Genetic Epidemiology

III.7

Heike Bickeböller

7.1 Introduction

Genetic epidemiology both emanates from and combines the scientific disciplines of human genetics and epidemiology as well as biometry. Strong interdisciplinary relationships exist among others with the fields of molecular genetics and medicine/medical care. Some overlap exists with molecular epidemiology (see Chap. III.6 of this handbook).

Genetic epidemiology is essential in the field of human genetics as it aims to detect the genetic origin of phenotypic variability in humans (Vogel 2000). In particular, genetic epidemiological studies unravel the genetic components that contribute to the development or the course of a disease, or in general terms to a *phenotype*, i.e. the observed trait.

Genetic epidemiology is the subdiscipline of epidemiology devoted to diseases/phenotypes with genetic components and to their respective genetic risk factors. The aims are (1) the description of genetically influenced phenotypes or diseases in populations and families, (2) the identification of genetic risk factors associated with the frequencies of phenotypes in the population and/or leading to familial aggregation, and (3) the modelling of the role of these genetic risk factors in populations and families (Khoury et al. 1993). Thus, both population-based and family designs are complementary and play a central role in genetic epidemiological studies. In contrast to classic epidemiology, the three main complications in genetic epidemiology are dependencies, use of indirect evidence and complex data sets: Genetic epidemiology is highly dependent on the direct incorporation of family structure and biology. The structure of families and chromosomes leads to major dependencies between the data and thus to customized models and tests. In many studies only indirect evidence can be used, since the disease-related gene, or more precisely the functionally relevant DNA variant of a gene, is not directly observable. In addition, the data sets to be analyzed can be very complex.

Genetic epidemiology is also a highly specialised subdiscipline of biometry and mathematical population genetics. The field has made major biometrical contributions to human genetics or, relying on earlier biometrical work (since the field was only recently endowed with the name) such as the description of the central Hardy–Weinberg equilibrium (HWE) (Hardy 1908; Weinberg 1908) and the development of statistical methods including segregation analysis, linkage analysis, association analysis, simulation methods and computer algorithms for all major study designs implemented.

The International Society of Genetic Epidemiology describes the field as a marriage between the disciplines of genetics and epidemiology (IGES 2003). It emphasises the need to join the fields. Genetics tends to focus on the genotype-phenotype correlation neglecting the environment. Epidemiology tends to focus on environmental risk factors as well as demographic factors (e.g. age, sex, ethnicity) and familial aggregation as a first step towards genetic risk factors. However, a full understanding of the etiology of complex traits may only be achieved by considering both, genetics and environment, and thereby explaining how genes are expressed in the presence of different environmental contexts. This chapter is solely devoted to methods dissecting the genotype-phenotype correlation with a binary phenotype (affected/unaffected). It is not covering the important subjects of quantitative phenotypes and gene-environment interaction.

Section 7.2 presents an overview of major study designs and types of analysis. Section 7.3 introduces the most important genetic models. Sections 7.4–7.6 will cover the three major types of analysis, i.e. segregation, linkage and association analysis.

7.2 Study Types

Genetic epidemiological investigations are usually triggered by epidemiological studies that demonstrate a *positive family history* as a risk factor for disease indicating putative genetic or shared environmental factors. Often the goal of initial studies is to estimate the *relative risk for relatives of affected individuals* in relation to the general population in order to support the genetic hypothesis.

To further investigate familial aggregation, a *segregation analysis* may be carried out in pedigrees. The aim of such an analysis is to determine whether a *major gene* is influencing a given phenotype in these families and if so to estimate the parameters of the underlying genetic model. All methods for segregation analysis are based on probability calculations for observed phenotypes conditional on hypothetical genetic model parameters and on family structure, i.e. *genealogies*. Parameter estimation is often based on likelihood-ratio tests in order to select the most plausible model nested within a hypothetical general model.

The primary cause of a so-called *monogenic disease* such as cystic fibrosis is a mutation within a single gene that segregates according to Mendelian laws (see below). The predisposing variants, i.e. the alleles carrying the risk, of this *major gene* are usually rare in the population. For *complex or multifactorial diseases*, Mendelian subforms such as the subform of breast cancer caused by the major gene BRCA1, genetic and non-genetic susceptibility factors, or risk modifying genes can exist. For rare monogenic diseases and rare Mendelian subforms of complex diseases, segregation analysis and subsequent further analyses perform well. However, complex diseases in general require more sophisticated methods of analysis than monogenic diseases. For example in Alzheimer's disease at least three major genes and several susceptibility genes confering moderate risk (*oligogenes*) exist. Oligogenes as genetic risk factors can be frequent in the population. *Polygenic* effects at many loci across the whole genome may contribute to disease, each with a minor effect.

If there is sufficient evidence for the existence of genetic factors contributing to a (complex) disease, the next step will be to locate or to identify susceptibility genes in order to quantify the genetic influence and to understand the underlying genetic model and pathway to the phenotype. For this purpose, measures of correlation between a *genetic marker* and the (unknown) *disease locus* are used. A genetic marker is a DNA segment with multiple alleles for which the localisation on the chromosome is known and the alleles can be determined. In general, methods assume Mendelian segregation of the marker (see Sect. 7.3.2). The most frequently used markers are multiallelic *restriction fragment length polymorphisms (RFLPs)* or *microsatellites* and biallelic *single nucleotide polymorphisms (SNPs)*. Usually the frequency of the most common variant must be less than 99% before a marker is termed a *polymorphism*.

For the analysis of complex diseases with genetic marker data we can distinguish two major approaches. Both investigate the genotype-phenotype correlation. Depending on the context, either the first or the second approach is more efficient (Clerget-Darpoux and Bonaïti-Pellié 1992). The first approach is a *genome scan,* i.e. the systematic coarse grid search of the whole genome with a map of genetic markers, with the objective to locate a region harbouring a susceptibility gene. A typical study would investigate approximately 350 markers with an average distance of 10cM (centiMorgan, see Sect. 7.3.3.) along the genome in families. The other approach is to investigate *candidate genes* (or candidate gene regions). Thereby, the focus is set on genes for which their function on the pathway to the phenotype can evidently be assumed. The most prominent example of a candidate gene system is the HLA (human leucocyte antigen) complex on chromosome 6. HLA is involved in immune resistance and is thus a natural candidate gene region for all autoimmune diseases.

The aims of a candidate gene investigation are to find evidence of any contribution of the candidate gene to the disease and to model its influence on the disease. The genotypes of the relevant functional component of the candidate genes are not always observed. We therefore need to use the information on genetic markers that lie in close proximity to the candidate gene in question. In general, nonparametric approaches are to be preferred, since they need fewer assumptions about the underlying genetic model.

There are two types of information that describe the correlation between a genetic marker and the susceptibility locus of a disease. (The correlation is maximized, when the genetic marker is identical to the functional variant for susceptibility):

- *Linkage (cosegregation at the family level)*: The common segregation of a marker and a disease is investigated. Inheritance is characterised by the transmissions of DNA segments from parents to offspring. If the transmissions at the marker locus and at the disease locus from one parent to a child are not independent, then this is denoted as linkage. Under linkage relatives with a similar disease status (e.g. both affected) are more similar at the marker locus than to be expected under independence.
- *Linkage disequilibrium (association at the population level)*: Linkage disequilibrium (LD) is present, if the probability for the existence of a specific marker allele together with a specific disease allele in a population gamete differs from the product of individual probabilities. Certain marker alleles of affected in-

dividuals will be more frequent or less frequent than in a randomly selected individual from the population.

Linkage analysis in families uses the concept of linkage. *Association analysis* in populations or families uses the concept of linkage disequilibrium. Some designs and corresponding statistical methods are capable of integrating both types of information into the analysis.

Detailed information on diseases used as examples in this chapter may be found in the standard reference of human genetics of McKusick (1998) or its online version, Online Mendelian Inheritance in Man (OMIM 2000).

Genetic Models 7.3

Fundamental to all investigations regarding genetic hypotheses is the assumption or the development of the genetic model. In the context of the parametrization of genetic models, some necessary genetic terminology (Thompson 1986) will be introduced. Only binary phenotypes are considered here. Quantitative phenotypes including threshold models creating a binary phenotype from a latent quantitative phenotype will not be considered.

Terminology 7.3.1

The *genome* is the complete collection of an individual's genetic material present in every cell. This material consists of *chromosomes*, i.e. long strands of DNA. A *gene* is a piece of a chromosome coding for a function that can be seen as the inheritable unit. The *locus* is the position of a piece of a chromosome along the chromosome. Thus, the locus might denote the position of e.g. a gene, a gene complex or a marker. The different variants of a gene are called *alleles*. Often the term gene is also used for each single variant of a gene.

The human genome is *diploid*, i.e. chromosomes are all paired *(homologous chromosomes)* with the exception of the sex-linked chromosome in males. Each human somatic cell contains 22 *autosomal* chromosome pairs and 1 pair of sex chromosomes. The autosomal chromosomes of a pair contain the same gene with possibly different alleles at the same gene location. During *meiosis*, a diploid set of chromosomes is reduced to a *haploid* chromosome set of a germ cell, the *gamete*. In this chapter we will exclusively consider the analysis of autosomes.

A pair of alleles of an individual at a locus is called *genotype*. If the two alleles are identical, the individual is called *homozygous* at the locus, otherwise *heterozygous*. Two copies of a gene are called *identical by descent (IBD)* if both copies are the same allele *and* they are copies of the same gene in a common ancestor. An individual is *homozygous by descent (HBD)* when its gene pair is IBD. When considering several loci simultaneously, the multilocus alleles which are inherited from the same parent are called *haplotype*.

7.3.2 Mendelian Single Locus Model

Mendelian segregation (Mendel 1865) is the simplest and most applied model for the *mode of inheritance*. It applies to a single locus. An individual randomly and independently inherits one allele from father and mother respectively. Each parent randomly and independently passes on a copy of one of two alleles to each of his/her offspring (binomial distribution with probability 0.5). All segregation events from parents to offspring are independent. The segregation process implies that copies of some alleles are frequently present in offspring and other alleles are lost in subsequent generations (genetic drift).

Consider the phenotype affected/unaffected of a certain disease. Let S denote a susceptibility gene with n alleles $S_1, S_2, \dots, S_n$. The distribution of allele frequencies $P(S_r)$ in the population is denoted by:

$$p_r = P(S_r) \,, \quad r = 1, \dots, n \,.$$

For *ordered genotypes* the origin of inheritance (father or mother) is distinguished, for *unordered genotypes* not. There are four possible ordered genotypes and three unordered genotypes. Usually unordered genotypes are used. Under Hardy–Weinberg equilibrium (HWE), the (unordered) genotype frequencies are given by

$$P\left(S_r S_s\right) = 2p_r p_s = p_r^2 \quad \text{for } r = s$$

$$P\left(S_r S_s\right) = 2p_r p_s \quad \text{for } r \neq s \,.$$

HWE assumes random mating. Thus the frequencies are yielded by independence of the corresponding allele frequencies, while combining two ordered genotypes for heterozygotes. The maintenance of HWE in a population can be derived by applying Mendelian segregation to each possible parental mating type (see e.g. Khoury et al. 1993).

The *penetrance* describes the relation between genotype and phenotype. It is the conditional probability that an individual with a given genotype will be affected:

$$f_{rs} = P(\text{affected} | S_r S_s) \,, \quad r, s = 1, \dots, n \,.$$

For classical monogenic diseases, the disease is caused by a single major gene. The penetrances of the different genotypes will only take on the values 0 or 1. Often a locus S is assumed to be *biallelic*, i.e. to have only two different alleles. Let S_1 denote the 'susceptibility' allele (mutation) and S_2 the 'normal' allele (wild type). For a classical *dominant disease* all carriers of the susceptibility allele will become affected such that $f_{11} = f_{12} = f_{21} = 1$ and $f_{22} = 0$. For a classical *recessive disease* only homozygous carriers of the susceptibility allele will become affected such that $f_{11} = 1$ and $f_{12} = f_{21} = f_{22} = 0$.

Many classical hereditary diseases follow a Mendelian mode of inheritance. Often the prevalence of classical Mendelian diseases is below 1 in 1000 live births. Prominent examples are Chorea Huntington (autosomal dominant gene, CFTR, on chromosome 4) and cystic fibrosis (autosomal recessive gene, Huntingtin, on chromosome 7). Many different mutations of the gene CFTR (cystic fibrosis transmembrane regulator) cause cystic fibrosis. The gene Huntingtin causing Chorea

Huntington contains a variable number of CAG trinucleotide repeats. This number is low for unaffected individuals and high for those affected. Thus both genes are characterized by allelic heterogeneity. However, the assumption of a biallelic locus with two groups of alleles (susceptibility, normal) worked well in identifying these two genes as causes of a Mendelian hereditary disease, even though it is clear that the true inheritance is much more complicated. The aim in statistical genetics is not to specify a completely correct model in the first place, but to address the scientific question adequately with a parsimoneous mathematical model. If this model is too simple then extended or new biologically motivated models need to be implemented.

The relation of genotype to phenotype is not straightforward for many diseases. Individuals with a susceptibility genotype can stay unaffected (*incomplete penetrance*) and individuals with a non susceptibility genotype can become affected (*phenocopies*). In general terms, given different genotypes, penetrances at a specific gene locus may all be different. It may often be assumed, that the origin (father or mother) of an allele has no influence on a disease, i.e. $f_{12} = f_{21}$. For the general single locus mode of inheritance with susceptibility allele S_1 we assume $1 \geq f_{11} \geq f_{12} = f_{21} \geq f_{22} \geq 0$. For a recessive mode of inheritance we assume $f_{12} = f_{21} = f_{22}$, and for a dominant mode of inheritance we assume $f_{11} = f_{12} = f_{21}$.

Linkage 7.3.3

For the joint inheritance at two loci, it may not generally be assumed that there is independent Mendelian segregation, owing to crossover events and recombinations. Gametes are formed during meiosis. In this process, homologous chromosomes are arranged next to each other and partly overlap. A chromosome breakage and a *crossover* (or *crossing over*), i.e. an exchange between homologous chromosome segments, can occur. A *recombination* between loci A and B occurs when a gamete will have a haplotype other than the combination genes that occurred in the parents, due to crossovers between the loci.

Consider the formation of gametes during meiosis displayed in Fig. 7.1. Between very distant loci A and B (see Fig. 7.1a) a crossover is likely to result in a recombination of the haplotypes A_1B_1 and A_2B_2 to give the new haplotypes A_1B_2 and A_2B_1. If the two loci A and B are very close (see Fig. 7.1b) this is very unlikely. In fact, the *map distance* is defined as the expected number of crossovers between two loci (Haldane 1919). Since the map distance is an expectation, this distance measure is additive. Thus, for three (ordered) loci A, B and C the map distance between A and C is given by the sum of the map distances between A and B and between B and C. The map unit is Morgan, M, named after T.H. Morgan, 1866–1945. Often centiMorgan, cM, are given. The total length of the human autosomal genome is approximately 35 M. Single chromosome lengths are between 0.5 M and 3 M. As a very rough guide 1 cM corresponds to 1 Million base pairs in the physical map.

By genotyping it is possible to observe recombinations between two loci, but not crossovers. Figure 7.1a shows recombination due to a single crossover. For a double

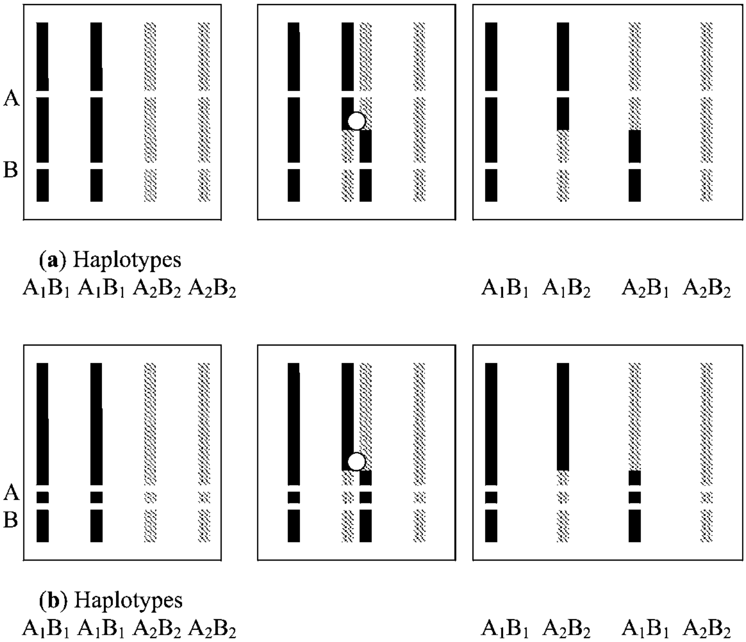
A
B
(a) Haplotypes
A_1B_1 A_1B_1 A_2B_2 A_2B_2
A_1B_1 A_1B_2 A_2B_1 A_2B_2
A
B
(b) Haplotypes
A_1B_1 A_1B_1 A_2B_2 A_2B_2
A_1B_1 A_2B_2 A_1B_1 A_2B_2

Figure 7.1. Formation of gametes during meiosis from one parental pair of chromosomes with a single crossover. *Left*: parental chromosome pair, *middle*: crossover event (crossover point denoted by the *circle*), *right*: gametes for offspring formation. At the two loci *A* and *B* the parent is double heterozygous A_1A_2 and B_1B_2. (a) The crossover occurred between locus *A* and *B*. The *two middle* gametes show recombination. (b) The crossover occurred above locus *A* and *B*, so that the gametes do not show recombination

crossover, i.e. two chromosomal exchanges between the loci *A* and *B*, no recombination would be observed. In mathematical terms a *recombination* between loci *A* and *B* can be defined as an uneven number of crossovers between them.

The *recombination rate* θ, i.e. the ratio of the number of recombinant gametes to the total number of gametes formed, is used as a measure of *genetic distance* between two loci. If loci are on different chromosomes or far away on the same chromosome they segregate independently during the formation of gametes. This results in $\theta = 0.5$, and the loci are designated unlinked. By definition there is *linkage* between the loci if $0 \leq \theta < 0.5$ and no linkage if $\theta = 0.5$. If loci are closer to each other, recombination is less likely. Complete linkage, i.e. complete co-segregation, implies no recombination and thus $\theta = 0$.

In Fig. 7.2 a double heterozygous parent with haplotypes A_1B_1 and A_2B_2, and a double homozygous parent with haplotype A_3B_3 are considered. For the double heterozygous parent a meiosis can create the non-recombinant haplotypes A_1B_1 and A_2B_2 or the recombinant haplotypes A_1B_2 and A_2B_1. In order to determine recombination a parent homozygous even at one locus is not informative. Given that recombination is present, each of the two recombinant haplotypes occurs

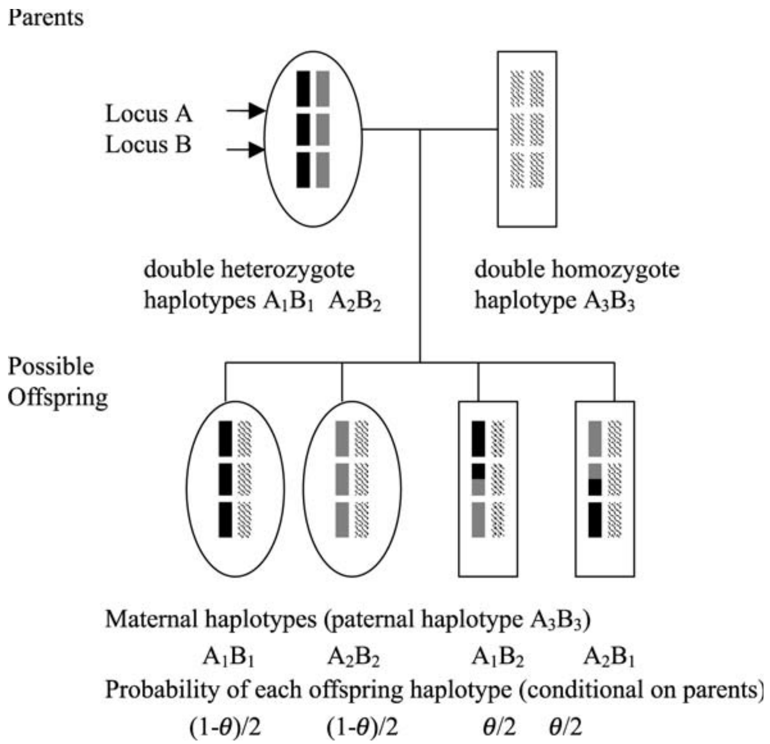
Parents
Locus A
Locus B
double heterozygote
haplotypes A_1B_1 A_2B_2
double homozygote
haplotype A_3B_3
Possible
Offspring
Maternal haplotypes (paternal haplotype A_3B_3)
A_1B_1 A_2B_2 A_1B_2 A_2B_1
Probability of each offspring haplotype (conditional on parents)
$(1-\theta)/2$ $(1-\theta)/2$ $\theta/2$ $\theta/2$

Figure 7.2. Formation of recombinant and non-recombinant haplotypes by meiosis

with probability 0.5. Given that no recombination is present, each of the two non-recombinant haplotypes occurs with probability 0.5. For $\theta = 0.5$ there is independent segregation so that all four possible haplotypes are equally likely.

If the distance between loci is small, i.e. $\theta \leq 0.1$, a recombination corresponds to a crossover. If three ordered close loci A, B and C are considered, $\theta_{AC} \approx \theta_{AB} + \theta_{BC}$. In contrary to the map distance in Morgan, recombination distances are not additive. A recombination between A and B and one between B and C corresponds to an even number of crossovers for the interval A to C and thus will not result in a recombination between A and C. With the help of so-called *mapping functions*, recombination distances can be translated into Morgans. In the majority of chromosomal regions recombination rates for women are higher than for men, which is most often neglected in genetic epidemiological studies.

To fully describe the segregation process in a (chromosomal) pedigree along a complete chromosome, it is sufficient to denote the paternal or maternal origin by 0 or 1 respectively for each meiosis in a pedigree along the whole chromosome (see Fig. 7.3). A crossover event is present at a particular position when the parental origin switches at a particular meiosis.

The potential informativity of a single marker chosen from an existing marker map (without consideration of the disease locus) is determined by its genetic

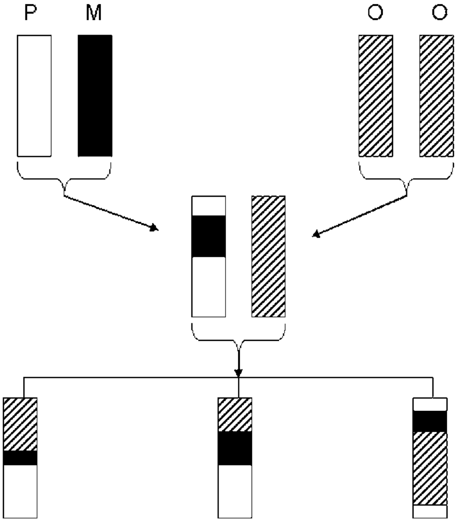

Figure 7.3. Three-generation chromosomal pedigree. The pedigree describes the grandparental inheritance of the haplotypes of three grandchildren which they inherited from the left grandparent. Consider the 'left' grandparent with its two chromosomes denoted by *P* (*white*) for paternal and *M* (*black*) for maternal. The segregation of these two chromosomes from the grandparent to its offspring (with a double crossover) and to its three grandchildren can be followed. For each meiosis the chromosomal segments are yielded by the crossover process. Chromosomal segments in the grandchildren are in part inherited from the chromosomes *O* (*dashed*) of the other grandparent. At each position along the chromosome and for each offspring chromosome, the paternal or maternal inheritance can be denoted by 0 or 1. This is true in the parent generation and in the child generation. Thus, the inheritance can be completely described by a vector of 0's and 1's with dimension equal to the number of meioses considered

variability, i.e. allele distribution, and by the relation of the marker's (laboratory) phenotype to its corresponding genotype. In this laboratory context, phenotype denotes the observed measure of the marker's true underlying genotype. An example of such an observed measure for a genotype is the length determined by a gel electrophoresis for an RFLP marker instead of the exact base sequence. Two measures of marker informativity are *heterozygosity H* and *polymorphism information content PIC* (Botstein et al. 1980). For a locus with n alleles, these are defined as follows:

$$H = \sum_{r \neq s}^{n} p_r p_s \,,$$

$$PIC = 1 - \sum_{r=1}^{n} p_r^2 - \sum_{r=1}^{n-1} \sum_{s=r+1}^{n} 2 p_r^2 p_s^2 = 2 \sum \sum p_r p_s (1 - p_r p_s) \,.$$

In HWE, the heterozygosity H equals the probability that a random individual is heterozygous. *PIC* denotes the probability that one of two randomly mating individuals is heterozygous and the other has a different genotype (Weiss 1993; Ott 1999).

Linkage Disequilibrium

7.3.4

Linkage and linkage disequilibrium (LD) are concepts that need to be distinguished. Linkage describes the co-inheritance at two loci and can only be observed in families. Linkage is independent of the specific alleles. LD describes the relation between alleles at two loci in a population.

Consider the frequencies of specific alleles at two loci S and M in a population. They can be in *linkage disequilibrium* (or gametic disequilibrium, LD). LD is present if the probability for the presence of specific S and M alleles in one gamete is not equal to the product of the individual probabilities at the single loci.

Let S denote the locus with n alleles $S_1, S_2, \dots, S_n$ and allele frequencies

$$p_r = P(S_r) \ , \quad r = 1, \dots, n \ ,$$

and M a locus with m alleles $M_1, M_2, \dots, M_m$ and allele frequencies

$$q_i = P(M_i) \ , \quad i = 1, \dots, m \ .$$

A common measure of LD is the difference of the haplotype probability from its expectation under no association. For two biallelic loci it is denoted by D or δ. For multiallelic markers the parameter δ_{ir} is often used to define the linkage disequilibrium between M_i and S_r as

$$\delta_{ir} = P(M_i S_r) - P(S_r)P(M_i) \ , \quad i = 1, \dots, m \ ; \quad r = 1, \dots, n \ .$$

Linkage disequilibrium or LD is present if $\delta_{ir} \neq 0$ for at least one pair of alleles M_i, S_r.

Linkage equilibrium is present if

$$\delta_{ir} = 0 \quad \text{for all} \quad i = 1, \dots, m \ ; \quad r = 1, \dots, n \ .$$

Under linkage equilibrium the allele distribution at locus M is independent of the specific S allele present.

Linkage disequilibrium can also be described by the coupling frequencies c_{ir} defined as

$$c_{ir} = P(M_i | S_r) \ , \quad r = 1, \dots, n \ ; \quad i = 1, \dots, m \ ,$$

i.e. by the conditional probabilities that a gamete with S_r also has allele M_i. Other parametrizations of LD are also commonly used.

LD may also be considered between multiple loci. Of course the parametrization is more complicated in this case.

Linkage disequilibrium can have different origins (Suarez and Hampe 1994). At linked loci complete LD can be caused by a recent mutation at one locus. However, LD is also possible without linkage between the loci. In extreme cases the loci can even lie on different chromosomes. One important mechanism for the development of LD at unlinked loci is *population stratification*. This is often a result of the admixture of subpopulations (e.g. through immigration) with different allele distributions in the subpopulations. Non-random mating (e.g. by religion or social status) can also be such a cause.

The following formula describes a population genetics model for the degradation in generation time of an existing LD at generation time 0, δ_0, during g generations caused by recombination/linkage (Maynard Smith 1989):

$$\delta_g = (1 - \theta)^g \delta_0 \, .$$

An LD may not necessarily be caused by linkage (possible mechanisms given above), but in the presence of tight linkage it can stay strong during many generations. Without tight linkage LD will degrade rapidly. Thus LD provides indirect evidence for linkage.

As a conlusion of this section two additional widely used measures of LD for fine-scale mapping with biallelic marker data will be introduced (Devlin and Risch 1995).

Consider markers A and B, each with two alleles A_1, A_2 and B_1, B_2. In a haplotype, let the first position denote the allele at marker A, the second position the allele at marker B. The haplotype probabilities are listed in Table 7.1. Rows and columns portray the marginal probabilities. The LD as the difference of the haplotype probability from its expectation under no association can be calculated by

$$D = \pi_{11} - \pi_{1+}\pi_{+1} = \pi_{22} - \pi_{2+}\pi_{+2} = \pi_{11}\pi_{22} - \pi_{21}\pi_{12} \, .$$

D is an absolute measure of LD. Its value is 0 if marker A and marker B are not associated.

Table 7.1. Haplotype probabilities for two biallelic markers A and B

	Marker B		
Marker A	Allele B$_1$	Allele B$_2$	
Allele A$_1$	π_{11}	π_{12}	π_{1+}
Allele A$_2$	π_{21}	π_{22}	π_{2+}
	π_{+1}	π_{+2}	

Maximum and minimum possible values of D depend on the allele frequencies in the population. Thus, define $D_{\max}$ and $D_{\min}$ by

$$D_{\max} = \min(\pi_{1+}\pi_{+2}, \pi_{+1}\pi_{2+})$$

$$D_{\min} = \min(\pi_{1+}\pi_{+1}, \pi_{+2}\pi_{2+}) \, .$$

Rescaling D relative to its maximum and minimum results in the relative measure D' *(Lewontin's D', Lewontin 1964)*:

$$D' = \begin{cases} \dfrac{D}{D_{\max}} = \dfrac{\pi_{11}\pi_{22} - \pi_{21}\pi_{12}}{\min\left(\pi_{1+}\pi_{+2}, \pi_{+1}\pi_{2+}\right)} & D > 0 \\ \dfrac{D}{D_{\min}} = \dfrac{\pi_{11}\pi_{22} - \pi_{21}\pi_{12}}{\min\left(\pi_{1+}\pi_{+1}, \pi_{+2}\pi_{2+}\right)} & D < 0 \,. \end{cases}$$

Since in essence LD is a correlation between the marker and the susceptibility gene in populations, the correlation coefficient can also be used as an LD measure (Hill and Robertson 1968), denoted by Δ:

$$\Delta = \frac{\pi_{11}\pi_{22} - \pi_{21}\pi_{12}}{\sqrt{(\pi_{1+}\pi_{2+}\pi_{+1}\pi_{+2})}} = \frac{D}{\sqrt{(\pi_{1+}\pi_{2+}\pi_{+1}\pi_{+2})}} \,.$$

Another LD measure based on odds ratios and motivated by the epidemiological measure 'attributable risk' (cf. Chap. I.2 of this handbook) is

$$\delta = \frac{\pi_{11}/\pi_{21} - \pi_{12}/\pi_{22}}{\pi_{11}/\pi_{21} + 1} = \frac{\pi_{11}\pi_{22} - \pi_{12}\pi_{21}}{\pi_{+1}\pi_{22}} = \frac{D}{\pi_{+1}\pi_{22}} \,.$$

Segregation Analysis

7.4

The aim of *segregation analysis* is to find evidence for the existence of a major gene for the phenotype under investigation and to estimate the corresponding mode of inheritance. Therefore, if possible, the pattern of inheritance over several generations within the structure of larger families is investigated. Sometimes segregation analysis has to be carried out on the basis of many small families.

Consider a Mendelian single locus model for a major gene with the susceptibility allele S_1 and the normal allele S_2. For classical Mendelian diseases the penetrances $P(\text{affected}|\text{genotype})$ take on only the values 0 and 1. For such diseases, the genotype-phenotype relation is so obvious that the discrete genotype translates into a discrete disease phenotype. Thus, families in which such a Mendelian disease gene segregates display very characteristic disease patterns. For example, in the case of autosomal dominant diseases generations should not be skipped by the disease, an affected individual married to an unaffected individual should produce an approximate $1:1$ ratio of affected to unaffected offspring and the distribution of the trait among sexes should be almost equal (Tamarin 1986). Ratios like the above mentioned are called *segregation ratios*.

The simplest types of segregation analyses are based on tests for segregation ratios hypothesising a particular mode of inheritance. To illustrate the principle, consider first a rare autosomal dominant disease and a random sample of matings. Matings between an affected and an unaffected individual will usually be of the $S_1S_2 \times S_2S_2$ *mating type*. Since the susceptibility allele is rare,

$S_1S_1 \times S_2S_2$ matings for affected-unaffected couples can be neglected as a first approximation. Thus, for the moment consider only $S_1S_2 \times S_2S_2$ families and suppose that r of n offspring are affected. In the offspring generation the expected segregation ratio between affected and unaffected is 0.5. A test can be built on the binomial distribution, considering n as the number of trials, q as the probability for a single child to be affected, and r as the observed number of affected children. If the null hypothesis of $q = 0.5$ is not rejected, it may be concluded that the data are compatible (more precisely not inconsistent) with an autosomal dominant disease pattern. For further test procedures see e.g. Sham (1998).

More generally, the probability distribution of the six possible mating types ($S_1S_1 \times S_1S_1$; $S_1S_1 \times S_1S_2$; $S_1S_1 \times S_2S_2$; $S_1S_2 \times S_1S_2$; $S_1S_2 \times S_2S_2$; $S_2S_2 \times S_2S_2$) can be formed according to the parental genotypes. For each given mating type, the distribution of genotypes and phenotypes in the offspring may be determined, on which tests can then be built. However, families are most often sampled according to recruitment criteria and not randomly, yielding an oversampling of families enriched for disease. Therefore, for a test procedure to be valid the probability distributions need to be corrected for *ascertainment bias*. Consider again the binomial distribution for the number of affected offspring in a sibship of a particular mating type. We assume ascertainment for families with 'at least one affected offspring'. The binomial distribution for the number of affected offspring could be corrected for ascertainment by considering a truncated binomial distribution assuming at least one affected offspring per family. Unfortunately, the ascertainment process could imply that families with more affected children have a higher chance of being part of the sample. Proper ascertainment correction maybe complicated and the ascertainment criteria should be known as precisely as possible. Moreover, mathematical assumptions must be made about the ascertainment sampling process in order to estimate genetic parameters or to test genetic models. In general, misspecification of the ascertainment process might cause serious bias in the estimation of genetic parameters (see e.g. Shute and Ewens 1988).

The ascertainment process is often parametrized by the *ascertainment probability*

$$\pi = P(\text{proband}|\text{affected}) ,$$

i.e. the probability that an individual will be part of the family data set given that the individual is affected. The following types of selection have been defined by the parameter π: If $\pi = 1$, this is called *truncate selection*. For truncate selection an individual will be recruited if he/she is affected. Families without affected members are not recruited. If $\pi \to 0$, we speak of *single selection*. In single selection families with r affected children are recruited with probability $r\pi$ and almost all families have only one affected child. Multiple selection is defined by $0 < \pi < 1$.

For *extended pedigrees* with many individuals and several generations a numerical procedure is needed for all probability calculations. Let L denote the likelihood for the observed phenotypes Y, given a genetic model M and the pedigree structure. L can be calculated by summing over all possible genotypic constellations $g_i, i = 1, \dots, N$, where N denotes the number of individuals in the pedigree:

$$L(Y) = \sum_{g_1} \sum_{g_2} \cdots \sum_{g_N} P(Y|g_1 g_2 \cdots g_N) P(g_1 g_2 \cdots g_N) \ .$$

It is assumed that the phenotype of an individual is independent of the other pedigree members given its genotype.

Widely used in segregation analysis is the Elston–Stuart algorithm (Elston and Stuart 1971), a recursive formula for the computation of the likelihood L given as

$$L = \sum_{g_1} \sum_{g_2} \cdots \sum_{g_N} \prod_{j=1}^{N} f(g_j) \prod_{k=1}^{N_1} P(g_k) \prod_{m=1}^{N_2} \tau(g_m|g_{m1} g_{m2}) \ .$$

The notation for the formula is as follows: N denotes the number of individuals in the pedigree. N_1 denotes the number of *founder* individuals in the pedigree. Founders are individuals without specified parents in the pedigree. In general, these are the members of the oldest generation and married-in spouses. N_2 denotes the number of *non-founder* individuals in the pedigree, such that $N = N_1 + N_2$. $g_i, i = 1, \dots, N$, denote the genotype of the ith individual of the pedigree. The parameters of the genetic model M fall into three groups: (1) The genotype distribution $P(g_k), k = 1, \dots, N_1$, for the founders is determined by population parameters and often Hardy–Weinberg equilibrium is assumed. (2) The transmission probabilities for the transmission from parents to offspring $\tau(g_m|g_{m1}, g_{m2})$, where m_1 and m_2 are the parents of m, are needed for all non-founders in the pedigree. It is assumed that transmissions to different offspring are independent given the parental genotypes and that transmissions of one parent to an offspring are independent of the transmission of the other parent. Thus, transmission probabilities can be parametrized by the product of the individual transmissions. Under Mendelian segregation the transmission probabilities for parental transmission are $\tau(S_1|S_1S_1) = 1$; $\tau(S_1|S_1S_2) = 0.5$ *and* $\tau(S_1|S_2S_2) = 0$. (3) The penetrances $f(g_i), i = 1, \dots, N$, parametrise the genotype-phenotype correlation for each individual i.

This recursive formula works well on *simple pedigrees* of arbitrary size. Computations on *complex pedigrees*, i.e. pedigrees with marriage and inbreeding loops (such as consanguineous marriages) are often only possible with approximation methods.

Segregation analysis is a successful tool for monogenic diseases. Major problems in segregation analysis for complex diseases result in essence from the fact that the relationship of genotype to phenotype is not a straightforward 1 to 1 function or n to 1 function, i.e. the genotype does not unambiguously determine the phenotype. This unclear relationship is such a critical issue that some use this as a definition of complex diseases. Further, several genetic factors are assumed to have an influence on complex diseases. The penetrance can be incomplete and phenocopies can exist. In addition the penetrance can depend on other non-genetic factors such as age, gender and exposure factors for example.

For the definition of the phenotype, problems arise when specifying and applying diagnostic criteria. In addition, many complex diseases show a large phenotypic variation, which might be characterised by severity, by different diseases or by different co-occurrence of diseases.

Genetic heterogeneity is a further problem (Evans and Harris 1992). The same phenotype can be caused by different genes (locus heterogeneity). Different phenotypes can be caused by different alleles at the same locus (allelic heterogeneity). Owing to modifying factors different phenotypes can segregate within a family (intra-family heterogeneity). When different phenotypes segregate in different families, but the phenotype is constant within one family, this might indicate that a gene segregates in one family and not in the other (inter-family heterogeneity). In addition, there are further types of genetic heterogeneity such as genomic imprinting, where the penetrance of a heterozygous genotype depends on the (paternal oder maternal) origin of the susceptibility allele.

In the presence of heterogeneity the formation of homogeneous subgroups is a means to arrive at a clearer genotype-phenotype relation und thus, to identify a possible Mendelian subform of the disease. Homogeneous subgroups can be defined by e.g. clinical phenotypes, severity of the disease, age of onset of the disease, family history or ethnicity.

An example of a highly successful segregation analysis for a complex disease is breast cancer (Newman et al. 1988). The families were ascertained through a population-based large epidemiological programme in San Francisco and Detroit. The ascertainment criteria for index cases were women with breast cancer, Caucasian, diagnosis before the age of 55, histologically confirmed primary tumour, becoming incident in a specified period. No selection on positive family history was taken. The personal interview of the index case on her nuclear family, i.e. mothers and sisters, regarding breast cancer was considered sufficiently reliable. 1579 nuclear families were recruited and one large extended pedigree. This sample also included some rare cases of male breast cancer.

Complex segregation analysis was applied to these breast cancer families using the programme POINTER which is based on the so-called ‘unified’ model (Lalouel et al. 1984). This model is called ‘unified’, since it unifies the so-called ‘mixed’ model (Morton and MacLean 1974) and the concept of transmission probabilities mentioned above. The Mendelian ‘mixed’ model assumes an underlying normal distribution for each of the three genotypes S_1S_1, S_1S_2, S_2S_2, of the major factor which differ in their means. The disease status for breast cancer is considered as resulting from an underlying quantitative trait (as a mixture of the three normal distributions) by exceeding a certain threshold. In this threshold model, individuals affected with breast cancer have exceeded the threshold deterministic for disease, individuals not affected by breast cancer have a value for the quantitative trait below the threshold. Thus, for a predisposing genotype the mean is shifted in comparison to the distribution for the wildtype heterozygotes such that more individuals will exceed the threshold. In addition, this model assumes an additive effect of the major factor, a polygenic component and an environmental component. The following parameters have to be estimated for the mixed model: the allele frequency p of the susceptibility allele S_1, the displacement between the means for S_1S_1 and S_2S_2, the dominance parameter defined as the difference in the means for S_2S_2 and S_1S_2 and the heritability H, defined as the proportion of variance due to the polygenic component (see Sect. 7.2). Further parameters for the unified model are the transmission probabilities. In this case, the values 1, 0.5 and 0 for a Mendelian major gene as major factor should not be rejected. Evaluation of models with direct modelling (and estimation) of the transmission probabilities allows the identification of the major factor as a major Mendelian gene. In the above mentioned breast cancer study transmission probabilites were estimated close to the values required by Mendelian segregation and a model without a Mendelian inheritance factor could be rejected.

Several parameters need to be prespecified (as input parameters) for complex segregation analysis. For the breast cancer families the ascertainment probability was assumed within the bounds $\pi = 0.01–0.27$, the lower bound corresponding to almost single ascertainment and the upper bound corresponding to the mean proportion of breast cancer cases in the families. Liability classes for the population based liabilities were estimated from cumulative incidences in the general population of the regions under investigation. These are 0.0010 for women until age 15 and all men regardless of age, 0.0045 for women aged 16–40 years, 0.0283 for women aged 41–55 years and 0.0819 for women older than 55 years. These necessary parameters could be well estimated, since a large epidemiological study had been carried out in the region.

Complex segregation analysis requires many likelihood ratio comparisons between different assumed models for the estimation of parameters and the acceptance of a most parsimonious model. In the breast cancer study example the autosomal dominant major gene was postulated, since the general single locus model with three penetrance parameters and the dominant single locus model with only two penetrance parameters resulted in a comparable fit. In a first step we investigate whether the data are consistent with a major gene model and in a second step we consider with which mode of inheritance for the major gene the data are consistent. Important for the avoidance of false-positive results is the investigation into whether an identified major factor is really Mendelian by the use of transmission probabilities. In the breast cancer family data evidence for an autosomal dominant transmission was given both in the 1579 nuclear families and the one large extended family. These results were supported by the same qualitative results even under sensitivity analysis for the ascertainment probability and by the well-defined liability parameters based on prior studies. Thus, segregation analysis may be successful even for complex diseases. In the breast cancer example an autosomal dominant rare gene with high penetrance could be postulated for early onset breast cancer as a result of the segregation analysis.

If there are no major genes with high penetrances, but only a few genes with a moderate effect on the disease, segregation analysis will not be a valuable tool. It should be mentioned that many diseases are studied nowadays by linkage and association analyses without segregation analyses which normally would have been carried out prior to this.

7.5 Linkage Analysis

In *linkage analysis* the co-segregation between marker and disease is investigated in related individuals. The aim is to find evidence for linkage and often to estimate the recombination rate. Sometimes exclusion of linkage is possible.

The classical linkage analysis method is the *lod score method* (Morton 1955). This is a test for linkage between a susceptibility gene locus and a marker locus (null hypothesis $H_0 : \theta = 0.5$ versus alternative $H_1 : \theta < 0.5$) in combination with the estimation of the recombination rate. For a detailed description see Ott (1999).

For the lod score method, the mode of inheritance M_0, that is the parameter of the genetic model at the susceptibility locus, and the marker allele distribution, have to be known. The mode of inheritance may be estimated by segregation analysis. Let $L(\theta, M_0)$ denote the likelihood for the observed phenotypes at a particular value for θ conditional on M_0, on the marker allele distribution and on the given pedigrees. As in the usual notation, the underlying conditioning is sometimes left out. The lod score function ('log odds') is the log likelihood ratio

$$Z(\theta) = LOD(\theta) = \log_{10} \frac{L(\theta, M_0)}{L(0.5, M_0)}$$

as a function of θ. $Z(\theta)$ compares the likelihood under linkage with recombination rate θ with the likelihood under no linkage, i.e. $\theta = 0.5$. $Z(\theta)$ will be maximized over all possible values for θ, i.e. $0 \leq \theta \leq 0.5$. If $Z_{\max} > 3$ then evidence for linkage exists. The recombination rate will be estimated by $\theta_{\max}$, the θ-value corresponding to $Z_{\max}$. If $Z_{\max} < -2$ linkage can be excluded. The limits 3 and -2 are based on a sequential Wald test, such that the a posteriori probability for linkage when rejecting the null hypothesis is 95% for a single alternative θ (Morton 1955). As logarithms of base 10 are used, the limits correspond to stopping limits of 1000 and 0.01 in the sequential testing procedure yielded by setting $\alpha = 0.001, \beta = 0.01$.

Let us determine the likelihood $L(\theta)$ for linkage between two loci A and B for a sibship of size n. The genotypes are observed directly. Thus, no underlying genetic model needs to be considered. The genotypes of the mother are A_1A_2 and B_1B_2 and the genotypes of the father are A_1A_1 and B_1B_1. Only the double heterozygous mother is informative for linkage. In general, it is not known which allele combinations of the mother are the result of the grandpaternal and the grandmaternal meiosis, i.e. which allele combinations form the haplotypes in the grandparental gametes. The so-called *phase* for the mother could be either composed of the haplotypes A_1B_1 and A_2B_2 (phase I with probability P_{I}) or by the haplotypes A_1B_2 and A_2B_1 (phase II with probability P_{II}).

Assume that the phase is known to be phase I, for example when the grandparents pass on this information. Let n_x and n_y denote the number of meioses from the mother to the n children, which are non-recombinants n_x or recombinants n_y, respectively. Then the likelihood $L(\theta)$ is

$$L(\theta) = \binom{n_x + n_y}{n_x} (1 - \theta)^{n_x} \theta^{n_y} .$$

For an unknown phase, phase I and phase II both have to be considered. If P_{II} is the true phase, then n_x children show recombination and n_y children show non-recombination. Under the assumption of linkage equilibrium phase I and phase II are both equally likely. Thus

$$L(\theta) = \binom{n_x + n_y}{n_x} \left[P_{\mathrm{I}}(1-\theta)^{n_x}\theta^{n_y} + P_{\mathrm{II}}\theta^{n_x}(1-\theta)^{n_y}\right]$$

$$= \binom{n_x + n_y}{n_x} \left[\frac{1}{2}(1-\theta)^{n_x}\theta^{n_y} + \frac{1}{2}\theta^{n_x}(1-\theta)^{n_y}\right] .$$

For the sibship in Fig. 7.4 let us now determine the likelihood $L(\theta)$, the lod score function $Z(\theta)$, $Z_{\max}$ and $\theta_{\max}$. The notation in this pedigree is motivated by an autosomomal dominant susceptibility gene S with a rare susceptibility allele S_1 and a normal allele S_2. Thus, the affected father and all affected siblings have genotype S_1S_2. In the pedigree the marker M is segregating with three alleles M_1, M_2 and M_3. The mother of the sibship of size 6 is homozygous and thus uninformative for linkage. She will not be considered further.

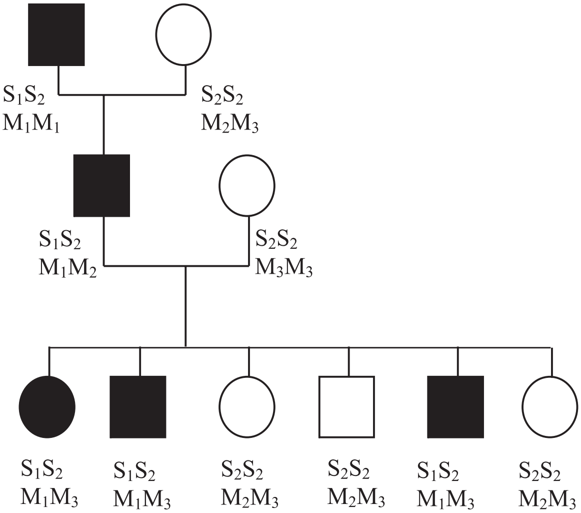

Figure 7.4. Pedigree with a sibship of size 6 with marker information and with genotype information concerning the susceptibility locus, owing to the clear-cut rare autosomal dominant mode of inheritance

As a result of the genotyped grandparents, the father's haplotypes are known: S_1M_1 and S_2M_2. Thus the phase is known and the likelihood is

$$L(\theta) = \binom{6}{0} (1-\theta)^6\theta^0 = (1-\theta)^6 .$$

The lod score function is

$$Z(\theta) = LOD(\theta) = \log_{10} \frac{L(\theta)}{L(0.5)} = \log_{10} \frac{(1-\theta)^6}{(0.5)^6}$$
$$= 6\log_{10}(1-\theta) + 6\log_{10} 2$$
$$= 6\log_{10}(1-\theta) + C\,,$$

where C denotes a constant independent of θ. The maximum of the lod score function is $Z_{\max} = 1.8$ for $\theta_{\max} = 0$. This corresponds to complete linkage as supported by no observed recombinations.

Missing information on grandparental genotypes in Fig. 7.4 results in an unknown phase. Then the lod score function would be

$$Z(\theta) = LOD(\theta) = \log_{10} \frac{L(\theta)}{L(0.5)} = \log_{10} \frac{0.5\theta^6 + 0.5(1-\theta)^6}{(0.5)^6}$$
$$= \log_{10}\left(\theta^6 + (1-\theta)^6\right) + 5\log_{10} 2\,.$$

In this case, the maximum of the lod score function is $Z_{\max} = 1.5$ for $\theta_{\max} = 0$. Due to the uncertain phase, the maximum lod score is reduced. However, the estimate for the recombination rate stays at $\theta = 0$.

In Fig. 7.4, assume now that the second affected child has the genotype M_2M_3 (and the genotype S_1S_2). With the phase as indicated in the figure, one recombination needs to be taken into account now. Thus

$$Z(\theta) = LOD(\theta) = \log_{10} \frac{L(\theta)}{L(0.5)} = \log_{10} \frac{6\theta(1-\theta)^5}{6(0.5)^6}$$
$$= \log_{10}\theta + 5\log_{10}(1-\theta) + 6\log_{10} 2\,.$$

With one recombination the maximum of the lod score function is $Z_{\max} = 0.63$ for $\theta_{\max} = 1/6 = 0.17$. Now linkage is estimated as not complete and $Z_{\max}$ is markedly reduced.

If in Fig. 7.4 the genotypes of the father and his parents are unknown, the father's genotype can be inferred as either M_1M_1 or M_1M_2. If HWE can be assumed, the likelihood of the recombination rate $L(\theta)$ can be calculated as a function of the marker allele frequencies in offspring. A detailed calculation will show that in this case, a rare marker allele M_1 will result in a high lod score, a more common marker allele M_1 will result in a lower lod score.

The likelihood $L(\theta)$ can be computed for more complex pedigrees with the help of the Elston–Stuart algorithm (Elston and Stuart 1971):

$$L = \sum_{g_1}\sum_{g_2}\cdots\sum_{g_N}\prod_{j=1}^{N} f(g_j)\prod_{k=1}^{N_1} P(g_k)\prod_{m=1}^{N_2} \tau(g_m|g_{m1}g_{m2}\theta)\,.$$

The notation is provided in the previous section with the extension that $g_j, j = 1,\ldots,N$, now refers to the haplo-genotypes, i.e. to the genotypes formed by the haplotypes of the underlying susceptibility locus S and the marker M. The recombination rate θ is now part of the transmission probabilities of the haplo-genotypes, since they describe the formation of gametes as recombinants or non-recombinants.

Consider now a set of K families. Since the segregation process is independent in different families, the total likelihood is given by the product of the individual likelihoods. Thus, the logarithm of the total likelihood, $l(\theta)$, is given by the sum of the individual logarithms of the likelihood $l_i(\theta)$:

$$l(\theta) = \sum_{i=1}^{K} l_i(\theta) \ .$$

The lod score method and its extensions have been very successful in localizing major susceptibility genes, especially for rare monogenic diseases (e.g. cystic fibrosis). However, the analysis of complex diseases poses many difficulties (Lander and Schork 1994). Often the mode of inheritance is unclear. Hence, the preassumption of parameters for the mode of inheritance in the lod score analysis is very critical. Often several modes of inheritance are 'tried out' (Terwilliger and Ott 1994). A false model can lead to false negative tests, and thus the erroneous exclusion of chromosomal regions, which indeed harbour susceptibility loci. The use of *LOD* scores for exclusion mapping should be considered with caution. Maximizing *LOD* scores over several models or the whole range of recombination rates increases the *a posteriori* false positive rate (Risch 1991), i.e. linkage is inferred erroneously. False assumptions of marker allele frequencies can also lead to false positive results (Ott 1999). Where possible, marker allele frequencies should be estimated for a given study or population, since allele frequencies are often not well known and may differ from population to population. The estimation of the recombination fraction itself after concluding for linkage can be biased, i.e. the true location of the susceptibility gene might be many centi Morgans apart from the estimated location. Even a combined segregation and linkage analyses with parallel estimation of the necessary parameters does not lead to meaningful and significant results, owing to the flatness of the likelihood function.

The localization of the BRCA1 gene for breast cancer is an example of a successful lod score analysis (Hall et al. 1990), which was based on the segregation analysis described in the previous section (Newman et al. 1988). In this analysis, cumulative *LOD* scores were calculated by ascending average age-of-onset for breast cancer cases in the families. By this procedure, linkage could be demonstrated for early-onset families.

The difficulties in employing a parametric linkage analysis become more and more important with the degree of complexity of a disease. In order to avoid the necessity of critical assumptions about the underlying genetic model, so-called *non-parametric methods* or *model-free methods* have been developed. The aim of model-free methods is to provide evidence for linkage without specifying parameters of the underlying mode of inheritance and without estimating the recombination rate (Elston 1998; Lander and Schork 1994).

Many of these methods are based on the *identity-by-descent (IBD)* status. For example consider a patient and one of his/her siblings (Penrose 1953). Their *IBD* status can take on the values 0, 1 or 2, according to the number of marker alleles that have been transmitted to both siblings from exactly the same (grandpaternal or grandmaternal) copy of a parent's gene and are thus identical (see Fig. 7.5).

Allele sharing methods are based on the fact, that in the presence of linkage relatives with a similar disease status (e.g. both affected) are more frequently similar at the marker locus – in the sense of *IBD* – than to be expected under independent segregation. Relatives with a different disease status (e.g. discordant sibs: affected, unaffected) are less frequently similar at the marker than under no linkage. The aim of these methods is to provide evidence for linkage and not to estimate the recombination rate.

In the *affected-sib-pair* (ASP) *method* (Day and Simons 1976), affected sib pairs are classified according to the *IBD* status. The classical χ^2-test compares the observed number of (independent) sibling pairs with 0, 1 or 2 marker alleles *IBD* with the expected number assuming independent segregation. If marker and disease locus are unlinked, the probability for 0, 1, or 2 marker alleles *IBD* is 0.25, 0.5 or 0.25, respectively. If the observed *IBD* distribution significantly differs from the expected distribution, this indicates linkage. It is possible to use the χ^2-goodness-of-fit-test to test for a hypothesised genetic model, taking the derived numbers under the model as expected.

When considering only affected individuals, the method is robust towards incomplete penetrance. Other allele sharing methods also incorporate unaffected individuals in the analyses as well as different pairs of relatives other than siblings. The literature is extensive and more powerful methods than the original ASP method have been developed (e.g. Holmans 1993; Whittemore and Tu 1998).

The determination of the *IBD* status assumes that the marker is sufficiently polymorphic and that the parents are genotyped for the marker (or neighbouring loci and other relatives yield the missing information). If it is not possible to determine *IBD* unambiguously (see Fig. 7.5) it needs to be estimated. Sometimes the IBS status is used instead. IBS is the number of marker alleles that are identical in the pair of individuals (*"identity by state"*) without considering ancestry, taking

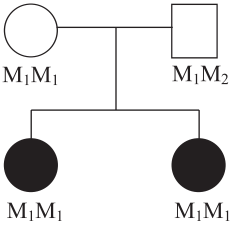

Figure 7.5. An affected sib-pair with parents. Marker genotypes are given. The IBS status is 2. The *IBD* status cannot be unambiguously determined, $P(IBD = 1) = P(IBD = 2) = 0.5$, since the mother transmits the grandpaternal allele M_1 or the grandmaternal allele M_1 with equal probability

on the values 0, 1, or 2. However, such methods are not robust towards imprecisions in the marker allele frequencies.

Association Analysis

The aim of *association studies* is to show evidence for association or linkage disequilibrium in a population. Linkage disequilibrium results in an association between marker alleles and alleles of a susceptibility gene, such that certain marker alleles will be present more often in affected individuals than in a random sample of individuals from the population.

In classic *case-control studies* marker allele frequencies or genotype frequencies in a group of unrelated affected individuals are compared to those in a group of unrelated unaffected individuals. Numerous associations have been identified with case-control studies, e.g. associations of autoimmune diseases (e.g. diabetes, multiple sclerosis) with the HLA system. A further example is the association of apolipoprotein E (APOE) allele $\varepsilon4$ with Alzheimer's disease (Corder et al. 1993). The APOE $\varepsilon4$ allele frequency is approximately 35% in Alzheimer's patients, but only approximately 15% in the older population not suffering from dementia. If a positively associated marker allele is frequent in a population, such as APOE $\varepsilon4$, then it is by itself not a good predictor for disease status and the proportion of homozygotes for the allele is high. Linkage analysis methods are in general not very powerful in this situation.

Besides the usual limitations of classical case-control studies in epidemiology (cf. Chap. I.6 of this handbook), case-control studies to investigate linkage disequilibrium in genetic epidemiology must take a particular form of confounding into account, i.e. population stratification: cases and controls must originate from the same homogeneous (including ethnically homogeneous) source population. This is especially difficult to achieve, to assess or test for in genetic epidemiology. If individuals stem from subpopulations with different allele frequencies, and this is not taken into account, then linkage disequilibrium can be simulated. This means that stratified populations can evoke linkage disequilibrium without linkage. The detection of such linkage disequilibrium detracts from the identification of susceptibility genes. It must be considered as an annoyance to this aim and as such be evaluated as false positive. The technical term for such false positive results is *spurious association*.

If an association is found that is not considered spurious, this may have two causes (Lander and Schork 1994):

- The positively (or negatively) associated allele is the susceptibility allele itself. If so, this association is expected to occur in all populations harbouring this allele.
- The positively (or negatively) associated allele is in linkage disequilibrium with the susceptibility allele at the disease locus. If this is the case, then different associations can occur in different populations due to different haplotype frequencies of the allele combinations of both marker and susceptibility locus.

In the first case, both marker and disease locus are identical. Thus, $\theta = 0$ and linkage disequilibrium is complete. In the second case marker and disease locus are in general very close to each other. For this reason association studies are highly valuable for the investigation of candidate genes.

When an association is identified, it can be quantified with the help of odds ratios (Woolf 1955) (see also Chap. I.2 of this handbook). With rare diseases the odds ratio approximates the relative risk. In addition, parameters of the underlying genetic model may be estimated e.g. with the likelihood ratio method (Thomson 1983; Risch 1983).

As mentioned above, uncontrolled stratification of populations may result in spurious associations. For case-control studies, there are essentially two methods for taking the existence of subpopulations into account during statistical testing. Both methods require a set of additional markers along the genome to be genotyped. In the *genomic control* method (Devlin and Roeder 1999) a variance inflation factor is used to adjust the test statistic, taking into account correlations between individuals in subpopulations. The *structured association* method (Pritchard et al. 2000) initially aims at directly identifying the population structure and assigning individuals to a subgroup. Association is subsequently investigated by testing against the null hypothesis of independent association within subgroups.

Association studies with internal controls, or so-called *family based association studies*, are intended by design to avoid possible bias through inadequate controls and population stratification. The concept of *internal controls* was developed by Falk and Rubinstein (1987). For the original design nuclear families with at least one affected child have to be recruited. The two parental alleles not transmitted to the affected child are used as internal controls (Fig. 7.6).

This design has the important property that information is used on both, linkage and association between a marker and the susceptibility gene.

For a biallelic marker, the data resulting from this study design may be presented in different ways in a 2 × 2 contingency table (Table 7.2) and analysed with statistical tests (Terwilliger and Ott 1992; Schaid and Sommer 1994). All test procedures test for association ($H_0 : \delta = 0$ vs. $H_1 : \delta \neq 0$) and most for linkage as well. In principal, the tests investigate whether certain alleles are transmitted from the parents to an affected child more often than not.

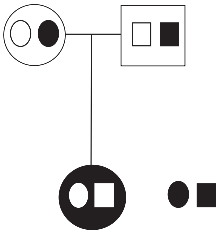

Figure 7.6. Nuclear family with one affected child. Alleles transmitted from the parents to the affected children are denoted in *white*. Alleles not transmitted from the parents to the affected child are denoted in *black*

The test procedures can be classified according to several criteria. Tests can be separated on the basis of whether genotypes or haplotypes are analysed. Haplotype-based analyses compare the transmitted allele with the non-transmitted allele. Genotype-based analyses compare the genotype transmitted to the affected child with the artifical genotype constructed from the two non-transmitted alleles. Another criterion is whether the tests are procedures for matched samples or not. The samples are indeed matched since one transmitted and one non-transmitted allele together describe the segregation from a single parent to an offspring.

Let M denote a biallelic marker with alleles M_1 and M_2; let M_1 be positively associated with the disease. Let genotypes, which are homozygous or heterozygous for M_1, be denoted with M_1 positive. Then, N families with an affected child can be presented in a 2×2 table as in Table 7.2. The traditional χ^2-test for independence can be applied to unmatched samples (Tables 7.2b and 7.2d) and the McNemar-test can be applied to matched samples (Tables 7.2a and 7.2c). The table for unmatched samples uses the marginal table of the corresponding table for matched samples.

The original *Haplotype Relative Risk* (HRR) *method* (Falk and Rubinstein 1987) is a genotype-based analysis for unmatched samples. The odds ratio in Table 7.2b, also called Haplotype Relative Risk (HRR), is a measure of association that is never more extreme, i.e. farther away from 1, than the estimator RR for the relative risk in a classic non-family based approach (Knapp et al. 1993). For $\theta = 0$, HRR = RR.

The most commonly used test is the *Transmission/Disequilibrium Test* (TDT) (Spielman et al. 1993). The TDT is a haplotype-based analysis of the matched sample (Table 7.2c). The test statistic is

$$TDT = (b - c)^2/(b + c) .$$

The TDT compares whether the M_1 allele is more often transmitted to an affected child (b) than the M_2 allele (c) from heterozygous parents, or visa versa. The test only considers M_1M_2 parents, since homozygous parents are not informative for preferential transmission of either allele. This separation can be made due to the matched analysis. In addition matching reduces the variance of the test statistic, thereby yielding a higher power.

The literature on family-based association analysis is vast (see e.g. Whittaker and Morris 2001). Important extensions of the above methods allow the application to multiallelic markers, to tightly linked loci and to quantitative traits. In addition, the design also allows for other types of nuclear families, such as sibships with affected and unaffected individuals (Spielmann and Ewens 1998). If a particular mode of inheritance is suspected specialized versions of the TDT or related likelihood methods may yield higher power (Schaid 1999).

If a candidate gene is to be investigated in detail, then a haplotype analysis will be carried out considering several biallelic polymorphisms (SNPs) in the same gene. The first step in a haplotype analysis is the estimation of the haplotype frequency in a population or the estimation of the most probable haplogenotype (haplotype pair) in an individual. For cases and controls see Excoffier

Table 7.2. 2 × 2 contingency table for family-based association methods based on N families with one affected child and both parents. Consider a biallelic marker with alleles M_1, M_2 and let M_1 be positively associated with the disease. Genotypes, which are homozygous or heterozygous for M_1, are denoted with M_1 positive. Capital letters denote genotype counts, small letters denote allele counts. A, B, C and D are defined as in Table 2a. a, b, c and d are defined as in Table 2c. N and $2N$ denote the total number of transmitted genotypes and alleles, respectively, to the affected child from the $4N$ parental alleles

(**a**) Genotype-based analysis for matched samples

	Non-transmitted genotype		
Transmitted genotype	M_1 positive	M_1 negative	Total
M_1 positive	A	B	$A + B$
M_1 negative	C	D	$C + D$
Total	$A + C$	$B + D$	N

(**b**) Genotype-based analysis for unmatched samples (HRR method)

	M_1 positive	M_1 negative	Total
Transmitted genotype	$A + B$	$C + D$	N
Non-transmitted genotype	$A + C$	$B + D$	N
Total	$2A + B + C$	$2D + C + B$	$2N$

(**c**) Haplotype-based analysis for matched samples (TDT)

	Non-transmitted allele		
Transmitted allele	M_1	M_2	Total
M_1	a	b	$a + b$
M_2	c	d	$c + d$
Total	$a + c$	$b + d$	$2N$

(**d**) Haplotype-based analysis for unmatched samples

	M_1	M_2	Total
Transmitted allele	$a + b$	$c + d$	$2N$
Non-transmitted allele	$a + c$	$b + d$	$2N$
Total	$2a + b + c$	$2d + c + b$	$4N$

and Slatkin (1995), for family samples see Rhode and Fürst (2001) and Qian and Beckmann (2002). In the second step linkage disequilibrium is investigated on the basis of the estimated haplotypes or haplotype frequencies. Some of the implemented LD measures have already been described above (Devlin and Risch 1995).

Conclusions

This chapter could only introduce the basic methods of genetic epidemiological studies. Important topics had to be completely left out, such as quantitative phenotypes or gene-environment and gene-gene interaction. Others could only be mentioned, such as genome-wide linkage analysis. Some topics of general epidemiology interest are also not covered in this chapter, such as study designs appropriate for the discussed study types (cf. Chap. I.7 of this handbook), power, multiple testing and (genotyping) errors.

In addition, the area of genetic epidemiology is rapidly evolving. At the moment, most developments are made in the area of association analysis where the current technological need is highest. Initial progress has been made considering haplotype tagging SNPs as being representative for the genetic information in a LD block across a chromosomal region. Progress is needed in the area of genome wide scans using SNP chips in case-control samples as the corresponding technology is available and will be used.

References

Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32:314–331

Clerget-Darpoux F, Bonaïti-Pellié C (1992) Strategies based on marker information for the study of human diseases. Ann Hum Genet 56:145–153

Corder EH, Saunders AM, Strittmatter WJ, Schmechel DE, Gaskell PC, Small GW, Roses AD, Haines JL, Pericak-Vance MA (1993) Genetic dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late-onset families. Science 261:921–923

Day NE, Simons MJ (1976) Disease susceptibility genes – their identification by multiple case family studies. Tissue Antigens 8:109–119

Devlin B, Risch N (1995) A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics 29:311–322

Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55:997–1004

Elston RC (1998) Methods of linkage analysis – and the assumptions underlying them. Am J Hum Genet 63:931–934

Elston RL, Stewart J (1971) A general model for genetic analysis of pedigree data. Hum Hered 21:523–542

Evans DGR, Harris R (1992) Heterogeneity in genetic conditions. Q J Med 84: 563–565

Excoffier L, Slatkin M (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 12:921–927

Falk CT, Rubinstein P (1987) Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. Ann Hum Genet 51:227–233

Haldane JBS (1919) The combination of linkage values and the calculation of distances between the loci of linked factors. J Genet 8:299–309

Hall JM, Lee MK, Newman B, Morrow JE, Anderson LA, Huey B, King MC (1990) Linkage of early-onset familial breast cancer to chromosome 17q21. Science 250:1684–1689

Hardy GH (1908) Mendelian proportions in mixed populations. Science 28:49–50

Hill WG, Robertson A (1968) Linkage disequilibrium in finite populations. Theor Appl Genet 38:226–231

Holmans P (1993) Asymptotic properties of affected-sib-pair linkage analysis. Am J Hum Genet 52:362–374

IGES (2003) International Genetic Epidemiology Society (http://www.biostat.wustl.edu/~genetics/iges/) Accessed June 03, 2004

Khoury MJ, Beaty TH, Cohen BH (1993) Fundamentals of genetic epidemiology. Oxford University Press, New York

Knapp M, Seuchter SA, Baur MP (1993) The haplotype-relative-risk (HRR) method for analysis of association in nuclear families. Am J Hum Genet 52:1085–1093

Lalouel JM, Rao DL, Morton NE, Elston RL (1984) A unified model for complex segregation analysis. Am J Hum Genet 26:484–603

Lander ES, Schork NJ (1994) Genetic dissection of complex traits. Science 265:2037–2048, Published erratum, Science 1994, 266:353

Lewontin RC (1964) The interaction of selection and linkage. I. General considerations; heterotic models. Genetics 49:49–67

Maynard Smith J (1989) Evolutionary genetics. Oxford University Press, New York

McKusick VA (1998) Mendelian inheritance in man. Catalogs of Human Genes and Genetic Disorders 12th edn. Johns Hopkins University Press, Baltimore

Mendel GJ (1865) Versuche über Pflanzenhybriden. Verhandlungen des Naturforschenden Vereins, Brünn

Morton NE (1955) Sequential tests for the detection of linkage. Am J Hum Genet 7:277–318

Morton NE, MacLean CL (1974) Analysis of familial resemblance. III. Complex segregation analysis of quantitative traits. Am J Hum Genet 26:484–503

Newman B, Austin MA, Lee M, King MC (1988) Inheritance of human breast cancer: evidence for autosomal dominant transmission in high-risk families. Proc Natl Acad Sci, USA 85:3044–3048

Online Mendelian Inheritance in Man, OMIM$^{\text{TM}}$(2000) McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD) (http://www.ncbi.nlm.nih.gov/omim/) Accessed June 03, 2004

Ott J (1999) Analysis of human genetic linkage, 3rd edn. Johns Hopkins University Press, Baltimore

Penrose LS (1953) The general sib-pair linkage test. Annals of Eugenics 18:120–144

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000) Association mapping in structured populations. Am J Hum Genet 67:170–181

Qian D, Beckmann L (2002) Minimum-recombinant haplotyping in pedigrees. Am J Hum Genet 70:1434–1445

Rhode K, Fürst R (2001) Haplotyping and estimation of haplotype frequencies for closely linked biallelic multilocus genetic phenotypes including nuclear family information. Human Mutation 14:289–295

Risch N (1983) A general model for disease-marker association. Ann Hum Genet 47:245–252

Risch N (1991) A note on multiple testing procedures in linkage analysis. Am J Hum Genet 48:1058–1064

Schaid DJ (1999) Likelihoods and TDT for the case-parents design. Genet Epidemiol 16:250–260

Schaid DJ, Sommer SS (1994) Comparison of statistics for candidate-gene association studies using cases and parents. Am J Hum Genet 55:402–409

Sham P (1998) Statistics in human genetics. Wiley, New York

Shute NC, Ewens WJ (1988) A resolution of the ascertainment sampling problem. II. Generalizations and numerical results. Am J Hum Genet 43:374–386

Spielman RS, Ewens WJ (1998) A sibship test for linkage in the presence of association: The sib transmission/disequilibrium test. Am J Hum Genet 62:450–458

Spielman RS, McGinnis RE, Ewens JW (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52:506–516

Suarez BK, Hampe CL (1994) Linkage and association. Am J Hum Genet 54:554–559

Tamarin RH (1986) Principles of genetics, 2nd edn. Prindle, Weber & Schmidt, Boston

Terwilliger JD, Ott J (1992) A haplotype-based 'haplotype relative risk' approach to detecting allelic association. Hum Hered 42:337–346

Terwilliger JD, Ott J (1994) Handbook of human genetic linkage. Johns Hopkins University Press, Baltimore

Thompson EA (1986) Genetic epidemiology: a review of the statistical basis. Statistics in Medicine 5:291–302

Thomson G (1983) Investigation of the mode of inheritance of the HLA associated diseases by the method of antigen genotype frequencies among diseased individuals. Tissue Antigens 21:81–104

Vogel W (2000) Genetische Epidemiologie oder zur Spezifität von Subdisziplinen der Humangenetik. Med Genet 4:395–399

Weinberg W (1908) über den Nachweis der Vererbung beim Menschen. Jahreshefte des Vereins für vaterländische Naturkunde in Württemberg 64:368–383

Weiss KM (1993) Genetic variation and human disease. Principles and evolutionary approaches. University Press, Cambridge

Whittaker JC, Morris AP (2001) Family-based tests of association and/or linkage. Ann Hum Genet 65:407–419

Whittemore AS, Tu IP (1998) Simple, robust linkage tests for affected sibs. Am J Hum Genet 62:1228–1242

Woolf B (1955) On estimating the relation between blood group and disease. Ann Hum Genet 19:251–253