Chapter 1

Genes, Pedigrees and Genetic Models

1.1 DNA, alleles, loci, genotypes, and phenotypes

The *DNA* in the nuclei of cells of an individual consists of about 3×10^9 base pairs (bp). This *DNA* is packaged into *chromosomes* each of which has a linear DNA sequence in a twisted double-helical structure. There are 46 chromosomes in the nucleus of each normal human cell, 22 pairs of *autosomes* and a pair of *sex chromosomes*. Of the two chromosomes of a pair, one derives from the DNA of the mother of the individual and the other derives from the DNA of the father. In this book, we will restrict attention to the autosomes, which contain the majority of the DNA coding for the proteins and affecting the characteristics of individuals. Similar approaches would apply to the sex chromosomes, but the details differ. There is additional DNA in the mitochondria, which are located in the cytoplasm of the cell; mitochondrial DNA is maternally inherited.

Any small segment of the DNA of the chromosome is known as a *locus*. Typically, a locus used to refer to the segment of DNA coding for some functional protein, but it is now used to refer to any position characterized by a specific DNA sequence, or by specific forms of variation in the sequence. These loci exhibiting observable variation in the DNA are *DNA marker loci*, and a *locus* simply indicates a particular position on a particular one of the pairs of chromosomes. The DNA at a locus may come in a variety of forms, or *alleles*. Any individual has two chromosomes of a given pair, and thus has two (possibly identical) alleles at each locus. The unordered pairs of alleles that an individual has is the individual's *genotype* at this locus. If the locus is one relating to a functional gene, the resulting potentially observable characteristic of the individual is the *phenotype*. A locus exhibiting non-negligible variation in a population is known as a *polymorphism*, or the locus is said to be *polymorphic*. Classically, the frequency of the the most frequent genotype should

be less than 99% for a locus to qualify as a polymorphism.

For example, the DNA which codes for the antigens that determine an individual's ABO blood type is at a certain position on Chromosome 9. This is a chromosome in mid-range size; chromosomes are numbered in approximately decreasing size order. This position is the *ABO locus.* There are three major alleles at the human ABO locus, A, B, and O, although these allelic types can be subdivided. The ABO locus is polymorphic in almost every human population. There are thus six genotypes; AA, AO, BB, BO, OO, and AB. However there are only four phenotypes (ABO blood types), type-A, type-B, type-O and type-AB. Individuals with genotype AA or AO have type-A blood type; individuals with genotype BB or BO have type-B blood type. For each of the phenotypes O and AB, there is a single corresponding genotype.

A genotype for which the two alleles are the same, such as AA, BB or OO are known as the *homozygous* genotypes. The individual is a *homozygote* or is *homozygous* at this locus. Where the two alleles are different (AO, BO or AB), the individual is a *heterozygote* or is *heterozygous* at this locus. Where a heterozygous genotype exhibits the same phenotype as one of the two homozygotes, the allele carried by this homozygote is said to be *dominant* to the other allele. At the ABO blood type locus, for example, individuals of genotype BO have type-B blood. The B allele is dominant to the O allele; the O allele is *recessive* to the B allele. Likewise, the A allele is dominant to the O allele; the O allele is recessive to the A allele. Individuals of genotype AB have type-AB blood, distinct from the phenotypes of both the AA (type-A) and BB (type-B) genotypes. The alleles A and B are said to be *codominant.*

Initially, genetic markers used in genetic analysis were blood type or enzyme markers such as the ABO locus. The first DNA markers were restriction fragment polymorphisms or RFLPs (Botstein et al., 1980). These often had several alleles, or at least two alleles with substantial frequency. These were followed by current microsatellite markers, where alleles correspond to different numbers of tandem repeats of a small number (2, 3, or 4) of base pairs. Microsatellite markers are often highly polymorphic, with 10 or more alleles observed with non-negligible frequency in any given population. These have become the markers of choice for genetic mapping, but statistically have several disadvantages all due to the high degree of polymorphism. Mutation rates at some microsatellite markers are high, and typing errors also more frequent. Accurate estimation of population allele frequencies is harder, and inferences can be sensitive to allele frequency assumptions. The newest DNA markers are single-nucleotide polymorphisms or SNPs. These measure variation at a single base of DNA. Although there are many SNPs in the human genome, perhaps as many as 1 per 500 bp, or several million in total, most have only two alleles. In the future, genetic mapping analyses may be based on a much larger number of much less informative markers with consequent additional challenges.

1.2 Mendel's laws and meiosis indicators

Mendel's First Law (1866) states that each individual has two "factors" (or genes) controlling a given characteristic, one being a copy of a corresponding gene in the father of the individual, the other a copy of a gene in the mother of the individual. Further, a copy of a randomly chosen one of the two is copied to each child, independently for different children and independently of genes contributed by the spouse. The probabilistic process of the random choice of genes to be copied is known as Mendelian *segregation.* The biological process forming the chromosomes of the gamete (sperm or egg) cell is known as *meiosis* At a single locus, the *segregation* of genes is fully specified by *meiosis indicators*

$$\begin{aligned} S_i &= 0 \quad \text{if copied gene is parent's maternal gene} \\ &= 1 \quad \text{if copied gene is parent's paternal gene} \end{aligned} \tag{1.1}$$

where $i = 1, ..., m$ indexes the meioses (parent-child links) in the pedigree. Mendel's First Law then simply states that the indicators S_i are independent, and

$$\Pr(S_i = 0) \;=\; \Pr(S_i = 1) \;=\; \frac{1}{2}.$$

For multiple loci, j, $j = 1, \ldots, L$, we must specify the segregation of genes at each locus:

$$\begin{aligned} S_{i,j} &= 0 \quad \text{if copied gene at meiosis } i \text{ locus } j \text{ is parent's maternal gene} \\ &= 1 \quad \text{if copied gene at meiosis } i \text{ locus } j \text{ is parent's paternal gene.} \end{aligned} \tag{1.2}$$

Contrary to Mendel's second law (Mendel, 1866), which in effect stated that $S_{i,j}$ are independent for different loci j, the segregation of alleles at loci on the same chromosome are dependent. The collection of alleles at loci on a chromosome in the maternal [paternal] gamete, is the maternal [paternal] *haplotype* of the offspring individual.

The word "gene" is overused in modern genetics, often referring to the locus (as in "the *ABO* gene"), or to an allele predisposing the individual to a particular disease or trait (as in "the cystic fibrosis gene"). Here we reserve the word "gene" for Mendel's original "factors"; the gene is the entity transmitted from parent to offspring. The meiosis indicators $S_{i,j}$ have also attracted a variety of names and notations. Karlin and Liberman (1979) used them in the theoretical analysis of meiosis patterns at loci on a chromosome (Chapter 5). Their first use in the computation of probabilities of gene descent in pedigrees is due to Donnelly (1983), who called them *switches.* Thompson (1994*c*) retained the notation of Donnelly (1983), but called them *segregation indicators.* Lander and Green (1987) use the phrase *inheritance vectors* while Sobel and Lange (1996) use *descent graphs.* Together with a defined pedigree structure, the meiosis indicators do indeed determine the inheritance or descent patterns of genes in a pedigree (section 3.6). However, in considering the indicators alone we prefer the name *meiosis indicators.*

For later convenience we define the following notation

$$
\begin{aligned}
S_{\bullet,j} &= \{S_{i,j}; i = 1, \ldots, m\}, \quad j = 1, \ldots, L \\
S_{i,\bullet} &= \{S_{i,j}; j = 1, \ldots, L\}, \quad i = 1, \ldots, m
\end{aligned} \tag{1.3}
$$

where m is the number of meioses in the pedigree, and L the number of loci along the chromosome. The m vectors $S_{i,\bullet}$ are *a priori* independent, but the components $S_{i,j}$ are dependent. The pattern of dependence depends on the process of meiosis, which will be considered further in Chapter 5. However, under (untrue) assumptions of absence of genetic interference in meiosis, there is a simple conditional independence structure. Suppose the loci are ordered $1, \ldots, L$ along a chromosome. Then given $S_{i,j}$, $(S_{i,1}, \ldots, S_{i,j-1})$ is independent of $(S_{i,j+1}, \ldots, S_{i,L})$. Or, $(S_{i,j})$ is first-order Markov over j.

1.3 Pedigrees: the conditional independence structure

A *pedigree* is a specification of the genealogical relationships among a set of individuals. A convenient form of this specification is to identify the father and the mother of each individual. Individuals at the top of the pedigree, whose parents are unspecified, are the *founders* of the pedigree; other individuals are *non-founders*. Individuals in the pedigree, and without offspring, are referred to as *final* individuals; unless there are data on such an individual, he contributes no information. Relationships among individuals are defined relative to the specified pedigree; thus, by definition, the founders are unrelated.

The pedigree of Figure 1.1 will be used extensively in examples throughout this book. It is a true pedigree structure, and derives from a study of Werner's syndrome, a rare recessive trait (Goddard et al., 1996). The pedigree has five founders and ten non-founders. In accordance with standard notation, male individuals are shown as squares, and females as circles. The form in which the pedigree is shown here is a *marriage node graph*. Individuals who together produce offspring are connected to a single *marriage node*, which is in turn connected to the resulting offspring. This pedigree was ascertained because the final individual was known to be the offspring of a first-cousin marriage, and was affected by a rare recessive trait. Typically affectation status for a trait of interest is depicted by shading, as in Figure 1.1. It was later discovered that each of the parents of the final individual is also the offspring of a marriage between first cousins.

The meiosis indicators determine the descent of genes in a pedigree. Figure 1.2 gives an example, using the pedigree of Figure 1.1. The meiosis indicators shown under each individual are for the paternal and maternal meiosis to that individual. For easier visualization males and paternal meioses or genes are shown to the left, and females and maternal meioses and genes to the right. For example it is seen that the paternal gene of the final individual is the same gene (labeled "8") as the maternal gene of his maternal grandfather. The gene does not descend from

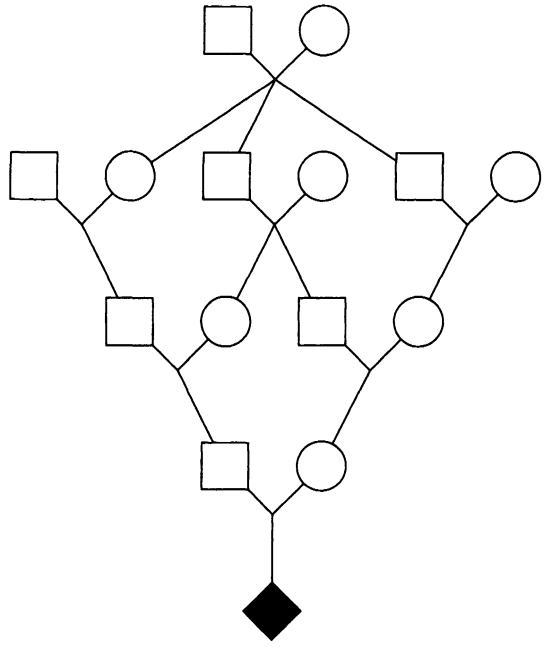

FIGURE 1.1. *An example pedigree from Goddard et al. (1996)*

grandparent to grandchild (the mother does not carry this gene), but both are copies of the same founder gene.

The meiosis indicators $\mathbf{S}$ determine the descent of founder genes. Together with the allelic type of each founder gene at each locus (the founder haplotypes), $\mathbf{S}$ determines the complete configuration of genotypes $\mathbf{G}$ on the pedigree. The independence of meioses then leads to a straightforward conditional independence structure for genotypes on a pedigree. Individuals receive copies of their parents' genes, and, together with their spouses, segregate genes to their offspring.

This conditional independence leads to one formulation of the probability of phenotypic data $\mathbf{Y}$ observed on a pedigree. The probability of the genotypes $\mathbf{G} = \{G_i\}$ is

$$\Pr(\mathbf{G}) = \prod_{\text{founders } i} \Pr(G_i) \prod_{\text{nonfounders } i} \Pr(G_i | G_{M_i}, G_{F_i}) \tag{1.4}$$

where G_i is the genotype of individual i, and M_i and F_i are the parents of individual

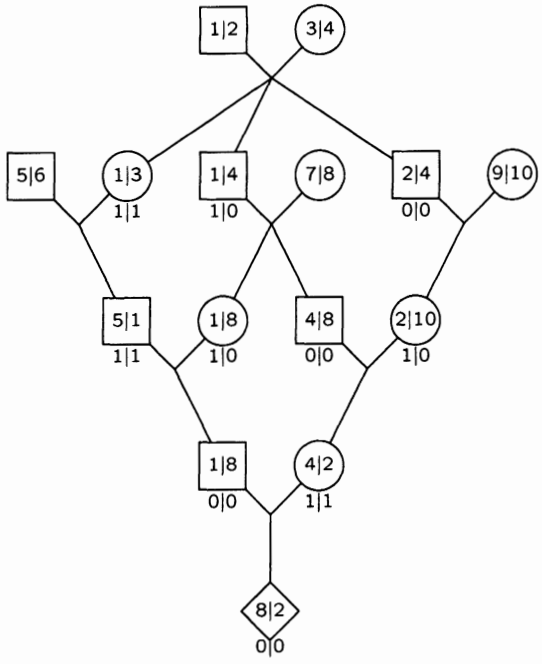

FIGURE 1.2. *Meiosis indicators $S_{\bullet,j}$ determine descent of founder genes, at any given locus j. The indicators $S_{i,j}$ are shown under the offspring individual, while the resulting labeled founder genes are shown within each individual*

i. The data on individuals are determined by their underlying genotypes:

$$\Pr(\mathbf{Y}) = \sum_{\mathbf{G}} \left(\prod_{\text{observed } i} \Pr(Y_i \mid G_i) \right) \Pr(\mathbf{G}) \tag{1.5}$$

The genotype G_i of individual i is the multilocus genotype: that is, a pair of haplotypes over all the relevant genetic loci. The phenotype Y_i may be a multivariate phenotype, with qualitative and/or quantitative components.

Two alternative views of the conditional independence structure are shown in Figure 1.3: this pedigree is slightly modified from our usual example, in order to have an individual with two spouses. As can be seen from equations (1.4) and (1.5), the conditional probability of a genotype G_i of individual i, given the genotypes of all other pedigree members, and given the data $\mathbf{Y}$, depends only on the data Y_i on individual i, and on the genotypes of parents, spouse(s), and offspring (Figure 1.3(a)). At a finer scale, provided paternal and maternal genes of individuals are distinguished, we may consider the dependence structure among

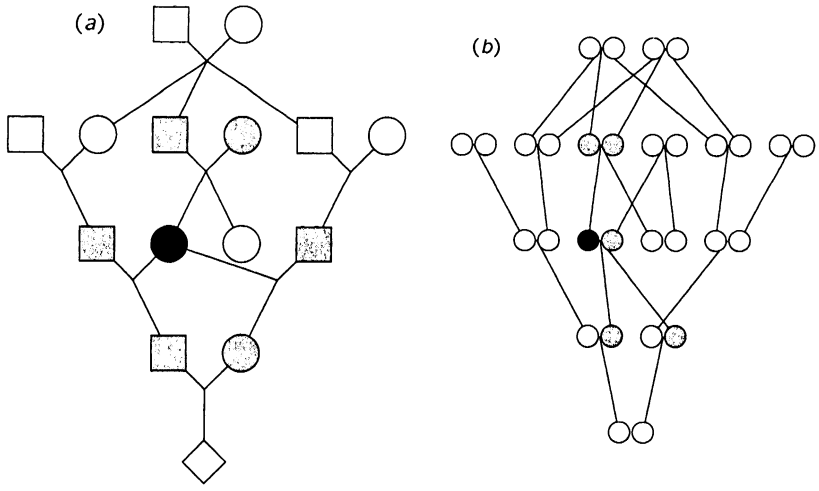

FIGURE 1.3. *The conditional independence neighborhood structure on a pedigree: (a) the individual neighborhood, and (b) the haplotype neighborhood. The reference individual (a) or haplotype (b) is dark shaded. The individuals [haplotypes] defining the local dependence structure for the reference individual [haplotype] are light shaded*

haplotypes (Figure 1.3(b)). Here, for example, for a paternal haplotype of a female individual i, the dependence is on the data Y_i, the maternal haplotype of i, the two haplotypes of the father of i, and the maternal haplotypes of the children of i. These are the haplotypes that segregate to, with, or from the paternal haplotype of i. The set of possible states of a haplotype neighborhood is smaller than of the genotypic neighborhood, since there are fewer haplotypes than multilocus genotypes, the latter being pairs of haplotypes; however there are more haplotype neighborhoods in the pedigree. Either can be more computationally efficient to consider in computing the probability of data on the pedigree (equation (1.5)).

1.4 Models, parameters, and inferences

Considered as a function of the parameters θ of the genetic model the probability $\Pr(\mathbf{Y})$ of equation (1.5) is the likelihood function $L(\theta)$. Broadly, there are three classes of parameters. First, there are *population parameters*, such as allele frequencies and allelic associations within and among loci. These index the probability distributions of founder genotypes and haplotypes in equation (1.4). Some examples of the estimation of these parameters from population data will be considered in Chapter 2. Other parameters such as those of assortative mating also enter into the probabilities of founder genotypes. Phenotypic correlations between

spouses impose genotypic dependencies which influence the probability distributions for offspring data. However, assortative mating will not be considered further in this monograph.

Second, there are *transmission parameters* which index the probability distributions of the meiosis indicators, and hence the probability of offspring genotypes, conditional on those of parents (equation (1.4)). Most importantly, there are parameters such as recombination frequencies, which characterize the dependence in meiosis among loci on a chromosome. The estimation of recombination frequencies will be addressed in Chapter 4, and linkage analysis more generally will be addressed throughout the monograph. The other major class of transmission parameters are those characterizing any deviations from Mendelian segregation. In some approaches to segregation analysis (Elston and Stewart, 1971), a test of Mendelian segregation proportions is performed. Theoretically, each heterozygote should transmit each allele with equal probability, and the probability distribution of the meiosis indicators should not depend upon the allelic types of the genes. However, caution is necessary in interpreting the results of such tests. Apparent distortion may result from selection; offspring individuals surviving to be typed, or to reproduce, may not be a random sample of those resulting from meiosis. In the case of crop plants, or domestic animals, there may be very strong artificial selection on certain loci, which affects the apparent segregation at loci on the same chromosome. In the case of studies of data on human pedigrees, similar apparent distortions can result from the ascertainment of pedigrees, or of parts of a pedigree, in which a particular trait is segregating. This ascertainment also leads to distorted segregation patterns at linked marker loci showing any allelic associations with the trait locus. Indeed, some tests for linkage in the presence of trait-marker allelic associations are tests of apparent segregation distortion in the meioses to affected offspring. Ascertainment is an important topic in the analysis of data on human pedigrees, and there is a large literature from Weinberg (1912) to Karunaratne and Elston (1998). However, it is outside the scope of this monograph.

Third, there are *penetrance parameters* indexing the relationship between genotype and phenotype. These enter only into $\Pr(\mathbf{Y} \mid \mathbf{G})$ (equation (1.5)). The probability that an individual carrying a certain allele is affected by a trait is known as the penetrance of the allele, which includes the degree of dominance. Another penetrance parameter is the probability of phenocopies (individuals exhibiting the phenotype of a genetic trait, but not having the predisposing genotype). More generally, penetrance parameters may characterize allelic and genotypic contributions to a quantitative trait, and the effects of individual environmental effects and covariates. Important covariates include gender and age; many complex traits are age and gender dependent. Also, different genotypes predisposing to the same disease may have different effects on age of onset. Additionally, there may be interaction effects, between alleles at different loci contributing to a given trait (epistasis), between multiple traits affected by alleles at a single locus (pleitropy), or between genetic effects and environmental covariates. The focus of this monograph is on Mendelian traits, such as DNA markers. We shall not consider the broad spectrum of parameters indexing the relationship between genotype and complex phenotypes.

The primary focus of this monograph is methods for inference from genetic data on pedigrees. We shall focus on inferences about the parameters of genetic models; that is, on segregation and linkage analysis. As data become increasingly available on a genomic array of markers, we focus on genetic mapping and the analysis of genetic maps. However, with a given genetic map, the probability of data $\Pr_{\gamma}(\mathbf{Y})$ (equation (1.5)) provides a likelihood for an hypothesized pedigree structure among individuals. Thus, pedigree validation and relationship estimation, using a genomic array of linked DNA markers, are methodologically analogous to segregation and linkage analysis. Other inference questions also require the computation of a probability $\Pr(\mathbf{Y})$, of phenotypes observed on some members of a pedigree structure. For example when both genetic model and pedigree structure are assumed correctly known, data provide information for the inference of ancestral origins of alleles (Geyer and Thompson, 1995), or of phenotypic risks for individuals.