of the primers and any tendency for the template to form complex folded structures.

PCR amplification is very useful in generating large quantities of a specific DNA sequence. The main limitation of the technique is that the DNA sequences at the ends of the region to be amplified must be known so that primer oligonucleotides can be synthesized. There are many applications in which this requirement is met. In population genetics, for example, PCR can be used to amplify different alleles present in natural populations.

PROBLEM 1.3 PCR was used to amplify five alleles (designated *a*–*e*) of the gene *Rh3* coding for a light-sensitive protein in the eye of *Drosophila simulans*, a species of fruit fly closely related to *D. melanogaster*. The resulting DNA fragments were sequenced (Ayala et al. 1993). The data show the nucleotide present at each of 16 polymorphic nucleotide sites found in the first 500 nucleotide sites in the amino acid coding region of the gene; the remaining 484 nucleotide sites were monomorphic in this sample. Any nucleotide site that is an exact multiple of three is at the third position of a codon. In this region of the gene:

(*a*) what proportion of polymorphic nucleotide sites are in third positions of codons? What can you infer from this observation?

(*b*) what proportion of nucleotide sites are polymorphic?

(*c*) why is the binomial standard error $\sqrt{(\hat{p}\,\hat{q}/n)}$ not appropriate for the estimate in part (*b*)?

Nucleotide site in gene

Allele	132	142	162	192	198	201	207	240	246	351	354	372	375	405	417	483
a	T	C	T	A	C	C	T	C	C	T	C	G	G	T	T	A
b	T	C	C	T	A	C	C	T	C	C	T	G	G	T	T	T
c	C	T	C	C	C	C	C	T	C	T	T	T	G	C	T	A
d	C	T	C	C	C	C	C	T	T	C	T	G	A	C	T	T
e	C	T	C	C	C	T	C	T	T	T	T	G	G	C	C	A

ANSWER (*a*) Among the 16 polymorphic sites, only site 142 is not an exact multiple of three, hence $\frac{15}{16}$ = 94% of the polymorphic sites are in the third codon position. The inference is that many of the nucleotide polymorphisms are silent (synonymous) in that they do not alter the amino acid sequence of the polypeptide. (In fact, all 16 are silent polymorphisms, including the C → T change in 142, which alters the codon from CUA → UUA, both of which code for leucine.) (*b*) A total of $\frac{16}{500}$ = 3.2% of the nucleotide sites are polymorphic in this region of the gene. (*c*) The binomial standard error is not appropriate in this case because the nucleotides within a gene are not independent samples; they are genetically closely linked.

With PCR, the need to know the sequences at the ends of the fragment to be amplified (in order to be able to synthesize the complementary oligonucleotide primers) may seem like a serious limitation, but even this difficulty can be circumvented in a clever method for studying polymorphism known as **amplified fragment length polymorphism (AFLP)**. The method is outlined in Figure 1.14. The first step (part A) is to digest genomic DNA with a restriction enzyme; this example uses the enzyme *Eco*RI, whose restriction site is 5′–GAATTC–3′, and the stars represent the positions of two adjacent restriction sites. (The size of the DNA fragment between the sites, relative to the size of

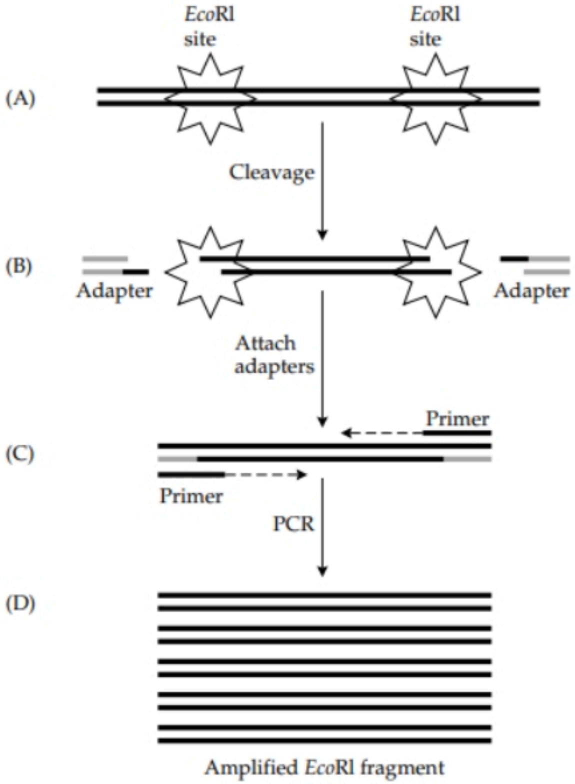


FIGURE 1.14 Amplified fragment length polymorphisms (AFLPs) make use of primer adapters to amplify restriction fragments produced with a particular restriction enzyme, in this case *Eco*RI. (A) Part of the DNA molecule in a chromosome showing the positions of two *Eco*RI restriction sites. (B) After digestion with the restriction enzyme, the fragment is mixed with double-stranded adapters that have single-stranded overhangs complementary to the single-stranded overhangs produced by the restriction enzyme, and then the hybridized adapters are ligated onto the restriction fragment using an enzyme. (C) Primers that are complementary to the adapter sequences are then used to amplify the restriction fragment by means of the polymerase chain reaction. (D) Many copies of the restriction fragment are produced. Normally, the DNA from a single individual produces many different amplified fragments, and any fragment that can be amplified from some individual but not others is an AFLP.

the sites themselves, is not shown to scale.) Digestion yields a large number of restriction fragments flanked by what remains of an *Eco*RI site on each side. *Eco*RI digestion yields a 5′ overhang at each end, with the sequence 5′-AATT (Figure 1.9). These overhangs are long enough that, at low temperature, they can pair with complementary 3′ overhangs of special primer *adapters* (Figure 1.14B), which are attached to the restriction fragments using the enzyme DNA ligase. The resulting fragments (C) are ready for amplification by means of PCR using oligonucleotides that are complementary to the adapters. Note that the same adapter is ligated onto each end, and so a single primer sequence will anneal to both ends and support amplification. There are nevertheless a number of choices concerning the primer sequence. A primer that matches the adapters perfectly will amplify all fragments, but this often results in so many amplified fragments that they are not well separated in the gel. Since a PCR primer must match perfectly at its 3′ end to be elongated, additional nucleotides added to the 3′ end reduce the number of amplified fragments. These primers will amplify only those fragments that, by chance, have a complementary nucleotide immediately adjacent to the *Eco*RI site.

Single Nucleotide Polymorphisms

The ultimate level for the study of genetic polymorphisms is that of the DNA sequence itself, and the smallest unit of polymorphism is the **single nucleotide polymorphism** or **SNP**. A SNP is said to be present at a particular nucleotide site if the DNA molecules in the population frequently differ in the identity of the nucleotide pair that occupies the site. Consider, for example, a single nucleotide site in the protein-coding strand of DNA. Some DNA molecules in a population may have a T (thymidine) nucleotide at this site, whereas other DNA molecules in the same population may have a C (cytosine) nucleotide at the same site. This difference constitutes a SNP. The SNP defines two "alleles" for which there could be three genotypes among individuals in the population, namely, homozygous with T at the corresponding site in both homologous chromosomes, homozygous with C at the corresponding site in both homologous chromosomes, or heterozygous with T in one chromosome and C in the homologous chromosome. The word "allele" is in quotes because any paricular SNP need not be in a coding sequence, or even in a gene. The human genome is thought to contain at least 10 million SNPs, or about one in every 300 base pairs. More than 4 million SNPs have been identified, at which the alternative alleles are both relatively common, and the density among these is about one SNP site every 1000–3000 bp in protein-coding DNA, and about one SNP site every 500–1000 bp in noncoding DNA. How SNPs are used in studies of human population genetics is discussed in Chapter 10.

Synonymous and Nonsynonymous Polymorphisms

One inevitable limitation of protein electrophoresis is that it can detect only those nucleotide polymorphisms in a coding sequence that result in one amino acid being replaced with another in the protein. (In fact, it can detect only the subset of these that alter the protein's charge under the conditions of electrophoresis.) Nucleotide polymorphisms that result in amino acid replacements are known as **nonsynonymous polymorphisms**. There is also a large class of **synonymous polymorphisms** that are present in coding regions but that do not result in an amino acid replacement. These are quite common because the genetic code allows many synonymous nucleotide substitutions. For example, the nucleotide triplet codons TTA, TTG, CTT, CTC, CTA, and CTG all code for the amino acid leucine, and the codons CGU, CGC, CGA, CGG, AGA, and AGG all code for the amino acid arginine. (Only two of the 20 amino acids in the genetic code have no synonymous codons.) Many polymorphisms are also present in noncoding regions of the genome, such as in the region upstream of a coding sequence, in the region downstream of a coding sequence, or in introns that interrupt a coding sequence. Synonymous and noncoding polymorphisms may have subtle effects on the organism, and the polymorphisms may therefore be affected by natural selection; the polymorphic alleles are synonymous or noncoding only in the sense that they all code for the same amino acid sequence.

An example of extensive synonymous polymorphism in *Drosophila* is illustrated in Figure 1.15 for alleles of the gene coding for alcohol dehydrogenase. This gene has an electrophoretic polymorphism that is widespread in natural populations with two predominant alleles, slow (*Adh-S*) and fast (*Adh-F*). The molecular difference is that the codon for amino acid number 193 in *Adh-S* is AAG (lysine) whereas in *Adh-F* it is ACG (threonine). The enzymes differ not only in electrophoretic mobility. The product of the fast

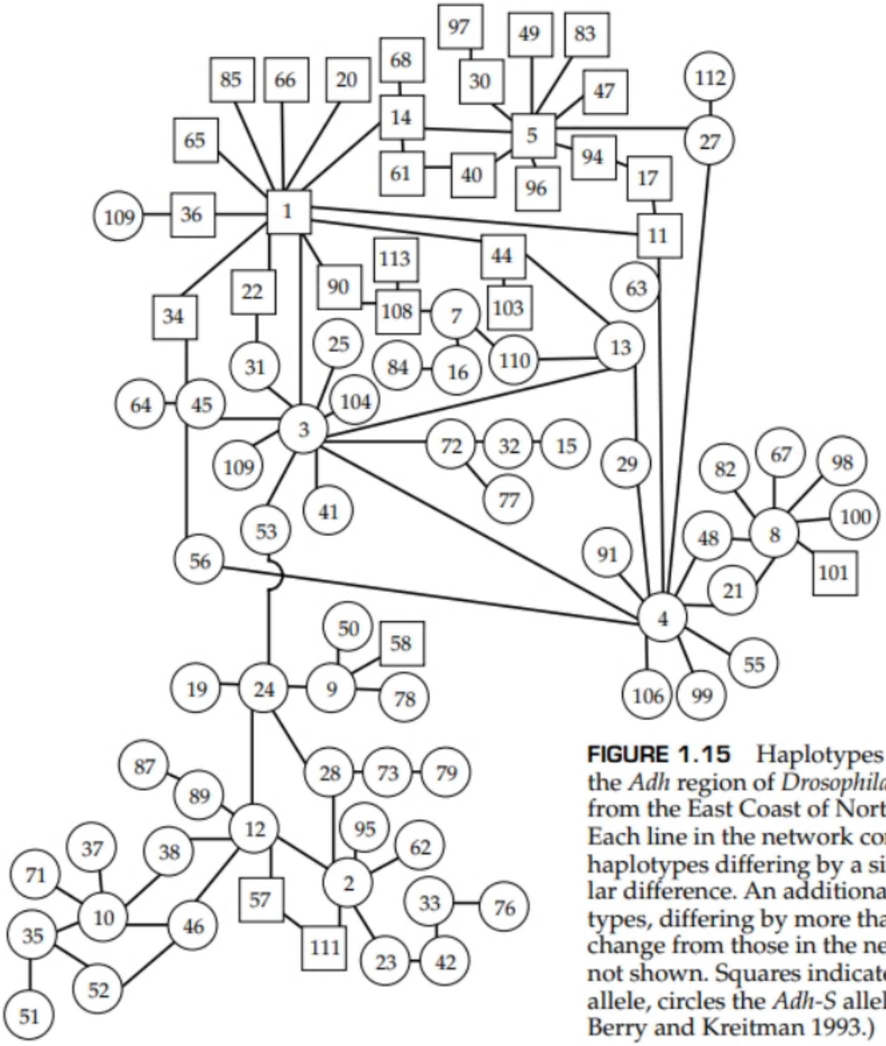


FIGURE 1.15 Haplotypes of alleles in the *Adh* region of *Drosophila melanogaster* from the East Coast of North America. Each line in the network connects two haplotypes differing by a single molecular difference. An additional 20 haplotypes, differing by more than one change from those in the network, are not shown. Squares indicate the *Adh-F* allele, circles the *Adh-S* allele. (From Berry and Kreitman 1993.)

allele has a greater enzymatic activity and is also synthesized in greater amount than that of the slow allele.

The data in Figure 1.15 are derived from studies of the *Adh* region of 1533 flies isolated from 25 populations throughout eastern North America (Berry and Kreitman 1993). A total of 113 haplotypes were identified. A **haplotype** is a unique combination of allelic states of genetic markers present along a single chromosome. The less genetic recombination there is across a region of the genome, or the stronger the selection, the more strongly differentiated the haplotypes are likely to be. The extreme examples are mitochondrial DNA, chloroplast DNA, and the Y chromosome, in which no recombination normally takes place. (See Chapter 4 for a discussion of mitochondrial DNA and chloroplast DNA evolution.)

In Figure 1.15, the haplotypes indicated with squares are *Adh-F* and those with circles are *Adh-S*. The number inside each symbol is the relative abundance of the haplotype (1 being the most frequent, 2 the next most frequent, and so forth). A straight line connecting two haplotypes indicates that they differ by a single nucleotide change. Figure 1.15 includes 93 haplotypes related to at least one other by a singe change; the other 20 haplotypes observed in the study include additional changes. The main point of the *Adh* example is that natural populations contain a great abundance of different types of nucleotide-sequence variation that does not affect amino acid sequence.

Segregating Sites and Nucleotide Mismatches

DNA sequence data contain more information about genetic variation than does protein electrophoresis, because nucleotide polymorphisms are detected even if they are synonymous polymorphisms or are in noncoding DNA, and also because each nucleotide in a set of aligned sequences can be considered individually. (Sequences are said to be *aligned* if corresponding subunits in each sequence, in this case nucleotides, derive from the corresponding subunit in an ancestral sequence.)

By analogy with estimates from protein polymorphism, genetic variation in DNA sequences is conveniently assessed according to the number of nucleotide sites that are polymorphic in a sample of sequences, as well as according to the number of nucleotide sites that are heterozygous. In a set of aligned sequences, the number of *segregating sites* (nucleotide sites that are polymorphic in the sample) is symbolized S and defined as

$$S = \text{Number of nucleotide sites that differ among the aligned sequences} \quad (1.7)$$

The analog of heterozygosity for DNA sequences is the number of nucleotides that differ along any pair of aligned sequences. These are **nucleotide mismatches**, and we will use Π to denote the average number of nucleotide mismatches for every possible pairwise comparison among the aligned sequences. For n sequences, there are $n(n-1)/2$ possible pairwise comparisons. The value of Π for any sample is therefore defined as

$$\Pi = \frac{\text{Total number of nucleotide mismatches}}{\text{Total number of pairwise comparisons}} \tag{1.8}$$

The estimates of both S and Π have a variance due to both random sampling and population history, which are considered in Chapter 4.

To relate S and Π to underlying parameters that affect genetic variation, such as $\theta = 4N\mu$ where again N is the size of an idealized population and μ the mutation rate, one needs a model of how the DNA sequences evolve. One of the simplest models is the *infinite-sites model*, in which the sequence is assumed to consist of a very large (infinite) number of nucleotide sites with no recombination, in which each nucleotide substitution occurs at a different site and is selectively neutral. After a sufficiently long time, the population eventually reaches steady state at which both S and Π are constant, but the individual sequences are slowly turning over owing to recurrent mutation and random loss. At this steady state, it can be shown that the expected values of S (Watterson 1975) and Π (Kimura 1968) for a sample of n sequences are given by

$$E(S) = \theta\left(1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \ldots + \frac{1}{n-1}\right) \tag{1.9}$$

$$E(\Pi) = \theta \tag{1.10}$$

These expressions are derived and discussed later in this book, but we introduce them now to show how they help to tie together some basic measures of DNA sequence variability. The values of the sum of reciprocals in Equation 1.9 are given for small values of n in Table 1.2. For values of n larger than about 20, the sum of the reciprocals equals approximately $0.577 + \ln(n-1)$ (Nei 1987). It is important to emphasize that, in Equations 1.9 and 1.10, while $\theta = 4\mathrm{N}\mu$, the value of μ corresponds to the mutations rate across the entire sequence. In other words, μ equals the average rate of mutation per nucleotide multiplied by the number of nucleotides in each sequence being compared.

At steady state in the infinite-sites model, Π gives one estimate of θ (see Equation 1.10) and S gives another estimate through Equation 1.9 as

$$\theta = S \bigg/ \left(1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \ldots + \frac{1}{n-1}\right) \tag{1.11}$$

Testing for equality between these two estimates is one of many ways to detect departures from steady state in the infinite-sites model (Tajima 1989). For example, certain types of selection (or recent population growth) result in an excess of rare alleles, and so the estimate of θ from Equation 1.10 will be smaller than that from Equation 1.11. Likewise, other types of selection (or a recent decrease in population size) result in too few rare alleles, in which case the estimate of θ from Equation 1.10 will be greater than that from Equation

1.11. This and other tests for departures from selective neutrality are discussed in greater detail in Chapter 4.

TABLE 1.2 Sums of Reciprocals

n	$\Sigma(1/i)^*$	n	$\Sigma(1/i)$
2	1.000	12	3.020
3	1.500	13	3.103
4	1.833	14	3.180
5	2.083	15	3.252
6	2.283	16	3.318
7	2.450	17	3.381
8	2.593	18	3.440
9	2.718	19	3.495
10	2.829	20	3.548
11	2.929	21	3.598

Note: $\Sigma(1/i)$ runs from $i = 1$ to $i = n - 1$.

The data in Problem 1.3 are typical and can be used to exemplify the calculations. There we considered $n = 5$ sequences each of length 500 nucleotides, and found that the observed number of segregating sites was $S = 16$. Hence the estimate of θ from Equation 1.10 is

$$\theta = 16 \Big/ \left(1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4}\right) = 7.68 \qquad (1.12)$$

The average number of nucleotide mismatches Π is the average number of nucleotide differences between all possible pairs of sequences in the sample. For the data in Problem 1.3, the sample of 5 sequences allows $(5 \times 4)/2 = 10$ possible pairwise comparisons. The pairwise comparisons may be considered for each nucleotide in turn. For the polymorphic sites in Problem 1.3, the number of pairwise mismatches totals 6 ($= 2 \times 3$) for six sites (sites 132, 142, 246, 351, 405, and 483), it totals 4 ($= 1 \times 4$) for nine sites (sites 162, 198, 201, 207, 240, 354, 372, 375, and 417), and it totals 7 for one site (site 192). Among the 484 monomorphic nucleotides in Problem 1.3, the number of pairwise mismatches is 0. The average number of pairwise mismatches per pairwise comparison is therefore

$$\Pi = \left(\frac{(6 \times 6) + (4 \times 9) + (7 \times 1) + (0 \times 484)}{10}\right) = 7.90 \qquad (1.13)$$

For the data in Problem 1.3, therefore, the estimate of $\theta = 7.68$ (Equation 1.12, based on S in Equation 1.11) is very close to that of $\theta = 7.90$ (Equation 1.13, based on Π in Equation 1.10). The agreement is very good, but on the other hand the sample size is very small. In this kind of comparison, the values of S and Π depend on the length of the sequences being analyzed, and this length will differ from gene to gene and from one study to the next. How the effects of differing sequence lengths can be removed is examined in Chapter 4.

Estimates of segregating sites and nucleotide mismatches can also be carried out with restriction-site data in the form of restriction fragment length polymorphisms (RFLPs). The simplest way to proceed is to analyze the restriction sites in turn. Each monomorphic restriction site is regarded as identifying six adjacent monomorphic nucleotides (or four monomorphic nucleotides, if the enzyme has a four-base restriction site). Each polymorphic restriction site is regarded as identifying five monomorphic nucleotides and one segregating nucleotide (or three monomorphic sites and one segregating site, if the enzyme has a four-base restriction site). In other words, each restriction site polymorphism is assumed to result from segregation of a single nucleotide difference in the restriction site. Pairwise comparisons to estimate Π are also carried out under this assumption. The reasoning is illustrated in the following problem.

PROBLEM 1.4 Restriction-site variation was studied in the region of the gene for alcohol dehydrogenase (*Adh*) in a population of *D. melanogaster* descended from animals trapped at a Dutch fruit market in Groningen (Cross and Birley 1986). The region contained a total of 23 sites for five restriction enzymes, each having a six-base restriction site. A total of 16 sites were cut in all flies in the sample. The accompanying table documents the presence (+) or absence (−) of each of the seven polymorphic sites in a sample of 10 chromosomes. Assuming that only one nucleotide is altered for each restriction site that is lost, estimate the value of θ based on the number of segregating nucleotide sites S and on the average number of nucleotide mismatches Π. Do these estimates seem to be equal for these data?

*Bam*HI	*Hin*dIII	*Pst*I	*Xho*I	*Pst*I	*Eco*RI	*Eco*RI
+	−	−	+	+	−	−
+	−	−	−	−	+	+
−	−	+	−	−	+	−
−	+	−	+	−	+	+
−	+	−	+	−	+	+
−	+	−	+	−	+	+
−	+	−	+	−	+	+
−	+	−	+	−	+	+
−	+	−	+	−	+	+
−	−	−	+	−	+	−

ANSWER Consider first the segregating sites. The 16 monomorphic restriction sites identify 16 × 6 = 96 monomorphic nucleotide sites, whereas the 7 polymorphic restriction sites identify 7 × 5 = 35 monomorphic nucleotide sites and 7 × 1 polymorphic (segregating) sites, assuming only one nucleotide is altered for each restriction site that is lost and that all DNA molecules that are missing a particular restriction site have the same haplotype of nucleotides at that site. Altogether, there are 138 assayed nucleotide sites of which $S = 7$ are segregating sites. Because $n = 10$, the denominator in Equation 1.11 equals 2.829 (from Table 1.2). The estimate of θ based on S is therefore $\theta = 7/2.829 = 2.47$. For estimating Π, there are $(10 \times 9)/2 = 45$ pairwise comparisons, and a restriction site with i "plus" and $(10 - i)$ "minus" means that the segregating nucleotide site results in $i \times (10 - i)$ pairwise mismatches. Therefore, the total number of mismatches for each of the restriction sites, from left to right, equals 16, 24, 9, 16, 9, 9, and 21, respectively, totaling 104. Therefore, the estimate of $\Pi = 104/45 = 2.31$, which is also the estimate of θ based on Π. The estimates $\theta = 2.47$ and $\Pi = 2.31$ are in very good agreement. However, the sample size is too small to generalize this conclusion.

1.7 UTILITY OF GENETIC POLYMORPHISMS

Whether studied through allozymes or nucleotide sequences, natural genetic variation has many uses. Genetic variation provides a set of built-in markers for the genetic study of organisms in their native habitats, including organisms for which domestication or laboratory rearing is unfeasible or for which conventional genetic manipulation is impossible.

Genetic polymorphisms are useful in investigating the genetic relationships among subpopulations in a species. The principle is that alleles are shared among subpopulations because of migration, and therefore similarity in allele frequencies among subpopulations can be used to estimate the rate of migration (see Chapter 6). Within subpopulations, alleles are shared because of common ancestry. For example, the Ainu people of Northern Japan have numerous Caucasoid-like features, including their facial features, light skin, and hairy bodies, yet their genetic polymorphisms clearly show them to be more closely related to other Mongoloid groups (Watanabe et al. 1975). Among the most informative alleles, the Ainu people possess the *D*(*Chi*) allele of transferrin protein and the Di^a allele of the Diego blood group, both of which are virtually restricted to Mongoloid populations. Conversely, the Ainu people lack several alleles that are polymorphic in Caucasoids.

From a practical point of view, genetic polymorphisms are useful in human populations as genetic markers that may be genetically linked to harmful genes that cause disease (see Chapter 10). In kinships with a family history of the disease, the genetic markers can be used to determine which members of the kindred are likely to be carriers of the harmful gene. The markers can also be used in early diagnosis of persons likely to be affected. RFLPs and other types of DNA polymorphisms that are linked to disease genes have also demonstrated their utility as probes for identifying recombinant DNA clones containing the defective genes. The nearby genetic markers enable the defective gene and its function to be identified, thus serving as a first step in the search for effective treatments.

Particularly useful in population genetics are DNA markers with a large number of alleles of moderate frequency. In most organisms, many regions of the genome have multiple alleles consisting of a short sequence of bases repeated in tandem. Multiple alleles result because the number of copies of the repeated sequence may differ from one chromosome to the next. The genotypes are even more variable because each genotype carries two alleles. One of the practical applications of the use of such polymorphisms is in **DNA fingerprinting**, in which the alleles in the DNA from a suspect are matched with those from a crime-scene sample. The examination of a sufficient number of such highly variable regions provides a basis for distinguishing one person from another because no two people (with the exception of identical twins) have the same genotype. Genetic variability of this sort is used in determining paternity as well as in criminal investigations.

DNA fingerprinting has also been applied to studies of the natural mating systems of plants and animals because, with the large number and high specificity of DNA types, close relatives can be detected in populations. In behavioral studies, DNA typing can determine whether organisms that perform mutually altruistic acts are genetically related. Polymorphisms of other types can also be informative about mating systems. For example, the observed frequencies of genotypes can be used to estimate the amount of self-fertilization in populations of monoecious plants or hermaphroditic animals.

From the standpoint of evolutionary biology, sequences of genes and patterns of polymorphism can be used to make inferences about evolutionary history and about the evolutionary process. There is in fact an international project called the DNA barcoding project that aims to catalog unique DNA sequences that can be used to identify species of virtually any organism by means of an ever-expanding database. The sequences of macromolecules contain within themselves a record of their evolutionary history. Organisms with a shared ancestry usually have similar gene sequences. Conversely, similarity in sequence can be regarded as a measure of shared ancestry. As an index of shared ancestry, sequence similarity provides a means of inferring the ancestral relationships among a group of organisms (*molecular phylogenetics*, discussed in Chapter 7). The rates and patterns of change in sequence within species and between closely related species also contain a record of evolutionary forces at work. Population genetics has evolved from a relatively data-poor field to a relatively data-rich field, and numerous new methods of data analysis and hypothesis testing have been developed.

SUMMARY

1. Galton studied mostly continuous traits, including height and weight, which are measured on a quantitative scale, whereas Mendel studied discrete variation, including round versus wrinkled peas, resulting from segregation of the alleles of a single gene.
2. In natural populations, most continuous traits are multifactorial, which means that they are determined by the combined effects of multiple genetic and environmental factors.
3. Multifactorial traits require special methods to study their genetic basis, whereas simple Mendelian variation is the rule for genes and their products.
4. For a simple Mendelian trait controlled by two alleles in a population undergoing random mating, there is a simple relation expected between the allele frequencies (p, q) and the genotype frequencies (p^2, $2pq$, q^2).
5. Protein polymorphisms are frequent in most natural populations, and many alleles for protein variants are common.

6. Widespread polymorphism casts doubt upon the classical theory of genetic variation, which posited that most genetic variation was due to rare, harmful alleles maintained by recurrent mutation. Other types of data also cast doubt on the balance hypothesis, which held that most polymorphisms were maintained owing to a greater fitness of the heterozygous genotype.
7. Neither the classical nor the balance model considered that many genetic polymorphisms might have virtually no effect on the ability of the organism to survive and reproduce (the neutral theory).
8. The ultimate level of genetic variation consists of differences in the nucleotide sequence of DNA at corresponding positions in the chromosomes of different individuals. The human genome consisting of approximately 3 billion nucleotide pairs is estimated to contain about 10 million single-nucleotide polymorphisms.
9. Many single-nucleotide polymorphisms occur in parts of the genome that do not code for proteins. In protein-coding regions, nonsynonymous polymorphisms change the amino acid sequence, whereas synonymous polymorphisms do not change the amino acid sequence.
10. Nucleotide variation along a set of aligned DNA sequences can be quantified according to the number of segregating nucleotide sites and the average number of mismatches between a random pair of sequences. Comparison of these measures is one of many ways to make inferences about the historical evolutionary forces that have been acting upon a gene.
11. Genetic polymorphisms are important in almost every aspect of modern biology including human genetics, molecular and cellular biology, plant and animal breeding, and wildlife management and conservation. Applications to human populations range from the identification of genetic risk factors for complex diseases, to the implication of guilt in criminal cases through DNA fingerprinting.

PROBLEMS

1. Shown here is a table of a phenotypic values measured on each member of a random sample of 100 individuals. Estimate the mean, variance, and standard deviation in the population from which the sample was drawn.

82	80	106	102	82	94	74	123	93	110
102	112	105	97	125	96	89	105	111	97
106	116	127	140	117	94	130	82	79	80
91	114	81	128	73	130	95	94	98	109
99	96	109	71	90	95	107	92	112	110
87	101	113	117	97	80	139	108	107	103
120	86	90	67	88	87	120	124	112	107

120	101	104	97	72	106	113	88	120	99
106	75	100	82	98	126	103	118	120	104
106	73	88	142	89	96	112	95	99	79

2. Assuming that the sample in Problem 1 was drawn from a normal distribution, what number of individuals in the sample is expected to have a phenotypic value exceeding the mean plus one standard deviation? What number of individuals in the sample is expected to have a phenotypic value smaller than the mean minus one standard deviation? How do these predictions compare with the observations?

3. Assuming that the sample in Problem 1 was drawn from a normal distribution, what number of individuals in the sample is expected to have a phenotypic value exceeding the mean plus two standard deviations? What number of individuals in the sample is expected to have a phenotypic value smaller than the mean minus two standard deviations? How do these predictions compare with the observations?

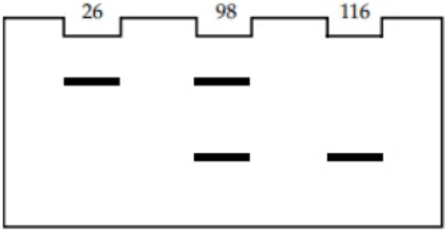
4. A typical means for using a computer to generate normally distributed random numbers is to choose 12 uniform random numbers and add them up. After scaling the sum by a constant that depends on the mean and the variance of the uniform distribution, the result represents a sample from a normal distribution. Why does this approach work?

5. One statement of the central limit theorem is that the sum of independent, identically distributed random variables has a limiting normal distribution. If the random variables that are added are not independent, but instead exhibit a positive correlation in successive draws, how would you expect the sum to deviate from the normal distribution predicted by the central limit theorem?

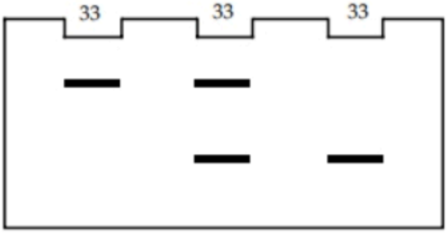
6. Protein electrophoresis is carried out using blood samples from wild mice for a gene with two alleles whose product runs either fast or slow in the gel. Among 100 females, 64 have only a fast band, 4 have only a slow band, and 32 have both a fast and a slow band. Among 100 males, however, 80 have only a fast band and 20 have only a slow band, with no apparent heterozygous genotypes. What mode of inheritance might explain the absence of heterozygous males?

7. Many proteins exist in an active form only as dimers, with the two subunits held together by hydrogen bonds or covalent cysteine bridges. If an enzyme is active only as a dimer, then in a population with two alleles encoding fast and slow migrating forms of the protein, what kind of banding pattern would you expect from tissue from a heterozygous genotype?

8. The diagram shows the result of electrophoresis of tissue samples from 240 Canadian geese, *Branta canadensis*, stained for the enzyme aldehyde oxidase. The samples were placed in the depressions ("wells") shown at the top of the gel, and as electrophoresis took place the proteins migrated in a downward direction. The numbers are the number of individuals in the sample showing each banding pattern. Estimate the allele frequencies of the fast and slow alleles, and calculate the expected genotype frequencies assuming Hardy-Weinberg proportions. Do the observed results seem to fit these expected values?



9. The accompanying gel diagram summarizes the results of electrophoresis of 99 individuals from a population of pointed phlox (*Phlox cuspidata*). Does this population appear to be in Hardy-Weinberg proportions?



10. A gene with two alleles in a population has genotype frequencies P, Q, and R for the genotypes AA, Aa, and aa, respectively, Show that, if the population is in Hardy-Weinberg proportions, then the expected relation between the genotype frequencies is $PR = Q^2/4$.

11. How many copies of a fragment of DNA should be present after 30 rounds of PCR, assuming perfect efficiency?

12. Four sequences of a 1200 bp DNA fragment from a gene gave the following counts of pairwise differences: 4, 7, 5, 3, 6, 5. What is the estimate of the number of pairwise mismatches for this sample?

13. The nucleotide sequences shown here are the complete set of polymorphic nucleotides found in a region of 5 kb in a sample of six chromosomes from corn, *Zea mays*. The polymorphic nucleotides are not adjacent, as shown here, but scattered throughout the 5 kb. What is the number of segregating sites S and the average number of pairwise mismatches Π among these sequences? Estimate θ from both S and Π. Do these estimates seem to be consistent with one another? If this result were found for a larger sample, what might it suggest about the selective forces acting on the sequence polymorphisms?

```
GCCTT  TATGG  CCTGT  ATGAG
ACTAT  TAAGG  CTTGT  TTGAT
ACCAC  TGTCG  CCCGT  ACGCG
GTCAT  TGTGG  TCCTC  TTGAG
GCTTT  TATGA  CCTTT  ATAAG
ACCAT  CATGA  CCTTT  ATAAT
```

14. In forensic applications of genetics, if the DNA fingerprint taken from a crime-scene sample and that of a suspect do not match, the confidence one has in the conclusion is much greater than if the DNA fingerprints do match. What conclusion would be drawn, and why can one have confidence in this conclusion?