Genetic linkage studies in the pseudoautosomal region of the human sex chromosomes

Dissertation
zur
Erlangung des Doktorgrades (Dr. rer. nat.)
der
Mathematisch-Naturwissenschaftlichen Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von
Antònia Flaquer Massanet
aus
Artà, Mallorca

Bonn, 06 Februar 2009

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn

1. Referent: Professor Dr. med. Thomas F. Wienker

2. Referent: Professor Dr. rer. nat. Wolfgang Alt

Tag der Promotion: 06 Februar 2009.

Diese Dissertation ist auf dem Hochschulschriftenserver der ULB Bonn unter http://hss.ulb.uni-bonn.de/diss_online elektronisch publiziert.

Erscheinungsjahr: 2009.

CONTENTS

Preface

The importance of human genetics research is clear in order to understand human nature, diseases and development of effective disease treatment. Human genetics covers a variety of overlapping fields including cytogenetics, molecular genetics, population genetics, human biology, medical genetics and statistics. In fact, there is a long historical link between genetics and statistics. Two of the most important pioneers of modern statistics, Karl Pearson and Ronald Fisher, were involved in genetics at some point in their careers. In the first issue of *Biometrika* (1901), the editorial stated:

"The biologist, the mathematician, and the statistician have hitherto had widely differentiated fields of work ... Patient endeavour to understand each other's methods, and to bring them in harmony for united ends and common profits – this is the only method by which we can earn for biometry a recognized place in the world of Science ..."

Despite this strong historical citation, genetic statistics is often regarded as a difficult area by both geneticists and statisticians. Geneticists are often taken aback by novel statistic methods, and some statisticians are put off by the technical terminology of genetics. Genetic research requires geneticists and statisticians to work together and to learn, at least to some extent, the other person's field.

Chapter 1 starts with an introduction of genetic epidemiology and the aim of this work. The first section of chapter 2 provides an introduction to the basic molecular genetical mechanisms that are required as a background for understanding the statistical methods in the following chapters. The next sections are intended as a brief introduction to linkage analysis. They present parametric versus non-parametric analysis and twopoint versus multipoint analysis. An in-depth understanding of the linkage methods relies on some standard algorithms enabling one to compute multipoint likelihoods. Two of these algorithms are presented in chapter 3, *Elston-Stewart algorithm* and *Lander-Green algorithm*. Accurate linkage maps are crucial for the success of gene mapping projects. How genetic maps are constructed is the subject of chapter 4. While linkage analysis relies on segregation information in families, association studies focus on differences at the population level. The concept of association is introduced briefly in chapter 5. The topic of this work are the pseudoautosomal regions in the field of genetic linkage as is presented chapter 6. In addition, in chapter 7 a new genetic map is constructed for the pseudoautosomal regions using the techniques of linkage. Finally, chapter 8 provides a summary and the conclusions from this work.

Chapter 1

INTRODUCTION TO GENETIC EPIDEMIOLOGY

Genetic Epidemiology is a broad discipline combining aspects of statistics, population genetics, classical epidemiology, and human genetics. The basic goal of genetic epidemiology is to understand the role of specific genes, specific environmental factors, and interactions between genes and environment in determining a particular trait of interest. This trait can be either a binary trait such as a particular disease (schizophrenia, breast cancer) or a quantitative trait (serum cholesterol levels, height). It differs from epidemiology that explicitly genetic factors and similarities within families are taken into account. On the other hand, it can be distinguished from medical genetics by considering population rather than single patients or individuals.

Once a gene is found, this should lead to a much clearer understanding of the disease and new more targeted therapies can be tested. While most scientists agree that effective gene therapies are a long way off, some drugs based on molecular interventions are becoming a reality in contemporary medicine. Such work is likely to take many years, but it should be a good start into the right direction. It is already in practice for several years the so called genetic diagnosis. The diagnosis of genetic diseases, or a predisposition to a disease, is done using genetic diagnosis, or genetic tests. By examining the genes or proteins in a patient's cells, one can determine if the patient carries variant genes (e.g. genes that cause cystic fibrosis) or is predisposed toward a specific disease (like certain cancers). These tests apply to adults, children, and even embryos. To date, with the genetic diagnosis in embryos more than two hundred genetic disorders can be reliably diagnosed, most of them so-called monogenic diseases.

Before researchers identify and finally sequence the gene responsible for a disease, it must be first mapped, located in the Genome. Chromosome maps are a natural way of organizing genetic data about chromosomes. Existing chromosomes maps can be broadly divided into two categories: *Physical Maps* and *Genetic Maps*. Although both of them make reference to the same biological entity, namely chromosomes, they differ substantially in the types of genetic experiments conducted and the types of genetic data collected. These maps provide essential tools for understanding the organization and function of the genome. This work will focus in the direction of Genetic Mapping. The two major techniques used for Genetic Mapping are linkage and association analysis. Linkage is a method that allows to determine regions of chromosomes that are likely to contain a risk gene, and rule out areas where there is a low chance of finding a risk gene. Linkage works by using markers, which are well-character-

ized regions of DNA. Linkage studies relies on the cosegregation of stretches of DNA in families rearranged by recombination events. Researchers are searching for a marker that is consistently present in those that are affected, and is not present in non-affected relatives. When this marker is found, the marker and the disease-causing gene are said to be linked, and are assumed to be very close together. After using linkage to get an idea where risk genes may be located, association studies allow to test candidate genes, or very small genetic regions, to see if they are associated with having the disease. Association studies also require the use of DNA from many individuals. However, association studies do not necessarily use families. Rather, they look at DNA from affected individuals compared to DNA of non-affected individuals. Once a gene is located in a chromosome and suspected to being involved in a certain disease, then it is referred as a *candidate gene*.

Linkage and association analysis are two standard methods in genetic epidemiology. These methods have succeed in the past to allocate genes for autosomal and X-linked diseases but have shown some weakness to detect genes in the pseudoautosomal regions to be linked to a disease. The pseudoautosomal regions are two regions of near sequence identity at the tips of X and Y chromosomes. The pseudoautosomal regions behave like autosomes, in the sense of pairing and crossing over during meiosis. In contrast to the autosomes the recombination activity is extremely high and different between males and females. To date 29 genes have been reported to be located in these regions. Possible connection with clinical disorders such as short stature, asthma, psychiatric disorders and leukemia have been suggested but only one pseudoautosomal gene, SHOX (Short Stature Homeobox) has been associated clearly with disease, short stature of Turner syndrome.

This work focuses mainly on statistical methods for linkage analysis and genetic mapping in the pseudoautosomal regions. Some points about association analysis in these regions are also presented. The last chapters are dealing exclusively with estimation of genetic maps and methodological developments to account for the special characteristics of the pseudoautosomal regions. In addition, the construction of a new genetic map for the pseudoautosomal regions is presented.

Chapter 2

INTRODUCTION TO LINKAGE ANALYSIS

2.1 Transmission and recombination

All the cells of a human are derived from a single cell called the *zygote* which is formed by the union of two *gametes*, the ovum and the sperm. Each gamete contributes a half (or *haploid*) set of the 23 chromosomes, so that the zygote receives a full (or *diploid*) set of 23 pairs of chromosomes. Normal gametes contain 22 *autosomes* and a *sex* chromosome. The 22 autosomes are numbered in order of decreasing length from 1 to 22 (except that chromosome 21 is slightly shorter than chromosome 22). There are two types of sex chromosomes, X and Y. The sex chromosome in a normal ovum is always X, but in a normal sperm is equally likely to be X or Y. If a zygote contains a Y chromosome it will normally develop into a male, otherwise it will normally develop into a female. Figure 2.1 illustrates the karyogram of a female and male zygote.

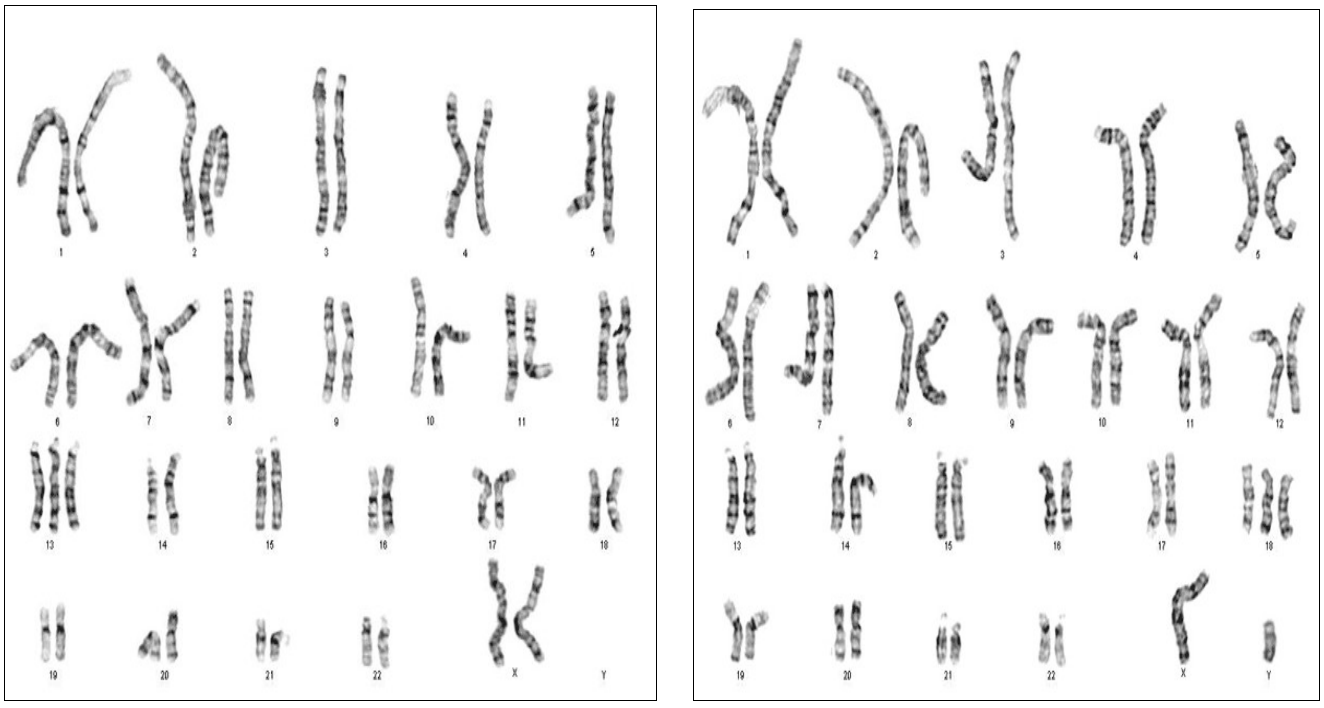


Figure 2.1 *represents the 23 paired chromosomes. Left side a female karyogram and the right side a male karyogram. Images from the Centre for Genetics Education (CGE), Royal North Shore Hospital in Sydney.*
Female karyogram: http://www.genetics.com.au/images/factsheets/fs29-2.gif
Male karyogram: http://www.genetics.com.au/images/factsheets/fs30-2.gif

The 23 pairs of chromosomes in the zygote are duplicated every time a cell division occur. The only exceptions are the gametes, which are produced by the sex organs (testes and ovaries). Gametes are produced by a special form of cell division called *meiosis*. Two chromosomes are said to be *homologous* if they pair (synapse) during meiosis. Two homologous chromosomes are not only similar in length, but are also similar in sequence. This similarity between homologous chromosomes means that diploid organisms, such as human, have two copies of every gene, except of those on the X and Y chromosomes. Meiosis gives rise to daughter cells which contain only a haploid set of 22 autosomes and one sex chromosome. This ensures that the union of two gametes will produce the set of 23 pairs of chromosomes. In addition, meiosis involves the rearrangement of genetic material by the event of crossover. This rearrangement is achieved by the exchange of genetic material between a chromosome of parental origin and the corresponding homologous chromosome of maternal origin. This exchange produces new chromosomes which consist of alternating segments of paternally and maternally DNA. The 23 new pairs of chromosomes segregate independently into four daughter cells. These daughter cells develop into gametes that ensures the survival of the genetic material into the next generation. The process of meiosis is crucial for the production of gametes. The process of meiosis begins with a regular diploid cell containing 22 autosome pairs and one pair of sex chromosomes. All pairs of chromosomes are duplicated to form two sister strands *(chromatids)* connected to each other at a region called *centromere*. The chromosomes then form pairs, resulting in four chromatids known as a *tetrad or bivalents*. At this stage the non-sister chromatids adhere to each other in a semi-random fashion at regions called *chiasmata* (meaning a cross, singular *chiasma*). Figure 2.2 illustrates the process of crossover during meiosis for one pair of autosomal chromosomes.

Each chiasma represents a point where crossing over between two non-sister chromatids can occur. Crossovers do not occur entirely at random, as they are more likely further away from the centromere, and it is unusual to find two crossovers very close to each other, this phenomenon is termed *interference*. During meiosis the presence of one chiasma usually reduce the formation of a second chiasma nearby, resulting in *positive interference*. The average number of crossovers for a chromosome depends of the length of the chromosome, ranging from just over 1 for chromosome 21 to nearly 4 for chromosome 1. On average, there are more crossovers in females than in males. During meiosis in a female cell the two sex chromosomes(XX) behave like an autosomal and could recombine on their entire length. During male meiosis however the two sex chromosomes (XY) only could crossover on two smalls region called the *pseudoautosomal regions.* The pseudoautosomal regions are discussed in more detail in chapter 6. The rest of the Y chromosome, that is left exclusively to male individuals, is not ho-

mologous to X and thus can not recombine.

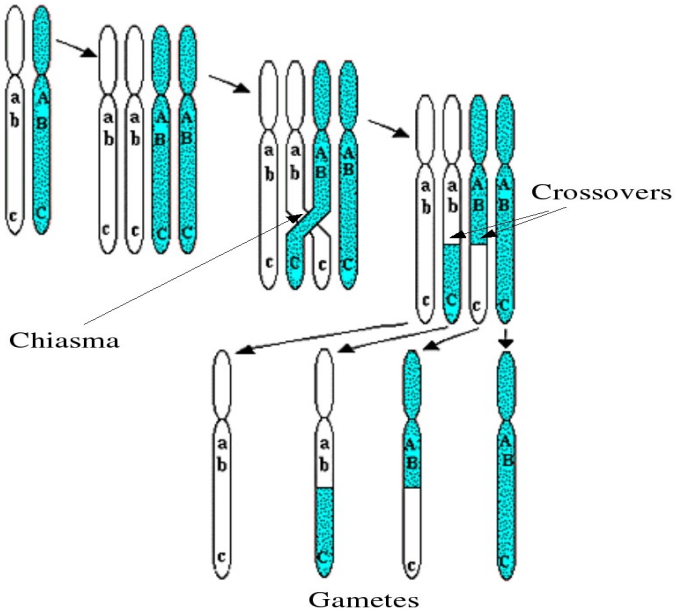


Figure 2.2 *A pair of two autosomal chromosomes are duplicated resulting in two pairs of chromosomes. The four chromosomes may crossover leading to exchange of DNA fragments in the adjacent homologous chromosomes regions resulting in the phenomenon of genetic recombination. Finally, one of the four resulting chromosomes is randomly picked to be incorporated in the new gamete. Figure from the Access Excellence (AE) the national health museum* *http://www.accessexcellence.org/RC/VL/GG/crossing.php*

The remarkable structure of DNA allows accurate copies of it to be made. This accuracy is crucial because changes in the DNA may disrupt a coding sequence and lead to the production of a protein that has a harmless effect on the cell. Nevertheless, changes in the DNA do occur from time to time, and such *mutations* introduce diversity and evolution into the population.

The further apart two segments are from each other on a chromosome, the greater the probability is that a crossover will occur between them. Hence, the greater is the distance, the higher the probability is for a recombination to be observed between the segments. The probability for a recombination, termed the *recombination fraction*, abbreviated by Θ, can be utilized as a stochastic measure for the

distance between two genes. If the segments were located very close to each other, they would almost never be separated by a crossover, hence Θ would approximate 0. If at the other extreme, the segments were situated very far apart from each other or at different chromosomes, could recombine randomly (Θ=1/2).

The existence of crossovers during the process of meiosis, is not only important for the rearrangement of genetic material from one generation to the other. It makes also possible to locate genes using the probability of co-segregation of two genes. This sequence of distribution and combination forms the basis of a number of test statistics that will be introduced later.

At this point I will introduce some fundamental genetic terminology. Definitions and terminology are the same as used by Jurg Ott in his book "*Analysis of human genetic linkage"* (1999). Heritable characters are determined by *genes,* where different genes are responsible for the expression of different characteristics. In modern terminology, a gene is a specific coding sequence of DNA, the unit of transmission, recombination and function. Genes consists of coding regions (exons) with intervening sequence of noncoding DNA (introns). Each individual carries two copies of each gene, one of which was received from the mother and the other from the father. A gene can occur in different forms or states called alleles, each potentially having a different physical expression. Any heritable quantity that follows the mendelian laws is generally called a *locus* (plural, *loci*), where a gene is special type of locus, that is, a locus with a function (a gene product). Well-characterized loci with a clear mendelian mode of inheritance may serve as genetic *marker loci* (or *marker* for short). As in the case of genes, loci may also occur in different allelic forms. In the true sense of the word, "locus" refers to the position of a gene or any other mendelizing unit than to these quantities themselves. The relative frequencies in the population of the different alleles at a locus are called *gene frequencies* (this term is a reminder of the alternative meaning of "gene"). A locus is called polymorphic when its most common allele has a population frequency of less than 95% (lately a criterium of 99% is used). At a given locus, the pair of alleles in an individual constitutes that individual's *genotype*. When two alleles have identical forms at a particular autosomal locus, the individual is called *homozygote* for that locus. The opposite of homozygote is *heterozygote*, an individual who has two different allele of a particular locus. Males outside the pseudoautosomal regions possess only one allele on the X chromosome. At these loci the individual is called *hemizygote.* However, except for rare and disputed exceptions, the Y chromosome is void of genes (a genetic *dummy*), and thus males are hemizygous for the X chromosomal alleles only, effectively. Finally, when the two alleles in a homozygous genotype are known to be copies of the same ancestral allele (identical by descent, IBD). that genotype is termed *autozygous*, otherwise it is called *allozy-*

gous.

2.2 Estimation of the recombination rate and genetic maps

A genetic map is a sequence of genetic loci with distances between adjacent loci reflecting crossover activity. Physical and genetic maps are essential for linkage and association studies. The physical position is known for the most of the markers from the assembled sequence of the human genome, however estimation of genetic maps still pose some difficulties.

The genetic map distance (in units of Morgans) between two loci is defined as the expected numbers of crossovers occurring between them on a single chromatid during meiosis. Since each chromosome consists of a tetrad, and each crossover involves two chromatids, the genetic map distance between two loci is also equal to half the average number of crossovers between them for the tetrad as a whole. The genetic map length of an entire chromosome is equal to half the average number of crossover that occur in a tetrad in a meiosis. Another commonly used unit map distance is the centi-Morgans (cM), which is defined as 1/100 of a Morgan. 1 cM correspond approximately to 1000 000 base pairs (1000 Kbp).

A mathematical relationship that converts recombination fraction (Θ) into genetic distance (m) is called a *mapping function*. In the last decades, several mapping functions have been proposed. These mapping functions are discussed in more detail in chapter 4. Table 4.3 lists some commonly used mapping functions and their inverses.

On the other hand, is there a correspondence between physical and map distance? The total length of the human genome can be assumed to be approximately 3300 *cM*. On average, 1cM corresponds to a 0.88 Mb. However, the actual correspondence varies for different chromosomal regions due to chiasma incidence is not uniformly distributed along a chromosome. In the genome exists recombination hot spots and cold spots. Hot (or cold) spots of recombination are genomic regions exhibiting much higher (or lower) rates of recombination than the genome average. In human, chiasmata are generally more frequent in females than in males meiosis. In each linkage map females distances are greater than male distances, providing evidence for a relative increase in female recombination across the human genome. This general trend has been also observed in studies in which male:female levels have been analyzed, females maps are always significantly longer than males maps. One exception to

this rule are the pseudoautosomal regions in the sex chromosomes. In these regions males exhibit a much higher recombination rate than females.

Three different type of methods could be used to estimate recombination rates for the construction of genetic maps: Three generation families, sperm typing in males and unrelated individuals. Sperm typing studies can of course only estimate male recombination rates. Whereas, three generation studies allow estimates of sex-specific recombination rates but only generate maps above the megabase scale. Using unrelated individuals, a very high resolution is reached but only sex-average recombination rates can be estimated.

Three generation families

Genetic maps of three generation families allows to allocate the order of loci as well as estimate sex-averaged and sex-specific recombination rates. This approach is done using the techniques of linkage analysis. The first genetic map of the human genome was created in 1987 by the *Centre d'Etude du Polymorphisme Humaine* (CEPH). It was followed by the *Genethon* map (Weissenbach et al., 1992), *Marshfield* map (Broman et al., 1998) and the *deCODE* map (Kong et al., 2002). The last map constructed using this technique is the *Rutgers* map (Kong et al., 2004). Although this last map is based on 2000 meioses, its resolution is still limited. Generally the possible resolution depends on the number of meiosis. For example, the estimation of a recombination rate of 0.5% with 95% confidence interval of width 0.25% requires 12.000 informative meioses.

Single sperm typing

In 1988, Arnheim and colleagues (Li et al., 1988) reported the extraordinary finding that unique DNA sequences could reliably be amplified from isolated sperm cells. In sperm typing studies, alleles at a single haploid cell can be typed and the haplotypes are determined directly without using pedigree analysis. The method allows for studying recombination on a fine scale since a larger number of meioses can be analyzed using sperm from one male. The typing of large numbers of single sperm cells from a single volunteer provides genetic information which is simple to interpret though very tedious to obtain by single cell PCR. Genetic distances and gene order can be derived directly by counting the number of cells with each allelic combination. Therefore, ignoring individual variability of recombination rate estimates including few men but many sperm cells leads to underestimation of standard errors. In addition, genetic maps from sperm typing studies could be biased since not all sperm represent viable

gametes. A genetic map for the human pseudoautosomal region was constructed using this technique (Lien et al., 2000). To date, it does not exist a map based on sperm typing for the whole human genome.

Unrelated individuals

Haplotypes from unrelated individuals also bear information about recombination rates. They reflect linkage disequilibrium (LD) on the population level since LD describes the present of a non random association between two or more alleles at distinct loci. The LD is reduced if recombination occurs during transmission into the next generation. Since other evolutionary forces influence the LD too, recombination rate estimation from haplotypes requires some extra assumptions like population size, population structure, mutation and selection. Recently, statistical methods based on coalescence theory have made feasible to estimate recombination rates using unrelated individuals (McVean et al., 2004). They enable a very fine resolution in densely typed regions by taking historic recombination events into account. Maps derived from unrelated individuals compared with those with pedigrees display a strong concordance for the majority of the regions with only some discrepancies at the end of the chromosomes. Two disadvantages are that the coalescence model assumes a finite neutrally evolving population with constant population size which can not be easily validated with empirical data, and only sex-averaged maps can be created. This technique has been applied by HapMap-Consortium for the human genome at the kilobase scale (phase2, 2005).

Several attempts have been made to integrate multiple types of mapping data. LDB (the Genetic Location Database, Collins et al., 1996) was the first attempt to integrate multiple types of genetic mapping data (genetic linkage, radiation hybrid and physical maps) to provide a superior estimate of the physical map position of genetic markers. This often gave improved ordering of markers, and this information could then be used to construct finer grained genetic maps for linkage analysis. Most recently, David Duffy used weighted regression to obtain smoothed local recombination rates to interpolate between markers with known genetic distances (Duffy, 2006) .

2.3 Genetic markers for linkage studies

In order to analyze and estimate the recombination frequencies it is fundamental to have the availability of the so-called genetic markers with a specified and known location in the genome. A genetic marker is a special DNA locus with at least one base being different between at least two individ-

uals. The conditions of being detectable and having a known location in the human genome are the requisites for a locus to serve as a marker. However, some qualities are especially desirable for a genetic marker, it should have a low mutation frequency and it must be polymorphic, i.e., several distinguishable variants must exist in the population with sufficient frequency.

The last quality is one of the most important, to be useful as a marker for linkage, a locus should be highly polymorphic. An ideal index for marker informativeness should therefore measure not only the number of possible alleles occurring at the locus, but also the frequencies of the alleles. The first index is the probability that a randomly selected individual from the population under random mating is heterozygous at the locus. This index, called *average heterozygosity (H),* is defined as:

$$H = 1 - \sum_{i=1}^{n} p_i^2 \tag{2.1}$$

where p_i is the frequency of the ith allele at the locus, and n is the total number of alleles. Another popular index most useful in the context of linkage analysis in dominantly inherited diseases is called *polymorphism information content (PIC) value* (Botstein et al., 1980):

$$PIC = 1 - \sum_{i=1}^{n} p_i^2 - \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} 2\, p_i^2\, p_j^2 \tag{2.2}$$

The type of genetic markers utilized is a rapidly changing topic. However, the two types that are most often seen today are microsatellites and single nucleotide polymorphisms.

In the non-coding parts of DNA, large areas can be found where parts of the sequence are tandemly repeated. Because the number of repeats units varies between individuals, these can be used as genetic markers. Depending on the size of the single repeated unit, these are called *variable number of tandem repeats (VNTRs)* or *microsatellites*, which is used synonymously for *short tandem repeat (STRs).* In contrast to other repeat polymorphism, STRs consists of shorter sequences of typically two to four nucleotides. STRs have the advantage of being highly polymorphic, although there are usually only one or two highly frequent alleles. Concerning to the informativeness, that means that the large number of alleles leads to high informativity. It has been discussed whether STRs have biological function. On the other hand, it has been assumed that they have no physiologic function but are useful in the areas of forensic DNA profiling and linkage analysis.

The most abundant form of variation in the human genome, about 90%, are the *single nucleotide polymorph (SNP)*. Basically, these are variations at only a single base, meaning that one base is substituted by another. Because of their short length, SNPs can easily be typed by PCR methods. Although most SNPs are found in non-coding regions, some of them are located in genes or in the promotor of genes and can be directly viewed as candidate variation for disease. Because SNPs are nearly always diallelic, they are inherently less informative than STRs. Currently, there are several groups offering high throughput of DNA SNP microarrays, called "GeneChips" genotyping. The biggest ones, Affymetrix and Illumina provide GeneChips containing more than one million SNPs covering the whole genome. The two platforms vary in the genome coverage of their maps, the extent of missing data and in their accuracy, but both have costs that are more than 100 times lower than what was available only a few years ago.

2.4 Pedigrees for linkage

Because recombinations events can be recognized only on the basis of haplotypes passed from parents to children, linkage analysis cannot be carried out with unrelated individuals and requires observations on relatives. Therefore, for linkage analysis, researchers collect phenotypic information on members on family pedigrees. A *pedigree* may be defined as a set of relatives with known relationship among individuals (see Figure 2.3).

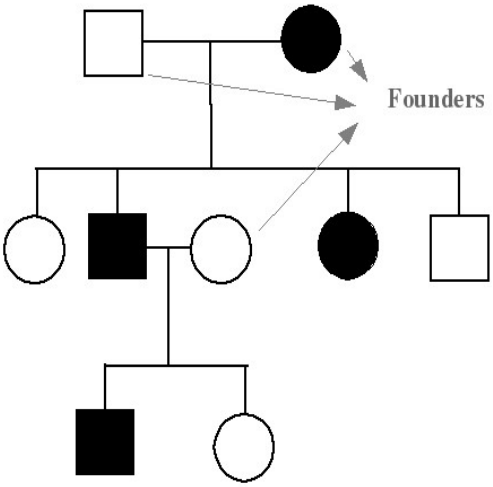


Figure 2.3 *Illustration of a pedigree. The three founders in the pedigree are marked by an arrow, the rest of individuals are non-founders. Males are represented by a square and females by a circle. When the figure is filled in black represents that the corresponding individual is affected by the phenotype under study.*

Pedigree members fall into two classes: *founders* are individuals whose parents are not in the pedigree whereas *non-founders* have their parents in the pedigree. For statistical reasons, each non-founder is assumed to have both parents represented in the pedigree. Founders are generally assumed to be unrelated and drawn randomly from the population. A *nuclear pedigree* is a family formed by father, mother, and their offspring. Those pedigrees including other relationships are called *extended pedigrees*.

2.5 Test for linkage

Linkage analysis is the method typically used to determine the genetic location of a disease gene or to map a gene in the genome. The aim is to identify a piece of DNA of known location that is co-segregated by all family members affected by the disorder being studied, and is not inherited by any of the unaffected family members. Once this piece of DNA is found, one knows that the disease gene must be located somewhere close by. The main idea is to determine the location of the disease gene and then the gene itself. Linkage analysis methods rest on the biological event of crossover, and hence on the recombination fraction. In the simplest case where two loci are considered, the problem reduces to test the hypothesis of linkage versus the hypothesis of no linkage, and then the estimation of the recombination

fraction between these two loci. If the number of recombinant and non-recombinant gametes in a random sample of gametes can be counted, then an estimation of the recombination fraction, θ, is simply the proportion of gametes that are recombinant, and then a test for linkage is the test whether this proportion is equal to ½ (the null hypothesis of no linkage) or less than ½ (the alternative hypothesis of linkage). The hypotheses can be written in the following manner:

$$H_0 : \theta = \frac{1}{2} \quad vs \quad H_A : 0 \leq \theta < \frac{1}{2} \tag{2.3}$$

The conditions where the number of recombinant and non-recombinant gametes can be counted directly are easy to observe in experimental organisms but rarely so in humans.

Linkage analysis involving two loci is known as *twopoint analysis (*also referred as *twolocus* or *singlepoint analysis*). Usually, one of the loci is well-known and already mapped and the other one is not yet mapped. The second locus could be a so far not located marker, with the objective of determine its position. The second locus could be also a disease susceptibility gene i.e., genes which are responsible for the development of a certain disease, with the objective of map it on the genome. The other type of analysis, *multipoint analysis* (also referred as *multilocus analysis*), is carried out for a set of markers that are linked to each other against a disease locus or a new marker. The marker order and the inter-marker distances are assumed to be known. The disease locus or the new marker is then placed in each marker interval and left and right of the most left and most right markers, respectively. Next, the likelihood is computed for all possible positions of the trait in the considered chromosomal regions, and the most likely position is chosen. The inter-marker distances are given by the genetic map, and it is fundamental to use of a reliable map, in multipoint analyses.

2.5.1 The genetic model of hereditary diseases

There are two main linkage analysis approaches: *parametric (or model-based)* and *nonparametric (or model-free)* methods. Parametric methods require specification of genetic parameters, such as penetrance; disease-allele frequency; phenocopy and mutation rates describing the mode of disease inheritance (presented in the next section). The statistical method in parametric models employs the LOD score, based on calculation of the likelihood of the observed pedigree data, given assumed parameter values. Complex diseases are not caused by a single gene, but by multiple genetic and environmental factors, and thus it is much harder to specify a genetic model. Therefore, alternative nonparametric

models linkage (NPL) methods have been developed.

In linkage analysis, the parameter of primary interest is the recombination fraction in the case of twopoint analysis, or a set of recombination fractions (or genetic map) in a multipoint analysis. These are the only parameters that appear in the likelihood function for simple pedigrees and simple co-dominant loci. However, in order to deal with more complex pedigrees and loci, it is helpful to introduce additional parameters.

2.6 Parametric linkage analysis and likelihood

As a matter of definition, there is a clear distinction between *phenotype* and *genotype*. The genotype of an individual is usually defined as the set of alleles present in the individual at loci under consideration and is, in principle not observable. The phenotype of an individual is defined as the observed characteristics of the individual that are influenced by the locus under consideration. The relationship between phenotype and genotype is specified by a set of parameters known as *penetrance* (2.4). Formally, penetrance is the conditional probability of being affected with the disease under consideration (phenotype) given a specific genotype. If we assume a disease-causing genotype with one disease-causing allele *D* and one allele not causing the disease *d*, the probabilities of being affected depending of the genotypes (penetrances) can be expressed as follows:

$$f_0 = P(affected \mid dd) \qquad f_1 = P(affected \mid Dd) \qquad f_2 = P(affected \mid DD) \tag{2.4}$$

Specially, f_0 is termed the frequency of *phenocopies,* the probability of being affected without carrying a particular disease-causing allele at the locus under consideration.

2.6.1 The pedigree likelihood

Parametric linkage analysis is a special form of the likelihood ratio test. The likelihood principle as stated by Edwards (1972, p. 30) says:

> *"Within the framework of a statistical model, all the information which the data provide concerning the relative merits of two hypotheses is contained in the likelihood ratio of those hypotheses on the data. ...For a continuum of hypotheses, this principle asserts that the likelihood function contains all the necessary information."*

The main point lies into compute the likelihood of a genetic model (including its parameters) given family tree data "L(Model | data)", that is proportional to the observed family data given the genetic model.

$$L(Model \mid Data) \propto L(Data \mid Model) \tag{2.5}$$

The constant of proportionality could be selected arbitrarily, in the following, it will be set to unity. In this case, the observed data refers to the individual genotypes and phenotypes as well as the family relationship between the individuals. The genetic model need to be defined by the penetrances, the disease allele frequency, and the allele frequencies at the loci. In addition, the recombination frequencies between the loci are also part of the model. For the calculation of the likelihood in a pedigree, it is necessary to treat founders and non-founders differentially. The founders are supposed to be independent from each other and drawn at random from the population and non-founders are treated as founder's dependents.

Denoting the genotype and phenotype of the *i*th individuals in a pedigree by g_i and x_i respectively, the conditional probability of x given g can be written as:

$$P(x \mid g) = \prod P(x_i \mid g_i) \tag{2.6}$$

where the product is taken over all pedigree members. Each of the conditional probabilities of a phenotype given a genotype is a function of the penetrance parameters. Since genotypes g are not directly observed, the likelihood of a set of pedigree data is just the unconditional probability of the pedigree phenotypes x, and this can be expressed as the sum of the joint probabilities of x and g over all possible values of g:

$$P(x) = \sum P(x \mid g) P(g) \tag{2.7}$$

where the summation is taken over all possible values of g. The genotypes of the individuals in a pedigree are interrelated to each other by genetic transmission from parent to offspring. Since θ determines the probabilities of haplotype transmission from parent to offspring and hence the probability distribution of offspring genotype conditional on parental genotypes, it is known as a *transmission parameter.* The parameters that determine the probability distribution of genotypes in the founding members of a pedigree are known as *population parameters.* In the most general case, the population parameters con-

sists of the frequencies of all possible ordered genotypes at the loci concerned. Consequently, if the analysis concerns two loci, one with m_1 and the other with m_2 alleles, then the number of haplotypes is m_1m_2 and the number of possible ordered genotypes is $H(m_1m_2)^2$. For example, when two loci have 5 and 3 alleles, respectively, there are 15 possible haplotypes and 225 ordered genotypes. So, the number of possible genotypes is increasing rapidly with the number of alleles at the loci.

If the n members of a pedigree are ordered such that the f founders precede the n-f non-founders, then the joint probability of the genotypes of all members can be written as a product as follows:

$$\begin{aligned} L(g) &= P(g_1)P(g_2|g_1)P(g_3|g_1,g_2)\ldots P(g_n|g_1,g_2,\ldots,g_{n-1}) \\ &= P(g_1)\ldots P(g_f)P(g_{f+1}|g_{f+1,f},g_{f+1,m})\ldots P(g_n|g_{n,f},g_{n,m}) \end{aligned} \tag{2.8}$$

where $g_{j,f}$ and $g_{j,m}$ denote the genotypes of the father and mother of jth nonfonder member of the pedigree. The likelihood of a pedigree is a multiple summation of products each involving n penetrance parameters, f population genotypes frequencies (for the f founders) and n-f conditional transmission probabilities (for the n-f non-founders) over all possible combinations of genotypes for the n individuals in the pedigree. Denoting the penetrance, population frequencies and the transmission probability of genotype g_i by pen(x_i | g_i), pop(g_i) and tran($g_i|g_{i,f}$, $g_{i,m}$) respectively, this multiple summation can be written as:

$$L = \sum_{G_1}\ldots\sum_{G_n}\prod_{1}^{n} pen(x_i|g_i)\prod_{i=1}^{f} pop(g_i)\prod_{i=f+1}^{n} tran(g_i|g_{i,f},g_{i,m}) \tag{2.9}$$

where G_i represents all possible genotypes for individual i. For two loci with m_1 and the m_2 alleles, each person can have $(m_1m_2)^2$ possible ordered genotypes, so that there are $(m_1m_2)^{2n}$ possible combinations of ordered genotypes for the entire pedigree of n members. The likelihood is then a sum of $(m_1m_2)^{2n}$ terms, each term being a product of $2n$ probabilities. The number of terms in the summation increases with the number of loci and the alleles included in the analysis. Efficient algorithms are therefore necessary for multipoint linkage analysis of general pedigree data.

The formulation of the pedigree likelihood as a multiple summation of products between population, transmission and penetrance parameters, over all combinations of possible genotypes, was proposed by Elston and Stewart (1971), who also suggested a recursive algorithm that greatly reduces the computational time. The so-called Elston-Stewart algorithm has been extended to deal with complexities such as consanguineous matings (Ott, 1974; Lange and Elston, 1975; Cannings et al., 1978). This

method is presented in detail in chapter 3.

The above remarks for the calculation of the likelihood refer in particular for autosomal loci. For loci situated on the X-chromosome, outside the pseudoautosomal regions, some special characteristics have to be taken into account. The penetrance parameters for X-chromosomal loci in females are the same as defined for autosomal loci. For hemizygous males, considering a locus with *m* alleles, that means *m* different genotypes consisting of one allele each, thus *m* independent penetrance parameters. Considering a diallelic locus, located at the X-chromosome, to each male would correspond two penetrance parameters and to each female three penetrance parameters. Also the genotype frequencies in females are the same that the ones regarding to autosomal loci. For males, the genotype frequencies of a locus situated on the X-chromosome would be identical to the allele frequencies. The probabilities of transmission for X-chromosome have to be differentiated between women and men. When a woman inherits her maternal haplotype, it can be treated in the same way as autosomal loci, since the mother of the woman carries two X-chromosomes. The transmission probability, outside the pseudoautosomal region, for the paternal inherited haplotype is always 1, since the father of the woman has only one X-chromosome. He will pass the entire haplotype to his daughter without recombinations. A male offspring inherits always the X-chromosome from his mother, which undergoes crossover and recombination, of course.

When more than one pedigree is considered, each family is supposed to be independent from each other. However, they have to come from the same population. Thus, the likelihood of a sample of *t* pedigrees $L_{(\text{all pedigrees})}$ is therefore the product of the likelihoods for each simple pedigree contained in the sample:

$$L_{(all\ pedigrees)} = \prod_{k=1}^{t} L_{(pedigree_k)} \tag{2.10}$$

2.6.2 Estimation of the parameters

When dealing with parametric linkage analysis all parameters of the genetic model must be known and specified. In some cases however, these parameter are unknown and need to be estimated. A way to estimate the unknown parameters is using the likelihood function on the basis of the observations. In statistical terminology, observations are random variables. Any function of random variables, which does not depend on unknown parameters, is termed a *statistic.* Various statistical methods of parameter estimation exits, in the statistical analysis of human pedigree data, maximum likelihood estimator (*MLE*) is usually the method of choice for estimating model parameters. In some situations, a closed-form solution can be found; others require a numerical solution using iterative methods. Statistical geneticists have employed several iterative schemes. Among these are Newton-Raphson scoring, the simplex method, quasi-Newton methods, simulated annealing, and the EM (*expectation-maximization)* algorithm. The EM algorithm is a numerical method for finding the MLE of parameters and is applicable when the problem can be formulated as one of incomplete data. In other words, it is applicable to situations where the estimation can be made much easier if certain additional pieces of data are available. First, the unknown parameters are assumed to take an initial set of plausible values. Then, based on these initial values, the expected values of the missing data are calculated. These expected values are imputed for missing data, so that together with the available data, a complete data set is obtained. This is known as the *expectation step,* since expected values are imputed for missing data. From the complete data set, maximum likelihood estimates of the parameters are obtained, and these constitute improved estimates of the parameters. This is known as the *maximization step,* since maximum likelihood estimates are obtained from the complete data. The improved parameter estimates are used in another expectation step to give an improved set of values for the missing data. The new imputed values are then combined with the observed data and subjected to another maximization step to give a set even more accurate parameter estimates. This procedure of alternating expectation and maximization step is repeated until the changes in parameter estimates are small and tolerable for the purpose of the study. In the case of pedigree data, the MLE can be used for the estimation of the parameters such as allele frequencies, haplotype frequencies, penetrances and recombination frequencies. In most cases, the recombination frequency between the putative disease locus and one or a set of marker loci is estimated.

2.6.3 The LOD Score

In parametric linkage analysis, the likelihood function is calculated using the observations on the pedigree members to test for the presence of linkage. In the case of twopoint linkage analysis the recombination fraction θ between two loci need to be estimated, so the likelihood function will depend on the unknown parameter θ. In the framework of multipoint linkage analysis the likelihood will depend on a vector of $\theta's$, representing the recombination fractions to a group of relatively close loci. For the calculation of the likelihood, in both cases, all the other parameters regarding to the genetic model must be known and specified.

Twopoint Analysis

Twopoint analysis is performed considering one marker locus and a disease locus, or two marker locus. In the case of marker-disease, the specification of the disease model it is necessary . This specification is given by the disease allele frequency and the penetrances at the disease locus. The allele frequencies of the marker locus need to be specified, too. A test of linkage concerning two marker loci is performed in the same way, with the calculation of the likelihood. In this case, of course, it is easier because it is not necessary to define the disease model. The calculation of the likelihood, which will depend on θ, does not give whether the two locus are linked or not. To test for linkage between two loci, one needs to calculate the so-called likelihood ratio $L(\theta)/L(\theta=1/2)$ where $L(\theta=1/2)$ is the likelihood under the null hypothesis (H_0), which considers that both loci are unlinked to each other. $L(\theta)$ represents the likelihood under the alternative hypothesis (H_A), i.e. for $\theta<1/2$. The likelihood ratio indicates how much higher the likelihood of the data is under linkage than under the absence of linkage. It is usually convenient to work not with the likelihood ratio but with its logarithm (to base 10), the resulting score is named *LOD score* (Barnard, 1949) and is denoted by $Z(\Theta)$:

$$Z(\theta)=\log_{10}\frac{L(\theta)}{L(\theta=1/2)} \tag{2.11}$$

In addition, it is necessary to define the MLE of the recombination fraction from the pedigree data, in fact for $0\le\hat{\theta}\le 1/2.$ The reason for this restriction is that recombination fractions bigger than ½ make no sense in terms of genetics. Then one needs to compute the likelihood for multiple values of θ, or to use the EM-algorithm described above to obtain the estimate of θ, $\hat{\theta}$. Using the MLE $L(\theta=\hat{\theta})$ in the numerator of the Equation (2.11) the so-called *maximum LOD score* is obtained:

$$\hat{Z} = Z(\hat{\theta}) = \log_{10} \frac{L(\theta = \hat{\theta})}{L(\theta = 1/2)} \tag{2.12}$$

The maximum LOD score, $\hat{Z}$, is the statistic for measuring linkage as well as the one to use for testing the hypothesis $H_0 : \theta = 1/2 \quad vs \quad H_A : \theta < 1/2$. Under H_0, hypothesis of no linkage, the statistic $2 \cdot \hat{Z} \cdot \ln(10)$ is asymptotically distributed as a Chi-square distribution with one degree of freedom (χ_1^2) and may be used to test the statistical significance of the differences of likelihood between the two hypotheses. The degrees of freedom used for Chi-square test is the difference in number of estimated parameters in the two models under study. In this situation one parameter is estimated under H_A, whereas under H_0 all parameters are fixed. Positive LOD scores indicate evidence in favor of linkage, and negative LOD scores indicate evidence against linkage. The calculation of the $\hat{Z}$ is not only used to carry out the linkage test, in the same time provides an estimator of the recombination fraction between the marker locus and the disease locus. This is very helpful to determine the exact position of the disease gene.

Multipoint Analysis

In parametric multipoint analysis, linkage relationships between a disease locus and a multitude of marker loci on a known map are investigated, so like in the case of twopoint parametric analysis here one needs to specify the genetic model at the disease locus and the allele frequencies of the marker loci. In addition one must specify the marker distances , and thus implicitly, the marker order. Multipoint analysis differ from twopoint analysis in the fact that the order of the marker loci now is very important. The number of parameters, haplotypes, and genotypes increases drastically with the number of loci considered. With multipoint analysis can also be tested whether a group of markers are linked to a further marker. In this case, of course, the specification of the disease model is not necessary.

In the multipoint analysis the likelihood ratio will be defined as *L(x)/L(x unlinked)* and the multipoint LOD score, *Z(x),* as:

$$Z(x) = \log_{10} \frac{L(x)}{L(x\ unlinked)} \tag{2.13}$$

The numerator indicates the likelihood for a genetic position *x* of the disease locus. The denominator describes the likelihood under H_0, i.e. in the case that the set of markers are linked, and the test locus is

not linked to any marker in this set of marker loci. Two situations need to be considered for the calculation of the likelihood for a concrete position of the disease locus. One is whether the disease locus is within the set of markers, and the other case would be whether the disease locus is situated outside the set of markers. If it lies within the set of markers, then it is inserted between two flanking markers. Then, the recombination fraction between the two markers needs to be split up in two intervals according to the position of the disease locus, i.e. first marker-disease locus and then disease locus-second marker. If the disease locus is outside of the set of markers, it must be attached to the appropriate end of the set of markers. Under H_0, no linkage, the recombination fraction will be ½, and it is unimportant whether the disease locus is added respect to the set of markers. For the calculation of the likelihood under H_A the MLE of the location of the disease locus must be determined. For that, the likelihood is maximized for each interval between two markers and the global MLE is determined. In the case where the disease locus is attached outside the set of markers, the likelihood is computed in the the same way as for the twopoint analysis. Then, the maximum LOD score $\hat{Z}$ is obtained using the MLE $L(x=\hat{x})$ in the numerator of the Equation 2.13:

$$\hat{Z} = \log_{10} \frac{L(x=\hat{x})}{L(x\,unlinked)} \tag{2.14}$$

$\hat{Z}$ is the statistic for the multi-marker analysis for testing the hypothesis H_0: disease locus unlinked vs H_A: disease locus linked to the set of markers. Here again, Under H_0, the test statistic $2 \cdot \hat{Z} \cdot \ln(10)$ asymptotically has a Chi-square distribution with one degree of freedom (χ_1^2) As in the case of twopoint analysis, the calculation of the $\hat{Z}$ permits to test on linkage and in the same time produces an estimate of the location of the disease locus.

The LOD score for several pedigrees is simply the summation of the LOD score for single pedigrees (see 2.10), considering *t* pedigrees one obtains

$$Z_{(all\ pdigrees)} = \log_{10} \frac{L(H_1)_{(all\ pedigrees)}}{L(H_0)_{(all\ pedigrees)}} = \log_{10} \frac{\prod_{s=1}^{t} L(H_1)_{(all\ pedigrees)}}{\prod_{s=1}^{t} L(H_0)_{(all\ pedigrees)}} = \sum_{s=1}^{t} \log_{10} \frac{L(H_1)_{(all\ pedigrees)}}{L(H_0)_{(all\ pedigrees)}} = \sum_{s=1}^{t} Z_{(pedigrees)} \tag{2.15}$$

2.6.4 Interpretation

When $\hat{Z}$ reaches or exceeds a certain critical value Z_0, the data is said to convey *significant evidence for linkage*. The critical value of the test is obtained on the basis of additional assumptions about the test distribution. The first of these assumptions follows the fact that under H_0, values of $2 \cdot \hat{Z} \cdot \ln(10)$ are distributed as χ^2 with one degree of freedom. Thus, given the type I error α determines the critical value of the test, $P[2 \cdot \ln(10) \cdot \hat{Z} > \chi^2_{2\alpha,1}] = \alpha$. The second assumption about the test distribution includes an attempt to use Bayesian arguments. The Bayesian procedure results in a-posteriori probability for linkage, i.e. probability for H_A conditionally on the observed data, and a-posteriori probability about the relationship between H_A and H_0. Following these assumptions, Morton (1955) proposed a critical value of Z_0=3, for an autosomal locus, and Z_0=2 for X-linked locus.

For autosomal loci a $\hat{Z}$ between 2 and 3 is taken as *suggestive evidence for linkage*. If the LOD score is below -2 for a certain region, then the location of a disease susceptibility gene can be excluded from that region. This strategy is called *exclusion mapping*.

In the past years the "LOD 3 criteria", has been in the focus of long discussions. In response to the frequent failure to replicate claimed localizations of disease susceptibility genes, Lander and Kruglyak (1995) proposed a series of thresholds:

- *Suggestive linkage* is a LOD score or p-value that would be expected to occur once by chance in a whole genome scan.
- *Significant linkage* is a LOD score or p-value that would be expected to occur by chance 0.05 times in a whole genome scan (i.e. the conventional p = 0.05 threshold of significance).
- *Highly suggestive linkage* is a LOD score or p-value that would be expected to occur by chance 0.001 times in a whole genome scan.
- *Confirmed linkage* is when a significant linkage observed in one study is confirmed by finding a LOD score or p-value that would be expected to occur 0.01 times by chance in a specific search of the candidate region.

2.6.5 Effects of model misspecification.

The likelihood method of linkage analysis requires the specification of a statistical model, in which the parameter of first interest is usually the recombination fraction between two loci, or the position of a locus relative to a set of fixed loci. Other parameters, such as allele frequencies and penetrances, are usually of secondary interest and considered as nuisance parameters. In linkage analysis these nuisance parameters are not usually jointly estimated with the primary parameters, but are specified before the analysis according to prior knowledge. The analysis is 'optimal' when the model is correctly specified. Model misspecification of any form is expected to have negative effects, which may include a reduction in the power to detect linkage, and biased estimates of the parameters of interest.

The first concern of researchers regarding model misspecification is that it may invalidate the statistical test for linkage. Fortunately, the likelihood ratio test (and the LOD score method) is quite robust in this regard. By deriving the exact probability distribution of the LOD score functions for nuclear family data, Clerget-Darpoux et al. (1986) found that misspecification of the disease locus parameters (i.e. disease allele frequency and penetrances) did not inflate the false positive rate of declaring linkage between the disease and the marker loci. However, the test for linkage is not entirely robust to model misspecification. When pedigrees ascertained for a disease contain founders with unknown marker phenotypes, Ott (1992) showed that misspecification of the allele frequencies of the marker locus could lead to an increased rate of false positives linkage findings between the disease and the marker.

Multipoint linkage analysis is much more sensitive to misspecification of the disease model (Risch and Giuffra, 1992). The reduction in power due to the misspecification of genetic model parameters may be substantial. Often, if there is a disease locus in an interval spanned by several markers, but the parameters of the disease locus are misspecified, the LOD score function is grossly deflated within the interval, and only becomes moderately positive outside the interval. The explanation lies in the mathematical properties of the multipoint likelihood function. In comparison to the twopoint analysis, there is less confounding between the position of the disease locus and the parameters of the disease model. This sensitivity of multipoint linkage analysis to model misspecification has led some researchers to favor twopoint analysis when there are uncertainties about the true disease model. This is somehow a negative view, as the situation with multipoint analysis can also be viewed in a positive

light. Multipoint data offers more information than twopoint data for the estimation of unknown parameters, when the disease model is uncertain.

2.7 Nonparametric analysis

Complex diseases are not caused by a single gene, rather by multiple genetic and environmental factors, and thus it is much harder to specify a genetic model. Therefore, alternative nonparametric linkage (NPL) methods based on allele sharing by relatives have been developed. The basic idea is to consider a pedigree with several affected members, it is likely that the affected individuals share the same disease alleles from one or a few founders. In this approach, suggested by Penrose (1935), is to discard the parametric method of linkage analysis, and to focus on the association between the sharing of diseases status and the sharing of marker allele by relatives (usually affected sib-pairs, ASPs).

2.7.1 Measures of allele-sharing by relatives

Allele sharing is central to nonparametric methods of linkage analysis. There are two different definitions of allele-sharing, *identity-by-state* (IBS) and *identity-by-descent* (IBD). Two alleles of the same form (i.e. having the same DNA sequence) are said to be IBS. If, in addition to being IBS, two alleles are co-segragated from the same ancestral allele, then they are said to be IBD. Figure 2.4 illustrates the distinction between IBD and IBS for a pair of full sibs. In addition, it also illustrates the existence of a close relationship between IBD status and recombinations events. This close relationship implies that IBD is more directly relevant than IBS for linkage analysis.

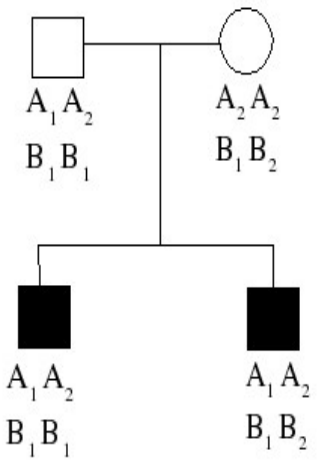


Figure 2.4 *Consider a nuclear family with the following genotypes at loci A and B, which are in extremely tight linkage. The two siblings have identical genotypes (A_1A_2) at locus A, and are therefore IBS. A_1 allele must be descended from the same parental A_1 allele. These two alleles are therefore IBD. The two allele A_2 are descended from the mother. Since she has two A_2 alleles, these are indistinguishable from each other, and it is impossible to determine whether the two A_2 alleles shared by the siblings were descended from the same A_2 allele or not, using the information of locus A alone. However, the mother transmitted B_1 to the the first sibling and B_2 to the second sibling, and so the two A_2 alleles transmitted from the mother to the two siblings could only be IBD if a recombination event occurred between A and B in one of the two meiosis. Since A and B are in extremely tight linkage, this is highly unlikely, so the two siblings are IBD for only one allele (A_1) at locus A.*

Affected sib pair test

The simplest and one of the most important test of the nonparametric linkage analysis is the so called *affected-sib-pairs* (ASP) analysis (Penrose, 1935). For this variant at least a pair of affected siblings is needed and the two parents. With the ASP method one determines how many alleles are shared between a pair of siblings, i.e whether the pair shares zero, one or two alleles IBD. These absolute frequencies are denoted by n_0, n_1 and n_2 respectively. The relative frequencies are the estimation for the probability z_i that an affected sib pair shares i alleles IBD:

$$\hat{z}_i = \frac{n_i}{n} \qquad for \quad i = 0,1,2 \tag{2.16}$$

where $n = n_0 + n_1 + n_2$ is the sample size. If locus x is unlinked to the disease locus (H_0), the expected dis-

tribution of alleles IBD for sib pairs equals the binomial probabilities, i.e. $z_0=1/4,\ z_1=1/2\ \ and\ z_2=1/4$ (see Appendix, A1). Thus, the hypothesis testing at locus *x* can be formulated as:

$$H_0:(z_0^0,z_1^0,z_2^0)=(\frac{1}{4},\frac{1}{2},\frac{1}{4})$$
$$H_A:(z_0,z_1,z_2)\neq(\frac{1}{4},\frac{1}{2},\frac{1}{4}) \qquad (2.17)$$

If H_A is true, the power of the test at locus *x* will depend on how much (z_0,z_1,z_2) deviates from (½, ¼, ½). This depends both on the genetic model and the recombination fraction between locus *x* and the disease locus. Obvious constraints for (z_0,z_1,z_2) are: $z_0+z_1+z_2=1,\ z_0\geqslant 0,\ z_1\geqslant 0\ \ and\ \ z_2\geqslant 0.$

There are several statistic tests, in order to examine whether the observed distribution of the alleles IBD deviate significantly from the distribution under H_0. For example, like in the case of parametric linkage analysis, a likelihood ratio test exists also in the context of the ASP analysis (Risch, 1990c). Here the parameters $(z_0,z_1,z_2)\ \ and\ \ (n_0,n_1,n_2)$ are the observed data. The test statistic T_{LR} (*likelihood ratio*) represents the likelihood of the observed data under H_A in relation to the likelihood under H_0:

$$T_{LR}=\log_{10}\frac{\hat{z}_0^{\,n_0}\hat{z}_1^{\,n_1}\hat{z}_2^{\,n_2}}{(z_0^0)^{n_0}(z_1^0)^{n_1}(z_2^0)^{n_2}}=\log_{10}\frac{\hat{z}_0^{\,n_0}\hat{z}_1^{\,n_1}\hat{z}_2^{\,n_2}}{(1/4)^{n_0}(1/2)^{n_1}(1/4)^{n_2}} \qquad (2.18)$$

Then, $2\cdot\ln(10)\cdot T_{LR}$ is asymptotically Chi-square distributed with two degrees of freedom and provides a test for linkage. There are only two free parameters because of the constrain of $z_0+z_1+z_2=1$.

Holmans (1993) and Faraway (1993) showed that the power could be increased by restricting the maximization of the IBD probabilities $(i.e.\ \ \hat{z}_0,\hat{z}_1,\hat{z}_2)$ to a set of inequalities:

$$z_1\leq\frac{1}{2}\qquad 2\cdot z_0\leq z_1\qquad z_0\geq 0 \qquad (2.19)$$

These three restrictions in a special plot form a triangle. A more powerful test for linkage can be obtained by restricting the alternative hypothesis to this region. The *triangle test statistic (TTS)* is asymptotically distributed as a mixture of Chi-square distribution with two degrees of freedom.

The *proportion test* (Day and Simons, 1976; Suarez et al., 1978) tests whether the proportion of affected sib pairs sharing two alleles IBD equals ¼ or whether this sharing is increased. The test statis-

tic T_{prop} is examined whether $\hat{z}_2$ deviates significantly from the expected value $z_2^0 = 1/4$ under H_0, where T_{prop} follows asymptotically a normal distribution for large n.

$$T_{prop} = \hat{z}_2 = \frac{n_2}{n} \tag{2.20}$$

Another affected sib pair test is the *mean test* proposed by Blackwelder and Elston (1985). The statistic used for the proportion of alleles shared IBD for a sib pair is given by:

$$T_{mean} = \hat{z}_1 + 2\hat{z}_2 = \frac{n_1}{n} + 2\frac{n_2}{n} \tag{2.21}$$

T_{mean} examines whether the the proportion of alleles IBD deviates significantly from the expected value 1/2 under H_0. T_{mean} is for sufficient large n also asymptotically normal distributed.

A large study to compare the power of ASP statistics was conducted by Blackwelder and Elston (1985). They compared three different scores, the T_{prop} test, the T_{mean} test and the T_{LR} test. Their conclusion was that although the most powerful test depends on the true mode of inheritance, the disease allele frequency and the recombination fraction, the T_{mean} test is generally more powerful.

In many applications, the collected sample will contain several sibships with more than two affected siblings. In practice, all possible pairs are formed, so for a sibship of size s, there are $s(s-1)/2$ possible pairs, though only $s-1$ of theses pairs are independent. If all possible combinations of sib pairs from the same family are used this can lead to an underestimation of the p-value (Daly and Lander, 1996). One suggestion is to use weighted LOD score of each affected sib pair, weighted with $2/s$, although it results in a conservative test (Meunier et al., 1997).

2.7.2 Methods for analyzing larger pedigrees

Within an extended pedigree tree, pairs of relatives other than siblings are suitable for nonparametric linkage analysis (Risch, 1990b). The general approach for nonparametric linkage analysis in extended pedigrees is the *affected pedigree member (APM)* method, developed by Weeks and Lange (1988, 1992). In its original version, the sum of the observed number of alleles shared IBS by each affected pair of relatives is compared to its expected value in the absence of linkage. Although it has been extended in several ways so that unaffected relatives could be included, the APM method had two ma-

jor weak points. First, the results are profoundly dependent of the allele frequencies at the marker loci. Second, the APM method is constructed on the concept of IBS information and is less powerful than methods based on the IBD distribution. An improved version of the APM method, *SimIBD*, is based on IBD rather than IBS (Davis et al., 1996) and computes an empirical p-value using conditional simulation. This method was found to perform poorly when analyzing sibships without typed parents (Davis and Weeks, 1997). Another alternative is the *nonparametric linkage (NPL)* approach, originally proposed by Kruglyak et al. (1996). It is based on the observed marker inheritance patterns of the affected individuals. The alleles IBD are determined using the genotypes of all loci of all the individuals from the same pedigree. Thereby the recombination fractions between the loci are considered. In this way the NPL score can be determined for different genetic positions. This allows the determination of a position of a susceptibility disease gene, in a given maker map. The computation of the NPL score will be described in detail in the next chapter through the procedure known as Lander-Green algorithm. The NPL score was first implemented in the program GENHUNTER. The significance levels of the NPL statistic calculated by GENHUNTER is based on the assumption of complete IBD information, and the test is conservative when this is not the case, resulting in some loss of power. This may make the NPL statistic less efficient than other methods for analyzing sib pair data, although it appears to be a powerful method of nonparametric analysis on extended pedigrees. The problem of the conservativeness of the NPL statistic has been addressed by Kong and Cox (1997). However, the Lander-Green algorithm scales highly unfavorable with the number of individuals in a pdigree, and thus large pedigrees cannot be analyzed under this approach.

Another method based on IBD sharing in affected relatives and for large pedigrees is the one developed by Curtis and Sham (1944) called *extended relative pair analysis (ERPA)*. ERPA calculates for each affected relative pair the prior IBD probabilities based on the degree of relationship, without taking into account the genotype data, and the posterior IBD including the degree of relationship as well as marker genotype information. The *weighted pairwise correlation statistic (WPC)* (Commenges, 1994; Commenges and Jacqmin-Gadda 1997) is also designed to perform nonparametric analysis on arbitrary large pedigrees. Its basics premise is that under linkage, the correlation of the residuals (the observed trait value minus its expected value) of a pair of relatives will increase with the number of alleles shared IBD. The WPC method can be applied to either quantitative or dichotomous traits.

2.8 Parametric versus nonparametric analysis

When the mode of inheritance is known, such as for monogenic Mendelian traits, model-based maximum likelihood methods applied to pedigrees with multiple generations are the preferred form of analysis, due to their high statistical power. However, model-free methods do have some advantages over model-based parametric analysis for the detection of linkage to complex traits. First, they do not require specification of the disease model, thereby evading the problem of multiple testing causing by analyzing a number of different models. Second, large pedigrees containing multiple affected members are usually rare for complex traits, especially those with late onset. This is due to relatively small recurrence risks for complex diseases. Small pedigrees, such us nuclear families or affected sib pairs, are relatively common and easier to collect.

Chapter 3

ALGORITHMS FOR THE LIKELIHOOD CALCULATIONS

The calculation of the likelihood is essential in linkage analysis, either parametric and nonparametric. In nonparametric linkage analysis, the likelihood is required to estimate the number of alleles shared IBD in affected individuals in large pedigrees, or in the situation of incomplete marker information. Only for pedigrees with a simple structure and with a very few genetic markers the likelihood can be calculated straightforward by hand. Larger pedigrees, multiple markers and untyped individuals increase drastically the number of possible genotypes in Formula 2.9. In these situations it is necessary to use efficient algorithms for the calculation of the likelihood. Thereby, the number of arithmetic operations is to be reduced as much as possible. Likelihood calculations are generally carried out recursively. The data is split into suitable subsets, and calculations are performed on one subset at a time, with the results attached to the next subset, and so on. This allows for large data sets to be processed in a sequential manner. Two types of recursive procedures are in common usage. The first one, the *Elston-Stewart algorithm* (Elston and Stewart, 1971) uses a recursion over members in a pedigree, a procedure in which all loci are considered at once. The second one, the *Lander-Green algorithm* (Lander and Green, 1987) uses recursion over loci, a procedure in which all pedigree members are considered jointly.

3.1 The Elston-Stewart algorithm

The Elston-Stewart algorithm is based on a particular way of handling the pedigree likelihood function, which is expressed as a multiple sum of products of penetrance, populations and transmission parameters over the all possible combinations of genotypes of the pedigree members (Equation 2.9). In the next section the algorithm is described in detail. It follows mostly the same formulations as presented by Strauch (2002).

3.1.1 Peeling from nuclear families

The key feature of the algorithm is to look at the edges of the pedigree for points where the computations should start, in order to limit the number of genotypes to be considered simultaneously. A pedigree without loops can be seen as a sequence of nuclear families with neighbors connected by a single individual i_1, this individual is called pivot and is one of the parents of the nuclear family. So, the

members of a nuclear family are connected through the pivot to the rest of the pedigree (Figure 3.1). The pivot's genotype is denoted by g_{i1} and the partial likelihood $L^*_{i1}(g_{i1})$ is computed. The subindex 1 indicates the first nuclear family. The star in the likelihood indicates that only the individuals of the nuclear family are used and only the genotypes of these members are considered. Then, the likelihood is computed in the way as defined in section 2.6.1. Founders contribute by the population genotype frequencies and non-founders by the transmission probabilities. In addition, the likelihood with the penetrance is multiplied for each individual. In this way $L^*_{i1}(g_{i1})$ is computed for each possible genotype g_{i1} of the individual i_1. The contribution of the likelihoods of the other members are now summarized in $L^*_{i1}(g_{i1})$, and thus the nuclear family is collapsed in its pivot.

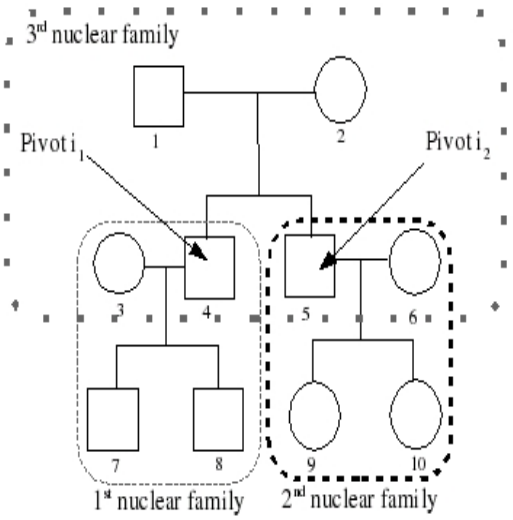


Figure 3.1 *illustrates the Elston-Stewart algorithm. The pivots in this pedigree are individuals 4 (i_1), who links the nuclear family including individuals 3, 4, 7 and 8 to the rest of the pedigree, and individual 5 (i_2) who links the nuclear family including 5, 6, 9 and 10 to the rest of the pedigree. First the likelihood for the first nuclear family, $L^*_{i1}(g_{i1})$* is computed. *This is the likelihood of individuals 3, 7 and 8 conditional on the possible genotypes of the pivot, individual 4. Then the likelihood of the second nuclear family $L^*_{i2}(g_{i2})$ for individuals 6, 9 and 10 conditional to the possible genotypes of its pivot, individual 5. Now the conditional independence of the two nuclear families is used to calculate the likelihood of individuals from 3 to 10 conditional on the possible genotypes of the individuals at the top of the pedigree $L^*_{i3}(g_{i3})$. Finally the three likelihoods are summed up to get the full pedigree likelihood.*

This simplifies the calculations for the next nuclear family which borders with the first nuclear family containing the pivot i_1. The second nuclear family has its own pivot i_2. Again the partial Likelihood $L^*_{i2}(g_{i2})$ is computed for each genotype g_{i2} of the pivot i_2, where all members of the second nuclear family are considered. $L^*_{i2}(g_{i2})$ will contain the contributions of the first and second nuclear families, these contributions are concentrated on pivot i_2. This is the way that the Elston-Stewart algorithm works con-

sidering all possible nuclear families within a pedigree. At the end, the likelihood for the entire pedigree is the summation of all partial likelihoods. The specification of the order of pivots in which the computations are carried out is known as *peeling sequence*, and in many situations the efficiency of the algorithm may be affected by the choice of such sequences (Kong, 1991). An example is illustrated by Figure 3.1. For simplicity the calculations are carried out without specifying maker alleles and affection status.

3.1.2 Pedigree with loops

A loop is said to exist in a pedigree when a path consists of a complete circle, leaving an individual by one line and returning to the same individual by a different line. Two kind of loops can be distinguished in pedigrees. A *consanguinity or inbreeding loop* contains at least two mates who are related (Figure 3.2 A); in contrast, a *marriage loop does not imply* mating of related individuals. A common marriage loop exists when, in two pairs of siblings, each sib of each pair is married to one sib of the other pair (Figure 3.2 B), or when several marriages form a closed circle (Figure 3.2 C).

The presence of loops in pedigrees poses computational problems in likelihood calculation that can be solved by creating an equivalent unlooped pedigree. The extension of the peeling algorithm proposed by Lange and Elston (1975) removes a loop by 'doubling' one of the participating individuals. This virtual doubling consists in fact in creating an exact copy of one of the individuals. The doubled individual is called a loop breaker. This approach requires: (1) identifying individuals participating in loops; (2) selecting a set of the loop breakers; (3) cloning each of them (copying the genetic and phenotypic information) in order to get an unlooped (post-doubling) pedigree.

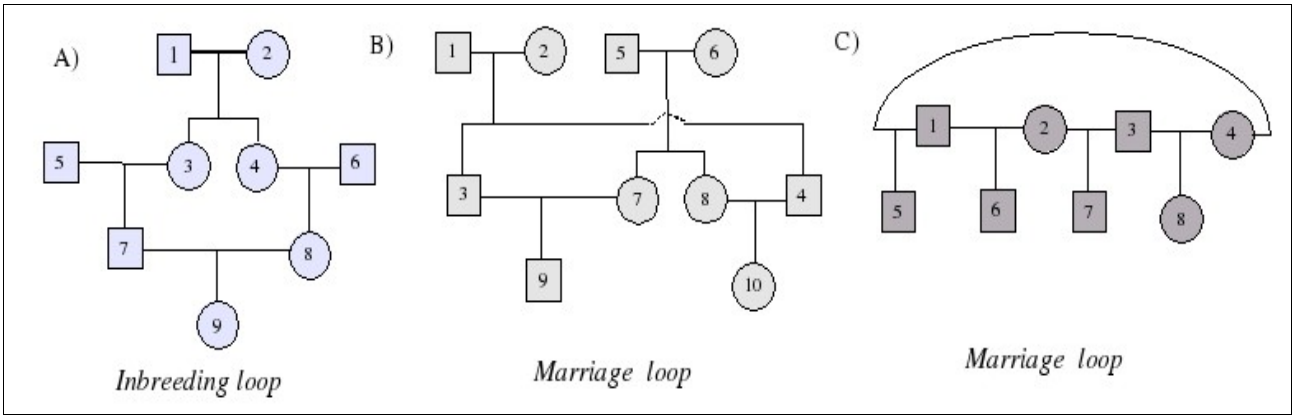


Figure 3.2 *A) illustrates a pedigree with an inbreeding loop. Any of the individuals 3, 4, 7 and 8 could be used as a loop breaker. B) shows a pedigree with a marriage loop, individuals 3, 4, 7 and 8 could be used as loop breaker and C) represents another kind of marriage loop where individuals 1, 2, 3 and 4 could be used as loop breaker. For pedigrees A and B the loop breaker will have once only the parents and once will have only children. For pedigree C the loop breaker will have children once with one partner and once with the other partner.*

Once the loops are broken it is again possible to compute the likelihood for each nuclear family by peeling. It has to be noted that the parameters contributed by the loop breaker, i.e. penetrances and genotype frequencies has to be counted only once. It is recommended to take a genotyped individual as a loop breaker. The selection of loop breakers has a significant impact on the computational efficiency of the likelihood computations.

3.1.3 Properties

The Elston-Stewart algorithm follows the principle of Equation 2.9, which formulates the computation of the likelihood. However, the factors of the individuals are arranged according to nuclear families, and the sums are as far as possible moved toward inside. This corresponds to the so-called peeling method, summing over all possible genotypes of all individuals (except the pivot) of each nuclear family. The Elston-Stewart algorithm is a procedure oriented towards genotypes. It observes all possible genotypes for each individual at each locus and then constructs all possible multi-genotype combinations. It is repeated for each single nuclear family. That makes the complexity of the algorithm to scale linear in the number of individuals, but exponentially in the number of loci. The Elston-Stewart algorithm allows to analyze big pedigrees, but due to the exponential increase in computation time and memory requirements with the number of loci, it can handle a limited number of multiallelic markers.

3.1.4 Implementation

The Elston-Stewart algorithm and its extensions have been implemented in many linkage analysis programs. The early implementations, LIPED (Ott, 1976) and LINKAGE (Lathrop et al., 1984) allow for computation of twopoint and multipoint LOD scores of small pedigrees using few markers. Their successors, such as later versions of LINKAGE, MENDEL (Lange et al., 1988) and FASTLINK (Cottingham et al., 1993) extend the capabilities considerably . The current version of FASTLINK improves the analysis of complex pedigrees by efficient loop-breaking algorithms. Another example of efficient optimizations is VITESSE (O'Connell and Weeks, 1995), which raises the computational boundaries of the Elston-Stewart algorithm and allows for the computation of multipoint LOD scores for several polymorphic markers with many unknown genotypes. All these programs can treat complex pedigrees with many members, but with a limited number of marker loci.

3.2 The Lander-Green algorithm

The Lander-Green algorithm (Lander and Green 1987; Kruglyak et al., 1995; Kruglyak et al., 1996) is based on a special representation of inheritance data. These authors point out that the inheritance of a genetic locus in a pedigree can be completely described by identifying from which grand parental chromosome each child derives its alleles at that locus. They proposed a unified approach to both parametric and nonparametric analysis. The algorithm is constructed in three main steps,

1. Enumeration of all possible *inheritance vectors* in the pedigree.
2. Iterating over inheritance vectors and markers to calculate the probability of the observed genotypes for each marker conditioned on a particular inheritance vector.
3. Finally, to define a statistic to test for linkage given the inheritance vector (which depends only on the nature of the trait).

In the next sections the algorithm is described in detail. It follows mostly the same formulations as presented by Ziegler and König (2006) in their book *"A statistical approach to genetic epidemiology: conceps and applications"*.

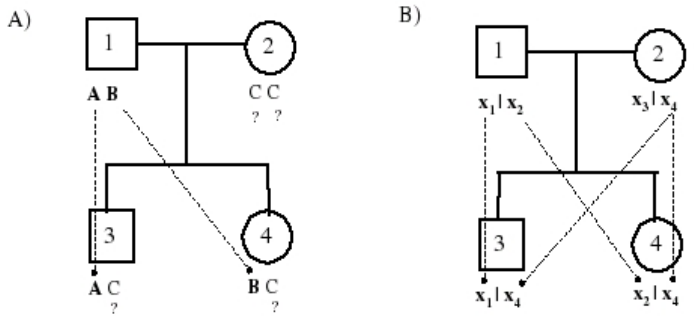
3.2.1 Representing and computing inheritance information

The inheritance vector for a single marker

The first part of the algorithm consists in determining the inheritance pattern for every single marker. For each meiosis in a family, it is possible to determine whether an individual passed the paternal or the maternal inherited allele on its descendants. So, the result of a meiosis can be represented by a bit, which will take the value "0" for the paternal transmission and "1" for the maternal transmission of the inherited allele. More specifically, the inheritance pattern at each locus *j* can be completely described by a binary *inheritance vector* $v(j)=(p_1, m_1, ..., p_i, m_i, ..., p_n, m_n)$, whose coordinates describe the outcome of the paternal and maternal meiosis given rise to the *n* non-founders in the pedigrees. Specially, $p_i = 0$ or 1, depending whether the grandpaternal or grandmaternal allele was transmitted during paternal meiosis giving rise to the *i*th non-founder; m_i carries the same information corresponding to the maternal meiosis. In this way, the inheritance vector specifies which of the $2f$ founder alleles are transmitted to each founder. For *n* non-founders in a pedigree it will be a total of $m=2n$ meiosis and a total of $2^m=2^{2n}$ possible inheritance vectors denoted by *V*. The inheritance vector permits to determine whether common alleles in two or several individuals are IBD or only IBS.

In practice, due to incomplete data, it is not feasible to determine the true inheritance vector at every point. Incomplete data is when some members are unavailable or/and genetic markers have a low heterozygosity, providing only partial information about the inheritance. In both cases the bit describing the meiosis remains undefined. As a consequence, one has to consider to represent partial information from a pedigree by a probability distribution over the possible inheritance vectors at each locus, that is $P(v(j) = w)$, for all inheritance vectors $w \in V$. The concept of inheritance vectors is illustrated in Figure 3.3. The probability $P_{posterior,\,j}(v)$ refers to the conditional probability for the inheritance vector *v* given the genotype information at a marker *j* and is denoted as $P_{Marker\ j}(v)$. It is identical to the relative likelihood of the vector *v*:

$$P_{Marker\,j}(v)=P(V_j=v\,|\,M_j)=\frac{P(M_j\cap V_j=v)}{P(M_j)}=\frac{L_{Marker\,j}(v)}{L_{Marker\,j}}=L_{rel,\,Marker\,j}(v) \qquad (3.1)$$



Inheritance vector	P_{prior}	$P_{posterior}$	*TRUE*
0000	1/16	0	0
0001	1/16	0	0
0010	1/16	1/8	0
0011	1/16	1/8	0
0100	1/16	0	0
0101	1/16	0	0
0110	1/16	1/8	0
0111	1/16	1/8	1
1000	1/16	1/8	0
1001	1/16	1/8	0
1010	1/16	0	0
1011	1/16	0	0
1100	1/16	1/8	0
1101	1/16	1/8	0
1110	1/16	0	0
1111	1/16	0	0

Figure3.3 *Illustration of the inheritance vector and its distribution. In pedigree (A) the phase is unknown and for each founder it is unknown which of the two alleles are from the father and which from the mother. Pedigree (B) is the same pedigree but where the phase is known with the paternal derived allele listed first. The table list the 16 possible inheritance vectors. Bit 1 (and bit 2) correspond to the paternal (and maternal) meiosis from which the son has given rise (x_1 and x_4). The Meiosis of the father and the mother, determining the genotype of the daughter, is represented by bit 3 and bit 4 respectively. The inheritance vectors are formed by '0' and '1' in the way defined in the above paragraph. P_{prior} denotes the distribution before any genotyping has been performed, in this case any vector has the same probability to occur, 1/16. In pedigree (A) one can see that the father (indv. 1) transmits allele A to his daughter and allele B to his son. This information corresponds to bit 1 and bit 3 of the inheritance vector, however, the parental origin of the two alleles of the father is unclear, giving two different possibilities to the vector, bit 1 and bit 3 could correspond to the constellation '0' and '1' or to the constellation '1' and '0'. The mother (indv. 2) is homozigous and it is impossible to known which of the two C alleles have been transmitted to each of the offspring, the mother meiosis in this case is uninformative. So, bit 2 and bit 4 could take any arbitrary value. Finally, conditionally on the genotype information 8 different inheritance vectors could be possible, which according to Mendel's law each has the same probability, 1/8. On the other hand, for pedigree B only one inheritance vector is possible, is the one corresponding to (0111).*

M_j summarizes the data for the jth marker, using the genotypes of all individuals of the pedigree at this marker. Here, V denotes the set of all possible 2^m inheritance vectors and $q_j(v)$ designates the probability of the observed marker data given the inheritance vector v:

$$q_j(v) = P(M_j \mid V_j = v) \tag{3.2}$$

An extreme situations occurs if the family is completely uninformative at a marker j. This happens if all typed individuals in the pedigree are homozygous with the same genotype or the marker is untyped, then the bits of the inheritance vector remain for all m meioses undefined. In the absence of any genotype information, according to Mendel's first law, all inheritance vectors are equally likely with a uniform distribution ($P_{uniform}$), which corresponds to the prior distribution of the inheritance vector:

$$P_{Marker\ j}(v) = \frac{1}{2^m} = P_{prior} \quad \forall \quad v \in V \tag{3.3}$$

As genotype information is added, the probability distribution is concentrated in some concrete inheritance vectors. Sometimes it is possible to reconstruct untyped individuals using the genotype information of relatives. In this way, some bits of the inheritance vector are fixed and this reduces the number of possible vectors. When it is not possible to reconstruct the genotype of an untyped individual, then one needs to consider all possible constellations of genotypes for that individual. All these compatible vectors are used to compute $q_j(v)$. Generally, this case is more complicated than the example illustrated in Figure 3.3 where all individuals are typed and each of the 8 compatible vectors has the same posterior probability.

The probability $q_j(v)$ at a locus j can be determined with the Lander-Green algorithm which treats all 2^m possible vectors separately. The algorithm follows six steps:

1) A graph is generated to create a specific inheritance vector $x=(x_1,...,x_{2f})$ at a given locus j.
2) The number of nodes is identical to the number of founder alleles, $2f$, where f denotes the number of founders.
3) Instead of the observed alleles in the founders, symbolical alleles are used 1,2,...,$2f$.
4) Each individual in the pedigree is represented by an edge connecting his/her two symbolical alleles based on the specific inheritance vector v. Note that two nodes are connected by several edges if the inherited parental genotypes of several offspring are identical.

5) The observed alleles are matched to the symbolical alleles, complying with Mendelian laws. If an individual is genotyped, the edge is marked with the two observed alleles. For these individuals, the observed allele has to appear at all edges that are connected to the respective node. This assignment determines the orientation of the genotype to its neighboring edge. It also determines the node of the second allele of the genotype.

6) If an assignment leads to an incompatibility, the orientation of the genotype has to be altered in the first step. If the altered orientation also leads to an incompatibility at any position in the graph, the inheritance vector has probability zero given the marker data.

These steps are taken for all possible inheritance vectors, and the probability $q_j(v)$ is determined by using all compatible inheritance vectors. However, instead of using equal weights for all compatible inheritance vectors, they are weighted by their allele frequencies:

$$q_j(v) = P(M_j \mid V_j = v) = \sum_C \prod_{l=1}^{2f} f^j(a(x_l)) \tag{3.4}$$

where C=*{all founder-alleles compatible with v and all individuals genotypes}*. $a(x_l)$ denotes the relationship of the symbolical allele x_l to the observed alleles of the founders l=1,...,2f. In the case of a heterozygous genotype a factor 2 is not applied because an exchange of a paternal and maternal inherited allele would correspond to a different assignment. Since the inheritance vector gives the sample transmissions, the conditional probability $P(M_j \mid V_j=v)$ contains, in difference to the partial likelihood of the vector v, only the genotype frequencies of the founders and not the probabilities of transmission. So, $q_j(v)$ differs from the partial likelihood by a factor 2^m, which corresponds to the inverse of the prior probability for each vector:

$$P_{Marker\,j}(v) = P(V_j = v \mid M_j) = \frac{P(M_j \cap V_j = v)}{P(V_j = v)} = \frac{L_{Marker\,j}(v)}{P_{prior}} = L_{Marker\,j}(v) \cdot 2^m \tag{3.5}$$

An example to understand the concept of the founder-graph algorithm is illustrated in Figure 3.4. It describes the pedigree presented in Figure 3.3 for the calculation of $P(M_j \mid V_j=v)$ and $P_{Marker\,j}(v)$. The algorithm is also applicable to more complex pedigrees with loops.

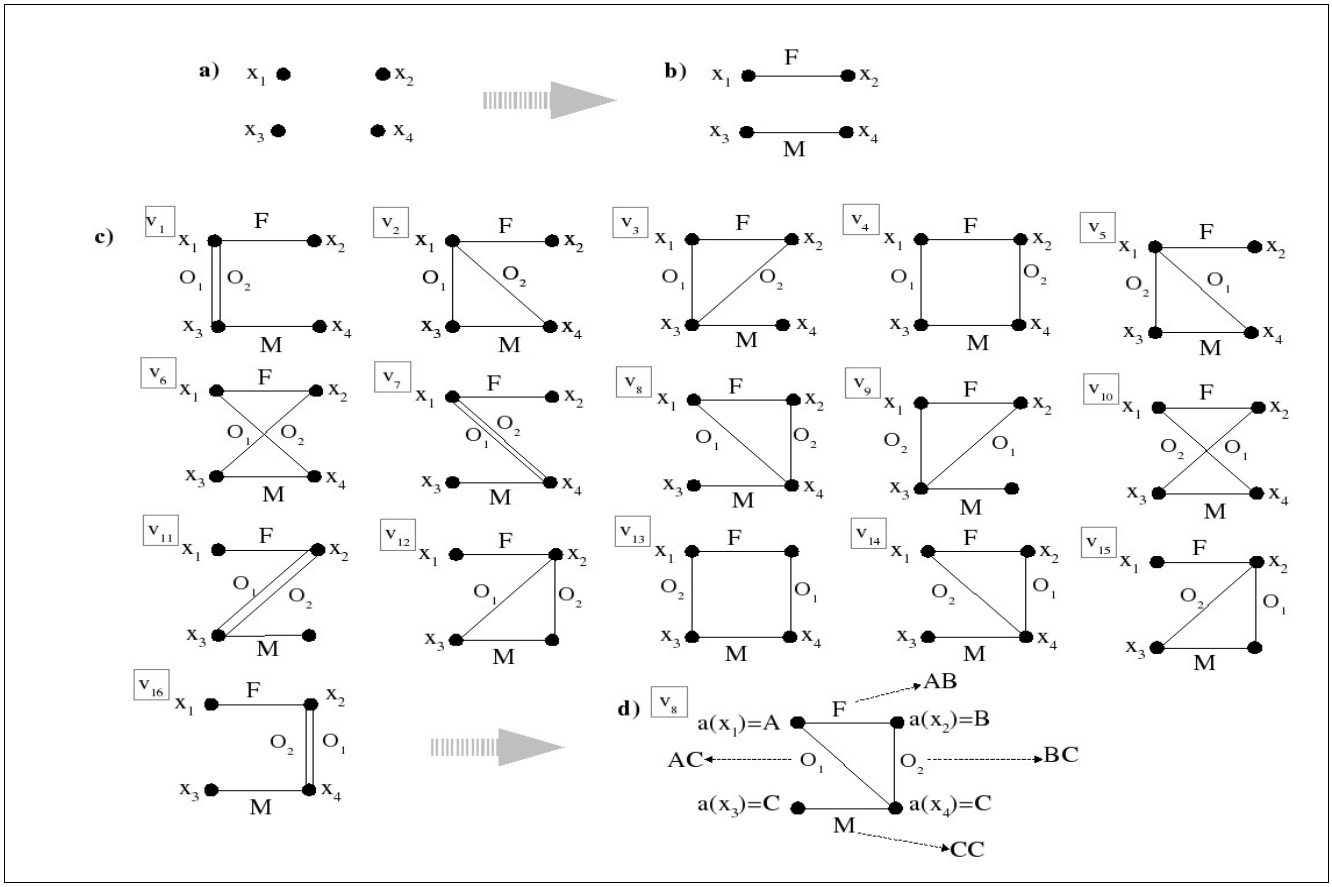


Figure 3.4 *Pedigree from Figure 3.3A has two founders. Therefore, there are 4 symbolical alleles x_1. a) shows the corresponding nodes for the symbolical alleles of the graph. The father has alleles x_1x_2 and the mother has alleles x_3x_4. b) displays the nodes and the parental edges denoted by M (mother) and F (father). c) shows all nodes and edges for the 16 theoretical possible inheritance vectors using the symbolical alleles x_1 to x_4. O_1 and O_2 along the edges denote offspring 1 and offspring 2. We already know from table of Figure 3.3 which inheritance vectors are compatible with the pedigree. d) displays the distribution of the symbolical alleles to the observed alleles, which is clear in this case considering the pedigree from Figure 3.3B. An edge connect allele x_1 and x_4 because they both pass through O_1 whose genotype is observed, then the allele x_4 is transmitted to O_2 jointly with allele x_2 from the father and so on. The probability of the marker j given vector v_8 is the product of the founder alleles for this assignment, i.e. $q_j(v)=f^j(A) \cdot f^j(B) \cdot f^j(C)^2$.*
For the pedigree in Figure 3.3A an analog performance need to be done for the other 7 compatibly vectors (see table from Figure 3.3). In this example, all eight compatible inheritance vectors have the same probability $P_{Marker\ j}(v)=1/8$. In general, the posterior probability is not the same for all compatible vectors; in particular if genotypes are missing.

The inheritance vector for multiple markers

The algorithm described above can be used to estimate the probability $P_{Marker\ j}(v)=P(V_j=v \mid M_j)$ of the inheritance vector at any marker position j for all inheritance vectors v. However, to extract the full information from the data set one should calculate the inheritance distribution conditional on the genotypes at all marker loci ($P_{complete}$), i.e., the inheritance distribution at an arbitrary chromosomal position x:

$$P_{complete,x}(v) = P(V(x) = v \mid M_{1,..}, M_k) \tag{3.6}$$

for markers $M_1,...,M_k$, which are ordered according to their order on the chromosome. To calculate $P_{complete,x}(v)$ one needs to use a Markov property. The Markov property relies on the assumption of absence of interference. With this assumption, the inheritance pattern of a pedigree at all positions can be described by a Markov chain along the markers of a chromosome. The observed states of the Markov chain correspond to the genotypes. The inheritance vectors are, however, hidden. Therefore, this Markov chain can be interpreted as a *Hidden Markov Model* (HMM), were the hidden states, i.e., the inheritance vectors, can be partly reconstructed using the observed states, i.e, the genotypes. This idea for connecting observed and hidden states has already been illustrated for the single marker case, where $q_j(v)$ has been calculated using the founder-graph approach. In the multi-marker case, one considers inheritance vectors v_j and v_{j+1} at marker j and j+1. If the inheritance vectors v_j and v_{j+1} are different, at least one transition has occurred between markers j and j+1. Specifically, the difference in a single bit between neighboring markers indicates a change in the inheritance patterns that is due to a recombination in the interval between the two markers. The Hamming distance $H(v_j, v_{j+1})$ measures the total number of recombinations for all meiosis m in a pedigree for inheritance vectors v_j and v_{j+1}. Thus, the H counts the number of bits being different between v_j and v_{j+1}. The transition probability from inheritance vector v_j at marker j to v_{j+1} at marker j+1 is given by:

$$\begin{aligned} T_{v_j, v_j+1}(\theta_{j,j+1}) &= P(V_{j+1} = v_{j+1} \mid V_j = v_j) \\ &= (\theta_{j,j+1})^{H(v_j, v_{j+1})} (1 - \theta_{j,j+1})^{m - H(v_j, v_{j+1})} \end{aligned} \tag{3.7}$$

where $\theta_{j,j+1}$ denotes the recombination fraction between markers j and j+1. These probabilities for all pairs (v_j, v_{j+1}) form the (2^mx2^m)transition matrix $\boldsymbol{T}(\theta_{j,j+1})$ from marker j to j+1 with j and j+1 denoting rows and columns, respectively. Due to the next relationship, the matrix $\boldsymbol{T}(\theta_{j,j+1})$ is symmetric:

$$P(V_{j+1} = v_{j+1} \mid V_j = v_j) = P(V_j = v_j \mid V_{j+1} = v_{j+1}) \tag{3.8}$$

Let's denote $\boldsymbol{p}_j^L$ the (2^mx1) vector of probabilities $p_j^L(v_j)$ at marker j given all observed genotypes of markers 1,..., j:

$$[\boldsymbol{p}_j^{\boldsymbol{L}}]_{v_j} = p_j^L(v_j) = P(V_j = v_j \mid M_1, \ldots, M_j) \tag{3.9}$$

The superscript L indicates that the probabilities at marker j are conditioned to all the left markers. For the first marker, one obtains

$$[\boldsymbol{p}_j^{\boldsymbol{L}}]_{v_1} = P(V_1 = v_1 \mid M_1) = P_{Marker_1}(v_1) \tag{3.10}$$

Furthermore, let $\boldsymbol{q}_j$ be the (2^mx1) column vector of probabilities $q_j(v_j)$ for all v_j at marker j:

$$[\boldsymbol{q}_j]_{v_j} = q_j(v_j) = P(M_j \mid V_1 = v_j) \tag{3.11}$$

Note that q_j is conditioned on $V_j = v_j$ in contrast to Equation (3.9) where the condition is on the markers $M_1,\ldots,M_j$. With the notations from above and "$\circ$" denoting the *Hadamard product* of element by element multiplication (i.e., if $\boldsymbol{a}$ and $\boldsymbol{b}$ are two vectors $\Rightarrow (\boldsymbol{a} \circ \boldsymbol{b})_v = \boldsymbol{a}_v \boldsymbol{b}_v$), $\boldsymbol{p}_j^{\boldsymbol{L}}$ can be calculated for $j=2,\ldots,K$ recursively (Lander and Green 1987) as:

$$\boldsymbol{p}_{j+1}^{\boldsymbol{L}} = \frac{(\boldsymbol{p}_j^{\boldsymbol{L}})^T \boldsymbol{T}(\theta_{j,j+1}) \circ \boldsymbol{q}_{j+1}}{(\boldsymbol{p}_j^{\boldsymbol{L}})^T \boldsymbol{T}(\theta_{j,j+1}) \boldsymbol{q}_{j+1}} \tag{3.12}$$

The validity of Equation (3.12) can be proven by writing $\boldsymbol{p}_{j+1}^{\boldsymbol{L}}$ element-wise (Appendix, A2)

Analogously to $\boldsymbol{p}_j^L$ we now define a (2^mx1) vector $\boldsymbol{p}_j^R$ comprising the probabilities $\boldsymbol{p}_j^R(v_j)$ for all inheritance vectors at marker j conditional on the observed data at markers $j,\ldots,k$.

$$[\boldsymbol{p}_j^{\boldsymbol{R}}]_{v_j} = p_j^R(v_j) = P(V_j = v_j \mid M_1, \ldots, M_k) \tag{3.13}$$

Here, the superscript R indicates that the probabilities at the marker j are conditioned to the "right". For the right most marker k, one specially obtains $[\boldsymbol{p}_k^{\boldsymbol{R}}]_{v_k} = P_{Marker\,k}(v_k)$. The recursion formula, which is analogous to Equation (3.12) is:

$$\boldsymbol{p}_{j-1}^{\boldsymbol{R}} = \frac{(\boldsymbol{q}_{j-1})^T \circ \boldsymbol{T}(\theta_{j-1,j})\, \boldsymbol{p}_j^{\boldsymbol{R}}}{(\boldsymbol{q}_{j-1})^T \boldsymbol{T}(\theta_{j-1,j})\, \boldsymbol{p}_j^{\boldsymbol{R}}} \tag{3.14}$$

In total, one obtains the following (2^mx1) column vector:

$$(P_{complete,\,j})_{v_j} = P_{complete,\,j}(v_j) = P(V_j = v_j \mid M_1, \ldots, M_k) \tag{3.15}$$

of the inheritance vectors at marker j given the genotypes at all markers as

$$\boldsymbol{P}_{complete,\,j} = \frac{(\boldsymbol{p}_{j-1}^{L})^T \boldsymbol{T}(\theta_{j-1,j}) \circ \boldsymbol{p}_j^{R}}{(\boldsymbol{p}_{j-1}^{L})^T \boldsymbol{T}(\theta_{j-1,j})\, \boldsymbol{p}_j^{R}} \tag{3.16}$$

The validity of Equation (3.16) is proven in Appendix A3. Instead of using Equation (3.16), $\boldsymbol{P}_{complete,\,j}$ can also be obtained by

$$\boldsymbol{P}_{complete,\,j} = \frac{(\boldsymbol{p}_j^{L})^T \circ \boldsymbol{T}(\theta_{j,j+1})\, \boldsymbol{p}_{j+1}^{R}}{(\boldsymbol{p}_j^{L})^T \boldsymbol{T}(\theta_{j,j+1})\, \boldsymbol{p}_{j+1}^{R}} \tag{3.17}$$

Equations (3.16) and (3.17) are the original formulations for calculating $\boldsymbol{P}_{complete,\,j}$. The recursive nature of these formulas is, however, somehow awkward. As with most algorithms, the recursion can be dissolved and a direct computation can be used instead (Kruglyak et al., 1995). This forward approach requires only the transmission matrices $\boldsymbol{T}(\theta_{j,j+1})$ and a posteriori probabilities of the inheritance vectors $q_j(v_j)$ at markers j=1,...,k. In the following, let $\mathbf{Q}_j$ = diag($q_j(v_j)$) be the 2^mx2^m diagonal matrix at marker j=1,...,k. Furthermore, let $\mathbf{1}$=(1,...,1)^T denote the 2^mx1 column vector. Then, the probability $\boldsymbol{P}_{complete,\,j}$ can also be written as:

$$\boldsymbol{P}_{complete,\,j} = \frac{\mathbf{1}^T \boldsymbol{Q}_1 \boldsymbol{T}(\theta_{1,2}) \boldsymbol{Q}_2 \ldots \boldsymbol{T}(\theta_{j-1,j}) \circ \boldsymbol{Q}_j \boldsymbol{T}(\theta_{j,j+1}) \ldots \boldsymbol{Q}_{k-1} \boldsymbol{T}(\theta_{k-1,k}) \boldsymbol{Q}_k \mathbf{1}}{\mathbf{1}^T \boldsymbol{Q}_1 \boldsymbol{T}(\theta_{1,2}) \boldsymbol{Q}_2 \ldots \boldsymbol{T}(\theta_{j-1,j}) \boldsymbol{Q}_j \boldsymbol{T}(\theta_{j,j+1}) \ldots \boldsymbol{Q}_{k-1} \boldsymbol{T}(\theta_{k-1,k}) \boldsymbol{Q}_k \mathbf{1}} \tag{3.18}$$

The Hadamard product could also be used on the right side of $\mathbf{Q}_j$ because $\mathbf{Q}_j$ is diagonal. The validity of Equation (3.18) is proven in Appendix A4.

The previous formulations for $P_{complete,\,j}(v_j)$ are formulated for chromosomal positions that are identical to genetic marker positions. The algorithm can be extended to compute $P_{complete,\,x}(v)$ at an arbitrary chromosomal position x that may be different from a marker position j. The idea is that a non-informative genetic marker at position x is added to the Markov chain (see Ziegler and König, 2006).

Information content

In studying a pedigree, it is useful to known how much of the total inheritance information can be extracted at each point of the genome by the typed markers. This information is carried by the measure called "*information content"* (IC) (Kruglyak and Lander 1995). IC provides a measure of how closely a given study approaches the goal of completely determining the inheritance outcome, and it shows where typing additional markers is most useful.

The classical information-theoretic measure of residual uncertainty in a probability distribution is its entropy, defined by:

$$E = -\sum_{w \in V} P(V(x)=w)\log_2 P(V(x)=w) \tag{3.19}$$

where $\log_2$ is used in order for the entropy to be measured in bits (Shannon 1948). The entropy of the probability distribution over inheritance vectors thus naturally reflects information content. In the absence of genotype data, the probability distribution is uniform over all 2^{2n-f} equivalent classes of inheritance vectors and the entropy is easily seen to be $E= 2n-f$ bits. If the inheritance vector is known with certainty, the probability distribution is completely concentrated on a single outcome. The entropy is thus $E=0$. The IC of the inheritance pattern at point x will be defined by:

$$I_E(x) = 1 - \frac{E(x)}{E_0} \tag{3.20}$$

where $E(x)$ is the entropy of the multipoint inheritance distribution at locus x and where $E_0= 2n-f$ bits is the entropy in the absence of genotype data. $I_E(x)=1$ indicates total informativeness at x, otherwise $I_E(x)=0$ indicates total uncertainty about inheritance in the pedigree at locus x. Entropy is an additive measure, it can be summed up over all pedigrees in the data set. Equation (3.20) is then used with total entropy to obtain the overall information content of a study.

3.2.2 Evaluation of the positions on the basis of the disease phenotypes

Once the inheritance distribution of $P(V(x)=v)$ at any genetic position x has been computed for a given set of markers, the phenotypes of the individuals have to be included. This can be done by specifying a scoring function $S(v, \phi^d)$ that depends on the inheritance vector v and the observed phenotype $\phi^d=(\phi_1^d, \ldots, \phi_n^d)$ for all the individuals in the pedigree. In this context it is possible to use parametric and nonparametric approaches. In the general case multiple inheritance vectors are possible. So, one can

generalize the scoring function by taking its expected value over the inheritance distribution:

$$\bar{S}(x,\phi^d)=\sum_{w\in V} S(w,\phi^d)P(v(x)=w) \tag{3.21}$$

The next two sections present parametric and nonparametric linkage analysis using the concept of inheritance vectors. The same formulation as used by Strauch (2002) are introduced.

Parametric linkage analysis

In parametric analysis the scoring function $S(v,\phi^d)$ is determined by the likelihood ratio *LR(v)* for the inheritance vector *v (*Strauch, 2002):

$$S(v,\phi^d)=LR(v)=\frac{P(\phi^d|v)}{\sum_{w\in V} P(\phi^d|w)P_{priori}(w)} \tag{3.22}$$

$P(\phi^d|v)$ is the likelihood at the disease locus given the inheritance vector *v*. The denominator designates the likelihood under H_0: disease locus and set of markers are unlinked. This corresponds, as shown above, to an uniform distribution of the inheritance vectors. The likelihood at the disease locus conditional no the inheritance vector can be expressed as follows:

$$P(\phi^d|v)=\sum_{g^d} P(g^d|v)Pen(\phi^d|g^d)=\sum_{g^d} P(g^d|v)\prod_{i=1}^{n} Pen(\phi_i^d|g_i^d) \tag{3.23}$$

where $g^d=(g_1^d,...,g_n^d)$ is a combination of genotypes at the disease locus from all the individuals in the pedigree. The sum is considered through all possible genotype combinations. $P(g^d|v)$ filters the genotypes, which are compatible with the inheritance vector *v*, and contains the disease allele frequency *f(m)*. The last term $Pen(\phi^d|g^d)$ is the product of penetrances $Pen(\phi_i^d|g_i^d)$ for each person *i,* it can be multiplied since the genotype of an individual does not affect the phenotype of other individuals. $P(\phi^d|v)$ will be determined using the peeling method, i.e. by the application of the Elston-Stewart algorithm. The expected value of the parametric scoring function $S(v,\phi^d)$ results in the likelihood ratio *LR(x)* for H_A: disease locus and the marker loci are linked vs H_0: disease locus and marker loci are unlinked (see Appendix A5).

$$\bar{S}(x,\phi^d) = \frac{L(x)}{L(x\,unlinked)} = LR(x) \tag{3.24}$$

The numerator of the equation designates the entire likelihood *L(x)*, as the probability for the observation of all markers and phenotypes, given that the disease locus is at position *x*. The denominator is the entire likelihood under the null hypothesis of no linkage between the set of markers and the disease locus. Using Equation (2.13) defined in the previous chapter for multipoint analysis one obtains:

$$\begin{aligned} Z(x) &= \log_{10} LR(x) = \log_{10} \frac{L(x)}{L(x\,unlinked)} \\ &= \log_{10} \frac{\sum_{w \in V} P(\phi^d|w)\, P(V(x)=w)}{\sum_{w \in V} P(\phi^d|w)\, P_{priori}(w)} \end{aligned} \tag{3.25}$$

As demonstrated in the previous chapter, LOD scores from different pedigrees can be added, one needs only to sum them up in order to obtain the LOD score of the entire sample.

Nonparametric linkage analysis

Using the context of the Lander-Green algorithm it is simple to conduct nonparametric linkage analysis, i.e., NPL analysis. This is because with the symbolical alleles the inheritance vector represents the different founder alleles, which has been inherited by each individual in a pedigree. So, it is possible to read directly from the inheritance vector, how many alleles IBD share two or several affected individuals. In addition, this has the great advantage that alleles IBD are easy to distinguish from alleles IBS. The scoring function $S(v,\phi^d)$ is a measure for the alleles shared between the individuals assigned for a certain vector *v*. The NPL score in a given genetic position *x* is again the expected value of the nonparametric scoring function relative to the inheritance distribution *P(V(x)=v)*.

Kruglyak et al. (1996) used two different definitions of $S(v,\phi^d)$ for nonparametric linkage analysis, S_{pairs} and S_{all}. The function S_{pairs} sums up the number of common founder-alleles shared IBD between two individuals given an inheritance vector *v* over all possible combination of pairs in the pedigree. The score function S_{pairs}, first suggested by Fimmers et al. (1989), is given by:

$$S_{pairs}(v,\phi^d)=\sum_{1\le i<i'\le n} s_{i,i'} \tag{3.26}$$

where $s_{i,i'}$ is one-forth the number of alleles shared IBD by the pair of individuals i and i'. It can be written using the following formula:

$$s_{i,i'}=\frac{1}{4}[\delta(a_{i,1},a_{i',1})+\delta(a_{i,1},a_{i',2})+\delta(a_{i,2},a_{i',1})+\delta(a_{i,2},a_{i',2})] \tag{3.27}$$

where $a_{i,1}$, $a_{i,2}$ and $a_{i',1}$, $a_{i',2}$ are the different founder-alleles of individuals i and i' respectively given for a concrete inheritance vector. δ is the Kronecker's delta function defined as $\delta=\{1 if\ i\neq i', 0\ \ if\ i=i'\}$. $s_{i,i'}$ takes the values 0, ¼ and ½, corresponding to zero, one and two alleles IBD respectively. The value ½ can only occur when a pair shares 2 alleles IBD, so it is only possible in the case of sibling pairs.

An alternative score function, S_{all}, was proposed by Whittemore and Halpern (1994b). The statistical power can be increased by considering larger sets of affected relatives, rather than just pairs. They proposed a statistic to capture the allele sharing associated with a given inheritance vector. In addition, one allele IBD that occur in more than two affected individuals, gives an strong evidence for linkage and gets a larger weight. Let a denote the number of affected individuals in a pedigree, and h a collection of alleles obtained by choosing one allele from each of these affected individuals, and let $b_i(h)$ denote the number of times that i-th founder allele appears in h (for i=1,...,2f). The score function S_{all} is defined as:

$$S_{all}(v,\phi^d)=\frac{1}{2^a}\sum_{h}\prod_{i=1}^{2f} b_i(h)! \tag{3.28}$$

where the sum is taken over the 2^a possible ways to choose h. In effect, the score is the average number of permutations that preserve a collection obtained by choosing one allele from each affected person. It gives sharply increasing weight as the number of affected individuals sharing a particular allele increases. For affected sib pairs, S_{all} and S_{pair}, provide the same results. Whittemore and Halpern showed by using simulations that the NPL considering S_{all} has a higher power with recessive inheritance diseases than S_{pairs}. However, for dominant inheritance S_{pairs} becomes more powerful than S_{all}. Obviously, the computation of S_{all} is substantially more complex than S_{pairs}, in particular, when the number of affected individuals in a pedigree is large. For the evaluation of the significance one assumes that the inheri-

tance sample of a pedigree could be clearly restored and only one possible inheritance vector exists. Then, for either approach a normalized score can be defined:

$$Z(v,\phi^d)=\frac{[S(v,\phi^d)-\mu]}{\sigma} \tag{3.29}$$

where μ *and* σ are the mean and standard deviation of $S(v,\phi^d)$ under H_0, i.e. P_{priori}. Under the null hypothesis of no linkage, the normalized $Z(v,\phi^d)$ has mean 0 and variance 1. For the combined scores among *t* pedigrees, one can take a linear combination:

$$Z=\sum_{s=1}^{t}\gamma_s Z_s \tag{3.30}$$

where Z_s denotes the normalized score for the *s*-th pedigree, and γ_s are weighting factors. The weighting factors should be chosen so that $\Sigma_s \gamma_s^2=1$, so that *Z* has mean 0 and variance 1 under the null hypothesis of no linkage. The statistic *Z* is referred as the NPL score.

Statistical significance

Suppose that an analysis involving one or more pedigrees yields an observed NPL statistic of Z_{obs}. For the significance level there are two simple approaches:

1. *Exact distribution.* It is straightforward to compute the exact probability distribution of the overall score *Z* under the null hypothesis of no linkage. Specifically, one can calculate the distribution for each pedigree by enumerating all possible inheritance vectors; the distribution for the collection of pedigrees is then obtained by convolving these distributions. One can then simply look up the exact value, $P(Z \geq Z_{obs})$.
2. *Normal approximation.* Under the null hypothesis of no linkage, the score *Z* will tend toward a standard normal variable as one studies many similar pedigrees. (This follows from the central limit theorem, since *Z* is an appropriately normalized sum of independent random variables). The significant level of an observation Z_{obs} can then be approximated by consulting a table of tail probabilities for the standard normal. Although less precise than the exact distribution, the normal approximation is useful in some settings.

3.2.3 Properties

The Lander-Green algorithm computes the likelihood indicated in Equation (2.9); it uses for it however, completely different logic than the Elston-Stewart algorithm. Marker loci and disease locus are considered independently. First, the inheritance vector is determined for all individuals in a pedigree, using first each marker and then all markers together. Finally, the phenotypes are included. The terms of the likelihood are not arranged according to individuals or nuclear families, but rather according to loci.

Overall, one can say that the Lander-Green algorithm is not genotype-oriented but rather to inheritance vector oriented. The genotypes from all individuals in the pedigree are considered at each locus, and from this the compatible inheritance vectors are constructed. According to this principle one locus is treated after the other.

The runtime and memory requirements of the Lander-Green algorithm scale exponentially with the number of individuals, but linear with the number of loci. A crucial measure for family size is the effective number of the bits of the inheritance vectors, thus $2n-f$. In a pedigree with f=3 founders and n=8 non-founders, there are 13 effective bits and 8192 distinguishable inheritance vectors. With 6 founders and 16 non-founders, i.e. 26 effective bits, and one needs to consider over 60 millions inheritance vectors.

Some alternative formulations, similar to the concept of inheritance vectors of Lander-Green, have been suggested by Sobel and Lange (1993), Thompson (1994), Whittemore und Halpern (1994a, 1994b) and Guo (1995).

3.2.4 Implementations

The original version of the Lander-Green algorithm was first implemented in the software package CRI-MAP (Lander and Green 1987). This was followed by the linkage analysis program MAPMAKER containing several components for special applications. MAPMAKER/EXP (Lander et al., 1987) permits the creation of genetic maps. MAPMAKER/HOMOZ (Kruglyak et al., 1995) accomplishes the so-called *homozygosity mapping*, it can effectively map recessive diseases in case of inbreeding. One looks for homozygous individuals having two copies of the allele from the same origin (*homozygosity by descent)*. MAPMAKER/QTL (Lander et al., 1987) handles quantitative traits. A nonparametric approach using affected sib pairs is implemented in MAPMAKER/SIBS (Kruglyak and Lander 1995), it can handle qualitative (categorical) and quantitative traits. Since then several improvements have been made on implementations of this algorithm towards more efficient use of computer re-

sources (runtime and fast memory). Kruglyak et. al. (1996) improved this result by recognizing that in ungenotyped founders, there is no way of distinguishing between the maternal and paternal genes. Thus inheritance patterns that differ only by phase changes in the founders are completely equivalent in the sense that they have the same probability independent of the genotypes. By treating such an equivalence class as one pattern, i.e. using founder reduction, the time (and space) complexity reduces by a factor of 2^f where f is the number of founders in the pedigree. This version of the algorithm was incorporated in the software package GENHUNTER. Idury and Elston (1997) put forward a version where they explore the regularity of the transition matrix by writing it as a Kronecker product of simple basic matrices further improving the runtime. Kruglyak and Lander (1998) suggested to use Fast Fourier Transform (FFT) for additive groups to reduce the complexity. By incorporating the founder reduction they obtained a version of the algorithm with the same time complexity as the Idury-Elston version with founder reduction. Further, Gudbjartsson et al. (2000) improved on the performance of the Lander-Green algorithm by a new technique called *founder couple reduction.* Symmetry between founder couples is used to further reduce the size of inheritance vectors and gain additional speed. Furthermore they implemented the single point calculation in a top down manner such that inconsistencies in marker segregation are rapidly detected. Their approach is based on the FFTs and is implemented in the software package ALLEGRO. The linkage software MERLIN (Abecasis et al., 2001) is based on the Idury-Elston algorithm but the inheritance vectors are represented as sparse binary trees.

3.3 Comparison of the algorithms

The Elston-Stewart algorithm and the Lander-Green algorithm have different requirements for computing time and fast memory usage. The complexity of the Elston-Stewart algorithm scales linear in the number of individuals but exponential in the number of markers. On the other hand, the complexity of the Lander-Green algorithm increases linearly in the number of markers but exponentially in the number of individuals in the pedigree. The Elston-Stewart algorithm is therefore appropriate for LOD score analysis with few markers and large pedigrees with none ore few loops. However, for true multipoint analysis the Lander-Green algorithm has to be applied, which can accommodate multiple markers at the same time. Although, here the pedigrees must not exceed a certain size. Within the framework of genome scans involving genetically complex disease, usually pedigrees of small to medium size are recruited. In this case, the Lander-Green algorithm is a good choice.

In conclusion of this chapter it should be noted that there are also different possibilities to compute the likelihood in an approximate manner. This offers the solution for large pedigrees with many

markers to be analyzed, or for pedigrees with several loops. The difficulty of such algorithms is that they do not guarantee a convergence of the likelihood to the accurate value. There exists a r*andom-walk* method (Lange and Sobel 1991), a *sequential-imputation* procedure (Kong et al., 1993; Irvin et al., 1994) as well as different *Monte-Carlo* methods (Sheehan 1989; Thompson and Wijsman 1990; Kong 1991; Thompson and Guo 1991; Guo and Thompson 1992). Another algorithm for approximate computation of the likelihood, based on pedigrees with loops, is given by Stricker et al. (1995). A *blocking Gibbs sampling* method (Jensen and Kong 1999) allows for linkage analysis with large pedigrees and a large number of loops.

Chapter 4

LINKAGE GENETIC MAPS

In chapter 3, methods for estimating the recombination fraction between loci (θ) have been introduced and discussed. One of the most important objectives of estimating θ is to make a genetic map. Traditionally, a linkage relationship among two genes is quantified using the estimated recombination fraction, $\hat{\theta}$, and its transform into a genetic distance by a mapping function. However, twopoint analysis may not be efficient for more than two loci. Multipoint analysis is more accurate. For example, information on intervals between loci *A* and *B* and between *B* and *C* is included in estimating genetic distance between *A* and *C* if the locus order is *ABC*.

4.1 Map and physical distances

The relationship between genetic distance and physical distance is confused. The ultimate physical distance between two genes can be quantified using the number of DNA base pairs between the two genes. The genetic distance is based on a statistical estimate based on crossover events. Genetic distance differ from species to species and can be different in the sexes of the same specie. Even within a single specie, the genetic distance could differ greatly according to genome location. In humans for example, crossovers are more frequent in females than in males meiosis. In each linkage map, females distances are greater than male distances, providing evidence for a relative increase in female recombination across the human genome. Usually, females genetic maps are significantly longer than males genetic maps. One exception to this rule are the pseudoautosomal regions in the sex chromosomes. In these regions it has been observed that males exhibit a much higher number of recombination than females.

The relationship between genetic distance and physical distance may vary greatly at the genome segment level. For example, seven genetic markers are shown on a genome segment (Figure 4.1). The markers may be evenly spread on the genetic map but not on the physical map (Figure 4.1A), or they could be spread evenly on a physical map but not on a genetic map (Figure 4.1B).

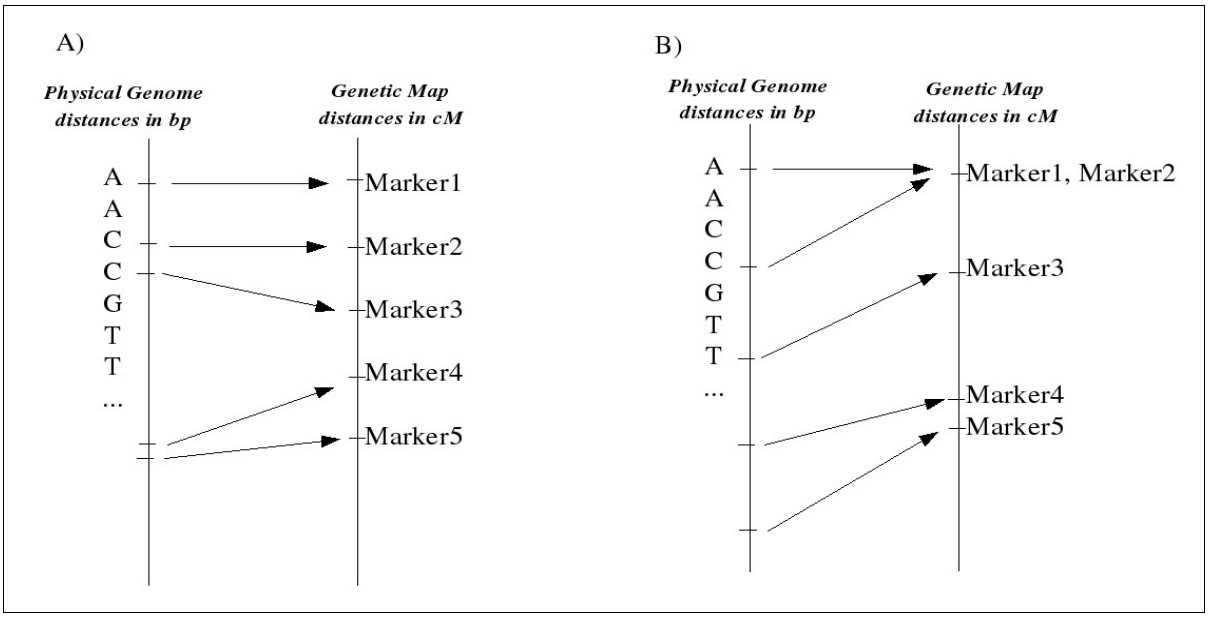


Figure 4.1 *illustrates two hypothetical situations to show the relationship between genetic distance (cM) and physical distance (bp).* ***A****: Equal genetic map distance may corresponds to different physical distances.* ***B****: Different genetic distances (even a distance of 0) may correspond to equal physical dis tance.*

4.2 Multipoint genetic maps models

For simplicity let's consider a three locus model with diallelic loci *A, B* and *C*. For three linked loci there are three possible recombinations fractions θ_{AB}, θ_{BC} *and* θ_{AC} (Figure 4.2).

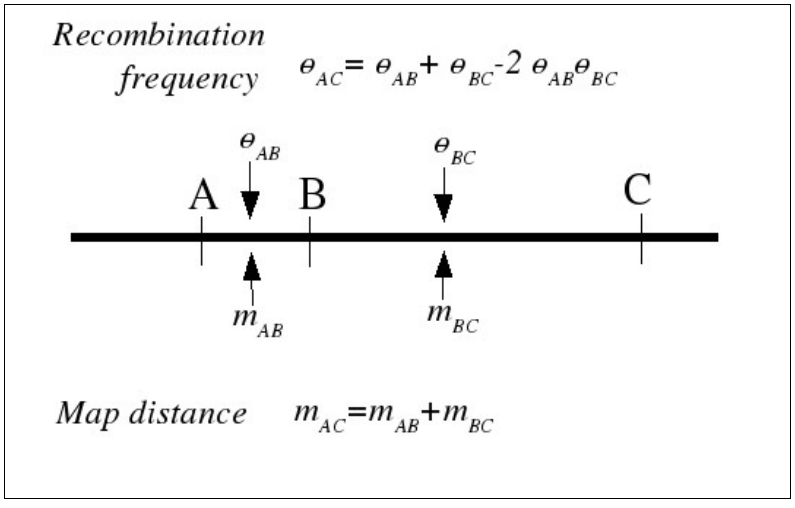


Figure 4.2 *Relationship among three ordered loci A, B and C are quantified using both pairwise recombination fraction and physical distance.*

If the three loci are in order *ABC* on the chromosome and crossovers are completely at random, then the relationship of the three recombination fractions are:

$$\theta_{AC} = \theta_{AB} + \theta_{BC} - 2\theta_{AB}\theta_{BC} \tag{4.1}$$

where $2\theta_{AB}\theta_{BC}$ is expected to be the double crossover frequency between *A* and *B* and *B* and *C* simultaneously. However, departure from this expectation has been observed. This departure is the interference defined on chapter 1. This phenomenon is quantified by adding a coefficient to Equation (4.1)

$$\theta_{AC} = \theta_{AB} + \theta_{BC} - C\,2\theta_{AB}\theta_{BC} \tag{4.2}$$

where *C* is defined as coefficient of coincidence and 1-*C* is defined as interference.

If there is no interference and crossover occurs randomly, then the expected double recombinant frequency will be $2\theta_{AB}\theta_{BC}$, *C*=1 and *Interference*=1-*C*=0. If crossovers in interval *AB* and *BC* are not independent, the observed double recombinant frequency may not be equal the expectation. If we use r_{12} to denote the true double recombinant frequency, the coefficient of coincidence is:

$$C = \frac{r_{12}}{2\theta_{AB}\theta_{BC}} \tag{4.3}$$

Interference is:
negative for $C > 1$ and *Interference*=1-C < 0
positive for $C < 1$ and *Interference* =1-$C > 0$
absence for $C = 1$ and *Interference* =1-$C = 0$
complete for $C = 0$ and *Interference* =1-$C = 1$
When there is absence of crossover interference, Equation 4.2 reduces to Equation 4.1.

High level interference has been a limiting factor in developing complete multipoint models. Many models have been developed assuming no interference or a low level of interference.

The recombination fraction for a large genome segment is not the sum of the small intervals within the large segment. However, if the expected number of crossover in each of the interval can be estimated, the expected number of crossovers within the large segment should be the summation over the intervals. The expected number of crossovers within a genome segment is used to define a genetic

distance. For example, $m_{AC}=m_{AB}+m_{BC}$ can be used to model the three loci in Figure 4.2, where m_{AB}, m_{BC} and m_{AC} are map distances between *A* and *B*, *B* and *C,* and *A* and *C,* respectively. The relation between number of crossovers and genetic distance is given by a so-called *mapping function* (section 4.3).

Configurations

Suppose three diallelic loci *A*, *B* and *C*, whit alleles *Aa, Bb,* and *Cc* respectively. There are four possible crossover configurations during meiosis. If a parent is heterozygous (*ABC/abc*) then one of the following will happen during meiosis (Figure 4.3):

(1) no crossover happens between *A* and *B*

(2) crossover happens between *A* and *B*, but not between *B* and *C*

(3) crossover happens between *B* and *C*, but not between *A* and *B*

(4) crossover happens in both between *A* and *B* and between *B* and *C*

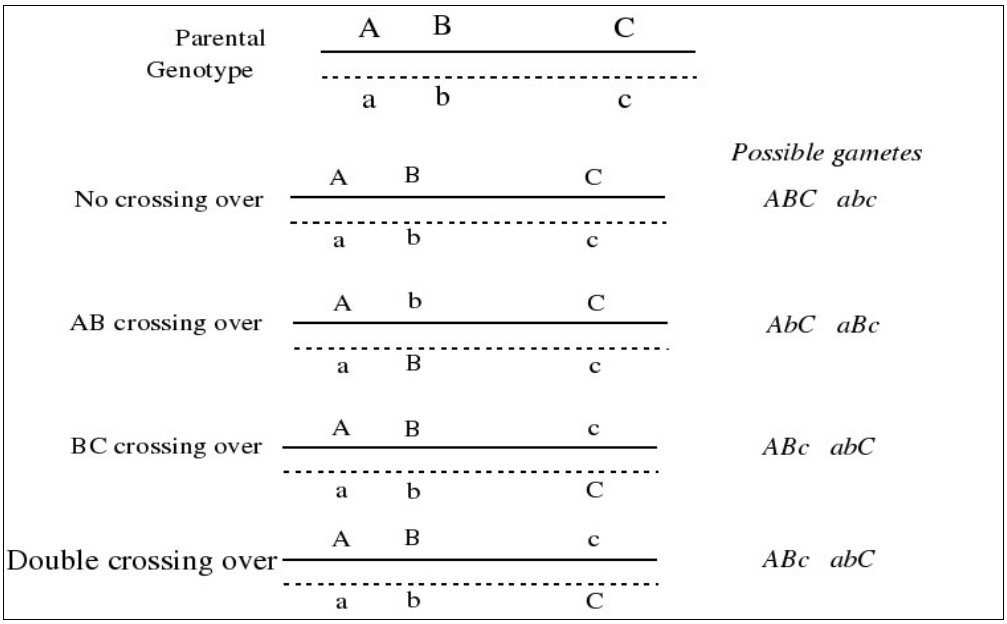


Figure 4.3 *represents gametes produced by a heterozygous parent (ABC/abc) for three loci A, B and C in order ABC.*

If the parent is heterozygous for the three loci, then the frequency of the four possible crossovers configurations can be observed by genotyping the three loci for a number of individuals in the progeny. Figure 4.3 shows the gametes produced by a heterozygous parent in association with the four configurations of crossover. Gametes *ABC* and *abc* are produced when no crossover happens during meiosis. A single crossover between *A* and *B* will produce gametes *Abc* and *aBC* and between *B* and *C* will pro-

duce *Abc* and *abC*. The double crossover will result in gametes *AbC* and *aBc*. If no crossover interference is assumed, then the probabilities of no crossover, *BC* crossover, *AB* crossover and double crossover are:

$$\begin{gathered}(1-\theta_{AB})(1-\theta_{BC})=1-\theta_{AB}-\theta_{BC}+\theta_{AB}\theta_{BC}\\(1-\theta_{AB})\theta_{BC}=\theta_{BC}-\theta_{AB}\theta_{BC}\\\theta_{AB}(1-\theta_{BC})=\theta_{AB}-\theta_{AB}\theta_{BC}\\\theta_{AB}\theta_{BC}\end{gathered} \tag{4.4}$$

respectively. When the crossover interference is taken into account then Equation 4.4 becomes:

$$\begin{gathered}(1-\theta_{AB})(1-\theta_{BC})=1-\theta_{AB}-\theta_{BC}+C\theta_{AB}\theta_{BC}\\(1-\theta_{AB})\theta_{BC}=\theta_{BC}-C\theta_{AB}\theta_{BC}\\\theta_{AB}(1-\theta_{BC})=\theta_{AB}-C\theta_{AB}\theta_{BC}\\C\theta_{AB}\theta_{BC}\end{gathered} \tag{4.5}$$

Double crossover

In some applications of genomic mapping, double crossover is ignored. In practice, gametes produced by double crossover have often been treated as potential experimental errors. In many cases, this is true. These double crossovers are either corrected after checking the data or discarded for the reason , they are expected to be quite rare. The genetic map built in this way is shorter in length. However, the treatment is complex because:

- Double crossovers are sometimes ignored in the likelihood function for the genetic map on which the lod score is based.
- The double crossover frequency strictly depend on locus order. So, the double crossovers identified using one locus order may entirely differ from those identified using another locus order.
- The expected double crossovers are usually treated the same as the unexpected double crossovers.

One should be cautious in using the double crossovers as a criteria to identify experimental errors. Any wrong doing may result in a biased genetic map. The biases could be an incorrect locus order, an underestimated genetic map, or an unrealistic LOD score for the locus order. Today, and in fact since many years already, locus order is determined by direct experimental methods, and ultimately by large scale sequencing. Thus, inferences on locus order based on linkage analysis is an outdated issue. However,

the estimation of recombination rates and thus map distances cannot be replaced by molecular techniques, and for this task linkage analysis still is indispensable.

Likelihood function

Likelihood function for the three locus model is the first step toward multipoint likelihood. Figure 4.4 shows the segregation of three loci in a pedigree. The offspring *F* is produced by a cross of two different homozygous parents *GF* and *GM*. *F* can produce eight types of gametes, whereas homozygous individual M produces only one type of gametes. the possible genotypes for *O* and their expected frequencies are listed in Table 4.2.

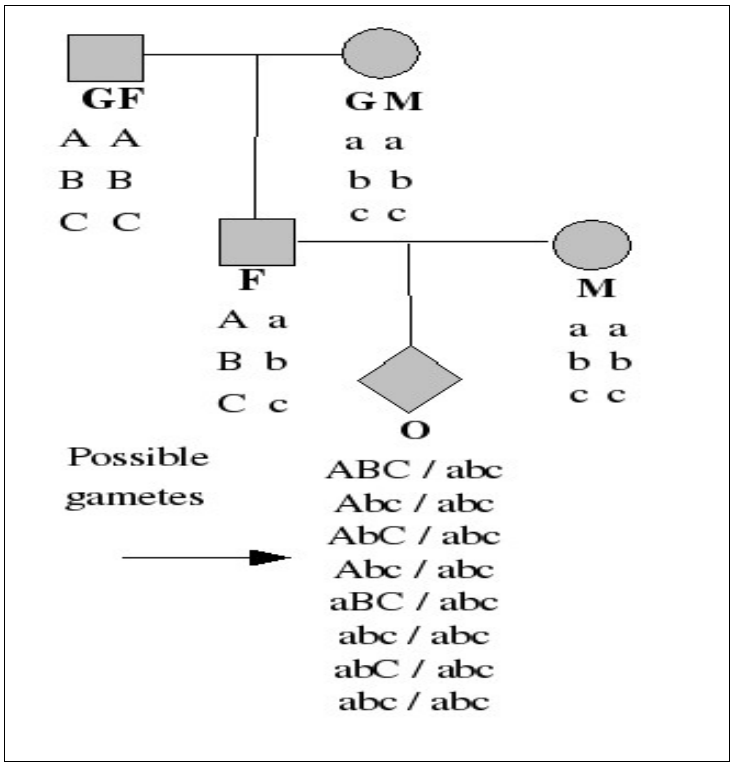


Figure 4.4 *illustrates the possible genotypes for the offspring "O" given the genotypes of the father, mother, grandfather and grandmother.*

Genotype	Observed counts	Notation	Double crossover	No double crossover
AaBbCc	f_1	p_1	$0.5(1-\theta_{AB}-\theta_{BC}+C\theta_{AB}\theta_{BC})$	$0.5(1-\theta_{AB}-\theta_{BC})=0.5(1-\theta_{AC})$
AaBbcc	f_2	p_2	$0.5(\theta_{BC}-C\theta_{AB}\theta_{BC})$	$0.5\theta_{BC}$
AabbCc	f_3	p_3	$0.5C\theta_{AB}\theta_{BC}$	0
Aabbcc	f_4	p_4	$0.5(\theta_{AB}-C\theta_{AB}\theta_{BC})$	$0.5\theta_{AB}$
aaBbCc	f_5	p_4	$0.5(\theta_{AB}-C\theta_{AB}\theta_{BC})$	$0.5\theta_{AB}$
aaBbcc	f_6	p_3	$0.5C\theta_{AB}\theta_{BC}$	0
aabbCc	f_7	p_2	$0.5(\theta_{BC}-C\theta_{AB}\theta_{BC})$	$0.5\theta_{BC}$
aabbcc	f_8	p_1	$0.5(1-\theta_{AB}-\theta_{BC}+C\theta_{AB}\theta_{BC})$	$0.5(1-\theta_{AB}-\theta_{BC})=0.5(1-\theta_{AC})$

Table 4.2 *Expected genotype frequencies for the offspring "O" from Figure 4.4. The expected frequencies are given twice, first considering the possibility that double crossovers may occur and then without considering the possibility of double crossovers.*

In linkage analysis, two situations have to be considered. When the three loci are located in a relatively small genome region, double crossover may have been ignored in some applications. This approach of ignoring double crossover certainly has potential problems when the target region is not small. A full three-locus models should include double crossover and crossover interference. Column 4 of Table 4.2 shows the expected genotype frequencies considering double crossover and crossover interference. If the coefficient of interference is set to one, then the events of crossover which occur in the segments are considered independent.

A log likelihood function for a three-locus model with locus order *ABC* is:

$$\begin{aligned} L(\theta_{AB},\theta_{BC}) &= (f_1+f_8)\,logp_1+(f_2+f_7)\,logp_2+(f_3+f_6)\,logp_3+(f_4+f_5)\,logp_4 \\ &= (f_1+f_8)\log(1-\theta_{AB}-\theta_{BC}+C\theta_{AB}\theta_{BC})+(f_2+f_7)\log(\theta_{BC}-C\theta_{AB}\theta_{BC}) \\ &\quad +(f_3+f_6)\log(C\theta_{AB}\theta_{BC})+(f_4+f_5)\log(\theta_{AB}-C\theta_{AB}\theta_{BC}) \\ &= (f_1+f_8)\log(1-\theta_{AC})+(f_2+f_7)\log(\theta_{AC}-\theta_{AB}) \\ &\quad +(f_3+f_6)\log(C\theta_{AB}\theta_{BC})+(f_4+f_5)\log(\theta_{AC}-\theta_{BC}) \end{aligned} \tag{4.6}$$

where f_i is the observed count, p_i is the expected genotype frequency (defined in Table 4.2). In general

the parameters θ need to be estimated. If in a pedigree the phase is known then the estimation of the recombination fraction is easy to obtain from the pedigree data. Given Table 4.2 and a population size of *N,* then the estimates of the recombination fractions are given by:

$$\begin{aligned}\hat{\theta}_{AB} &= (f_3 + f_4 + f_5 + f_6)/N \\ \hat{\theta}_{AC} &= (f_2 + f_4 + f_5 + f_7)/N \\ \hat{\theta}_{BC} &= (f_2 + f_3 + f_6 + f_7)/N\end{aligned} \tag{4.7}$$

When in a pedigree the phase is unknown, f_i can not be observed directly. Then the estimation of the parameters can be carried out by the recursive procedures of the Elston-Stewart or the Lander-Green algorithms which are explained in detail in chapter 3.

4.3 Mapping functions

Due to the non-additivity of the recombination fraction, they need to be converted into genetic distances. Mapping functions have been designed to solve this problem. These functions may apply to general or specific situations.

Ideally, genes or genetic markers are organized linearly on a map and their relative positions on the map can be quantified in an additive fashion. For example, considering six loci, *A, B, C, D, E* and *F* in order *ABCDEF*, the relationship can be quantified using:

$$\begin{aligned}m_{AF} &= m_{AE} + m_{EF} \\ &= m_{AD} + m_{DE} + m_{EF} \\ &= m_{AC} + m_{CD} + m_{DE} + m_{EF} \\ &= m_{AB} + m_{BC} + m_{CD} + m_{DE} + m_{EF}\end{aligned} \tag{4.8}$$

where m_{ij} is defined as the distance between loci *i* and *j* and it is derived from the expected number of crossovers between the two loci. If the expected number of crossovers is one in a genome segment, then the genetic distance between the two loci is 1 Morgan (M) or 100 centiMorgans (cM). For recombination fraction, θ_{ij} between two loci *i* and *j*, if a function

$$m_{ij} = F(\theta_{ij}) \tag{4.9}$$

exits for all pairs of loci and is a continuous function, then $F(\theta_{ij})$ is defined as a mapping function. Equation (4.9) has been commonly used to convert recombination fraction into genetic distance. For some mapping functions, the inverse of the function is:

$$\theta_{ij} = F^{-1}(m_{ij}) \tag{4.10}$$

and it is used to convert map distances to recombination fractions. This a a general formulation. However, there is not always a "closed form" of the inverse function.

Over several decades, a number of mapping functions have been developed. Table 4.3 list some commonly used mapping functions and their inverses.

Reference	***Map function*** $(m = F(\theta))$	***Inverse*** $(\theta = F^{-1}(m))$
Morgan (1928)	θ	m
Haldane (1919)	$-0.5\log(1-2\theta)$	$0.5(1-e^{-2\lvert m\rvert})$
Kosambi(1944)	$\frac{1}{2}\tanh^{-1}2\theta = \frac{1}{4}\log(\frac{1+2\theta}{1-2\theta})$	$\frac{1}{2}\tanh(2\mathrm{m}) = \frac{1}{2}\ \frac{1-e^{4\mathrm{m}}-1}{e^{4\mathrm{m}}+1}$
Carter & Falconer (1951)	$0.5(\tan^{-1}2\theta + \tanh^{-1}2\theta)$	no closed form
Rao et al (1977)	*	no closed form
Sturt (1976)	no closed form	$0.5\left[1-(1-\frac{m}{L})e^{\frac{m}{L}(1-2\mathrm{L})}\right]$
Felsenstein (1979)	$\frac{1}{2(k-2)}\log\frac{1-2\theta}{1-2(k-1)\theta}$	$\frac{1-e^{2(k-2)m}}{2\left[1-(k-1)e^{2(k-2)m}\right]}$
Karlin (1984)	$0.5\mathrm{N}\left[1-(1-2\theta)^{1/N}\right]$	$0.5\left[1-(1-\frac{2\mathrm{m}}{N})^{N}\right]$

Table 4.3 *List of commonly used mapping functions and their inverses.*

* $m = \left[p(2\mathrm{p}-1)(1-4\mathrm{p})\log(1-2\theta)\right]/6 + \left[8\mathrm{p}(p-1)(2\mathrm{p}-1)\tan^{-1}2\theta\right]/3 + (\left[2\mathrm{p}(1-p)(4\mathrm{p}+1)\tanh^{-1}2\theta\right]/3) + (1-p)(1-2\mathrm{p})(1-4\mathrm{p})\theta$

Morgan's function considers complete interference and uses the estimated recombination fraction as genetic distance, $\theta = m$. When a small segment is considered, the chance that double or multiple crossover occur in the segment is low. In such cases the estimated recombination fraction has the same expectation as the expected number of crossovers. So, Morgan's mapping function can be applied when small genome segments are considered. Contrary, *Haldane's function* assumes absence of interference. If interference is ignored the relationship among the pairwise recombination fractions among three ordered loci *A, B* and *C* can be quantified as in Equation (4.1). Equation 4.1 can be rewritten as:

$$1-2\theta_{AC}=1-2(\theta_{AB}+\theta_{BC}-2\theta_{AB}\theta_{BC})$$
$$=(1-2\theta_{AB})(1-2\theta_{BC})$$

If more than three loci are considered, then one can write:

$$1-2\theta_{l}=\prod_{i=1}^{l-1}(1-2\theta_{i}) \tag{4.11}$$

where l is the number of loci in the genome segment, θ_l is the recombination fraction between two markers flanking the whole segment and θ_i is the recombination fraction between two loci flanking a sub-segment. An additive function for Equation (4.11) is:

$$F(\theta)=c\log(1-2\theta) \tag{4.12}$$

where c is constant. Haldane derived his mapping function from Equation (4.12) by setting c=-1/2, which is:

$$m=F(\theta)=\begin{cases}\frac{-1}{2}\log(1-2\theta) & for\ 0\le\theta<0.5\\ \infty & for\ \theta\ge 0.5\end{cases} \tag{4.13}$$

Haldane's function is extensively used. When the recombination fraction is small, the map distances and recombination fraction are approximately equal. Haldane's function works for situations with absence of crossover interference. However, experimental evidence has been found to support that crossover interference exists and crossovers occur non-randomly in the genome. Taking into consideration interference, the relationship among pairwise recombination fractions for three ordered loci *A, B* and *C*, can be quantified using Equation (4.2) where *C* has been defined as the coefficient of coincidence and *C* -1 as interference. As it was previously discussed, recombination fraction can be considered a function of the expected number of crossovers or genetic distance, *F(m)*, for small segment flanked by *A* and *C* one can write:

$$\begin{aligned}\theta_{AB}&=F^{-1}(m)\\ \theta_{BC}&=\Delta m\\ \theta_{AC}&=F^{-1}(m+\Delta m)\end{aligned} \tag{4.14}$$

because recombination fraction and genetic distance are approximately equal for a short segment, Equation (4.2) can be rewritten, taking $\Delta \to -\infty$, as:

$$\begin{aligned} Equation(4.2): \quad & \theta_{AC} = \theta_{AB} + \theta_{BC} - 2C\,\theta_{AB}\theta_{BC} \\ & F^{-1}(m+\Delta m) = F^{-1}(m) + \Delta m - 2\mathrm{C}F^{-1}(m)(\Delta m) \\ & \Rightarrow \frac{F^{-1}(m+\Delta m) - F^{-1}(m)}{\Delta m} = 1 - 2\mathrm{C}F^{-1}(m) \\ & \Rightarrow \frac{dF^{-1}(m)}{dm} = 1 - 2\mathrm{C}\,\theta \\ & \Rightarrow \frac{dF(\theta)}{d\theta} = \frac{1}{1-2\mathrm{C}\theta} \end{aligned} \tag{4.15}$$

So, the mapping function is given by:

$$F(\theta) = \int_0^{\theta} \frac{1}{1-2\mathrm{C}u}\,du \tag{4.16}$$

if one sets C=1, then it becomes the Haldane's function. *Kosambi's function* takes $C = 2\theta$ and it is written as:

$$\begin{aligned} F(\theta) &= m \\ &= \int_0^{\theta} \frac{1}{1-4u^2}\,du \\ &= \begin{cases} \frac{1}{2}\tanh^{-1}2\theta = \frac{1}{4}\log\frac{1+2\theta}{1-2\theta} & for \quad 0 \le \theta < 0.5 \\ \infty & for \quad \theta \ge 0.5 \end{cases} \end{aligned} \tag{4.17}$$

Kosambi's function considers a moderate interference. The rationale behind this function is that the crossovers interference depends on the size of a genome segment. The interference is absent when a segment is sufficiently large (e.g., $C \to 1\ when\ \theta \to 0.5$). The interference increases as the segment decreases (e.g., $C \to 0\ when\ \theta \to 0.0$). The relationship between the size of the segment and the crossover interference is $C = 2\theta$. *Carter and Falconer function* and *Felsenstein's function* are also derived from Equation (4.16) by setting $C = 8\theta^3$ and $C = K - 2(K-1)\theta$, respectively (see Table 4.4 for the definition of *K).* The Carter and Falconer mapping function is:

$$
\begin{aligned}
F(\theta) &= m \\
&= \int_0^{\theta} \frac{1}{1-16u^4}\,du \\
&= \frac{1}{4}(\tan^{-1} 2\theta + \tanh^{-1} 2\theta)
\end{aligned}
\tag{4.18}
$$

which has no simple inverse. The Carter and Falconer function is commonly used when there is evidence of strong crossover interference. Felsenstein's functions is given by:

$$
\begin{aligned}
F(\theta) &= m \\
&= \int_0^{\theta} \frac{1}{1-2\left[K-2(K-1)\theta\right]}\,du \\
&= \frac{1}{2(K-2)} \log \frac{1-2\theta}{1-2(K-1)\theta}
\end{aligned}
\tag{4.19}
$$

where $-\infty < K < \infty$ is a parameter for crossover interference. If K=0, Equation (4.19) is the same as Kosambi's function and if K=1, then Equation (4.19) is the same as Haldane's function.

Rao's function is a weighted mean of the Morgan, Haldane, Kosambi and Carter and Falconer mapping functions. The Rao's mapping function is formulated as:

$$
\begin{aligned}
F(\theta) &= m \\
&= \left[p(2\mathrm{p}-1)(1-4\mathrm{p})\log(1-2\theta)\right]/6 \\
&+ \left[8\mathrm{p}(p-1)(2\mathrm{p}-1)\tan^{-1} 2\theta\right]/3 \\
&+ (\left[2\mathrm{p}(1-p)(4\mathrm{p}+1)\tanh^{-1} 2\theta\right]/3) \\
&+ (1-p)(1-2\mathrm{p})(1-4\mathrm{p})\theta
\end{aligned}
\tag{4.20}
$$

When p=0, p=0.25, p=0.5, and p=1, Equation (4.20) reduces to Morgan, Carter and Falconer, Kosambi and Haldane, respectively. The above discussed mapping functions are all based on Equation (4.16). and are summarized on Table (4.4). Those functions were obtained by setting different values for the coefficient of coincidence,C. Table (4.4) list their C values and some comments.

These are the most commonly used mapping functions, but as mentioned by Karlin (1984) there are two difficulties with deriving mapping functions from Equation (4.16):

- First, the mapping functions take into consideration only the genetic distance between markers regardless of the location of these markers on the chromosome. This consideration is not realistic when the mapping functions are applied to a large number of loci.

- Second, Karlin questions the existence of a global relationship between the marginal pairwise recombination fractions and the genetic distance. He suggest that a mapping function should include a distance and a parameter for the location of the segment on the genome.

References	Coincidence, *C*	Comments
Morgan (1928)	0	complete interference, absence of multiple crossover
Haldane (1919)	1	Absence of crossover interference
Kosambi (1944)	2θ	Crossover interference is a function of recombination fraction
Carter & Falconer (1951)	$8\theta^3$	Strong crossover interference
Felsenstein (1979)	$K-(K-1)2\theta$	K=1: absence of crossover interference K<1: Positive interference K >1: negative interference

Table 4.4 *List of mapping functions based on Equation (4.16).*

Karlin concludes that the mapping functions of Kosambi, Carter and Falconer, and Felsenstein (for K not in the range between 1 and 2) are not valid ones for a multilocus structure. He also concludes that the mapping function of Haldane, Sturt and Felsenstein (for $1 \le K \le 2$) are valid multilocus mapping functions if a global relationship between recombination fraction and genetic distance exits.

The *Sturt's function* was derived based on the assumption that there is one obligatory crossover and an additional crossover following a Poisson process, which is:

$$F^{-1}(m)=\theta = \begin{cases} \frac{1}{2}\left[1-(1-\frac{m}{L})e^{\frac{m}{L}(1-2L)}\right], & m<L \\ 0.5, & m\ge L \end{cases} \tag{4.21}$$

where L is the genetic length of the chromosome arm in Morgans. This mapping function can be applied to a genome segment representing a single chromosome arm.

Karlin's function considers that the number of crossovers in an interval follows a binomial distribution *B(N,p)* where N is the maximum number of crossovers in an interval and p is the probability of

crossover. So, Kalin's function is given by:

$$F(\theta) = m = \frac{1}{2} N\left[1-(1-2\theta)^{1/N}\right] \tag{4.22}$$

Commonly used mapping functions and the concept of crossover interference are largely derived from a three-locus model and certain assumptions regarding to the distribution of crossovers events over the genome. The mapping functions have often been used without considerations of experimental conditions and genetic configurations. The difference among the commonly used mapping functions are due to the assumptions of the crossovers distributions on the genome, crossovers interference and the length of the chromosome segment considered. If complete interference is assumed, then the genetic distance is numerically equal to the observed recombination fraction because multiple crossovers are ignored. If absence of interference is assumed, then genetic distance is much greater than the observed recombination fraction.

Mapping functions work only for specific conditions. There is no universal mapping function. Mapping functions do not estimate physical distance.

It is well known that multilocus models provides more information than twolocus models in linkage analysis. Here was presented the formulation for a three-locus model. The formulations are becoming more and more complex when the number of loci increases. A multilocus model is based on possible crossovers combinations among the loci (Figure 4.5). For three loci *ABC* there are 4 possible crossover combinations: 11, 10, 01 and 00, if "1" is used to denote that crossover happens in the segment and "0" to denote that crossover does not happen. For seven loci, there are 81 possible crossover combinations (Figure 4.5). There are 512 possible crossover combinations for 10 loci. In general, there are 2^{n-1} combinations for n loci. In this situations is when one need the recursive procedures such us as the Elston-Stewart algorithm or the Lander-Green algorithm which are discussed in detail in chapter 3.

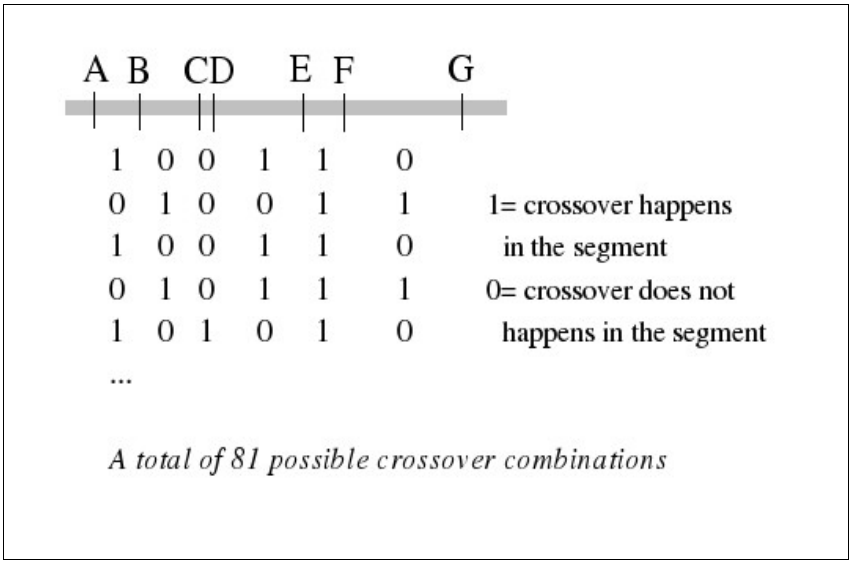


Figure 4.5 *illustrates some of possible crossover combinations considering 7 loci. "*1*"denotes the occurrence of crossover in the segment and "0" is used to denote that the cross-over do not happens.*

Chapter 5

INTRUDUCTION TO ASSOCIATION ANALYSIS

A major new area of genetic analysis that has arisen in the past years relates to association studies. Associations between diseases and single marker alleles have been reported for many years. However, the use of large scale association studies has come into play only with the human Genome Project and the availability of hundreds of thousands of markers. The increasing recognition of the important role of *linkage disequilibrium (LD)* in the genome and the use of it as a tool has motivated the recent research in the methodology for association studies.

In population genetics, LD is the non-independent occurrence of alleles at two or more (closely) loci. LD describes a situation in which some combinations of alleles occur more or less frequently in a population than would be expected from a random formation of haplotypes from alleles based on their frequencies. Non-independent occurrence between alleles at different loci are measured by the degree of LD.

When a disease mutation first occurs at a locus, it is on one of the two homologous chromosomes and therefore associated with all variants at loci close by on the same chromosome. The speed at which equilibrium is reached with respect to two or more loci has special importance. Consider two autosomal diallelic loci, with alleles *A* and *a* at one locus and alleles *B* and *b* at the other. Let the frequency of allele *A* in the population be p_A so that, under random mating, the genotypic frequencies at this locus are:

$$P(AA)=p_A^2$$
$$P(Aa)=2p_A(1-p_A)$$
$$P(aa)=(1-p_A)^2$$

Let the frequency of *B* be p_B so, analogously:

$$P(BB)=p_B^2$$
$$P(Bb)=2p_B(1-p_B)$$
$$P(bb)=(1-p_B)^2$$

There are four possible haplotypes, i.e. combinations of two alleles, one from each locus, on the same chromosome: *AB, Ab, aB, ab.* The haplotype frequencies in the absence of linkage disequilibrium (i.e. the alleles occur independently on haplotypes) and with LD are:

Haplotype	Haplotype frequencies without LD	with LD
AB	$p_A p_B$	$p_A p_B + \delta$
Ab	$p_A(1-p_B)$	$p_A(1-p_B) - \delta$
aB	$(1-p_A)p_B$	$(1-p_A)p_B + \delta$
ab	$(1-p_A)(1-p_B)$	$(1-p_A)(1-p_B) - \delta$
sum	1	1

In case that alleles *A* and *B* are associated, the frequency of haplotype *AB* would be $p_A p_B + \delta$, where δ is a measure of the strength of LD between the two loci. So, if allele *B* at locus2 predisposes to some disease phenotype, then if one ascertains a sample of affected individuals (cases) from the population, and a sample of unaffected individuals (controls), then allele *A* would be found more frequently in cases than in controls. In other words, there will be an association between allele A and the disease phenotype. In practice, one can test a large number of marker loci throughout the genome, or a set of polymorphisms in or around a candidate gene, in the hope that one of these marker loci would be close enough to a disease locus that some marker allele might be associated with the disease allele. Ideally one would like to find the individual base pair(s) in the DNA responsible for specific diseases, but there are over three billion base pairs in the human genome; the phenomenon of LD enables to test only a few hundred thousand, identified as SNPs, to interrogate about 85% of the genome in the mission to locate a disease mutation. It should be noted that the 85% refers to Caucasian population; current SNP platforms cover much less of the genome in other populations, especially older populations such as in Africa. that have undergone more recombination in their evolutionary history.

Association studies have increased precision in localizing a disease susceptibility locus and may have increased power compared to linkage analysis, particularly for genes with small individual effects (Risch and Merikangas, 1996). Association studies can be performed with samples of unrelated individuals, greatly simplifying the recruitment process, and thus enabling larger sample sizes to be studied.

Association between genetic markers and a disease phenotype may be missed, despite the presence of close linkage, if multiple, independent disease mutations are present (whether due to allelic heterogeneity or locus heterogeneity). However, LD may still be detectable if a large enough proportion of the sample has a common ancestral mutant allele. Allelic heterogeneity is less likely to be present in isolated populations with a small founder population. Linkage and association methods require very

large sample if there are multiple disease predisposing variants of modest individual effect, gene-gene interactions, gene-environment interaction, or allelic or locus heterogeneity.

Unfortunately, association between genetic markers and a disease can be present for reasons other than LD, including population stratification and chance, the last one leading to false positive associations. Population stratification is a confounding of the relationship between the marker and disease status due to the presence of population subgroups that do not intermarry. Among these population subgroups there must be differences in both disease and marker allele frequencies for confounding to occur. Population stratification leads to a 'fake' association between disease and genetic marker. Population stratification is of concern in any association study if, for example, individuals have been recruited from multiple locations or institutions. For this reason much of the statistical research for association analysis has focused on this problem.

There are several methods to detect and control for population stratification. The transmission disequilibrium test (TDT) is a case/control design that uses family controls when parental genotypes are known (Spielman et al., 1993; Ewens et al., 2005). Population stratification is controlled by comparing frequencies of alleles transmitted by a parent to an affected offspring to those that are not transmitted. One of the disadvantages of the TDT is the requirement of parental genotypes, eliminating one of the advantages suggested for association mapping.

Multiple testing, or chance, can always be a reason that association studies suffer from a high false positive rate. There are several methods available to adjust for multiple testing. An obvious way is to replicate association results in an independent data set. However, if several different associations are real, it is unlikely that the same genetic variants will have the strongest signals in independent samples, simply because chance fluctuations from sample to sample. More importantly, different alleles may be causative, or strongly associated with causative alleles, in different populations, because of genetic variation. The Bonferroni correction is one of the most commonly used methods to date to adjust for multiple testing. However, this method is known to be conservative, and so many researchers may have been tempted to ignore the multiple testing issue out of frustration due to the reduction of power. Anyway, other tests have been suggested to be less conservative such as the *false discovery rate* (FDR) proposed by Benjamini and Hochberg (1995). The FDR is the expected proportion of erroneous rejections among all rejections. When many of the tested hypotheses are rejected, indicating that many hypotheses are not true, the error from a single erroneous rejection is not always as crucial for drawing conclusions from the sample tested, and the proportion of errors is controlled instead. Thus one is ready to tolerate more errors when many hypotheses are rejected, but less when fewer are rejected. In many applied problems

it has been argued that the control of the FDR is the more appropriate response to the problem of multiple testing.

As stated by Elston and Spence (2006) whether an association study is designed for fine mapping as a follow-up to promising linkage results or for the investigation of polymorphisms within a candidate gene region, the analysis is complicated by the presence of numerous markers. A large number of markers must be evaluated to make disequilibrium mapping feasible, as we want to be sure that we have included a marker that is in LD with the disease locus. Because of the large number of markers that are required, SNPs, which are now relatively cheap to genotype, are the markers of choice. However, because of their diallelic nature, SNPs are individually relatively uninformative. To counter this, clusters of very closely spaced SNP markers, i.e. haplotypes, need to be analyzed together, so as to increase the information available at a particular location within the genome. The difficulty with utilizing haplotypes for association studies is that haplotypes are usually not known with certainty. Individuals who are heterozygous at more than one locus have ambiguous phase, and the probability of this event occurring increases with the number of markers being haplotyped. There are several methods available to help resolve the ambiguity. The most common used to resolve phase ambiguity is to use statistical approach to estimate haplotype frequencies. Numerical methods to compute maximum likelihood estimates of haplotype frequencies from random samples of unrelated individuals have been implemented, including the EM algorithm (Long et al., 1995; Hawley et al., 1995; Fallin et al., 2000). In the case of estimating haplotype frequencies given genotype information on each person, the 'missing' information is the phase between the marker alleles. Bayesian approaches have been implemented to approximate the maximum likelihood solution when the problem is too large for exact computations. In this approach, one has a choice of what prior distribution will be used. The uniform Dirichlet prior gives an equal weight to each of the possible alternative phases and so gives the same solution as the maximum likelihood EM algorithm in situations where exact computations can be performed. An alternative prior, called the 'pseudocoalescent' prior, gives greater weight to alternatives that produce similar haplotypes to those that have already been observed (Stephens et al., 2001). This prior will provide better estimates of haplotype frequencies than the equal weight Dirichlet prior in situations where this prior more accurately reflects reality, e.g. for tightly linked markers with inter-marker LD. In real data the two methods perform similarly when there is low inter-marker LD.

Once the haplotype estimation has been performed, the haplotype frequencies across populations or disease groups have been typically compared one haplotype at a time using the Cochran–Armitage trend test (Cochran 1954; Armitage 1955) or a logistic regression involving a one degree of

freedom test if the mode of inheritance is known (Schaid and Sommer 1993).

An important issue in association studies is the sample size necessary to detect an association, and hence the amount of genotyping that must be performed. For diseases where the loci have a small effect, or there are other competing causes of disease, very large sample sizes may be needed to detect an association at a particular locus.

This is the basis of association and LD mapping, which has been shown to work well either in the case of simple diseases in populations where there is likely to have been only one disease-predisposing allele at this locus and in the case of complex diseases.

Linkage or association?

There is considerable discussion in the literature about the power of linkage versus association studies in identifying disease susceptibility genes. In terms of association studies, current opinion varies as to whether family-based association studies are better than case-control studies. Use of SNPs versus haplotypes is also being greatly debated. Greenberg (1993) discussed the use of linkage and association approaches for localizing a gene influencing a trait when the locus under consideration increases susceptibility to the disease (as would be the case for elevated blood pressure) instead of being necessary and sufficient for the occurrence of disease (as would only be the case for a more simply inherited disease). He concluded that if the relative risk of having the disease given the associated allele is small, then the chance of finding linkage is correspondingly small. In this case, an association approach is much more powerful to detect genetic risk factors than linkage methods. However, there are also clear situations in which a linkage study is more suitable for localizing trait loci than an association study. In particular, if the disequilibrium between the marker and trait locus is low, association will not have sufficient power, and linkage is the better strategy for localizing genes.

Chapter 6

THE HUMAN PSEUDOAUTOSOMAL REGIONS

6.1 Introduction

Sex chromosomes have been for decades an issue of interest because of their distinctive patterns of transmission and their peculiar structure and function. The human sex chromosomes, X and Y, are morphologically and genetically distinct. In humans, females have two X chromosomes which are homologous and thus are of equal size and genetic content. Meiotic pairing and recombination can occur along their entire length. Males have one X and one Y chromosome which have two small regions of homology (identical sequence), located at the tips of the short and long arm of the X and Y chromosomes, as depicted in Figure 6.1. During male meiosis, pairing and crossover take place only in these

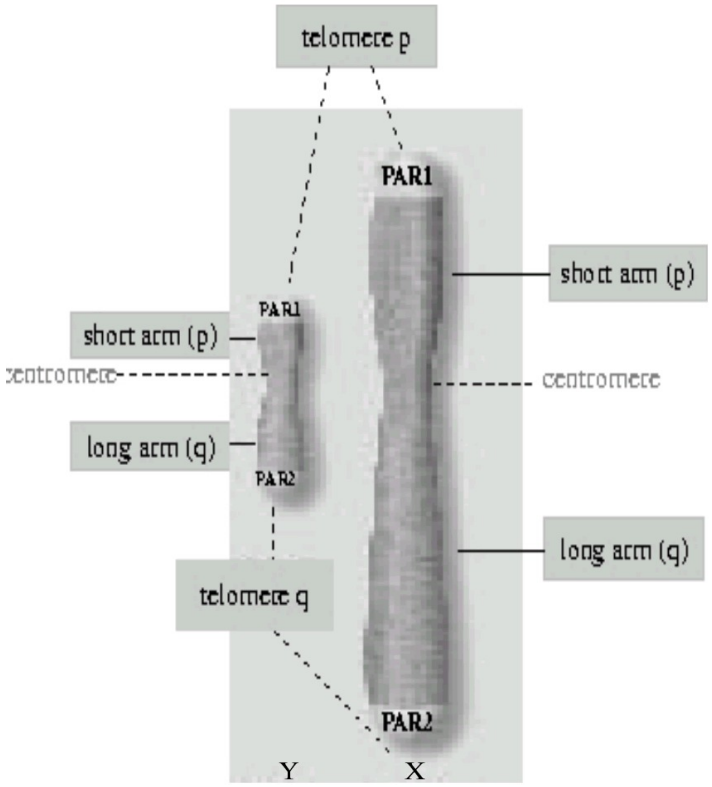


Figure 6.1 *Illustrates the structure of the human male sex chromosomes and the localization of the two homologous regions PAR1 and PAR2.*

these two small regions, which have been termed *pseudoautosomal regions* "PARs" (Burgoyne, 1982; Freije D et al., 1992). More precisely PAR1 refers to the tips of the short arms, Xp/Yp (Xp22.3-Yp11.3), and PAR2 refers to the tips of the long arms Xq/Yq (Xq28-Yq12). The physical lengths are approximately 2.7 Mb for PAR1 and 0.33 Mb for PAR2.

The human pseudoautosomal regions have attracted interest from topics in human genetics, cytogenetics, and evolutionary biology because of their special features. Pseudoautosomal nomenclature came up due to the unusual genetic behavior of these regions: markers close to the boundaries with the sex specific sequences behave as if they were tightly sex-linked, whereas, markers close to the telomeres act as if they were autosomal. The pseudoautosomal boundaries (PABs) are the interface between pseudoautosomal and sex-chromosome specific DNA sequences. PABs separate regions of intensive recombination from non-recombining regions on the Y and a moderately recombining region on the X chromosome.

A loss of PAR1 has been observed to be associated with male sterility leading to the theory that the existence of PAR1 is necessary for homologous X-Y chromosome pairing and the proper segregation of gametes (Gabriel-Robez et al., 1990; Mohandas et al., 1992). This is of interest because the region is physically relatively small. It is estimated that during male meiosis an obligate crossover occurs. Crossover activity in PAR1 is much higher in males than in females and also higher than for each of the autosomes. As a consequence of the elevated recombination rate in this region approximately one half of male children carry a recombinant PAR1 on their Y-chromosome whereas the other half inherit a non-recombinant PAR1 haplotype from their fathers. The rate of recombination in PAR2 is much lower than in PAR1 but still higher than the average of the remainder X-chromosome.

To date 24 genes have been identified in PAR1 and 5 in PAR2. Possible connections with clinical disorders such as short stature, asthma, psychiatric disorders and leukemia have been suggested, but only one pseudoautosomal gene, SHOX (Short Stature Homeobox), has been clearly associated with various short stature conditions and disturbed bone development. However, in systematic genome-wide linkage and association analysis, PARs have been largely neglected so far - a systematic "blind spot". The SHOX gene has been correlated to a disease via deletion mapping and not by linkage analysis.

This work focuses on statistical methods for genetic map construction, linkage and association analysis in the PARs. It summarizes the estimates of genetic maps, pseudoautosomal markers available on SNP-chips, and methodological developments which account for the special characteristics of the PARs in parametric and nonparametric linkage analysis as well as genetic association analysis. In addition a new genetic map for PAR1 and PAR2 is presented in chapter 7.

6.2 Evolutionary origin of the human sex chromosomes

The origin of the human sex chromosomes can be traced by comparing their sequence, gene content, and gene function in related species. It was first postulated by Muller in 1914 that X and Y chromosomes evolved from a pair of autosomes and have become separated in the course of millions of years (illustrated in Figure 6.2).

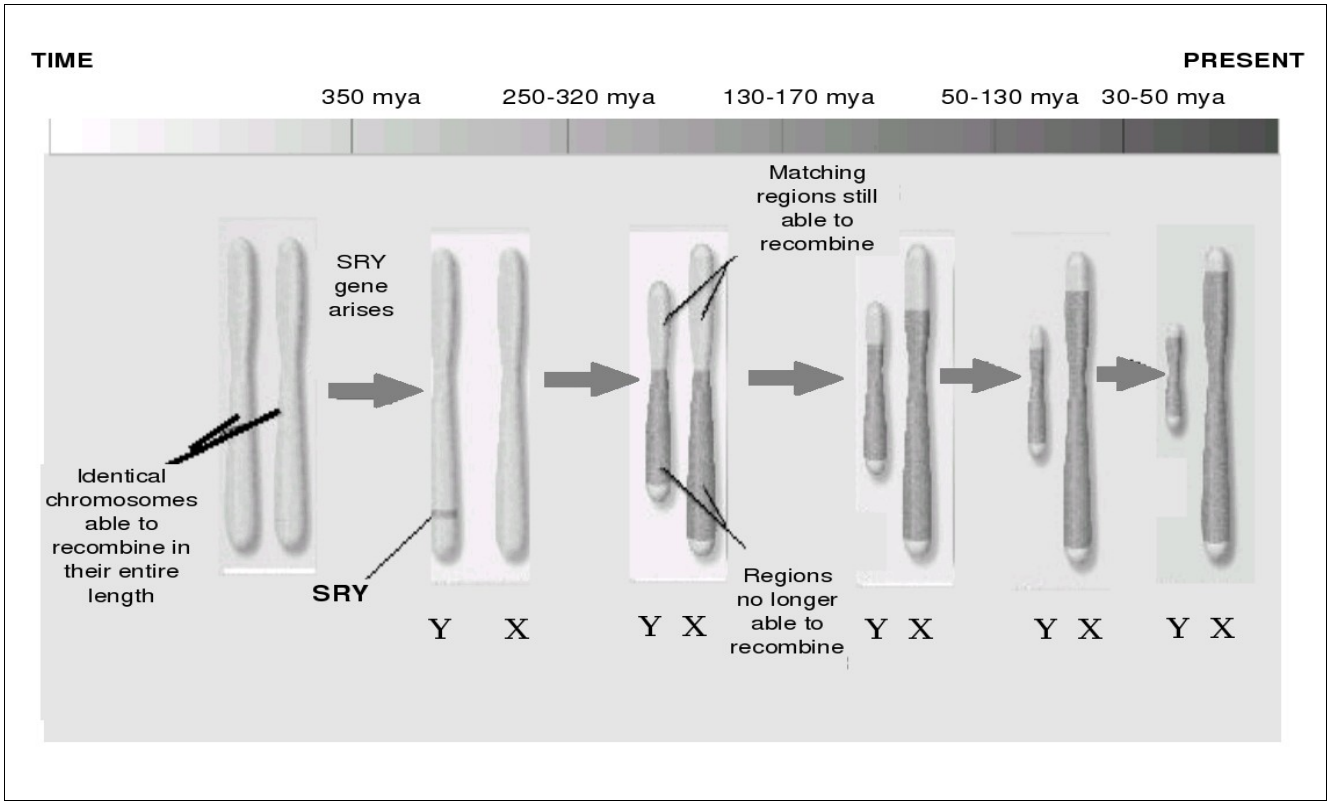


Figure 6.2 *The degeneration of the Y occurred in discrete episodes, beginning about 300 million years ago (mya) when an ancestor acquired the SRY gene on one of its autosomal chromosomes. Each of the episodes involved a failure of recombination to occur between the X and the Y chromosomes, resulting in subsequent decay of some genes in the non-recombining region. Figure adapted from: rediscovering biology "sex and gender". http://www.learner.org/channel/courses/biology*

The evolution of sex chromosomes emerged when one strand of a pair of autosomes obtained a male sex-determining gene, which has been identified as SRY (Sex-determining Region Y). Over time, additional genes with male specific functions accumulated in this sex-determining chromosome which gradually lost the ability to recombine with its counterpart. Subsequently, a degeneration of size started due to gene inactivation, mutation, deletion and insertion of junk. All evidence to date indicates that each time when a segment of the Y chromosome became separated, most of the genes in that segment were then inactivated by nonsense mutations and small deletions. Nevertheless, some few genes have refused to go away for hundred of millions of years, in some cases no clear reason for that could be found. In other cases, they have become specialized for spermatogenesis.

It is still an open question why the Y chromosome degenerated so quickly and why positive selection of male advantage genes did not work stronger against it. The theory that the Y could even disappear, is maintained by comparative studies in other vertebrates. With an approximated calculation of the average rate of loss of active genes from the human Y, Aitken and Graves (2002) predicted a complete loss of the Y chromosome in about 10 million years. Rather than predicting the extinction of the human species, Graves argues that this event could lead to the divergence of the human species in two distinct hominid species incapable of reproducing.

As in mammals the male is the heterogametic sex, this system of chromosomal sex determination is called X/Y system, as apposed to the W/Z systems realized in birds and (some) reptiles, where the male has the W/W and the females the W/Z chromosomal constitution.

Although, in the last years immense advances have been done in the understanding of the sex chromosomes, there are still a lot of questions unanswered. This enigma will however, be resolved within the next years by the availability of new informations as the recently published complete sequence of X and Y (Ross et al. 2006).

6.3 Evolutionary of the human pseudoautosomal regions

The first evidence of pairing between parts of the X and Y chromosomes dates about 70 years ago. In higher organisms it was first observed in the rat (Koller and Darlington 1934). In 1936, Haldane found evidence for partial sex linkage in human (Haldane, 1936), and by studying spermatocytes in meiotic prophase, pairing was demonstrated to occur in the short arms of X and Y chromosomes (Pearson and Bobrow 1970; Moses et al., 1975). It was claimed that a single obligatory crossover should be present in male meiosis between the X and Y chromosomes, restricted to PAR1 (Burgoyne 1982; Rouyer et al., 1986a; Page et al., 1987a). This crossover forms the chiasma that keeps the sex chromosomes

together during metaphase I of meiosis, and it was settled that it would therefore behave like an autosomal chromosomal segment. Absence of double recombinants in early studies suggested that only single recombinant events could occur in PAR1, but later studies reported a few double recombinants in PAR1 (Schmitt et al., 1993; Rappold et al., 1994). About 60 years after the discovery of PAR1, a second region of homology at the opposite ends of the X and Y chromosomes was observed, this new region was termed PAR2 (Freije et al., 1992). Since then PAR1 and PAR2 have provoked a great interest in researchers for their peculiarities. Comparative studies between the PARs of humans and others species have earned much interest in the last years with the objective of resolving the enigma of human evolution and human diseases (Graves et al., 1998; Gianfrancesco et al., 2001; Kohn et al., 2004; Yi and Li 2005; Bussell et al., 2005; Charlesworth D. 2005; Graves 2006; Graves et al., 2006). Whereas the human PAR1 is homologous to the pseudoautosomal region in several mammalian species, including great apes and Old World monkeys, the PAR2 sequence has a much shorter evolutionary history and it is specific to humans.

6.4 Genetic features of the pseudoautosomal regions

Genetic features of the PARs and the X and Y chromosomes are summarized in Table 6.1. Both regions, PAR1 and PAR2, although very small in size, display a higher gene density than the rest of the X-chromosome. In most genomic regions the recombination rate in females is higher than in males, but PAR1 and PAR2 represent a hot spot in males, exhibiting the highest recombination frequencies throughout the entire genome. In PAR1 male recombination activity is 10 to 20 times more frequent than in females (Page et al., 1987). In females, the recombination intensity in PAR1 is within the autosomal range. To date, no recombination event has been detected in females in PAR2.

PAR1 is necessary for homologous X-Y chromosome-pairing during male meiosis and, as with autosomes, undergoes one crossover event during this process. Although, PAR2 is not implicated in mediating male meiosis and undergoes a lower recombination activity than PAR1, it still represents a sixfold higher recombination activity when compared to the average of the autosomes. The sequence of PAR2 is completely known whereas in PAR1 six gaps with an estimated combined size of 370 kb could not be filled up to now.

Genetic Region	*Physical length (Mb)*	*Known protein coding Genes/Mb*	*Males recombination activity cM/Mb*	*Females recombination activity cM/Mb*	*Male/female quotient of genetic length*
PAR1	2.7	10	4.33-20.48	0.30-1.55	2.8-14.6
PAR2	0.33	15	6.06	(not detected so far)	-
X	165	6	(no homology)	1.21	-
Y	60	3	(no homology)	(not in females)	-
Autosomal range	46-245	3-23	0.80-2.40	1.40-2.80	0.57-0.85

Table 6.1 *Features of the pseudoautosomal region in comparison to the sex and autosomal chromosomes (Flaquer et al., 2008). Physical map lengths and known protein coding genes are taken from the ensembl data*base58. *Genetic map length for autosomes are based on the Rutgers map (Kong et al., 2004).*

29 genes lie within the human PARs (24 in PAR1, 5 in PAR2), and these genes exhibit 'autosomal' rather tan sex-specific inheritance. Table 6.2 illustrates the genes located in PAR1 and PAR2 and which diseases have been proposed to be associated to those genes. In addition, it is conceivable that further genes might reside within the gaps of PAR1. The pseudoautosomal regions display a higher gene-content (~10 genes per Mb in PAR1 and ~15 genes per Mb in PAR2) than the remainder of the X chromosome (~6 genes per Mb) or the Y chromosome (~3 genes per Mb). Due to an obligatory crossover in male meiosis, pseudoautosomal genes are exchanged frequently between X and Y chromosomes. To date, only one gene, SHOX (Short Stature Homebox), has been clearly associated to a disease, Thunder Syndrome, recently reviewed (Blascke et al. 2006).

Region	*Position*	*Gene*	*Suggested Disease*
PAR1	0.15	PLCXD1	None
	0.17	GTPBP6	None
	0.25	cM56G10.2	None
	0.29	PPP2R3B	None
	0.57	SHOX	turner syndrome (Blaschke and Rappold 2006)*
	0.92	bA309M23.1	None
	1.31	CRLF2	None
	1.38	CSF2RA	None
	1.52	IL3RA	None
	1.55	SLC25A6	None
	1.56	bA261P4.5	None
	1.57	CXYorf2	None
	1.59	ASMTL	None
	1.66	P2RY8	mental retardation (Cantagrel et al., 2004)
	1.76	CXYorf3	None
	1.79	ASMT	psychiatric disorders (Yi et al., 1993); epilepsy (Doherty et al., 2003)
	1.79	bB297E16.3	None
	1.91	bB297E16.4	None
	1.93	bB297E16.5	None
	2.37	DHRSX	None
	2.41	ALTE(ZBED1)	None
	2.54	Em:AC097314.2	None
	2.53	Em:AC097314.3	None
	2.63	CD99 (MIC2)	Ewing's Sacroma (Kreppel et al., 2006); sex cord-stromal tumors (Kommoss et al., 2000); Hodgkin's diseases (Kim et al., 2000); breast cancer (Byun et al., 2006)
PAR2	154.57	SPRY3	None
	154.71	SYBL1	Bipolar disorder (Saito et al., 2000); (Muller et al., 2002)
	154.81	IL9R	Asthma (Kauppi et al., 2000)
	154.81	Em:AJ271736.5	None
	154.82	CXYorf1	None

Table 6.2 *lists the homologous genes on the human X and Y chromosomes.* Position: distance from the telomere (Mb).
*The only gene clearly associated with a disease. Gene names as given by Ross et al., 2005.

6.5 Linkage and association analysis in the pseudoautosomal regions

The mode of inheritance for autosomal and X-linked Mendelian diseases can be modeled by classical segregation analysis with large samples of families. However, the segregation pattern of loci on the PARs depends on their location relative to the pseudoautosomal boundary (PAB). Two extreme situations for an affected father are depicted in Figure 6.3. In situation I, the susceptibility gene for a certain disease is close to the PAB of PAR1 and no recombination occurs between the gene and the sex-specific region. The dotted line indicates the position of a disease gene which could be at a) Y-PAR1 or b) X-PAR1. In the case that the disease locus is at Y-PAR1 all sons will be affected, due to

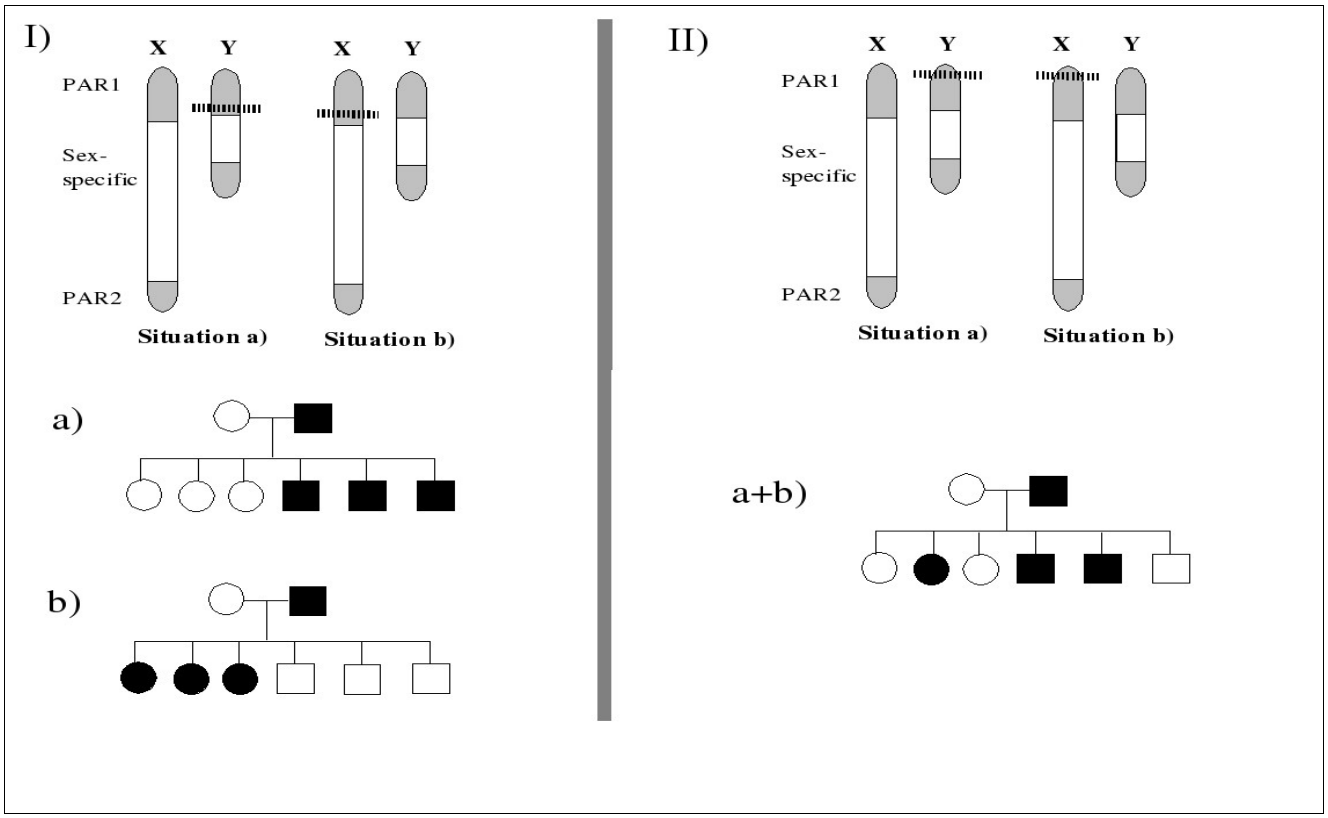


Figure 6.3 *illustrates pseudoautosomal segregation for dominant disease depending on the localization of the susceptibility gene (Flaquer et al., 2008). Case I) the susceptibility gene is located close to the PAB of PAR1. The way of segregation depends whether the susceptibility gene is at X-PAR1 or at Y-PAR1 . Case II) the susceptibility gene is located close to the PAR1 telomere and the way of segregation does not depend in which of the chromosomes is located.*

non-recombination activity all of them will inherit the locus from the father's Y-chromosome and all daughters will be unaffected since they receive the locus from the father's X-chromosome. If the disease locus is at X-PAR1 the inverse situation will hold, all daughters will become affected and all sons unaffected. In case II), the susceptibility gene is located close to the telomere and recombines with respect to loci within PAR1, following the same pattern of transmission like any autosomal dominant disease, where any offspring, male or female, get the the same chance of being affected or unaffected. Due to the small size of PAR2, it is thought that most of the susceptibility genes would follow the same pattern similar than in situation I)

Linkage analysis

In the last years researchers have been performing genome-wide studies to detect genes implicated in complex diseases. These days many methods to carry out multipoint analysis of genome-wide data, mainly parametric and nonparametric analysis, are widely available and implemented for autosomal and for X-linked markers. However, markers in PARs display a special behavior. Although, their transmission is similar to the autosomal markers, as despited in Figure 6.3, males are more likely to receive the allele locates on the father's Y chromosome and females are more likely to receive the allele located on the father's X chromosome. This occurs more markedly for markers near the boundary of the X-specific region. For that reason, an increased sharing of IBD between pairs of the same sex and a decreased sharing of IBD between pairs of opposite sex would be expected, independently whether a disease susceptibility gene is present or not. On the other hand, one has to consider the fact that in PARs the recombination events are much higher in male than in female meiosis.

To date only three analytical linkage methods have been suggested to deal specifically with the PARs (Ott 1986; Dupuis and Eerdewegh 2000; Strauch et al., 2004). Ott and Strauch et al. focused on parametric analytical approaches while Dupuis and Eerdewegh concentrated on nonparametric strategies.

The first method, proposed by Ott, covers different situations of sex-linked inheritance using parametric linkage analysis on a qualitative trait (dichotomous). In the case of two pseudoautosomal loci it is proposed to carry out linkage analysis as if these two loci were autosomal loci. Then, the male recombination fraction measures recombination frequency between the X and the Y chromosome while the female recombination fraction measures it between the two X's. In the scenario of linkage between an X-linked and a pseudoautosomal locus it is recommended to add a dummy allele to all males' hemizygous genotypes for X-linked markers, and to use a special definition of penetrances for the X-linked

locus and for the PAR locus. Both models could be carried out by adapting the input data structure to a program such as LIPED (Ott 1974, 1976) or LINKAGE (Lathrop et al., 1984).

Strauch et al. suggested a second approach following the same patterns as Ott. They proposed a recoding scheme for linkage programs that can only take codominant markers and a diallelic trait locus into account in an autosomal setting. This recoding scheme can be used for X-linked loci only or combined X-linked and pseudoautosomal loci, but it is required that the software offers modeling for imprinting. Models with imprinting comprise four penetrances; these four penetrances have been implemented into the program GENHUNTER-IMPRINTING (Strauch et al., 2000).

Both methods, Strauch et al. and Ott, were suggested considering only pseudoautosomal markers residing at PAR1, although, they could be easily applied to markers at PAR2, as well.

Dupuis and Eerdewegh presented a method for markers located at PARs to be applied in the field of nonparametric linkage analysis, using the ASP approach. The method takes into account the expected IBD sharing depending of the sex of the sib pair, the male recombination fraction between the marker and the sex specific region, and the presence or absence of a disease susceptibility gene in the PARs. For an autosomal marker unlinked to any disease locus, a sibling pair will share zero, one or two alleles IBD with probability ¼ , ½ and ¼ respectively. An increase in IBD sharing in affected sibs is taken as evidence for linkage to a disease-susceptibility locus. A similar reasoning is taken for the PARs. Under the hypothesis that there is not a disease susceptibility gene in the PARs, the probability that a pair shares zero or one alleles IBD from the maternal side in the PARs is ½. However, the probability of sharing zero or one alleles from the paternal side depends of the male recombination fraction between the marker and the X-linked regions and on the sex of the sib pair. A likelihood context is used to test for the presence of linkage in the PARs. In their paper, they show with real data how not taking the sex into account may lead to false-positives or false-negative results when an excess of sex-concordant affected sib pairs is present. This method looks quite robust although it does not deal with differences between male and female recombination rates in PARs. Unfortunately, the method is not implemented in any of the existing software programs for linkage analysis.

Association analysis

In association studies, besides confounding and bias from several sources (e.g. population structure, cryptic relatedness, and data errors), statistically significant association can only be found if the tested polymorphism has a variant that is causally related to the disease or if it is in strong linkage disequilibrium (LD) with a causal variant. Since LD on the population level only operates over short distances, a dense set of markers is to be used. As a simple rule at least one tagSNP every 5 kb is recommended to capture most of the common variation in an European sample (Consortium IH 2005). This would result in ~540 SNPs in PAR1 and ~60 SNPs in PAR2. However, in PAR1 this might not be enough since the LD in PAR1 might be lower than in autosomal regions of the same physical size, due to the high recombination rate in males in this region. It has not been analyzed so far, whether it would be an advantage to use special analytical methods to detect association with pseudoautosomal markers. Evidently, false positives results could arise if a causal variant is strongly associated with one sex, and the sex distribution is different among cases and controls.

The current situation regarding to linkage and association analysis in PARs is calling for new methods and their implementation. In most of the genome-wide studies, only the 22 autosomal chromosomes are analyzed excluding the sex chromosomes. However, some researches started to include the X chromosome in the genome-wide analysis but still excluding the pseudoautosomal loci from the analysis. Very few studies are using the PAR. Most of these few studies are analyzing pseudoautosomal loci using the general analytical approach described for autosomal loci.

6.6 Genetic maps for the pseudoautosomal regions

Several genetic maps have been proposed for the PARs, using the techniques of three generation pedigrees, single sperm typing, and unrelated individuals (Table 6.3). The first map was created in 1986 including only PAR1 (Rouyer et al., 1986a). This map was based on eight families and three microsatellites markers. One year later, a second map was published for PAR1 using forty-four families and 5 STR markers (Page et al., 1987). The first genetic map for PAR2 was suggested by Li and Hamer (1995) using a total of forty families and four markers. The estimated genetic length of PAR1 and PAR2 vary based on the existing genetic maps. In PAR1, estimated genetic length ranges from 12-55 cM in males and 0.8-4.18 in females.

	#markers	Male cM	Female cM	Mapping function	Sample	Reference
PAR1	5	48.5	5	Identity	8 CEPH families	Rouyer et al., 1986b
PAR1	5	49.9	4.18	Identity	44 CEPH families	Page et al., 1987
PAR1	11	49	4	Identity	38 CEPH families	Henke et al., 1993
PAR1	4[1]	38	-	Identity	2 sperm donors, 900 sperms cells	Schmitt et al., 1993
PAR1	9	55.3	-	Kosambi	4 sperm donors, 1912 sperms cells	Lien et al., 2000
PAR1	6	11.7	1.28	Kosambi	40 CEPH and 146 DECODE families	Matise et al. 2007
PAR1	1400	38[2]	3.8[2]	none	269 unrelated individuals	HAPMAP Consortium (2005)
PAR2	1[1]	2	0	Identity	40 CEPH families	Freije et al., 1992
PAR2	1	0.3	-	Identity	2 sperm donors, 900 sperms cells	Schmitt et al. 1993
PAR2	4	2	0	Identity	40 CEPH families	Li and Hamer 1995
PAR2	1	0.7	-	Kosambi	4 sperm donors, 1912 sperms cells	Lien et al., 2000
PAR2[4]	3	1.6	0	Identity	48 families	Matise et al., 2007
PAR2	140	0.7[3]	0	none	269 unrelated individuals	HAPMAP Consortium (2005)

Table 6.3 Estimates of genetic map length in male and female human pseudoautosomal regions (Flaquer et al., 2008) . In sperm typing studies male map distance are determined. In Freije et al. study one crossover in female was reported among 238 informative meioses but it could no be confirmed. [1] sex as a marker is not included here but used for map estimation. [2] Assuming a male/ female ratio of genetic length of PAR1 of 10. [3] two times the sex-averaged genetic length. [4] Own analysis based on data from the Matise et al. study (Matise et al. 2007). Hapmap did not use any mapping function, other method is used, described online supplementary information {consortium, 2005 #116}.

Because multipoint linkage analysis is the standard tool in the search for genetic variants that predispose to Mendelian and complex genetic diseases, and this method is generally more powerful than singlepoint methods, a genetic map is required. Misspecification of the genetic map has the potential to severely compromise the estimation and testing procedures used in multipoint linkage analysis. Specifically, inflation or deflation of the LOD score, loss of power to detect linkage or an increase in the false-positive rate, and bias in the disease locus position estimate are possible, depending on type and degree of misspecification (Daw et al., 2000; Fingerlin et al., 2006). Marker map misspecification

can arise from uncertainty due to the estimation process (sample error) or from over-simplified models of the biological process of recombination in this region. The number of meioses, and therefore the number of recombination events, used to infer inter-marker distances for genetic maps is often relatively small (<200). As a result, many current genetic maps have wide confidence bounds, particularly for dense maker spacing. Most notably, evidence for sex variation in recombination rates indicates that sex-averaged maps fail to incorporate relevant information about sex-specific genetic distances. The female:male genetic map distance ratio varies dramatically along chromosomes, and particularly in PARs. The biological mechanism(s) for theses differences is not yet well understood, but they are real and should be considered in multipoint linkage studies using sex-specific maps instead of sex-averaged maps.

Several attempts have been employed to integrate multiple genetic maps. Most recently, Duffy used a method of weighted regression to obtain smoothed local recombination rates to interpolate between markers in PARs using the genetic distances from Lien et al. (Duffy 2006). Figure 6.4 illustrates the relations among the published genetic maps for males and females in PAR1. The genetic maps for males proposed by Lien, Henke and Rouyer seem to be well in agreement for the first 750 kb. The discordance shown by the Schmitt's and Rutgers' map could be because of the most telomeric marker still being at a distance of approximately 750 kb from the telomere. After 750 kb, the maps display different tendencies, only the Rouyer's map and Henke's map display a similar trend.

The differences among these published maps could be due to a systematic map estimation bias, data errors, or an insufficient number of markers or meioses considered for the estimation of the recombination rate. These differences are less significant when comparing the genetic maps proposed for females. Based on these maps the estimated genetic length of PAR1 varies between 26-54 cM in males and between 2.8-6 cM in females. There is a very small number of genetic maps proposed for PAR2, and all of them used a poor number of genetic markers.

It is true that the international HapMap project used a much larger number of genetic markers (1400

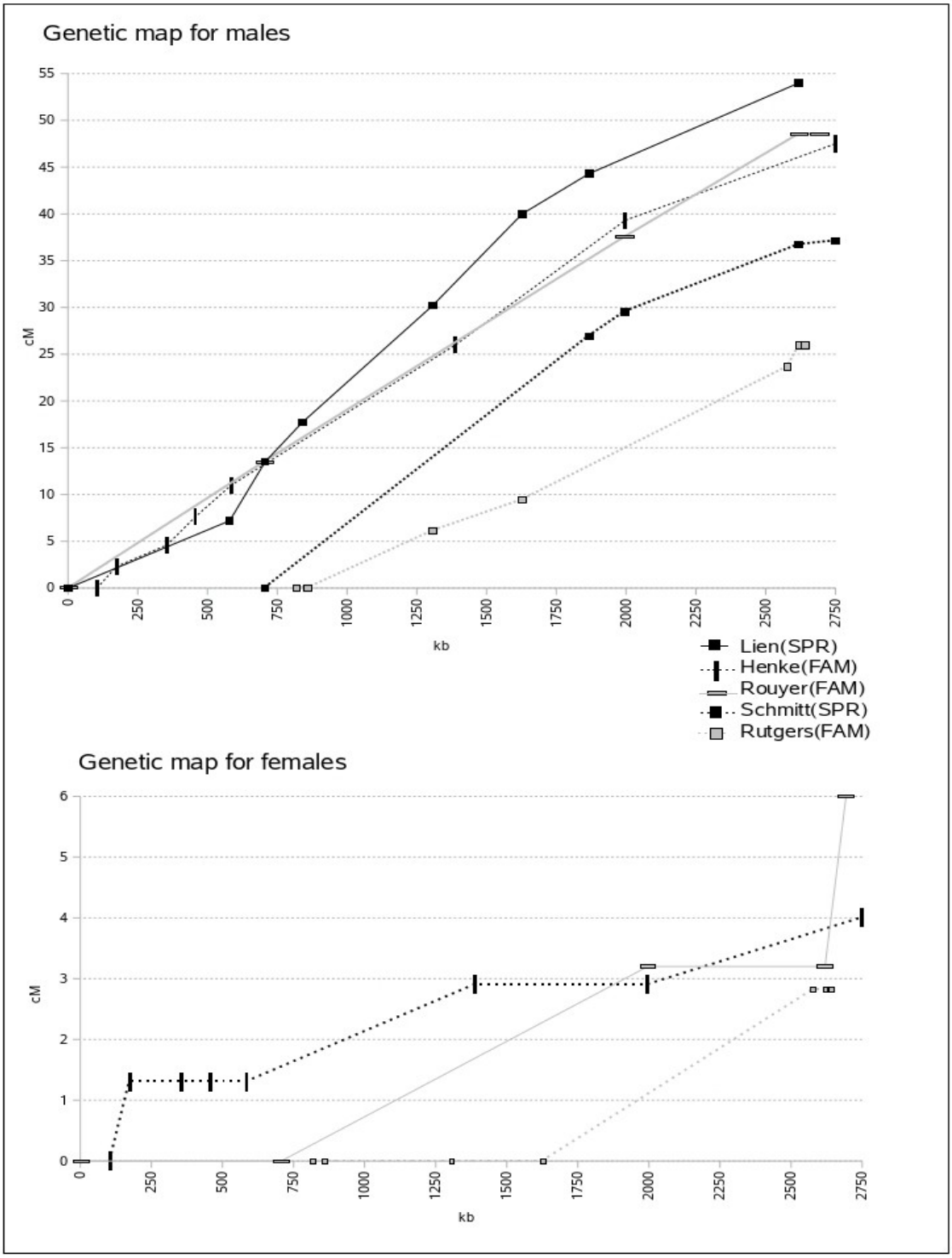


Figure 6.4 *Relations among the sex-specific genetic maps proposed so far for PAR1. Different techniques were used for the estimation of recombination rates, three-generation families (FAM) and sperm cell typing (SPR).*

SNPs in PAR1 and 140 SNPs in PAR2) to estimate a genetic map (HapMap consortium 2005). This map was based on the technique of unrelated individuals providing only sex-averaged distances. To be comparable with the other maps, we computed sex-specific map distances using a male/female map ratio of 10:1 (data not shown). HapMap estimates were lower in respect to the other maps, this could be because using the technique of unrelated individuals is very difficult to estimate accurate sex-specific recombination rates. With this technique a coalescent model of a finite neutrally evolving population with constant population size is assumed which cannot be easily confirmed with empirical data, and only sex-averaged recombination rates can be estimated. We do not present the map here because recently the new genetic map from HapMap phase 3 has become available where the sex chromosomes, and hence the PARs, have been excluded.

6.7 Coverage of the pseudoautosomal regions by Affymetrix and Illumina ANP arrays

Affymetrix and Illumina are the two big manufacturers of DNA SNP microarrays, called "GeneChips". Both of them offers several GeneChips that contain different number of SNPs covering the whole genome. The most of the GeneChips offers a poor coverage of the PARs. Table 6.4 lists the available GeneChips by Affymetrix and Illumina and the coverage offered at PAR1 and PAR2.

Company	GeneChips	PAR1	PAR2
Affymetrix	Genome-Wide Human SNP Array 6.0	391	32
	Genome-Wide Human SNP Array 5.0	155	0
	Mapping 500K Array Set	262	0
	Human Mapping 100K Set	19	0
	Mapping 10K 2.0 Array	5	0
Illumina	Human 1M	258	42
	HumanHap550	10	4
	HumanHap300-duo	2	0
	HumanHap240S	9	4
	Linkage V (6k)	18	5

Table 6.4 *Number of SNPs in PAR1 and PAR2 in the available GeneChips offered by Affymetrics and Illumina* (Flaquer et al., 2008).

In section 6.5 the suggestion has been put forward, to use for genetic association analysis at least one SNP every 5 kb to capture most of the common variation. This would result in 540 SNPs in PAR1

and 60 SNPs in PAR2. However, in PAR1 this might not be enough since the linkage disequilibrium might be much lower than in autosomal regions of the same physical size, due to the higher recombination intensity.

Chapter 7

A NEW LINKAGE MAP FOR THE HUMAN PSEUDOAUTOSOMAL REGIONS

A fundamental problem with constructing human genetic linkage map is that sufficient data is often missing. Human geneticists cannot simply count recombinants since typically the necessary information is lacking. Consequently, it is most often not possible to specify unambiguously where recombination events occurred. The reasons could be due to marker homozygosity in parents, missing genotypes, and other reasons for an uninformative situations in some of the loci of interest. Even when parents are heterozygous, it is often unknown which alleles at linked loci are in cis and which are in trans (i.e., the linkage phase is unknown). Another problem is that genotypes cannot always be uniquely inferred from phenotype. To address this problems, theoretical approaches based on the method of maximum likelihood considering all possibilities for the missing data is applied. These approaches uses the genetic distances to maximize the probability that the observed data would have occurred.

Accurate and comprehensive linkage maps are crucial for the success of several types of genetic studies. The physical position—and, hence, the order—of the vast majority of polymorphic markers can now be readily determined from the assembled sequence of the human genome. However, unless a given set of markers are all present on a single linkage map, specification of recombination-based genetic map distances for any large set of markers remains difficult. In theory, physical map distances can be used to interpolate and estimate linkage map distances. However, the existence of extreme variability in the genomic distribution of recombination necessitates a painstaking effort to identify and utilize appropriate region-specific estimates of the recombination intensity cM/kb, making such large-scale interpolation generally impractical. Accurate estimates of map distance cannot be obtained by any means other than linkage analysis using genotype data.

To help to address these issues, we have created a sex-specific genetic map using the technique of multipoint linkage analysis and three-generation CEPH pedigrees to estimate recombination rates, including the largest set, to the best of our knowledge, of genetic markers in PAR1 and PAR2.

7.1 CEPH pedigrees

Professor Dausset in Paris, Nobel laureate in 1980, proposed and implemented the idea that a human genetic map should be constructed on the basis of reference families of specified structure.

DNA from family members, isolated from permanent cell lines, would be made available to researches world-wide so that the genetics community could participate in this effort. A foundation in France supports this idea, and a set of originally 40 (later expanded to 64) such families from France and Utah formed what became known as the CEPH families (for Centre d'Etude du Polymorphisme Humain). Each family consists of one pair of parents, an average of 8.3 offspring and up to four grandparents (Dausset et al., 1990). With the advent of highly polymorphic markers, most gene mappers did not use all families but only a subset of them.

Genotypes for the pseudoautosomal regions were obtained from the CEPH genotype database and from the Department of Genetics from the Rutgers university (Kong et al., 2004). First we selected the CEPH families where the pseudoautosomal genotypes were available for three generations. In total, 22 and 6 genetic markers were available for PAR1 and PAR2 respectively for a total of 29 CEPH families. First, to identify genotypes that lead to non-Mendelian transmission and likely to be erroneous, as well as to search for problematic pedigrees the PEDCHECK program (O'Connell and Weeks 1998) was used. However, all these genotype data have been previously cleaned, either by the groups who determined the genotypes or by other groups who have used these data. Thus, no Mendelian inconsistencies were found and no problematic pedigrees were detected. Anyway, it is well known that some erroneous genotypes do not show up as Mendelian inconsistencies. Suppose for example that two parents with genotype 1/2 and 2/3 in a locus give birth to a child with genotype 1/2, but for some error the child appears in the database with genotype 1/3, the child is still consistent with the parents and this error will not be detected as a Mendelian inconsistency. To avoid these type of errors a procedure based on recombination events was used. It was argued whether multiple crossover events could occur within the small pseudoautosomal regions. Two studies reported a double crossover in male meiosis (Schmitt et al. 1993; Rappold et al. 1994) and to date no double recombinant has been reported in female meioses. In this study we found 1 double recombinant in a male meioses and 7 in female meioses. The 6 double crossovers in female meioses was regarded a warning to potential genotype errors. We took a closer look to those individuals and these events; they all were found in 7 different individuals within the same family (FAM-66). There was no clear-cut and obvious solution to this problem. The perfect solution would be to regenotype this family. For the time being we decided to exclude this family from analysis. Another warning was given with the detection of a triple recombination in 5 male meioses. After taking a closer look to those individuals it was evident that a genotype error occurred. We decided to blank the specific genotypes in all members of the pedigree. An example of how to detect genotype

errors using crossover events is depicted in Figure 7.1

At the end, 28 CEPH families were included in the analysis including the families: *13291-13294, 1331-1334, 1340, 1341, 1344, 1346, 1347, 1349, 1350, 1362, 1375, 1377, 1413, 1416, 1418, 1420, 1421, 1423, 1424, 12, 104* and *884. T*he characteristics of these families are summarized in Table 7.1.

28 CEPH families	***n***
Individuals	413
Founders	112
Nonfounders	301
Females	202
Males	211
Pedigree size	13-21
Generations	3

Table 7.1 *illustrates the characteristics of the 28 CEPH families used to estimate the genetic maps of PAR1 and PAR2.*

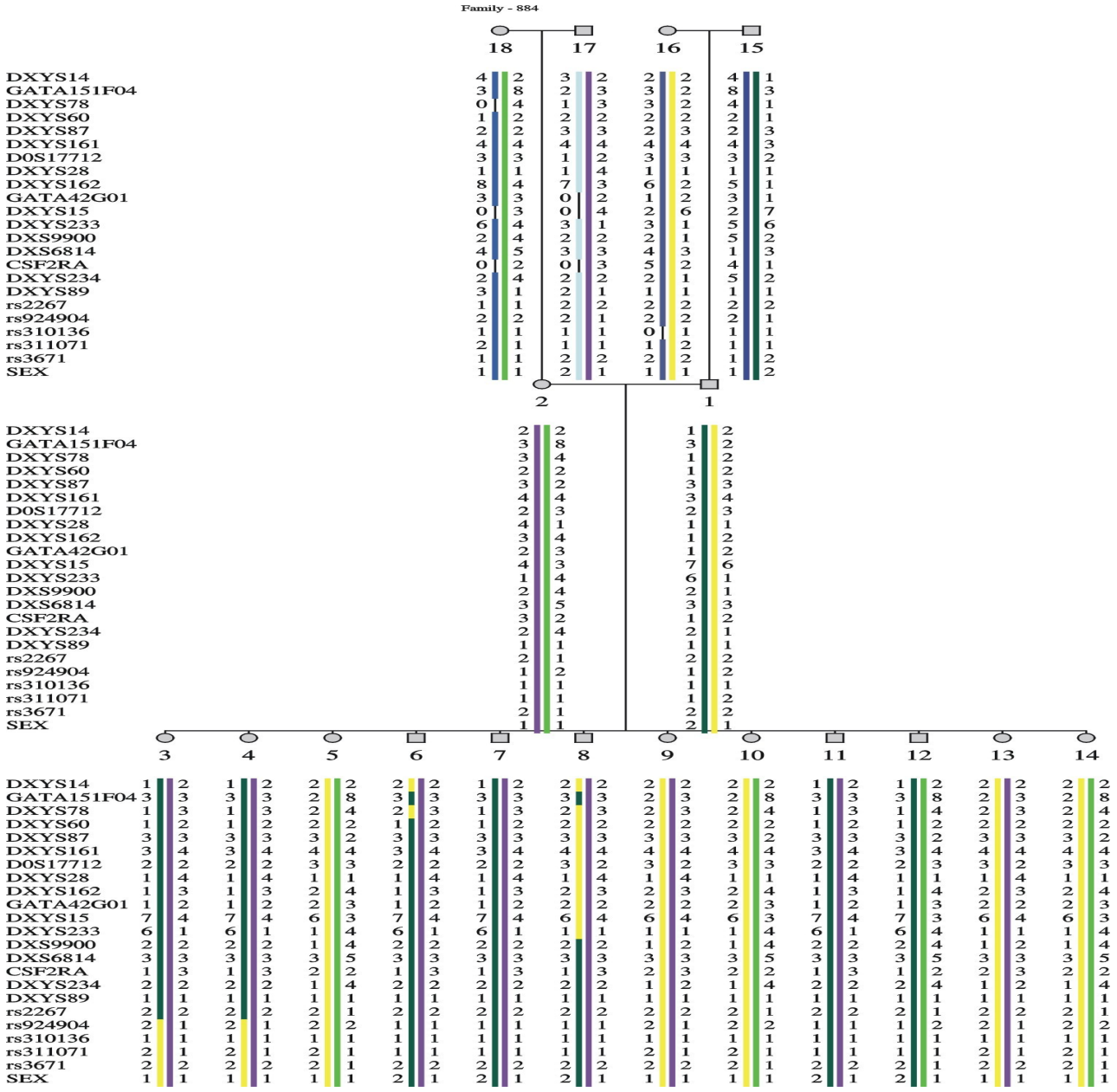


Figure 7.1 *illustrates how genotyping errors can be detected based on multiple recombination events. Because genotyping errors may lead apparent double recombinants within a short distance, it can significantly affect the overall recombination counts. Individual 6 has 3 recombinations events between markers DXYS14-GATA151F04, GATA151F04-DXYS20 and DXYS78-DXYS60. Because the second crossover event occurred in a very close locus from the first event, it has a high probability to be a genotype error. Having a close look to marker GATA151F04 one can see that for this marker individual 6 has genotype 3/3. The allele 3 (in dark green) is coming from the mother and it is were the crossover occurred. This allele 3 in the mother is likely to be correct because is present in 5 other siblings. So, the error should have occurred in individual 6 which inherited more likely the allele 2 from the mother instead of allele 3. This kind of errors are often in the databases when the difference between the real and assigned allele is only 1 or 2 bp. An analogous situation may have occurred in individual 8. In this situation the best solution would be to regenotype those individuals for the locus GATA151F04. For the time being genotypes of individuals 6 and 8 are changed to missing for locus GATA151F04. Using this technique now individuals 6 and 8 show a single recombination event, which would not give rise to objections.*

7.2 Genetic markers

All loci included to estimate the genetic map of PAR1 are listed in Table 7.2. These include the gene for the granulocyte-macrophage colony-simulating factor, CSF2RA, as well as other 16 microsatellite (STR) markers and 5 SNPs. Table 7.3 lists the genetic markers used to estimate the genetic map of PAR2. These include 3 SNPs and 3 STR markers, locus DXYS225 is closely linked to the SYBL1 gene and locus DXYS227 is very near to the IL9R gene. Heterozygosity percentages have been calculated from the genotypes of unrelated individuals from the 28 CEPH pedigrees (n=112). Probes from PAR1: *TSC0240426 ,TSC0240423* and *TSC0551442* and probes from PAR2: *TSC0268423* and *TSC0897419* were obtained from Tara Matise of the Rutgers University, Department of Genetics. The rest of the probes are publicly available from the CEPH database under www.cephb.fr.

A concerted effort was made to ensure the uniqueness of the markers in both sets. A comparison of marker name aliases and primer sequences were used to identify markers that were represented in data set more than once. Whenever possible, multiple lines of evidence, including comparison of physical positions, were sought to confirm an identified redundancy. These redundancies included only two markers that were identified by alias name. For this duplicated marker the one with the higher call rate was included in the analysis. One marker, DXYS20, matched more than 2 positions on contig(s) by e-PCR. The option Build from CRI-MAP was used to identify the most likely location. With this option the marker is placed in each possible interval between two flanking markers in the map. The resulting orders are then tested for compatibility with the database. Each order not excluded is subjected to a full maximum likelihood estimation. The order having the highest log10 likelihood is found, any order whose log10 likelihood is less than this one by more than a specified tolerance is eliminated. Using this method DXYS20 was placed between GATA151F04 and DXYS78 supporting the location given by Henke et al. (1993). At the end a set of 22 and 6 loci that, to the best of our knowledge, represent nonredundant loci within PAR1 and PAR2, respectively, where used to estimate a genetic map in PARs.

Locus	Probe	Enzyme	No. of Alleles	Heterozygosity (%)	No. of informative meioses (female/male)	location (bp)
DXYS14	CEB12	PvuII	4	79.8	364 (175/189)	4471
GATA151F04	GATA151F04B	pcr	10	63.6	105 (59/46)	35327
DXYS20	3cos-PP	TaqI	8	91.4	333 (182/151)	104476
DXYS78	cMS600	TaqI	4	98.4	395 (199/196)	173624
DXYS60	U7A	EcoRI	2	50.5	199 (69/130)	354978
DXYS87	P99	TaqI	3	60.8	267 (148/119)	456112
DXYS161	B6-Pol	TaqI	4	66.7	275 (130/145)	473000
D0S17712	UT708	pcr(n)	4	72.1	84 (32/52)	482571
DXYS28	pDP411a	TaqI	4	50.8	213 (106/107)	518276
DXYS162	AK1	TaqI	8	91.8	376 (180/196)	585000
GATA42G01	GATA42G01	pcr	5	67.5	132 (51/81)	635103
DXYS15	113D	TaqI	7	73.0	181 (85/96)	706747
DXYS233	AFMa284xc9	(AC)n	11	85.2	274 (150/124)	818694
DXS9900	GGAT3F08	pcr	5	69.2	114 (54/60)	1306914
CSF2RA	CSF2RA	TaqI	8	83.5	361 (180/181)	1389274
DXYS234	AFM273xb9	(AC)n	8	67.5	129 (50/79)	1627916
DXYS89	MS639	HinfI	4	50.1	203 (100/103)	1889000
rs2267	WIAF-2278	SNP	2	30.2	64 (38/26)	1895021
rs924904	TSC0240426	SNP	2	41.9	178 (64/114)	2578904
rs310136	TSC0240423	SNP	2	14.7	53 (22/31)	2621451
rs311071	TSC0551442	SNP	2	41.0	161 (71/90)	2642822
rs3671	WIAF-2434	SNP	2	17.2	47 (34/13)	2743668

Table 7.2 *Genetic markers included in the analysis to estimate the genetic map for PAR1.*

Locus	Probe	Enzyme	No. of Alleles	Heterozygosity (%)	No. of informative meiosis (female/male)	location (bp)
rs802488	TSC0440751	SNP	2	35.0	132 (45/87)	154698840
DXYS154	SDF1	pcr	8	76.3	118 (59/59)	154731517
DXYS225	LH1	pcr	2	51.7	28 (14/14)	154780869
DXYS227	LH2	pcr	4	69.0	22 (8/14)	154826627
rs963311	TSC0268423	SNP	2	49.0	258 (115/143)	154904883
rs188305	TSC0897419	SNP	2	48.0	245 (110/135)	154959677

Table 7.3 *Genetic markers included in the analysis to estimate the genetic map for PAR2.*

The physical position of the markers employed in each interval are represented in Figure 7.4. To improve the accuracy of the maps, the genomic sequence position of all genetic markers and hence the order of the loci were identified. Current sequence positions for 28 markers were readily identified from the Ensembl Genome browser (www.ensembl.org), the National Center for Biotechnology Information (NCBI) (www.ncbi.nlm.nih.gov) or/and from the dbSNP database (www.ncbi.nlm.nih.gov/SNP). At the end of PAR1 and thus effectively at the PAB, sex as a phenotypic marker was added to represent the SRY (sex determiner) gene. For this marker females were denoted as homozygous with alleles 1/1 and males as heterozygous with alleles 1/2.

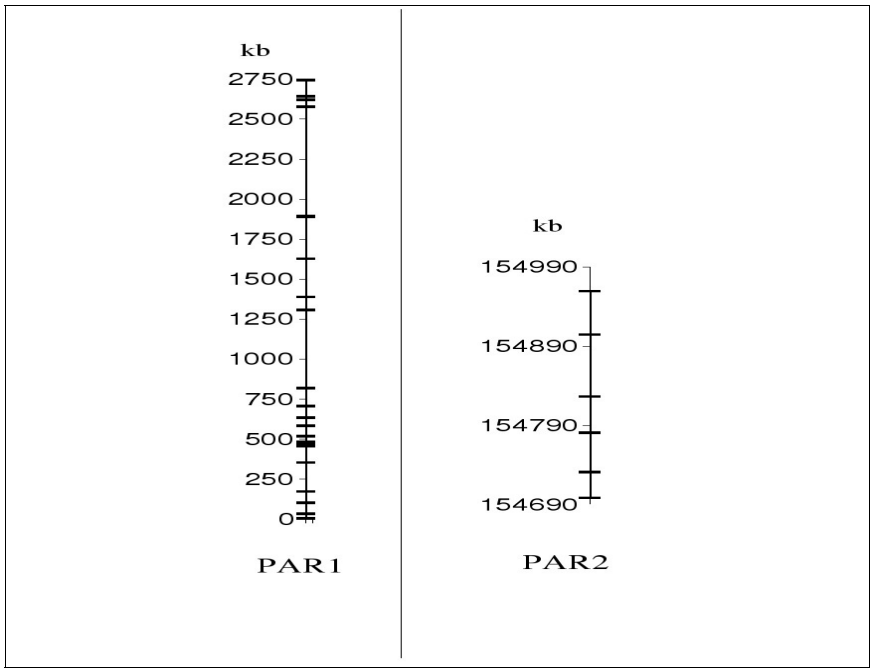


Figure 7.4 *illustrates the position and density of markers used to construct the genetic map in PAR1 and PAR2. Note that the graphs do not use the same scale*

7.3 Estimation of the genetic map for PAR1 and PAR2

Phenotypes at pseudoautosomal marker loci follow the same inheritance pattern as autosomal loci, becoming progressively more sex-linked as they become proximal to the pseudoautosomal boundary (PAB). Therefore genetic distances can be estimated with standard techniques and computer program for linkage analysis. Recombination rates are to be estimated for both sexes separately, since there is a marked difference in recombination frequency between males and females. SRY (sex) is included in the analysis as an additional two-alllelic marker in PAR1, homozygous for all females (XX) and het-

erozygous for males (XY). To establish the genetic maps the software CRI-MAP v2.4 which allows rapid, largely automated construction of multilocus linkage maps was used (Lander and Green 1987). CRI-MAP deduces missing genotypes where possible, and computes a likelihood based only on analysis of the known or deduced genotypes. The estimates of recombination fractions are based on the Lander-Green algorithm (chapter 3). Once the recombination fractions were estimated, to convert them into cM the identity function was applied. The identity function (1% recombination rate correspond to 1 cM) has been suggested to be the most suitable mapping function for the PARs (Flaquer et al. 2008).

The multilocus genetic map was estimated using a total of 23 genetic markers (including the sex-marker) in PAR1 and 6 genetic markers in PAR2. For each region 28 CEPH families were analyzed. In determining the haplotypes of 245 offspring, 112 crossover events in male meioses were observed and 12 crossover events in female meioses. A double recombination event in a male meiosis took place in family 13294 for individual 5. This double event was analyzed cautiously for a potential genotype error. It was decided to accept it after finding out that this event could be a replication from the double event found by Rappold et al. (1994) in the same individual. It was considered a replication after determining that the genotypes came from different labs and were genotyped by different methods.

The resulting genetic maps, for PAR1 and PAR2, are illustrated in Figure 7.3. Black dots represent the estimated genetic map for males, and gray dots for females. A high rate of recombination in female meioses is seen only at the Xp telomere within the first 100 kb. Within this telomeric region no difference in male and female recombination rates are observed, resulting in similar genetic distances for both sexes. After 100 kb, genetic distances are becoming increasingly different between males and females as they approach the PAB1. The estimated length of PAR1 is 55 cM for males and 6 cM for females. On the other hand, PAR2 does not seem to be significantly different in terms of recombination events between males and females, resulting in similar genetic maps. We detected one recombination event in male meioses and none in female meioses. These numbers are in the range suggested by Freije et al. (1994).

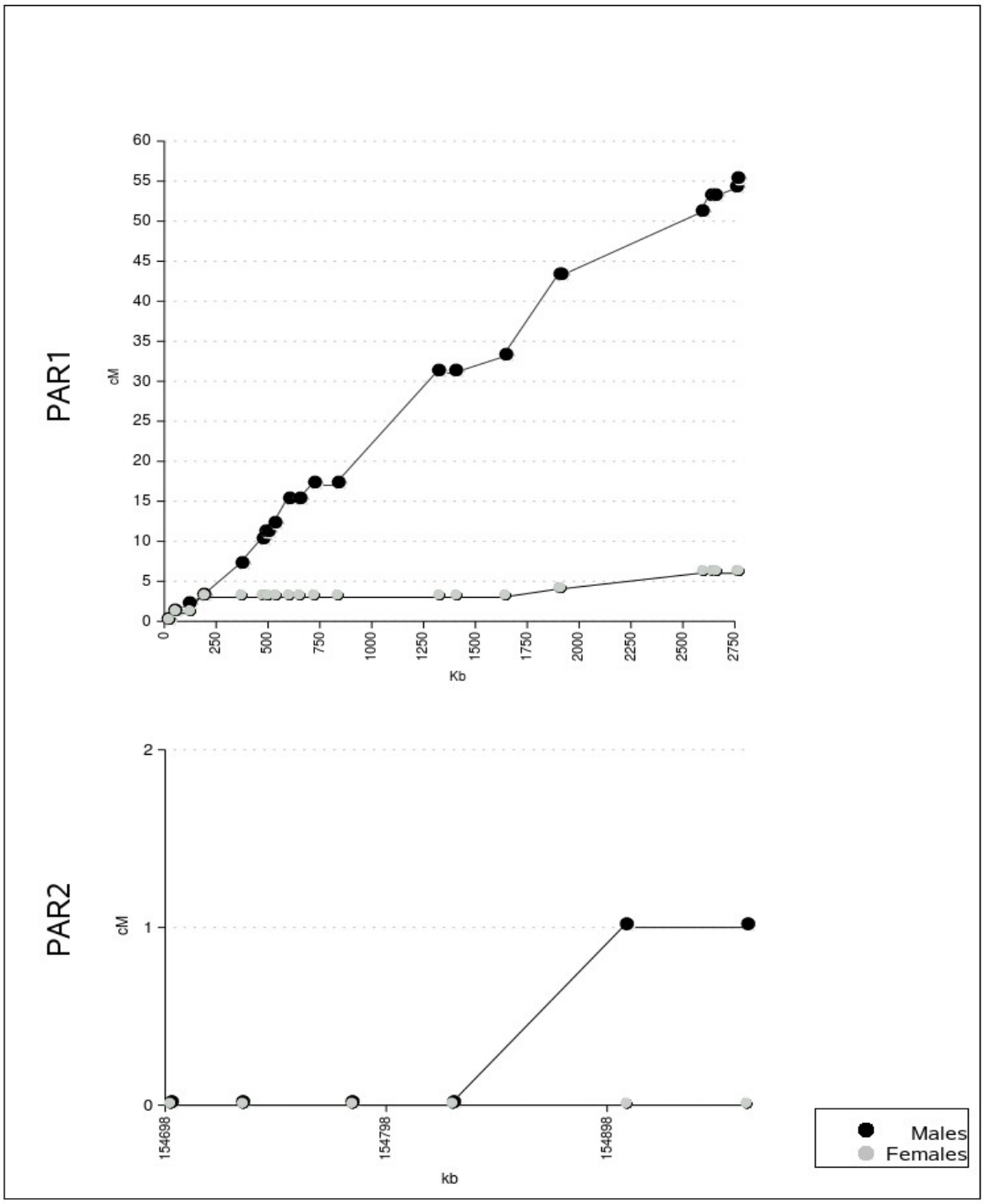


Figure 7.3 *illustrates the positions of the estimated genetic map (cM) in relation to the positions of the physical map (bp) of PAR1 and PAR2. Black lines represent the male map and the gray lines the female map.*

The resulting genetic length for PAR1 was compared with the genetic lengths provided by each of the previously published maps. We also compared the genetic length in three subintervals of PAR1, close to the telomere (~0.9-580 kb), in the middle of PAR1 (~580-1628 kb) and close to the PAB1 (~1628-2750 kb). Pairwise Z-tests (two-tailed) were used to statistically establish differences in recombination rates. P-values from the Z-tests after Bonferroni correction are illustrated in Table 7.4. When comparing the entire genetic length of PAR1 in males, large significant differences were detected with respect to the genetic lengths suggested by Schmitt and Rutgers ($P_{Schmitt}$=2.6×10^{-6}, $P_{Rutgers}$=4.9×10^{-11}). The significant difference shown in the genetic length suggested by Henke is at border line (P_{Henke}=0.04686). The genetic lengths proposed by Rouyer and by Lien did not show statistical differences with respect to the new genetic length although the Lien's map displayed a significant shorter genetic length close to the telomere and close to the PAB1 (P_{Lien}=3.6×10^{-6}, P_{Lien}=0.00460, respectively). When comparing the genetic length in females the Rutgers' map was the only one to show a slightly shorter length ($P_{Rutgers}$=0.04942).

The map obtained for PAR2 was not statistically compared with the other maps since only one recombination event was detected in male meiosis. This is in the same range as in the already published maps for PAR2.

			Lien		**Rouyer**		**Schmitt**		**Henke**		**Rutgers**	
	Length (~kb)	θ_{new}	θ	**P-value**	θ	**P-value**	θ	**P-value**	θ	**P-value**	θ	**P-value**
Males	0.9-2750 (entire PAR1)	0.55	0.54	0.39621	0.49	0.15836	0.37	**2.6x10^{-6}**	0.48	**0.04686**	0.26	**4.9x10^{-11}**
	0.9-580	0.17	0.07	**3.6x10^{-6}**	0.11	0.30521	n.c.	-	0.11	0.08592	n.c.	-
	580-1628	0.16	0.33	1.00000	0.22	1.00000	0.21	1.00000	0.21	1.00000	0.09	**0.04332**
	1628-2750	0.22	0.14	**0.00460**	0.16	0.37488	0.16	0.05515	0.16	0.18004	0.17	0.19864
Females	0.9-2750 (entire PAR1)	0.06	n.a.	-	0.06	0.50000	n.a.	-	0.04	0.17308	0.03	**0.04942**
	0.9-585	0.03	n.a.	-	0.00	0.15139	n.a.	-	0.01	0.31923	n.c.	-
	585-1628	0.00	n.a.	-	0.02	1.00000	n.a.	-	0.02	1.00000	0.00	1.00000
	1628-2750	0.03	n.a.	-	0.04	1.00000	n.a.	-	0.01	0.22349	0.03	1.00000

Table 7.4 *Comparison of recombination fractions between our new map and each of the published maps for PAR1.*

Θ_{new}: refers to the recombination fraction estimated in this work.
n.c.: not covered (genetic markers in this interval were not used).
n.a.: not available (map estimates based on sperm cells are only available for males).
P-values from Z-test (two-tailed) after Bonferroni correction for multiple testing. Hypothesis test for differences between two proportions (H_0: $\theta_{new} = \theta$ vs H_1: $\theta_{new} \neq \theta$).

The relation between the genetic and physical map of the male and female X chromosomes inferred from this study is represented in figure 4. The X-specific data is from the study, "A second-generation combined linkage-physical map of the human genome" (Matise el al. 2007).

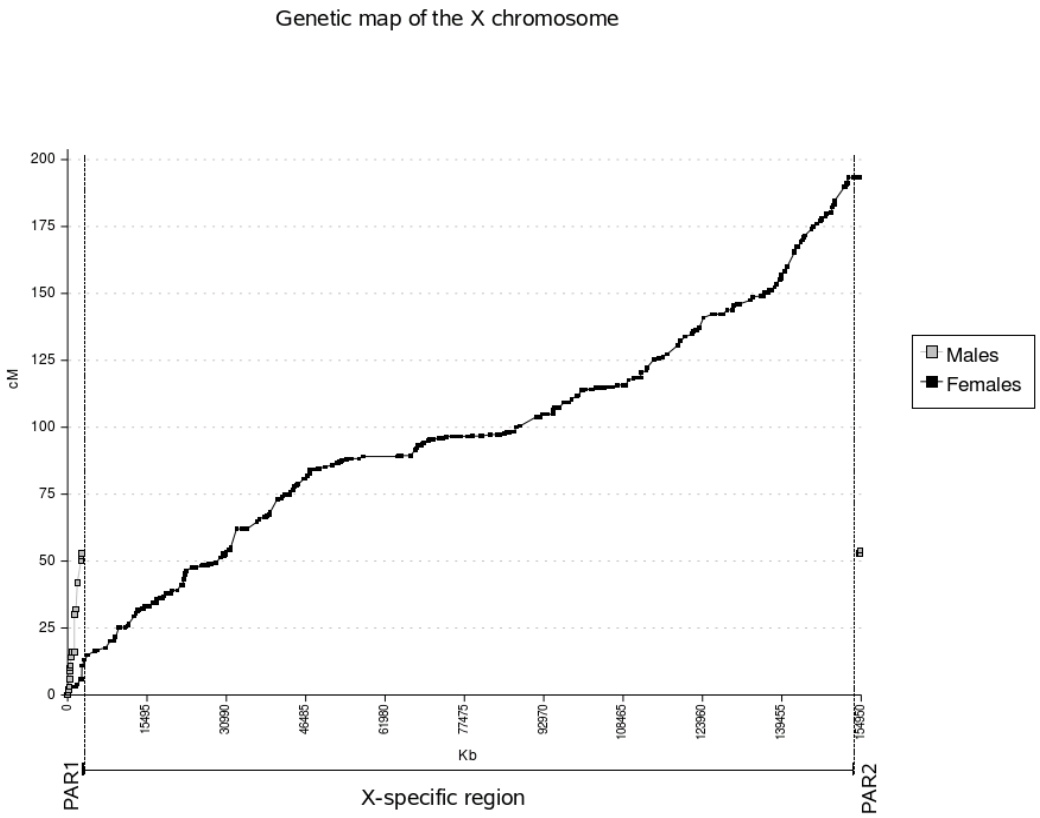


Figure 7.4 *represents the relation between the genetic and physical map of the human X chromosome. The X-specific data is from the Kong et al (2004). PAR1 and PAR2 are based on the estimated genetic map of this work.*

The human X chromosome exhibits some peculiarities compared to the autosomal chromosomes since females have two X chromosomes and males only one. Figure 7.4 represents the relation between the genetic and physical map of the human X chromosome based on this study. The genetic map of the male X chromosome can be only considered where recombination occur. The genetic length of the male X chromosome is very short because it can only recombine with the Y chromosomes in the PARs. the genetic map of the X chromosome in females is much longer because the two X chromosomes can recombine on their entire length and the genetic length is estimated to be about 193.3 cM.

Chapter 8

SUMMARY AND CONCLUSIONS

This work is concerned with the methods of linkage and association analysis as a techniques to map genes for complex diseases. The main intention is how suitable are these methods when a disease susceptibility gene is located in the pseudoautosomal regions (PARs) of the sex chromosomes.

In chapter 6 the two pseudoautosomal regions are presented in detail. The fact that in humans exist two different sex chromosomes, females have two X chromosomes and males have one X and one Y chromosome, incite to some special peculiarities that have to be carefully considered when using linkage and association methods.

One of the most important features of the PARs regions is the number of crossover events during meiosis. In female meioses, the number of crossover stands within the expected range. In male meioses, crossover activity increases drastically, resulting on average 7-fold higher than in females. These large differences in recombination rates are translated into different genetic map lengths for males and females. Because accurate and comprehensive genetic maps are crucial in multipoint linkage methods and they are constructed based on the number of crossovers during meiosis, it is one of the priorities to use two different linkage maps, one for males and one for females, when using multipoint linkage methods for the PARs.

Another peculiarity is the mode of inheritance. The segregation pattern of a disease susceptibility gene for a certain disease on the PARs depends on their location relative to the pseudoautosomal boundary (PAB). When a father carries a disease susceptibility gene located close to PAB and on his Y chromosome, only male offspring will be affected by the disease. If the disease-causing gene is close to PAB but on his X chromosome, only female offspring will be affected by the disease. In case that the disease susceptibility gene is close to the telomere it will follow the same pattern of transmission like any autosomal dominant disease, where any offspring, male or female, get the same chance of being affected. This feature should be taken into account specially in those methods based on allelic sharing, because the expected number of alleles shared IBD in PARs between an affected pair will depend on the sex of the pair.

Another property of the PARs is that due to the high recombination rate in males in this regions the LD is much lower than in autosomal regions of the same physical size. So, when testing for associa-

tion a larger number of loci should be considered. In addition false positive could arise if a causal variant is strongly associated with one sex, and the sex distribution is different between cases and controls.

The two PARs have drawn considerable interest from researchers in cytogenetics, cytology, evolutionary biology and developmental genetics. However, theses two regions have been widely ignored by the two genetic mapping approaches, linkage and association analysis methods. To date, very few analytical methods have been adapted to deal with the peculiarities of these regions, which may have led researches to ignore them. At least 29 genes are known to be located in the PARs, most of them of unknown function. Recent advances in large-scales LD analysis and GeneChip technologies for SNP genotyping have forwarded genome-wide association studies for complex diseases. However, the available GeneChips contain still a poor number of SNPs in the PARs.

Genetic maps are crucial for the success of gene mapping projects and for several other types of genetic studies. The physical position - and, hence, the order - of the practically all of polymorphic markers can now be readily determined from the assembled sequence of the human genome, and several large-scale genome wide linkage maps have been published. Nevertheless, the PARs have been largely neglected in genetic map construction. Several attempts have been done to construct genetic maps for the PARs, resulting in discrepancies. Apparently these discrepancies could be due to an insufficient number of markers or meioses considered for the estimation of recombination rates. Estimated lengths of existing maps for PAR1 varies between 26-54 cM in males and between 2.8-6 cM in females. Accurate estimates of meiotic map distance can be obtained by linkage analysis using genotype data in families and a considerable number of markers to cover the desired chromosomal segment. The difference between male and female genetic maps, chromosomal position and population under study, are a challenge to genetic map construction in diploid organisms in which sex is determined by a pair of different sex chromosomes.

To improve linkage analysis methods for complex diseases caused by pseudoautosomal genes, reliable genetic maps in these regions are crucial. Genetic maps increasing in length have been observed in telomeric and subtelomeric regions of different human chromosomes, reflecting a higher recombination rate per physical length unit in these regions. PAR1 exhibits these subtelomeric features in a particular way. Genetic maps in males are markedly longer than genetic maps in females, as inferred here from the comparison of male and female recombination events and from multipoint linkage data. The biological process for these differences is not yet well understood, but they exist and should be considered in genetic studies using sex-specific maps instead of sex-averaged maps. The human X

chromosome exhibits some peculiarities compared to the autosomes chromosomes since females have two X chromosomes and males only one. Based on our map, it is determined how genetically different, in genetic size, the female X chromosome is from the male X chromosome. The genetic map of the male X chromosome can be considered only where crossovers occur. The genetic length is therefore very short (56 cM) because the X chromosome can only recombine with the Y chromosome in the PARs. The genetic map of the X chromosome in females is much longer because the two X chromosomes can recombine on their entire length, the genetic length is estimated to be about 198 cM. For this reason, the human sex chromosomes cannot be treated like the autosomes when trying to map sex linked genes.

Comparisons between the new map and those already published for PAR1 revealed that our map provides statistically different estimates for genetic distances. The genetic maps proposed by Rouyer and by Lien were the only ones that did not display significant differences when comparing the whole PAR1. Nevertheless the latter showed a significant shorter genetic length close to the telomere and close to the PAB1. These differences could be due to the fact that close to the Xp telomere the Lien's estimates are based on only 3 genetic markers while the new map estimates are based on 11 genetic markers. The same conclusion holds close to the PAB1, this last interval is only covered with 2 genetic markers in Lien's map and with 7 markers were used for the new estimates. The differences found in the maps suggested by Rutgers and by Schmitt could be due to the lack of genetic markers at the first 750 kb failing to account for recombination events in this regions and resulting then in a much shorter genetic length. Therefore, it is concluded that the new sex-specific map presented in this work is based on the largest set of genetic markers in the PARs and thus represents the most accurate resource for obtaining genetic map information for these two regions.

To date, one of the most accurate and widely used genetic maps for the human genome is the Rutgers' map v2. However, this map shows a weakness in the PARs since a low number of genetic markers was considered in these two regions. To the improvement of genetic mapping projects the new map for the PARs presented in this work will be integrate in the Rutgers' map v.2. Although there is still a long way to go for understanding the exact mechanisms and the precise function(s) of the PARs, the new genetic map presented here could be a first step to mapping new pseudoautosomal genes responsible for complex diseases. It is hoped that researches will take these two regions into proper account when performing genome wide studies. So far, the PARs are effective gaps in genome wide analyses. In our view, in the context of such efforts, screening for pseudoautosomal linkage should not be neglected.

Appendix

- **A1) Tthe expected distribution of alleles IBD for sib pairs under H_0 follows a Binomial distribution with probabilities $z_0=1/4$, $z_1=1/2$ *and* $z_2=1/4$:**

Two sibling share an allele IBD if there is a founder in the pedigree that has transmitted one of its two alleles to both sibs. Consider a pedigree with a sib pair without inbreeding loops and where all founders are unrelated. If each of the two sibs gets one allele from the father, the probability is 0.5 that these two alleles are IBD. Similarity, the probability is 0.5 that the two alleles inherited from the mother are IBD. Let N be the total number of allele shared IBD by the sibs. Then N can be viewed as the number of successes in two independent experiments, where success means that the parent transmits two alleles IBD to the sib pair. Since the probability of success is 0.5, then N~Bin(2,0.5) where:

$$P(N=0)=(1-0.5)^2=0.25$$
$$P(N=1)=\binom{2}{1}0.5(1-0.5)=0.5$$
$$P(N=2)=0.5^2=0.25$$

«»

- **A2) Validation of Equation (3.13):**

$$(\boldsymbol{p}^{\boldsymbol{L}}_{\boldsymbol{j+1}})_{v_{j+1}} = \frac{[(\boldsymbol{p}^{\boldsymbol{L}}_{\boldsymbol{j}})^T \boldsymbol{T}(\theta_{j,j+1})]_{v_{j+1}} (\boldsymbol{q}_{\boldsymbol{j+1}})_{v_{j+1}}}{(\boldsymbol{p}^{\boldsymbol{L}}_{\boldsymbol{j}})^T \boldsymbol{T}(\theta_{j,j+1}) \boldsymbol{q}_{\boldsymbol{j+1}}} \qquad (3.12)$$

$$\doteq \frac{\sum_{w_j \in V} q_{j+1}(v_{j+1}) T_{w_j, v_{j+1}}(\theta_{j,j+1}) p^L_j(w_j)}{\sum_{w_j, w_{j+1} \in V} q_{j+1}(v_{j+1}) T_{w_j, w_{j+1}}(\theta_{j,j+1}) p^L_j(w_j)}$$

$$\doteq \frac{\sum_{w_j \in V} P(M_{j+1}|V_{j+1}=v_{j+1}) P(V_{j+1}=v_{j+1}|V_j=w_j) P(V_j=w_j|M_1,\ldots,M_j)}{\sum_{w_j, w_{j+1} \in V} P(M_{j+1}|V_{j+1}=w_{j+1}) P(V_{j+1}=w_{j+1}|V_j=w_j) P(V_j=w_j|M_{1,}\ldots,M_j)}$$

$$\doteq \frac{P(M_{j+1}|V_{j+1}=v_{j+1}) \sum_{w_j \in V} P(V_{j+1}=v_{j+1} \cap V_j=w_j|M_1,\ldots,M_j)}{\sum_{w_{j+1} \in V} P(M_{j+1}|V_{j+1}=w_{j+1}) \sum_{w_j \in V} P(V_{j+1}|w_{j+1} \cap V_j=w_j|M_1,\ldots,M_j)}$$

$$\doteq \frac{P(M_{j+1}|V_{j+1}=v_{j+1}) P(V_{j+1}=v_{j+1}|M_1,\ldots,M_j)}{\sum_{w_{j+1} \in V} P(M_{j+1}|V_{j+1}=w_{j+1}) P(V_{j+1}=w_{j+1}|M_1,\ldots,M_j)}$$

$$= \frac{P(M_{j+1}|V_{j+1}, M_{1,}\ldots,M_j) P(V_{j+1}=v_{j+1}|M_1,\ldots,M_j)}{\sum_{w_{j+1} \in V} P(M_{j+1}|V_{j+1}=w_{j+1}, M_{1,}\ldots,M_j) P(V_{j+1}=w_{j+1}|M_1,\ldots,M_j)}$$

$$= P(V_{j+1}=v_{j+1}|M_1,\ldots,M_{j+1})$$

The dots over the equalities stand for : "·" element-wise writing of vectors; "··" employ definition of probabilities; "···" use Markov property $P(V_{j+1}=v_{j+1} \mid V_j=v_j, M_1,\ldots,M_j) = P(V_{j+1}=v_{j+1} \mid V_j=v_j)$ because v_{j+1} depends on genotypes of markers 1 to j only via v_j ; "::" $P(M_{j+1} \mid V_{j+1}=v_{j+1}) = P(M_{j+1} \mid V_{j+1}=v_{j+1}, M_1,\ldots,M_j)$ because M_{j+1} depends only on v_{j+1}. «»

- **A3) Validation of Equation (3.17):**

$$(\boldsymbol{P}_{\boldsymbol{complete,j}})_{v_j} = \frac{[(\boldsymbol{p}_{\boldsymbol{j-1}}^{\boldsymbol{L}})^T \boldsymbol{T}(\theta_{j-1,j})]_{v_j} (\boldsymbol{p}_{\boldsymbol{j}}^{\boldsymbol{R}})_{v_j}}{(\boldsymbol{p}_{\boldsymbol{j-1}}^{\boldsymbol{L}})^T \boldsymbol{T}(\theta_{j-1,j})\, \boldsymbol{p}_{\boldsymbol{j}}^{\boldsymbol{R}}} \qquad (3.16)$$

$$= \frac{\sum_{w_{j-1}\in V} p_j^R(v_j)\, T_{w_{j-1},v_j}(\theta_{j-1,j})\, p_{j-1}^L(w_{j-1})}{\sum_{w_{j-1},w_j\in V} p_j^R(w_j)\, T_{w_{j-1},w_j}(\theta_{j-1,j})\, p_{j-1}^L(w_{j-1})}$$

$$= \frac{\sum_{w_{j-1}\in V} P(V_j=v_j|M_j,\ldots,M_k)\, P(V_j=v_j|V_{j-1}=w_{j-1})\, P(V_{j-1}=w_{j-1}|M_1,\ldots,M_{j-1})}{\sum_{w_{j-1},w_j\in V} P(V_j=w_j|M_j,\ldots,M_k)\, P(V_j=w_j|V_{j-1}=w_{j-1})\, P(V_{j-1}=w_{j-1}|M_1,\ldots,M_{j-1})}$$

$$= \frac{P(V_j=v_j|M_j,\ldots,M_k) \sum_{w_{j-1}\in V} P(V_j=v_j\cap V_{j-1}=w_{j-1}|M_1,\ldots,M_{j-1})}{\sum_{w_j\in V} P(V_j=w_j|M_j,\ldots,M_k) \sum_{w_{j-1}\in V} P(V_j=w_j\cap V_{j-1}=w_{j-1}|M_1,\ldots,M_{j-1})}$$

$$= \frac{P(V_j=v_j|M_j,\ldots,M_k)\, P(V_j=v_j|M_1,\ldots,M_{j-1})}{\sum_{w_j\in V} P(V_j=w_j|M_j,\ldots,M_k)\, P(V_j=w_j|M_1,\ldots,M_{j-1})}$$

$$= \frac{\frac{P(M_j,\ldots,M_k|V_j=v_j)\,P(V_j=v_j)}{P(M_j,\ldots,M_k)}\, P(V_j=v_j|M_1,\ldots,M_{j-1})}{\sum_{w_j\in V} \frac{P(M_j,\ldots,M_k|V_j=w_j)\,P(V_j=w_j)}{P(M_j,\ldots,M_k)}\, P(V_j=w_j|M_1,\ldots,M_{j-1})}$$

$$= \frac{P(M_j,\ldots,M_k|V_j=v_j,M_1,\ldots,M_{j-1})\, P(V_j=v_j|M_1,\ldots,M_{j-1})}{\sum_{w_j\in V} P(M_j,\ldots,M_k|V_j=w_j,M_1,\ldots,M_{j-1})\, P(V_j=w_j|M_1,\ldots,M_{j-1})}$$

$$= P(V_j=v_j|M_1,\ldots,M_k)$$

Using that $P(V_j=v_j \mid V_{j-1}=v_{j-1})=P(V_j=v_j \mid V_{j-1}=v_{j-1},M_1,\ldots,M_{j-1})$ as well as $P(M_j,\ldots,M_k \mid V_j=v_j) = P(M_j,\ldots,M_k \mid V_j,M_1,\ldots,M_{j-1})$. Further more, $P(V_j=v_j)$ is the a priori probability for inheritance vector v_j at marker j and given by $1/2^m$ for all inheritance vectors. «»

- **A4) Validation of Equation (3.19)**

$$\boldsymbol{P}_{\boldsymbol{complete,j}} = \frac{\mathbf{1}^T \boldsymbol{Q_1} \boldsymbol{T}(\theta_{1,2}) \boldsymbol{Q_2} \ldots \boldsymbol{T}(\theta_{j-1,j}) \circ \boldsymbol{Q_j} \boldsymbol{T}(\theta_{j,j+1}) \ldots \boldsymbol{Q_{k-1}} \boldsymbol{T}(\theta_{k-1,k}) \boldsymbol{Q_k} \mathbf{1}}{\mathbf{1}^T \boldsymbol{Q_1} \boldsymbol{T}(\theta_{1,2}) \boldsymbol{Q_2} \ldots \boldsymbol{T}(\theta_{j-1,j}) \boldsymbol{Q_j} \boldsymbol{T}(\theta_{j,j+1}) \ldots \boldsymbol{Q_{k-1}} \boldsymbol{T}(\theta_{k-1,k}) \boldsymbol{Q_k} \mathbf{1}} \qquad (3.18)$$

For proving the equality, let's first consider the denominator:

$$\mathbf{1}^T \boldsymbol{Q_1} \boldsymbol{T}(\theta_{1,2}) \boldsymbol{Q_2} \ldots \boldsymbol{T}(\theta_{j-1,j}) \boldsymbol{Q_j} \boldsymbol{T}(\theta_{j,j+1}) \ldots \boldsymbol{Q_{k-1}} \boldsymbol{T}(\theta_{k-1,k}) \boldsymbol{Q_k} \mathbf{1}$$

$$\begin{aligned}
&= \sum_{v_1,\ldots v_k \in V} q_k(v_k) T_{v_{k-1},v_k}(\theta_{k-1,k}) q_{k-1}(v_{k-1}) \cdot \ldots \cdot q_2(v_2) T_{v_1,v_2}(\theta_{1,2}) q_1(v_1) \\
&= \sum_{v_1,\ldots v_k \in V} P(M_k|V_k=v_k) P(V_k=v_k|V_{k-1}=v_{k-1}) P(M_{k-1}|V_{k-1}=v_{k-1}) \cdot \ldots \cdot \\
&\qquad \cdot \ldots \cdot P(M_2|V_2=v_2) P(V_2=v_2|V_1=v_1) P(M_1|V_1=v_1) \\
&= \sum_{v_1,\ldots v_k \in V} P(M_k|V_k=v_k) P(M_{k-1}|V_{k-1}=v_{k-1}) \cdot \ldots \cdot P(M_2|V_2=v_2) P(M_1|V_1=v_1) \cdot \\
&\qquad P(V_k=v_k|V_{k-1}=v_{k-1}) \cdot \ldots \cdot P(V_2 d=v_2|V_1=v_1) \\
&\doteq \sum_{v_1,\ldots v_k \in V} P(M_k|M_1,\ldots M_{k-1}, V_1=v_1,\ldots,V_k=v_k) \;\cdot\; P(M_{k-1}|M_1,\ldots M_{k-2}, V_1=v_1,\ldots,V_k=v_k) \\
&\qquad \cdot \ldots \cdot P(M_2|M_1, V_1=v_1,\ldots,V_k=v_k) \;\cdot\; P(M_1|V_1=v_1,\ldots,V_k=v_k) \;\cdot \\
&\qquad P(V_k=v_k|V_1=v_1,\ldots,V_{k-1}=v_{k-1}) P(V_{k-1}=v_{k-1}|V_1=v_1,\ldots,V_{k-2}=v_{k-2}) \\
&\qquad \cdot \ldots \cdot P(V_3=v_3|V_1=v_1,V_2=v_2) P(V_2=v_2|V_1=v_1) P(V_1=v_1) \frac{1}{P(V_1=v_1)} \\
&= \sum_{v_1,\ldots v_k \in V} P(M_1,\ldots,M_k|V_1=v_1,\ldots,V_k=v_k) P(V_1=v_1,\ldots,V_k=v_k) 2^m \\
&= \sum_{v_1,\ldots v_k \in V} P(M_1,\ldots,M_k \cap V_1=v_1,\ldots,V_k=v_k) 2^m \\
&= 2^m P(M_1,\ldots,M_k) = 2^m L_{all\ Markers}
\end{aligned}$$

The equality "·" holds because of a conditional independence argument: the marker data M_j depends on genotypes and inheritance vectors at other markers only via v_j so that $P(M_j|V_j)=P(M_j|M_1,\ldots,M_k\, V_1,..,V_k)$. In this step, the Markov property is used $P(V_j=v_j|\, V_{j-1}=v_{j-1}) = P(V_j=v_j|\, V_1=v_1,\ldots,V_{j-1}=v_{j-1})$. The numerator is identical to the denominator except for the sum over v_j from marker j:

$$\begin{aligned}
&[\mathbf{1}^T \boldsymbol{Q_1} \boldsymbol{T}(\theta_{1,2}) \boldsymbol{Q_2} \ldots \boldsymbol{T}(\theta_{j-1,j})]_{v_j} \; [\boldsymbol{Q_j} \boldsymbol{T}(\theta_{j,j+1}) \ldots \boldsymbol{Q_{k-1}} \boldsymbol{T}(\theta_{k-1,k}) \boldsymbol{Q_k} \mathbf{1}]_{v_j} \\
&= \sum_{v_1,\ldots,v_{j-1},v_j,\ldots,v_k \in V} P(M_1,\ldots,M_k \cap V_1=v_1,\ldots,V_k=v_k) 2^m \\
&= 2^m P(M_1,\ldots,M_k \cap V_j=v_j) \;=\; 2^m L_{all\ Markers}(v_j)
\end{aligned}$$

Finally one obtains,

$$\begin{aligned}
[\boldsymbol{P}_{\boldsymbol{complete,j}}]_{v_j} &= \frac{[\mathbf{1}^T \boldsymbol{Q_1} \boldsymbol{T}(\theta_{1,2}) \boldsymbol{Q_2} \ldots \boldsymbol{T}(\theta_{j-1,j})]_{v_j} \; [\boldsymbol{Q_j} \boldsymbol{T}(\theta_{j,j+1}) \ldots \boldsymbol{Q_{k-1}} \boldsymbol{T}(\theta_{k-1,k}) \boldsymbol{Q_k} \mathbf{1}]_{v_j}}{\mathbf{1}^T \boldsymbol{Q_1} \boldsymbol{T}(\theta_{1,2}) \boldsymbol{Q_2} \ldots \boldsymbol{T}(\theta_{j-1,j}) \boldsymbol{Q_j} \boldsymbol{T}(\theta_{j,j+1}) \ldots \boldsymbol{Q_{k-1}} \boldsymbol{T}(\theta_{k-1,k}) \boldsymbol{Q_k} \mathbf{1}} \\
&= \frac{2^m L_{all\ markers}(v_j)}{2^m L_{all\ markers}} \;=\; L_{rel,\,all\ markers}(v_j) \\
&= \frac{2^m P(M_1,\ldots,M_k \cap V_j=v_j)}{2^m P(M_1,\ldots,M_k)} \;=\; P(V_j=v_j|M_1,\ldots,M_k)
\end{aligned}$$

<< >>

- **A5) Validation of Equation (3.24)**

$$\bar{S}(x, \phi^d) = \frac{L(x)}{L(x\ unlinked)} = LR(x) \qquad (3.24)$$

$$\bar{S}(x, \phi^d) = \sum_{w \in V} LR(w)\, P(v(x) = w) = \frac{\sum_{w \in V} P(\phi^d|w)\, P(V(x) = w)}{\sum_{w \in V} P(\phi^d|w)\, P_{priori}(w)} = \frac{\sum_{w \in V} P(\phi^d|w) P(w|M_1, \ldots, M_k, x)}{\sum_{w \in V} P(\phi^d|w)\, P(w|M_1, \ldots, M_k, x\ unlinked)}$$

$$\doteq \frac{P(\phi^d|M_1, \ldots, M_k, x)}{P(\phi^d|M_1, \ldots, M_k, x\ unlinked)} = \frac{P(\phi^d \cap M_1, \ldots, M_k|x)}{P(\phi^d \cap M_1, \ldots, M_k|x\ unlinked)} \cdot \frac{P(M_1, \ldots, M_k|x\ unlinked)}{P(M_1, \ldots, M_k|x)}$$

$$= \frac{P(\phi^d \cap M_1, \ldots, M_k|x)}{P(\phi^d \cap M_1, \ldots, M_k|x\ unlinked)} = \frac{L(x)}{L(x\ unlinked)} = LR(x)$$

In the first equality of the second line "·" considers that one can replace the probability $P(\phi^d|w)$ by $P(\phi^d|w, M_1, \ldots, M_k, x)$ and/or $P(\phi^d|w, M_1, \ldots, M_k, x\ unlinked)$. The reason for this is that the disease phenotypes depend only on the inheritance vector of the concerned disease locus. Further, in the last steps the probability of the marker genotypes does not depend on the putative position *x* of the disease locus. <<>>

REFERENCES

[1] Abecasis GR, Cherny SS, Cookson WO, Cardon LR (2001) Merlin - rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 30:97–101

[2] Aitken RJ, Marshall Graves JAM (2002) The future of sex. Nature 415:963

[3] Armitage P (1955) Tests for linear trends in proportions and frequencies. Biometrics 11:375-386

[4] Barnard GA (1949) Statistical inference. J Roy Statistic Soc 61:734-747

[5] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing (1995) J Roy Statistic Soc of London, Series B 57:289-300

[6] Blackwelder WC, Elston RC (1985) A comparison of sib-pair linkage tests for disease susceptibility loci. Genet Epidemiol 2(1):85–97

[7] Blaschke RJ, Rappold G (2006) The pseudoautosomal regions, SHOX and disease. Curr Opin Genet Dev 16:233–239

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

[9] Broman KW, Murray JC, Sheffield VC, White RL, Weber JL (1998) Comprehensive human genetic maps: Individual and sex-specific variation in recombination. Am J Hum Genet 63:861–869

[10] Burgoyne PS (1982) Genetic homology and crossing over in the X and Y chromosomes of mammals. Hum Genet 61:85-90

[11] Bussell JJ, Pearson NM, Kanda R, Filatov DA, Lahn BT (2005) Human polymorphism and human-chimpanzee divergence in pseudoautosomal region correlate with local recombination rate. Gene 368:94

[12] Byun HJ, Hong IK, Kim E, Jin YJ, Jeoung DI, Hahn JH et al. (2006) A splice variant of CD99 increases motility and MMP-9 expression of human breast cancer cells through the AKT-, ERK-, and JNK-dependent AP-1 activation signaling pathways. J Biol Chem 281: 34833–34847

[13] Cannings C, Thompson EA, Skolnick MH (1978) Probability functions on complex pedigrees. Adv Appl Prob 10:26-61

[14] Cantagrel V, Lossi A-M, Boulanger S, Depetris D, Mattei M-G, Gecz J, et al. (2004) Disruption of a new X linked gene highly expressed in brain in a family with two mentally retarded males. J Med Genet 41:736–742

[15] Carter TC, Falconer DS (1951) Stocks for detecting linkage in the mouse, and the theory of their design. J Genet 50:307-323

[16] Charlesworth D, Charlesworth B, Marais G (2005) Steps in the evolution of heteromorphic sex chromosomes. Heredity 95:118-128

[17] Clerget-Darpoux F, Bonaïti-Pellié C, Hochez J (1986) Effects of misspecifying genetic parameters in lod score analysis. Biometrics 42(2):393–399

[18] Cochran WG (1954) Some methods for strengthening the common chi-squared tests. Biometrics 10:417-451

[19] Collins A, Frezal J, Teague J, Morton NE (1996) A metric map of humans:23,500 loci in 850 bands. Proc Natl Acad Sci USA 93:14771-14775

[20] Commenges D (1994) Robust genetic linkage analysis based on a score test of homogeneity: The weighted pairwise correlation statistic. Genetic Epidemiol 11:189-200

[21] Commenges D, Jacqmin-Gadda H (1997) Generalized score test of homogeneity based on correlated random effect models. J R Stat Soc B 59:157–171

[22] Consortium IH (2005) Hplotype map of the human genome. Nature 437:1299-1320

[23] Cottingham RW Jr, Idury RM, Schäffer AA (1993) Faster sequential genetic linkage computations. Am J Hum Genet 53:252–263

[24] Daly MJ, Lander ES (1996) The importance of being independent: sib pair analysis in diabetes. Nat Genet 14(2):131–132

[25] Dausset J, Cann H, Cohen D, Lathrop M, Lalouel J-M, White R (1990) Centre d'Etude du Polymorphisme Humain (CEPH): collaborative genetic mapping of the human genome. Genomics 6:575–577

[26] Davis S, Schroeder M, Goldin LR, Weeks DE (1996) Nonparametric simulation-based statistics for detecting linkage in general pedigrees. Am J Hum Genet 58:867–880
[27] Davis S, Weeks DE (1997) Comparison on nonparametric statistics for detection of linkage in nuclear families: single-marker evaluation. Am J Hum Genet 61:1431-1444
[28] Daw EW, Thompson EA, Wijsman EM (2000) Bias in multipoint linkage analysis arising from map misspecification. Genet Epidemiol 19:366-380
[29] Day NE, Simons MJ (1976) Disease susceptibility genes-their identification by multiple case family studies. Tissue Antigens. 8(2):109–119
[30] Doherty MJ, Glass IA, Bennett CL, Cotter PD, Watson NF, Mitchell AL, Bird TD and Farrell DF (2003) An Xp; Yq translocation causing a novel contiguous gene syndrome in brothers with generalized epilepsy, ichthyosis, and attention deficits, Epilepsia 44:1529–1535
[31] Duffy DL (2006) An integrated genetic map for linkage analysis. Behav Genet 36: 4-6
[32] Dupuis J, van Eerdewegh P (2000) Multipoint linkage analysis of the pseudoautosomal regions, using affected sibling pairs. Am J Hum Genet 67:462-475
[33] Edwards AWF (1972) Likelihood. Cambridge At The University Press, Cambridge
[34] Elston RC, Stewart JA (1971) General model for the genetic analysis of pedigree data. Hum Hered 21:523-542
[35] Elston RC, Spence MA (2006) Advances in statistical human genetics over the last 25 years. *Statist Med* 25:3049–80
[36] Ewens WJ, Spielman RS(2005) What is the significance of a significant TDT? Hum Hered 60:206-210
[37] Fallin D, Schork NJ (2000) Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. Am J H Genet 67:947-959
[38] Faraway JJ (1993) Distribution of the admixture test for the detection of linkage under heterogeneity. Genet Epidemiol 10:75–83
[39] Felsenstein J (1979) Excursions along the interface between disruptive and stabilizing selection. Genetics 93: 773-795
[40] Fimmers R, Seuchter SA, Neugebauer M, Knapp M, Baur MP (1989) Identity by descent analysis using all genotype solutions; in Elston RC, Spence MA, Hodge SE, MacCluer JW (eds): Multipoint Mapping and Linkage Based on Affected Pedigree Members: Genetic Analysis Workshop 6. Alan R. Liss, New York, pp 123-128
[41] Fingerlin TE, Abecasis GR, Boehnke M (2006) Using sex-averaged genetic maps in multipoint linkage analysis when identity-by-descent status is incompletely known. Genet Epidemiol 30: 384–396
[42] Flaquer A, Rappold GA, Wienker TF, Fischer C (2008) The human pseudoautosomal regions - a review for genetic epidemiologists. Eur J Hum Genet [Epub ahead of print]
[43] Freije D, Helms C, Watson MS, Donis-Keller H (1992) Identification of a second pseudoantosomal region near the Xq and Yq telomeres. Science 258:1784-1787
[44] Gabriel-Robez O, Ratomponirina C, Dutrillaux B, Carré-Pigeon F, Rumpler Y (1990) Deletion of the pseudoautosomal region and lack of sex chromosome pairing at pachytene in two infertile men carrying an X;Y translocation. Cytogenet Cell Genet 54:39-42
[45] Gianfrancesco F, Sanges R, Esposito T, Tempesta S, Rao E, Rappold G, Archidiacono N, Graves JAM, Forabosco A, D'Urso M (2001) Differential divergence of three human pseudoautosomal genes and their mouse homologs: implications for sex chromosome evolution. Genome Res 11:2095-2100
[46] Graves JAM, Wakefield MJ, Toder R (1998) The origin and evolution of the pseudoautosomal regions of human sex chromosomes. Hum Mol Genet 7:1991-1996
[47] Graves JAM (2006) Sex chromosome specialization and degeneration in mammals. Cell 10: 901-914
[48] Graves JAM, Koina E, Sankovic N (2006) How the gene content of human sex chromosomes evolved. Curr Opin Genet Dev 16:219-24
[49] Greenberg DA (1993) Linkage analysis of necessary disease loci versus susceptibility disease loci. Am J Hum Genet 52:135-143
[50] Gudbjartsson DF, Jonasson K, Frigge ML, Kong A (2000) Allegro, a new computer program for multi-

point linkage analysis. Nat Genet 25:12-13
[51] Guo SW (1995) Proportion of genome shared identical by descent by relatives: Concept, computation, and applications. Am J Hum Genet 56:1468–1476
[52] Guo SW, Thompson EA (1992) A Monte Carlo method for combined segregation and linkage analysis. Am J Hum Genet 51:1111–1126
[53] Haldane JBS (1919) The combination of linkage values and the calculation of distance between loci of linked factors. J Genet 8:299–309
[54] Haldane JBS (1936) Some natural populations of Lythrum salicaria. J Genet 32:393–397
[55] HapMap-Consortium, T.I. (2005) A haplotype map of the human genome. Nature 437:1299-1320
[56] Hawley ME, Kidd KK (1995) HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. J Hered 86:409-411
[57] Henke A, Fischer C, Rappold GA (1993) Genetic map of the human pseudoautosomal region reveals a high rate of recombination in female meiosis at the Xp telomere. Genomics 18: 478–485
[58] Holmans P (1993) Asymptotic properties of affected-sib-pair linkage analysis. Am J Hum Genet 52(2):362–374
[59] Kim SH, Shin YK, Lee IS, Bae YM, Sohn HW, Suh YH et al. (2000) Viral latent membrane protein 1 (LMP-1)-induced CD99 down-regulation in B cells leads to the generation of cells with Hodgkin's and Reed–Sternberg phenotype. Blood 95: 294–300
[60] Kommoss F, Oliva E, Bittinger F, Kirkpatrick CJ, Amin MB, Bhan AK, Young RH, Scully RE (2000) Inhibin-alpha, CD99, HEA 125, PLAP, and chromogranin immunoreactivity in testicular neoplasms and the androgen insensitivity syndrome. Hum Pathol 31:1055–1061
[61] Kreppel M, Aryee DN, Schaefer KL, Amann G, Kofler R, Poremba C et al. (2006) Suppression of KCM-F1 by constitutive high CD99 expression is involved in the migratory ability of Ewing's sarcoma cells. Oncogene 25: 2795–2800
[62] Idury RM, Elston RC (1997) A faster and more general hidden Markov model algorithm for multipoint likelihood calculations. Hum Hered 47:197-202
[63] Irwin M, Cox N, Kong A (1994) Sequential imputation for multilocus linkage analysis. Proc Natl Acad Sci USA 91:11684–11688
[64] Jensen C, Kong A (1999) Blocking-Gibbs sampling for linkage analysis in large pedigrees with many loops. Am J Hum Genet 65:885-902
[65] Karlin S (1984) Theoretical aspects of genetic map functions in recombination processes. Hum Pop Genet 209-228
[66] Kohn M, Kehrer-Sawatzki H, Vogel W, Graves JAM, Hameister H (2004) Wide genome comparisons reveal the origins of the human X chromosome. Trends Genet 20:598-603
[67] Koller PC, Darlington CD (1934) The genetical and mechanical properties of the sex chromosomes. 1. Rattus norvegicus. J Genet 29:159-173
[68] Kong A (1991) Analysis of pedigree data using methods combining peeling and gibbs sampling. In: Keramidas EM (Hrsg.) Computing science and statistics. Proc 23rd Symp on the Interface 379–384
[69] Kong A (1991) Efficient methods for computing linkage likelihoods of recessive diseases in inbred pedigrees. Genet Epidemiol. 8:81–103
[70] Kong A, Cox N, Frigge M, Irwin M (1993) Sequential imputation and multipoint linkage analysis. Genet Epidemiol 10:483–488
[71] Kong A, Cox NJ (1997) Allele-sharing models: LOD scores and accurate linkage tests. Am J Hum Genet 61:1179–1188
[72] Kong X, Murphy K, Raj T, He C, White PS, Matise TC (2004) A Combined Linkage-Physical Map of the Human Genome. Am J Hum Genetics 75:1143-1148
[73] Kauppi P, Laitinen T, Ollikainen V, Mannila H, Laitinen LA, Kere J (2000) The IL9R region contribution in asthma is supported by genetic association in an isolated population. Eur J Hum Genet 8: 788-792
[74] Kosambi DD (1944) The estimation of map distance from recombination values. Ann Eugen 12:172-175
[75] Kruglyak L, Lander ES (1995) Complete multipoint sib-pair analysis of qualitative and quantitative

traits. Am J Hum Genet 57:439-454
[76] Kruglyak L, Daly MJ, Lander ES (1995) Rapid multipoint linkage analysis of recessive traits in nuclear families, including homozygosity mapping. Am J Hum Genet 56:519–527
[77] Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES (1996) Parametric and Nonparametric Linkage Analysis: A Unified Multipoint Approach. Am J Hum Genet 58:1347-1363
[78] Kruglyak L, Lander ES (1998) Faster multipoint linkage analysis using fourier transforms. J Comp Bio 5:1-7.
[79] Lander ES, Green P (1987) Construction of multilocus genetic linkage maps in humans. *Proc Natl Acad Sci USA*, 84: 2363–2367
[80] Lander E, Green P, Abrahamson J, Barlow A, Daley M, Lincoln S, Newburg L (1987) MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics 1:174-181
[81] Lander E, Kruglyak L. (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet 11:241-247
[82] Lange K, Elston RC (1975) Extensions to pedigree analysis I. Likelihood calculations for simple and complex pedigrees. Hum Hered. 25(2):95–105
[83] Lange K, Weeks D, Boehnke M (1988) Programs for pedigree analysis: MENDEL, FISHER, and dGENE. Genet Epidemiol 5:471–472
[84] Lange K, Sobel E (1991) A random walk method for computing genetic location scores. Am J Hum Genet. 49:1320–1334
[85] Lathrop GM, Lalouel JM, Julier C, Ott J (1984) Strategies for multilocus linkage analysis in humans. Proc Natl Acad Sci USA. 81:3443–3446
[86] Li H, Gyllensten UB, Cui X, Saiki RK, Erlich HA, and Arnheim N (1988) Amplification and analysis of DNA sequences in single human sperm and diploid cells. Nature 335:414-417
[87] Li L, Hamer DH (1995) Recombination and allelic association in the Xq/Yq homology region. Hum Mol Genet 4:2013-2016
[88] Lien S, Szyda J, Schechinger B, Rappold G, Arnheim N (2000) Evidence for heterogeneity in recombination in the human pseudoautosomal region: high resolution analysis by sperm typing and radiation-hybrid mapping. Am J Hum Genet 66:557-566
[89] Long JC, Williams RC, Urbanek M (1995) An E-M algorithm and testing strategy for multiple-locus haplotypes. Am J H Genet 56:799-810
[90] Matise TC, Chen F, Chen W, De La Vega FM, Hansen M, He C, Hyland FC, Kennedy GC, Kong X, Murray SS, Ziegle JS, Stewart WC, Buyske S (2007) A second-generation combined linkage physical map of the human genome. Genome Research **A**:1783-1786
[91] Meunier F, Philippi A, Martinez M, Demenais F (1997) Affected sib-pair tests for linkage: type I errors with dependent sib-pairs. Genet Epidemiol 14(6):1107–1111
[92] McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P (2004) The fine-scale structure of recombination rate variation in the human genome. Science 304:581-584
[93] Mohandas TK, Speed RM, Passage MB, Yen PH, Chandley AC, Shapiro LJ (1992) Role of the pseudoautosomal region in sex-chromosome pairing during male meiosis: meiotic studies in a man with deletion of distal Xp. Am J Hum Genet 51:526-533
[94] Morgan TH (1928) The Theory of Genes. Yale University Press, New Haven, CT
[95] Moses MJ, Counce SJ, Paulson DF (1975) Synaptonemal complex complement of man in spreads of spermatocytes, with details of the sex chromosome pair. Science 187:363–365
[96] Morton BR (1995) Neighboring base composition and transversion/transition bias in a comparison of rice and maize chloroplast noncoding regions. Proc Natl Acad Sci USA 92: 9717–9721
[97] Muller HJ (1914) A gene for the fourth chromosome of Drosophila. J Exp Zool 17:325-335
[98] Muller DJ, Schulze TG, Jahnes E, Cichon S, Krauss H, Kesper K, Held T, Maier W, Propping P, Nothen MM, Rietschel M (2002) Association between a polymorphism in the pseudoautosomal X-linked gene *SYBL1* and bipolar affective disorder. Amer J Med Genet Part B Neuropsychiatr Genet 114:74-78
[99] O'Connell JR, Weeks DE (1995) The VITESSE algorithm for rapid, exact multilocus linkage analysis

via genotype set-recording and fuzzy inheritance. Nat Genet 11:402-408
[100] O'Connell JR, Weeks DE (1998) PedCheck: a program for identification of genotype incompatibilities in linkage analysis. Am J Hum Genet 63:259–266
[101] Ott J (1974) Estimation of the recombination fraction in human pedigrees: efficient computation of the likelihood for human linkage studies. Am J Hum Genet 26(5):588-597
[102] Ott J (1976) A computer program for linkage analysis of general human pedigrees. Am J Hum Genet 28:528–529
[103] Ott J (1986) Y-linkage and pseudoautosomal linkage. Am J Hum Genet 38:891-897
[104] Ott J (1992) Strategies for characterizing highly polymorphic markers in human gene mapping. Am J Hum Genet 51:283-290
[105] Ott J (1999) Analysis of the human genetic linkage. Third edition. Baltimore and London: the Johns Hopkins University press
[106] Page DC, Bicker K, Brown LG, Hinton S, Leppert M, Lalouel JM, Lathrop M, Nystrrm-Lahti M, Chapelle A de la, White R (1987) Linkage, physical mapping and DNA sequence analysis of pseudoautosomal loci on the human X and Y chromosomes. Genomics 1:243-256
[107] Penrose LS (1935) The detection of autosomal linkage in data which consists of pairs of brothers and sisters of unspecified parentage Ann of Eugen 6:133-138
[108] Pearson PL, Bobrow M (1970) Definitive evidence for the short arm of the Y chromosome associating with the X chromosome during meiosis in the human male. Nature 226:959–961
[109] Rao DC, Morton NE, Lindsten J, Hulten M, Yee S (1977). A mapping function for man. Hum Hered 27: 99-104
[110] Rappold GA, Klink A, Weiss B, Fischer C (1994) Double crossover in the human Xp/Yp pseudoautosomal region and its bearing on interference. Hum Mol Genet 3:1337-1340
[111] Risch N, Giuffra L (1992) Model misspecification and multipoint linkage analysis. Hum Hered 42: 77–92
[112] Risch N (1990b) Linkage strategies for genetically complex traits II. The power of affected relative pairs. Am J Hum Genet 46:229–241
[113] Risch N (1990c) Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs. Am J Hum Genet 46:242-253
[114] Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516-1517
[115] Ross MT, Grafham DV, Coffey AJ, Scherer S, McLay K, Muzny D, Platzer M, Howell GR, Burrows C, Bird CP et al. (2005) The DNA sequence of the human X chromosome. Nature 434:325-337.
[116] Ross MT, Bentley DR, Tyler-Smith C (2006) The sequences of the human sex chromosomes. Curr Opin Genet Dev 16:213-8
[117] Rouyer F, Simmler MC, Johnsson C, Vergnaud G, Cooke HJ, Weissenbach J (1986a) A gradient of sex linkage in the pseudoautosomal region of the human sex chromosomes. Nature 319:291-295
[118] Rouyer F, Simmler M-C, Vergnaud G, Johnsson C, Levilliers J, Petit C, Weissenbach J (1986b) The pseudoautosomal region of the human sex chromosomes. Cold Spring Harb Symp Quant Biol 51:221-228
[119] Rouyer F, de laChapelle A, Andersson M, Weissenbach J (1990) An interspersed repeated sequence specific for human subtelomeric regions. EMBO J 9: 505–514
[120] Saito T, Parsia S, Papolos DF, Lachman HM (2000) Analysis of the pseudoautosomal X-linked gene *SYBL1* in bipolar affective disorder: Description of a new candidate allele for psychiatric disorders. Am J Med Genet Part B Neuropsychiatr Genet 96: 317-323
[121] Schaid DJ, Sommer SS (1993) Genotype relative risks: methods for design and analysis of candidate-gene association studies. Am J H Genet 53:1114-1126
[122] Schmitt K, Vollrath D, Foote S, Fisher EM, Page DC, Arnheim N (1993) Four PCR-based polymorphisms in the pseudoautosomal region of the human X and Y chromosomes. Hum Mol Genet 2:1978
[123] Shannon CE (1948) A Mathematical Theory of Communication. Bell Syst. Tech J 27:379-423, 623-656
[124] Sheehan NA (1989) Image-processing procedures applied to the estimation of genotypes on pedigrees.

Technical Report 176, Department of Statistics, University of Washington, Seattle
[125] Sobel E, Lange K (1993) Metropolis sampling in pedigree analysis. Stat Methods Med Res 2:263–282
[126] Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J H Genet 52:506-516
[127] Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. Am J H Genet 68:978-989
[128] Strauch K, Fimmers R, Kurz T, Deichmann KA, Wienker TF, Baur MP (2000) Parametric and nonparametric multipoint linkage analysis with imprinting and two-locus-trait models: application to mite sensitization. Am J H Genet 66:1945-1957
[129] Strauch K (2002) Kopplungsanalyse bei genetisch komplexen Erkrankung mit genomischem Imprinting und Zwei-Genort-Krankheitsmodellen. Medizinische Informatik, Biometrie und Epidemiologie 87, Urban & Vogel, ISBN 3-860094-174-7
[130] Strauch K, Baur MP, Wienker TF (2004) A recoding scheme for X-linked and pseudoautosomal loci to be used with computer programs for autosomal lod-score analysis. Hum Hered 58:55-58
[131] Stricker C, Fernando RL, Elston RC (1995) An algorithm to approximate the likelihood for pedigree data with loops by cutting. Theor Appl Genet 91:1054-1063.
[132] Sturt E (1976) A mapping function for human chromosomes. Ann Hum Genet 40:147-163
[133] Suarez BK, Rice J, Reich T (1978) The generalized sib pair IBD distribution: its use in the detection of linkage. Ann Hum Genet 42(1):87–94
[134] Thompson EA, Wijsman EM (1990) The Gibbs sampler on extended pedigrees: Monte Carlo methods for the genetic analysis of complex traits. Technical Report no. 193, Department of Statistics, University of Washington.
[135] Thompson EA (1994) Monte Carlo likelihood in genetic mapping. Stat Sci 9:355-366
[136] Thompson EA, Guo SW (1991) Estimation of likelihood ratios for complex genetic models. IMA J Math Appl Med Biol 8:149–169
[137] Weeks DE, Lange K (1988) The affected-pedigree-member method of linkage analysis. Am J Hum Genet 42(2):315–326
[138] Weeks DE, Lange K (1992) A multilocus extension of the affected-pedigree-member method of linkage analysis. Am J Hum Genet 1992 50(4):859–868
[139] Weissenbach J, Gyapay G, Dib C, Vignal A, Morissette J, Millasseau P, Vaysseix G, Lathrop M (1992) A second-generation linkage map of the human genome. Nature 359: 794-801
[140] Whittemore AS, Halpern J (1994a) Probability of gene identity by descent: computation and applications. Biometrics 50:109–117
[141] Whittemore A, Halpern J(1994b) A class of tests for linkage using affected pedigree members. Biometrics 50:118-127
[142] Yi H, Donohue SJ, Klein DC, McBride OW (1993) Localization of the hydroxyindole-O-methyltransferase gene to the pseudoautosomal region: implications for mapping of psychiatric disorders. Hum Mol Genet 2:127–131
[143] Yi S and Li WH (2005) Molecular evolution of recombination hotspots and highly recombining pseudoautosomal regions in hominoids. Mol Biol Evol 22:1223-1230
[144] Ziegler A, König IR (2006) A statistical approach to genetic epidemiology: concepts and applications. Wiley-VCH Verlag GmbH & Co. kGaA