Chapter 2
Principles of Inheritance: Mendel's Laws and Genetic Models

It is difficult to overstate the impact of Mendel's research on the history of genetics; indeed, his research in genetics has been credited as one of the great experimental advances in biology (Fisher, 1965). Prior to the publication of his results on experimental hybridization in plants, the concept of inheritance of physical 'units' (later called genes) was accepted, and scientists had reported on many hybridization experiments in both animals and plants. Yet no one had set forth principles of inheritance which could be used as a universal theory to explain how traits in offspring can be predicted from traits in the parents. Mendel provided an explicit rule for how the genotypes of the offspring can be predicted from the genotypes of their parents, and he also established models for how genotypes were related to traits. This is nothing short of astonishing in view of the fact that genes and genotypes were not observed; rather their existence was inferred from the phenotypes that were observed. Needless to say, the underlying biology of cell division and the process of formation of sperm and egg cells was not then known; otherwise the derivation of Mendel's laws would be more straightforward.

Part of Mendel's success was due to his implicit introduction of the concept of a *genetic model*. A genetic model specifies a probability distribution for the trait, conditional on the underlying genotype at the hypothesized disease locus. Mendel's genetic models were very simple forms for dichotomous traits that lead to deterministic outcomes. Genetic models underlie most analyses used in statistical genetics. In order to formalize the process of localizing disease mutations and measuring their effect sizes, we often translate the problem to the framework of statistical hypothesis testing and estimation of parameters in the genetic model.

2.1 Mendel's Experiments

Mendel's work is known largely through a single research paper, 'Experiments in Plant Hybridization' published in 1865. It reported on eight years of experimentation with the garden pea. Mendel made several deliberate choices for his experiments which were crucial in enabling one to infer the laws of inheritance in his series of experiments, essentially examining very simple, now called Mendelian,

N.M. Laird, C. Lange, *The Fundamentals of Modern Statistical Genetics*,
Statistics for Biology and Health, DOI 10.1007/978-1-4419-7338-2_2,

forms of inheritance. In describing Mendel's experiments we use the terms gene and genotype to refer to the genetic locus underlying the traits, although the word gene came into use only after Mendel; following Mendel, we refer to the two alleles of a gene as A and a.

Mendel laid out several principles of good experimentation: using large enough samples of crosses, avoiding unintended cross fertilization, choosing hybrids with no reduction in fertility, etc. Here we focus only on those features of Mendel's experiments bearing on genetics. First is the importance of choosing simple, dichotomous traits for study which are easily recognizable and reproducible. (Mendel studied seven different dichotomous traits.) He called these 'constant differentiating characteristics', meaning that two forms of the trait, e.g., green or yellow pods, could be differentiated in plants, and that the same two forms appeared unchanged in offspring. Mendel excluded traits which produced 'transitional or blended' results in offspring, or quantitative traits generally. Using dichotomous traits enabled him to use simple genetic models to demonstrate laws of inheritance. It took many decades for scientists to develop models which allowed them to apply Mendel's laws to continuous traits.

Second was the use of self-pollinating plants which could also be cross-pollinated; both self-and cross-pollination were used in his experiments. See Fig. 2.1. Cross-pollination was used to form the first generation hybrid plants (called

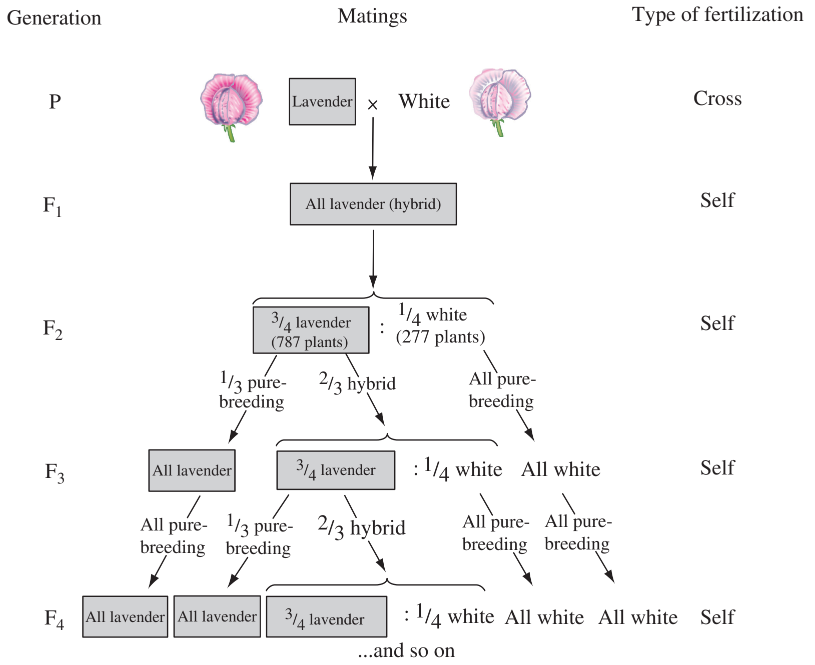

Fig. 2.1 Representation of Mendel's basic experimental design for the law of segregation. *Source*: Mange and Mange (1999)

F_1 in Fig. 2.1); self-pollination was used to develop the parental pure forms (called P in Fig. 2.1), and to infer the genotypes of subsequent crosses. Mendel started the hybridization with the mating of 'pure' forms (inbred forms of plants which always yielded the same form of the phenotype, e.g., plants always having either yellow pods or green pods); underlying the experiments was the implicit assumption that there were two genetic variants, say A and a, one for each of the two forms of each trait. The use of pure parental forms assured that the experiments always started with the mating of two homozygous parents, either AA or aa, so that the first generation crosses between two pure forms (F_1 hybrids) were always heterozygous Aa.

The result of crossing two different plants showed that only one of the two possible phenotypic forms (purple flowering plants in Fig. 2.1) was observed among the F_1 hybrids. This he termed the dominant form, and the form which disappeared among the first generation hybrids was the recessive. Implicitly, Mendel started with the simple genetic model for homozygotes:

$$\begin{aligned} P(\text{recessive form of trait}|aa) &= 1 \\ P(\text{recessive form of trait}|AA) &= 0 \end{aligned}$$

$$\begin{aligned} P(\text{dominant form of trait}|AA) &= 1 \\ P(\text{dominant form of trait}|aa) &= 0. \end{aligned}$$

Today we usually refer to dominant alleles rather than dominant forms of traits, but the general concept is the same. That is, the A allele is dominant because the Aa genotype has the same phenotype as the AA genotype. Note that these models are deterministic; given a genotype, the form of the trait is determined to be either recessive or dominant with probability 1.

It had already been shown by others that the mating of pure forms led to hybrids with only the dominant form of the trait, but Mendel's contribution was to insist on careful self breeding of successive generations in order to deduce their underlying genotype. He found that the offspring of F_1 hybrids, called F_2, had both recessive and dominant trait forms, in the ratio of 1:3, with the recessive form showing no evidence of contamination by the dominant form. The reappearance of the recessive form allowed him to conclude that the gene for the recessive form was present intact in the F_1 generation, although latent. From the results of the F_1 and F_2 generations we can conclude that

$$\begin{aligned} P(\text{dominant form}|Aa) &= 1 \\ P(\text{recessive form}|Aa) &= 0. \end{aligned}$$

Subsequent self fertilization over several generations of F_2 hybrids showed that (1) those plants manifesting the recessive form in the F_2 generation produced only recessive forms among their offspring, and (2) self fertilization of dominant form could be divided into 2 groups: $1/3$ produced only dominant offspring as in pure forms, but $2/3$ again produced both recessive and dominant forms in the same ratio seen in the F_2 generation of 1:3. These phenotypic ratios are idealized in Fig. 2.1. This led Mendel to deduce the following about the genotypes: $1/4$ of the F_2 hybrids were of the parental recessive form (aa), $1/4 = 3/4 \times 1/3$ were of the parental dominant form (AA), and $1/2 = 3/4 \times 2/3$ were the same as the F_1 generation. From this it follows that the genotypes AA, Aa, aa are in the ratio 1:2:1 in the F_2 generation. This allows us to infer Mendel's first law:

Mendel's First Law (Segregation): One allele of each parent is randomly and independently selected, with probability $\frac{1}{2}$, *for transmission to the offspring; the alleles unite randomly to form the offspring's genotype.*

In summary, the phenotypic ratio for Aa × Aa matings is 3:1 (for dominant to recessive forms) and genotypic ratios are 1:2:1. From Mendel's law of segregation, one can extend the results to a crossing of arbitrary genotypes, as is shown in Table 2.1. The law of segregation underlies the concept of *Mendelian transmissions* of alleles from one generation to the next generation; it is a fundamental and universal concept that forms the basis for many genetic analyses discussed in this book.

Mendel's second law concerns independent inheritance of different traits. We will not examine these experiments in great detail; they are fundamentally not different from the first set of experiments, although more complicated because of the large number of possible outcomes that can be observed when many traits are examined. In addition, as we discuss in the last section of this chapter, not all genes are transmitted independently, so that Mendel's second law is not always true. We now know that genes underlying several of his traits are on the same chromosome and they are not inherited independently. However, Mendel's sample sizes were not sufficiently large to pick up modest departures from independence.

To consider two traits, Mendel considered pure strains for each trait, say AABB and aabb, meaning that one parent always had dominant forms in each trait, and the other parent always had recessive forms for both traits. Experimental crossing gave rise to hybrids with Aa and Bb, which showed only dominant forms for both traits. However, the F_2 hybrids raised from F_1 seed showed four phenotypically different

Table 2.1 Distribution of offspring's genotype conditional upon parental genotypes

Father's genotype	Mother's genotype	Offspring's genotype		
		dd	dD	DD
dd	dd	1	0	0
dd	dD	$\frac{1}{2}$	$\frac{1}{2}$	0
dd	DD	0	1	0
dD	dd	$\frac{1}{2}$	$\frac{1}{2}$	0
dD	dD	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
dD	DD	0	$\frac{1}{2}$	$\frac{1}{2}$
DD	dd	0	1	0
DD	dD	0	$\frac{1}{2}$	$\frac{1}{2}$
DD	DD	0	0	1

plants: those with both dominant forms, plants with one dominant and one recessive form (2 kinds) and plants with two recessive forms, in the approximate ratio of 9:3:3:1 (see exercise 2 of Section 2.4). Subsequent self-pollination of the F_2 generation allowed him to deduce 9 genetic forms among the F_2 hybrids: AABB, AABb, AAbb, AaBB, AaBb, Aabb aaBB, aaBb and aabb in the ratio 1:2:1:2:4:2:1:2:1. These ratios exactly coincide with what one would expect if inheritance of the two traits is independent, for then, with F_2 hybrids,

$$\begin{aligned} P(AA \text{ and } BB) &= P(AA)P(BB) \\ &= (1/4)^2 = 1/16 = 1/(1+2+1+2+4+2+1+2+1), \\ P(AA \text{ and } Bb) &= P(AA)P(Bb) = (1/4)(1/2) = 1/8 = 2/16 \text{ etc.,} \end{aligned}$$

when describing the result of a double heterozygote mating.

Mendel's Second Law (Independent Assortment): *The alleles underlying two or more different traits are transmitted to offspring independently of each other; the transmission of each trait separately follows the first law of segregation.*

Fisher (1936) noted that many of Mendel's statistics were generally too close to their expectations, thus χ^2 statistics comparing observed numbers offspring with a given phenotype to those expected assuming his laws of segregation were true, were often too small, suggesting some data manipulation. This, and the lack of generality of his law of independent assortment (see exercise 3 of Section 2.4), has not diminished the value of his contributions. The lack of independent transmission of different genes is, in fact, fortuitous, as it provides the basis for mapping disease genes by linkage analysis, as will be described in Section 2.3, and in Chapter 11.

2.2 A Framework for Genetic Models

A *genetic model* describes the relationship, usually probabilistic, between an individual's genotype and their phenotype or trait. In Genetic Epidemiology, phenotypes will typically be affection status and we distinguish only between *affected* and *unaffected* subjects in the statistical analysis. Such binary traits can be coded by Y, where $Y = 1$ denotes affected and $Y = 0$ denotes unaffected. For other dichotomous traits such as those that Mendel used, this labeling is arbitrary. For complex diseases, e.g., Asthma, Chronic obstructive pulmonary disease (COPD), Obesity, etc., affection status is often defined by a set of *intermediate phenotypes* or *endophenotypes* which are quantitative measurements that can be more reproducible assessments of the disease features. They can also provide additional insight into the nature and severity of the disease. Standard intermediate phenotypes are body mass index (BMI) as an assessment of obesity, forced expiratory volume in one second (FEV1) for asthma, etc. In some cases, e.g., Alzheimer's disease, the phenotype affection status can be refined by selecting age-of-onset as the target phenotype in the statistical analysis. In general, the selection of the target phenotype is a key question in the planning of the study and the statistical analysis. The phenotype choice will depend on the disease, the possible study designs, statistical power considerations and the necessary adjustments for confounding factors. We will use the variable Y as the variable that describes the phenotype or trait of interest, whether dichotomous or measured.

An individual's genotype at a marker is given by the combination of their two alleles at that locus; we use the notation G to denote an individual's genotype. In the majority of scenarios that we will consider, the marker locus will have only two distinct alleles, e.g., alleles 'A' and 'a'. In the literature such genetic loci are called di-allelic or bi-allelic. Typically, the "small-letter" allele 'a' is assumed to be the more frequent allele of the two and is referred to as the wild type or normal allele. The less frequent allele is labeled with the capital-letter 'A' and referred to as the minor allele. This differs from Mendel's designation of the capital allele as representing the allele associated with the dominant form, because most of the genetic loci we study do not have any known associated dominant or recessive phenotypes, hence today the capital letter usually refers to the less common allele. Under the assumption that the genetic locus is bi-allelic, each of the two chromosomes has to carry either an 'a' or 'A' allele, and, consequently, only three different genotypes are possible: the two homozygous genotypes, AA and aa, and the heterozygous genotype Aa. Order does not matter, so Aa is the same as aA. Thus G can take on only three values in a di-allelic system. With three alleles, there are 6 possible genotypes, etc. Genotypes are inherently categorical but can always be recoded in the form of numerical or indicator variables, as we will discuss at the end of this section.

If the genetic locus is a Disease Susceptibility Locus (DSL), it is conventional to use the D/d designation, as opposed to A/a or B/b; the D-allele is then sometimes referred to as the *Disease Variant* or *Disease Susceptibility Allele*. In formulating genetic models for disease outcomes, we assume the DSL has a direct effect on the phenotype through some biological mechanism. Genetic models can either be deterministic, i.e., the genotype determines the phenotype exactly (*Mendelian Disease*, or, in most cases, probabilistic, i.e., the genotype influences the probability of disease. Conditional upon the individual's genotype G, the probabilistic effect of the locus on the phenotype Y is described by the *penetrance function* which is a set of conditional probabilities, or density functions for continuous phenotypes, which model the distribution of the phenotype/trait, i.e., $P(Y|G)$. If the genetic locus under consideration has no effect on the phenotype of interest, the penetrance probabilities for all three genotypes will be equal regardless of the individual's genotype, i.e., $P(Y|G = dd) = P(Y|G = dD) = P(Y|G = DD)$.

The specification of penetrance probabilities will depend on the type of the disease phenotype. If the phenotype of interest is dichotomous, the penetrance function specifies simple probabilities between zero and one for each genotype, with $P(Y = 1|G) + P(Y = 0|G) = 1$, for each G. When Y denotes disease status, the penetrance probability for $Y = 1$ defines the probability of disease conditional on the genotype of the individual. Mendel considered only two simple genetic models for dichotomous traits: recessive and dominant. The dominant model is

$$P(Y = 1|DD) = P(Y = 1|Dd) = 1 \text{ and } P(Y = 1|dd) = 0, \tag{2.1}$$

and the recessive is

$$P(Y = 1|DD) = 1 \text{ and } P(Y = 1|Dd) = P(Y = 1|dd) = 0. \quad (2.2)$$

Note that here D is the disease allele (the variant), and $Y = 1$ refers to disease, so that the two models are different. If disease is recessive, it requires two variants, but a dominant disease requires only one. However, if the dominant model holds for the disease outcome, then the recessive model holds for the non-disease outcome, $Y = 0$. This is why Mendel used the terms dominant and recessive to describe possible trait outcomes.

Apart from rare genetic disorders, deterministic models are not very reasonable. Variations of these basic models are constructed by considering stochastic versions which lead to *reduced penetrance* and *phenocopies*. A model is said to be of reduced penetrance if the probability of disease, $P(Y = 1|G)$, is less than 1 for values of G where it is one in the Mendelian models. That is, for the recessive model, $P(Y = 1|DD) = a$ for some $0 < a < 1$, and similarly for the dominant model. The Mendelian models are called *fully penetrant* in contrast to reduced penetrance models, because the probability of disease is either zero or one. The idea behind phenocopies is that the disease could also be caused by another genetic locus, or possibly some non-genetic variable, so that $P(Y = 1|G)$ is positive for those values of G where it is zero in 2.1–2.2. For the dominant mode, for example, $P(Y|dd) = b$ for some $0 < b < 1$. In other cases, the heterozygotes might be intermediate in disease risk between the two homozygotes, suggesting that the number of mutations influences disease risk. Figure 2.2 shows a possible choice for such a penetrance function which allows for both phenocopies and reduced penetrance. Probands with the genotype dd have a 10% chance of being affected. For probands with the genotype DD, the probability of being affected is 7 times higher.

One of the earliest non-Mendelian genes found was APOE for AD. Here there are two mutations giving rise to 3 major alleles (other alleles in the gene are very rare): $E2$, $E3$ and $E4$. The risk of late onset AD increases with an increasing number of $E4$ alleles, but having an $E2$ allele appears protective. Generally $P(Y = 1|G)$ is

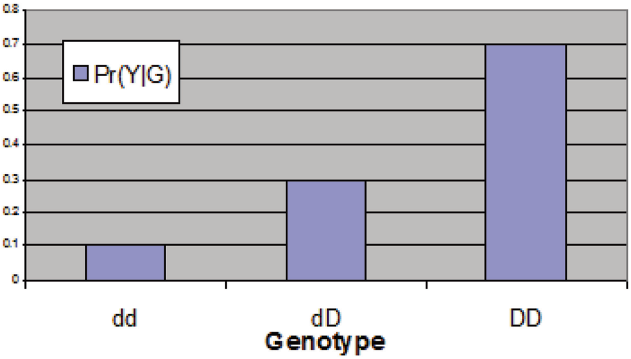

Fig. 2.2 Penetrance function for a dichotomous trait

a complex function of G, but never reaches 1 or 0 for any genotype at the APOE locus.

One publication from the popular press (Pamela McDonald, *The APOE Gene Diet: A Breakthrough in Changing, Cholesterol, Weight, Heart and Alzheimer's Using the Body's Own Genes*) lists the risk for AD as a function of selected APOE genotypes: 20% for 33, 50% for 24, 60% for 34 and 92% for 44. In reality, penetrance functions for AD as a function of APOE genotype are difficult to quantify because they also depend on sex and age. With six possible genotypes, large prospective samples will be required to quantify risk as a function of age and sex with much precision.

For quantitative traits, a natural choice for the penetrance function is a normal density, with a mean that depends upon the genotype while the variance does not. Thus we assume the density function of Y is given by $f(y|\mu_G, \sigma^2)$, where $f(y|\mu_G, \sigma^2)$ denotes the normal density with mean μ_G and variance σ^2; μ_G indicates that the mean depends on the genotype G. For other types of traits, e.g., age-of-onset, the penetrance probability can be selected to be trait-type specific density functions as are used in standard statistical models to describe the relationship between traits and a covariate. Figures 2.3 and 2.4 show examples of penetrance functions for a quantitative trait and for age-at-onset. Again, the notion that the D-allele is the risk allele is echoed in both figures, where we assume larger values of the quantitative trait are deleterious. In Fig. 2.3, the number of D-alleles is correlated with an increased likelihood for larger phenotypic values of Y. Figure 2.4 shows empirical survival curves for AD as a function of APOE genotype, estimated from a large study of individuals free of AD at age 60. Even with this large study, genotype groups have been combined because of sparse numbers at older ages and the low number of subjects with the 4/4 genotype.

Apart from recessive and dominant models for dichotomous traits, thus far we have specified only general probability models which allow the distribution of Y to depend upon G in some unspecified way. The term *Mode of Inheritance* refers to exactly how parameters of the distribution of Y depend on the number of disease alleles. Sometimes the term genetic model is used to describe only the mode of inheritance, and not the entire distribution, but we use genetic model to refer to the penetrance function specifying the entire distribution, and we generally use the mode of inheritance to indicate how the parameters of the penetrance function

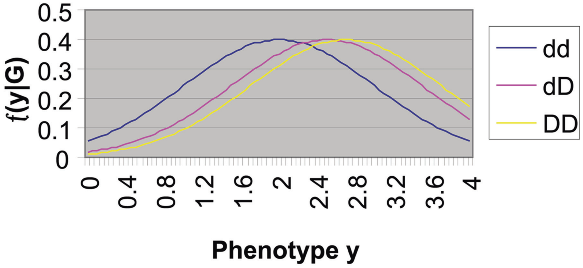

Fig. 2.3 Penetrance functions for a continuous trait

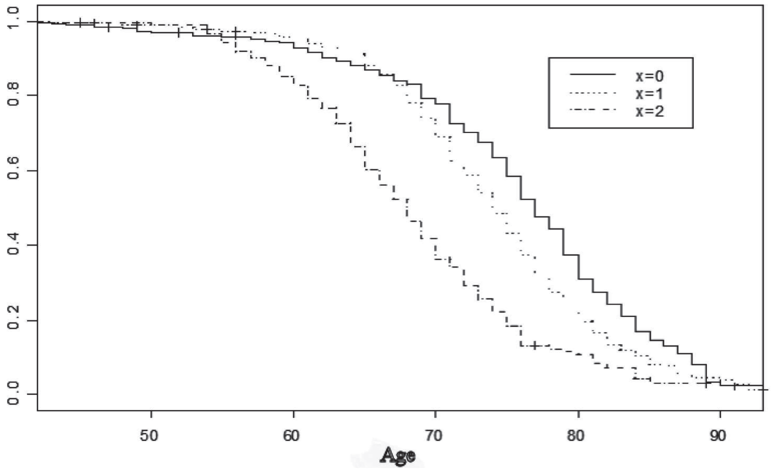

Fig. 2.4 Empirical survival curves for AD as a function of APOE genotype in the NIMH Genetics Initiative Alzheimer's Disease (AD) Sample. The genotype variable x counts the number of ε4-alleles at the locus

depend on the number of disease alleles. There are four modes of inheritance that are commonly used: *recessive*, *dominant*, *additive* and *codominant*. When only one copy of the disease allele is required to induce an effect on the disease phenotype, $\Pr(Y = 1|dD) = \Pr(Y = 1|DD)$, the mode of inheritance is called *dominant*. However, if 2 copies of the disease allele are required to elevate the disease risk, we speak of a *recessive model* or *recessive mode of inheritance*. Depending on the 'scale', with an *additive mode of inheritance* the penetrance probability of heterozygous genotype is mid-way between the penetrance probabilities of both homozygous genotypes, e.g., $P(Y = 1|Dd) = 0.5 * (P(Y = 1|DD) + P(Y = 1|dd))$ on the linear scale, or $P(Y = 1|Dd) = \sqrt{P(Y = 1|DD) * P(Y = 1|dd)}$ on the log (multiplicative) scale. The *codominant mode of inheritance* makes no assumptions about the relationship among the three penetrance functions, only that they are different. The *heterozygote advantage* model specifies that heterozygotes have the lowest (or highest for a heterozygote disadvantage model) risk of disease; it is occasionally used, especially in plant breeding. We do not use it since it is a special case of the more general codominant model.

Note that with dichotomous traits, $P(Y = 1|G)$ can be equivalently expressed as $E(Y|G)$, and likewise for the continuous trait, $\mu_G = E(Y|G)$. *Generalized Linear Models (GLM)* provide a convenient way to express the dependence of the trait mean on G without specifying the entire distribution of Y. A generalized linear model is similar to an ordinary linear regression model, except it allows the mean of Y to depend on covariates, X, in a non-linear way as:

$$g(E(Y|X)) = \beta_0 + X'\beta_1. \tag{2.3}$$

The link function, $g(\cdot)$, depends on the type of trait. For affection status outcomes, the logistic link:

$$\log[E(Y|X)/(1 - E(Y|X))] = \beta_0 + X'\beta_1, \tag{2.4}$$

or log(relative risk) link:

$$\log[E(Y|X)] = \beta_0 + X'\beta_1, \tag{2.5}$$

models are commonly used in epidemiological work; in genetics, linear models in the probabilities themselves are also commonly used.

Here X is a coding of the genotype that reflects the mode of inheritance; it can be a vector or a scalar, depending on the genetic model. By proper choice of X and link function $g(\cdot)$, all four modes of inheritance can be expressed by equation (2.3); β_0 is an intercept parameter, specifying $E(Y|G)$ when $X = 0$; β_1 gives the additional model parameters which specify how $E(Y|G)$ depends on G. Often the right-hand side of equation (2.3) is written as $X'\beta$ where β is a vector incorporating β_0 and β_1, and X is a vector with the first element always one; here we keep the parameters separate since a test of whether or not the gene affects the trait uses $H_0 : \beta_1 = 0$. Acceptance implies no relation between the gene and the trait. The coding of the genotype for each mode of inheritance is given in Table 2.2. From Table 2.2, we see that β_0 always specifies $E(Y|dd)$ and for the recessive model, it specifies $E(Y|Dd)$ as well. For the recessive, dominant and additive models, β_1 is a scalar and defines the 'effect size' in the chosen scale; for the codominant model, β_1 is a vector of length two that gives the effect of the DD and Dd genotypes compared to dd. Although more complex models can be constructed, these simple generalized linear models suffice for most analyses that we consider in detail.

Table 2.2 Coding the genotype (G) as X to specify the mode of inheritance

Recessive		Dominant	
X	G	X	G
1	DD	1	DD or Dd
0	dd or Dd	0	dd

Additive		Codominant		
X	G	$X1$	$X2$	G
2	DD	1	0	DD
1	Dd	0	1	Dd
0	dd	0	0	dd

2.3 The Biology Underlying Mendelian Inheritance

Today Mendel's Laws can be derived directly from our understanding of *Meiotic cell division* or *Meiosis*, which is the cell division that produces **gametes**, either sperm or ova; the union of a sperm and ova produces the fertilized egg cells (called *zygotes*). Meiotic cell division is in contrast to the standard cell division, *mitosis*, that serves the purpose of cell growth, development, repair and replacement of worn-out cells. While mitosis results in cells that are genetically identical (or clones), the purpose of meiosis is to introduce further genetic diversity by creating gametes, either egg cells or sperm cells, that are genetically different from the parent cells.

The nucleus of every cell contains two copies of each *chromosome* inherited from the parents, one maternal copy and one paternal copy. Such cells are called *diploid* because they have two copies of each chromosome (except for males who have one X and one Y for the sex chromosomes). Meiosis consists of two rounds of cell divisions, each following a meiotic division (Fig. 2.5) ending with four *haploid* cells containing only one copy of each chromosome.

In the beginning of the first meiotic division, both parental copies of the chromosome are duplicated; Fig. 2.5 illustrates the first meiotic division for a single parent in the top panel and the result of the second meiotic division in the bottom. Each parental chromosome is first duplicated as illustrated after the first arrow in the top panel. The duplicated chromosomes are called a pair of *sister chromatids*. The two duplicated chromosomes undergo a separation process; during this process, the arms of the chromosomes may overlap and segments of non-duplicate chromatids can be exchanged between the duplicated chromosomes, as illustrated after the second

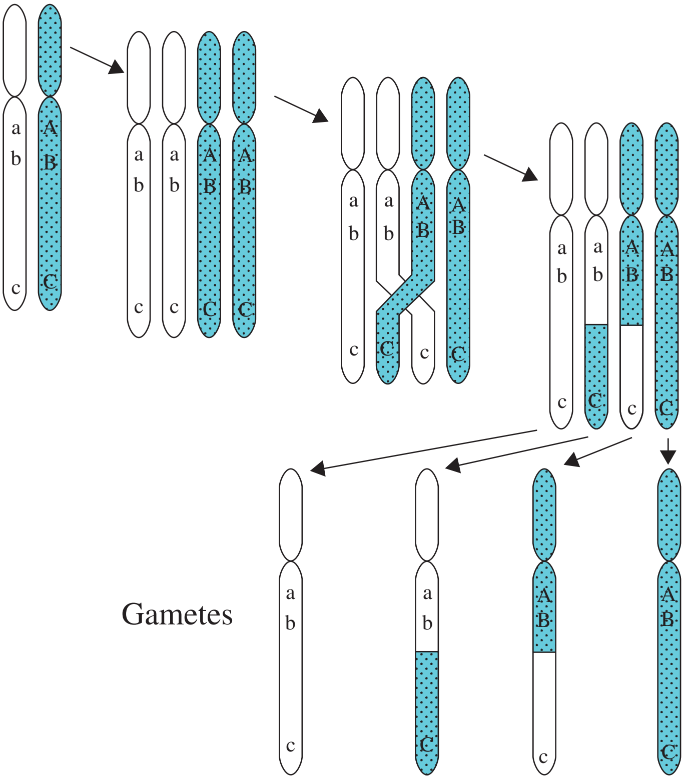

Fig. 2.5 Crossing-over and recombination during the formation of gametes (germ cells) or meiosis

arrow in Fig. 2.5. The exchange of material between two non-sister chromatids is called a *crossover event*. After the third arrow in Fig. 2.5, we see four chromatids. Two are identical to the one seen in the parent, but the other two are a mixture of the two chromosomes in the parent. Notice an important feature of crossing over: it allows each of the four gametes to be a mixture of the genetic material inherited from two grandparents, either maternal or paternal. Thus meiosis is not simply randomly choosing one of two parental chromosomes randomly but rather, it allows for creation of additional genetic diversity by mixing of grand parental information within a single chromosome. Each person inherits approximately $1/4$ of their genetic material from each of their four grandparents.

In the second meiotic division, the chromatids are separated and the final cell division forms two new cells around each chromosome, for a total of four haploid gamete cells. By crossing over, each gamete cell contains a different chromosome, however as a result of the first cell division, at each specific locus there are two gametes with the same maternal allele and two gametes with the paternal allele. A zygote requires one sperm and one ovum (egg cell); assuming that gametes unite randomly to form zygotes, it is then clear that the transmission of each parental allele occurs with probability $1/2$ since the two alleles are represented equally in the gamete cells.

Mendel's law of independent assortment states that alleles at different genetic loci are transmitted independently from one generation to the next. If they are on different chromosomes, this is naturally the case since each pair of chromosomes undergoes the process of meiosis independently. This creates a substantial amount of genetic variation, even without crossing-over; with crossing-over, the possible combinations are essentially infinite.

Crossovers are random events in the sense that they cannot be predicted with certainty; however they do not occur uniformly or independently along the chromosome. Rather, crossover rates can vary by sex, chromosomal region as well as chromosome number, individual and temperature. Crossing over is relatively rare at the centromere and at the ends of a chromosome. *Interference* can create dependencies in the occurrence of successive crossovers. For example, the occurrence of a crossover in a region decreases the chance of a second crossover in an adjacent region, nearly to zero if the regions are very close. Overall the entire genome, the average number of crossovers is about 55 in males, and about 50% greater in females. The average number of crossovers on a chromosome depends upon its length. Thus despite the fact that crossovers do not occur uniformly, they have served as a useful measure of distance for linkage mapping as described in Chapter 5.

Crossovers are inherently unobservable, so we use the concept of recombination to describe crossovers. If we obtain data at two or more loci on a parent and their offspring, then we can infer something about crossovers occurring between the loci provided the parent is heterozygous at the loci. Referring to Fig. 2.5, the parent is heterozygous at three locations, with alleles Aa, Bb and Cc. The set of alleles lying on the same chromosome is called the *haplotype*. Here the two haplotypes are ABC and abc. Note that these haplotypes have been inherited from the two parents of the parent, i.e., the grandparents of the offspring whose gametes are displayed. Suppose that the first gamete, abc, is inherited from the parent. There is no evidence of crossing over here because one parental chromosome is identical abc, and the other parental chromosome shares none of these alleles. In this case we say there is no *recombination* between either the A to B locus, or the B to C locus (or A to C either). Suppose the offspring inherits the second gamete, abC. In this case, the offspring's haplotype differs from either of the parent's haplotypes, thus a crossover must have occurred between the B and C locus, but not the A and B. Thus we say no recombination has occurred between A and B, but a recombination occurred between B and C.

There is not a one-to-one relationship between recombination events and crossing over because recombination refers only to what can be observed between the two specific loci, whereas crossing over refers to events that can occur anywhere in the interval. If no crossover has occurred between two loci (as between the A and B loci in Fig. 2.5) then we will not see a recombination. However, it is possible for two crossovers to occur in an interval; in this case, we may see no recombinant between two markers flanking the interval, i.e., there may be segments of grand-maternal material at the ends of the interval, with grand-paternal material in the middle. The formal definition of the recombination fraction θ is given by P(recombination occurs between two loci).

Crossovers between two loci very close to one another are rare. In this case, the probability of a recombination between the two loci is very small. For example in Fig. 2.5, considering loci A and B, among the four gametes, we observe two ab gametes and two AB gametes: thus among these gametes, the probability of A or a (or B or b) is always $\frac{1}{2}$ by Mendel's law of segregation, but P(A allele and B allele) $=$ P(a allele and b allele) $= \frac{1}{2}$ and P(A allele and b allele) $=$ P(a allele and B allele)$=$ 0. This is contrary to what we would expect by Mendel's law of independent assortment, which would specify a probability of $\frac{1}{4}$ for each of the four possible gametes.

Between loci B and C, the situation is different because we observe a recombination. Again, among the four gametes, P(B) $=$ P(b) and likewise for C and c, but now P(b and c) $=$ P(B and c) $=$ P(b and C) $=$ P(B and C) $= \frac{1}{4}$, which corresponds to independent assortment. In general, the distribution of gametes over many meioses will depend upon the number of crossovers between them. If the two loci are close, θ is small, and the alleles at two loci tend to be inherited together, so that the law of independent assortment does not hold.

The relationship between θ and the distribution of crossovers is given by Mather's law:

$$\theta = (1 - P_0)/2,$$

where P_0 is the probability of zero crossovers. Mather's law can be argued as follows. If there are no crossovers, $P_0 = 1$, and there can be no recombination. With probability $(1 - P_0)$, at least one crossover occurs. If at least one crossover occurs, then the probability of a recombination is $\frac{1}{2}$, regardless of the number of crossovers.

To see why, recall that crossovers cannot occur between sister chromatids, but only between non-sister chromatids. It is easy to see from Fig. 2.5 that one crossover will create two recombinant gametes and two non-recombinant gametes. With two crossovers, the same two non-sister chromatids can be involved in both crossovers (and the number of recombinant gametes is zero) or both sister chromatids of each pair cross over once with their non-sister chromatids, in which case all four gametes are recombinants. Since these two possibilities are equally likely, the average proportion of recombinants is $\frac{1}{2}$. The last possibility, that one sister chromatid crosses over twice with two different non-sister chromatids, gives 2 recombinant and 2 non-recombinant gametes. It is straightforward to argue the probability of a recombinant is also $\frac{1}{2}$ for three crossovers, and so on.

If two loci are very far apart, there are likely many crossovers between them; P_0 approaches one in the limit and the recombination fraction approaches $\frac{1}{2}$. The upper limit of θ corresponds to what we might expect if two loci are on different chromosomes, since by the law of independent assortment, if the parent is heterozygous at both loci, the four gametes will carry the four possible combinations, AB, Ab. aB, and ab with equal probability.

2.4 Exercises

1. Verify lines 1–3 of Table 2.1 using Mendel's first law.
2. Assume two genes with alleles A/a and B/b, controlling two different traits. Assuming that Mendel's second law holds (the alleles underlying the two different traits are inherited independently), and starting with the pure strains as in Mendel's experiments:

   (a) Verify the 1:2:1:2:4:2:1:2:1 ratios for the 9 possible genotypes inferred in the F_2 generation.
   (b) Verify the 9:3:3:1 ratio for 4 possible traits observed in the F_2 generation.

3. In the early 1900s, scientists William Bateson and R. C. Punnett studied inheritance in two genes of Sweet Peas: one affecting flower color (P, purple, and p, red) and the other affecting the shape of pollen grains (L, long, and l, round). Capital letters denote dominant forms, as in Mendel's paper. They crossed pure lines PP · LL (purple, long) × pp · ll (red, round), and self-fertilized the F_1 offspring Pp · Ll heterozygotes to obtain an F2 generation. The table below shows the counts of each phenotype in the F_2 plants.

	Number of progeny	
Phenotype (and genotype)	Observed	Expected from 9:3:3:1 ratio
purple, long (P/– · L/–)	4831	3911
purple, round (P/– · l/l)	390	1303
red, long (p/p · L/–)	393	1303
red, round (p/p · l/l)	1338	435
	6952	6952

(a) Verify the Expected column for testing goodness of fit to the 9:3:3:1 ratio.

(b) Show that the chi-square goodness of fit test exceeds significance.

Note: As a possible explanation for the lack of fit, Bateson and Punnett proposed that the F_1 had actually produced more P × L and p × l gametes than would be produced by Mendelian independent assortment. Because these genotypes were the gametic types in the original pure lines, the researchers thought that physical coupling between the dominant alleles P and L and between the recessive alleles p and l might have prevented their independent assortment in the F1. However, they did not know what the nature of this coupling could be.

(c) What is another possible explanation for lack of fit?

4. How many genotypes are possible with a 3-allele marker? With K alleles?
5. Early onset Alzheimer's disease is very rare; for illustrative purpose, assume it is 0.1% among adults aged 30-60. Rare variants in 3 genes, APP, PSEN1 and PSEN2 have been identified as causing early onset AD in a dominant fashion, with P(AD | any of the three variants) = 1. Early onset AD can also be caused by head injury; many other non-genetic factors have been suggested. In a series of 101 cases of early onset AD, only 7 (or approximately 7%) were found to have these variants in APP, PSEN1 or PSEN2; that is, the attributable risk due to the three rare variants is low. For simplicity, assume that the probability of variants in these 3 genes is so rare that we can assume P(no variant in any gene) ≈ 1. Let the disease allele D symbolize a variant in any one of the three genes, d is no variant, and Y = 1 means AD present.

   Estimate the probability of a phenocopy, $P(Y = 1|dd)$ (also known as phenocopy rate) for these genes combined, using the data given and Bayes Rule.
6. Consider a recessive Mendelian disease, where in the population, P(an individual has 2 disease variants) = 0.000001.

   (a) What is the probability that a randomly selected person is affected? Suppose that the randomly selected person is affected. What does that imply about the probability that their sibling is also affected (you can assume that having either one or two parents with two variants is so rare that you can ignore them)?

   (b) Now answer both of these questions assuming the penetrance is only $\frac{1}{2}$, i.e., $P(\text{disease} \mid 2 \text{ variants}) = \frac{1}{2}$, but the phenocopy rate is still zero.
7. Suppose we are dealing with a quantitative recessive trait, which is distributed as $N(\mu, 1)$ when there are two variants, and $N(0, 1)$ otherwise. Calculate the probability that a randomly selected person with two variants has a trait higher than a person with one or no variants, when $\mu = 0.5$, and when $\mu = 2$.
8. Suppose we observe a quantitative trait which seems to show variation in both the mean and the variance as a function of genotype. Give one example of a genetic model which allows for this.
9. One of the dichotomous traits that Mendel studied, length of plant stem, was actually dichotomized from the measured length. He selected plants with a 6–7'

long axis to have the dominant trait and plants with a $3/4'$ to $1.5'$ long axis to have the recessive trait. Mendel commented that in fact, "...the longer of the two parental stems is usually exceeded by the hybrid... Thus for instance, in repeated experiments, stems of $1'$ and $6'$ in length yielded without exception hybrids which varied in length between $6'$ and $7.5'$." What would be an appropriate (non-deterministic) Gaussian penetrance function model for axis length as a continuous trait? Mendel also noted that there is very little variation in stem height within genotype class. What does that imply about your Gaussian model?

10. Consider the Generalized Linear Model given in equation (2.3) Suppose you wish to include covariates, such as sex or age. Suggest how you might do that in the context of the GLM.
11. Verify the statement concerning two crossovers: If one paternal chromatid crosses over twice with two different maternal chromatids, this gives 2 recombinant and 2 non-recombinant gametes.

http://www.springer.com/978-1-4419-7337-5

The Fundamentals of Modern Statistical Genetics
Laird, N.M.; Lange, C.
2011, XIV, 226 p., Hardcover
ISBN: 978-1-4419-7337-5