Estimating relationships and relatedness from dense genome-wide data

Thesis submitted for the degree of

Doctor of Philosophy

at the University of Leicester

by

Meng Sun BSc MSc

Department of Health Sciences

University of Leicester

2015

Abstract

Estimating relationships and relatedness from dense genome-wide data

Meng Sun

Relationship and relatedness estimation from genetic markers is relevant to many areas, including genealogical research, genetic counselling, forensics, linkage analysis and association analyses in genetic epidemiology. Traditionally unlinked genetic markers (microsatellites) are used. But the problems which can be solved by such markers are limited. Linked genetic markers are not only available in much larger numbers, but also provide extra information which is not available from unlinked markers. It is desirable to exploit the increasing availability of dense genome-wide single nucleotide polymorphisms (SNP) data for estimating relationships and relatedness.

While Method of Moments (MoM) methods and other non-pedigree approaches only give a degree of pairwise relatedness, a pedigree likelihood approach can distinguish exact relationships. The pedigree likelihood approach also has advantages in that extra individuals can be considered jointly and extra data such as Y-chromosomal and mitochondrial SNPs can be incorporated with autosomal SNPs easily. In this thesis I firstly confirm that the increase in information obtained from large sets of linked markers substantially increases the number of problems that can be solved with pedigree approach. Furthermore, when two distant relatives do share genome segments through identity by descent (IBD), we usually have greater power to distinguish more distant relatives from unrelated pairs than was previously believed. Data on extra individuals always improve discriminatory power, but the position of the extra individuals in the pedigree dictates the extent of this increase of power. Linkage Disequilibrium (LD) is an issue for pedigree likelihood approach and it needs to be dealt with.

MoM methods are easy to use and are generally robust to the effect of LD, but they are only accurate for relatives up to second cousins. I propose using pedigree likelihood approach to estimate pairwise relatedness and find we can greatly improve the accuracy in detecting distant relatives.

Acknowledgement

Firstly I would like to express my sincere gratitude to my supervisors Prof. Nuala Sheehan and Prof. Mark Jobling for giving me the opportunity of doing this PhD, their supervision and support, especially to Prof. Sheehan for her excellent and patient guidance.

Thanks also go to my parents who have always supported my study since I was a child and encouraged me to pursue higher degrees.

I need to thank my wife Jie for supporting me doing this PhD and two daughters, Jasmine and Hannah, who behave well and have not taken too much time of mine although the time being with them is precious.

I am also grateful to everyone who has helped me in different stages of my education.

Contents

Table of Figures

Table of tables

List of abbreviations

AIM	Ancestry informative marker
Bp	Base pair
DNA	Deoxyribonucleic acid
GWAS	Genome-wide association study
HMM	Hidden Markov Chain
HS	Half sibling
HVSI	Hypervariable Segment
HWE	Hardy-Weinberg Equilibrium
IBD	Identity by descent
IBS	Identical by state
Kb	Kilobase
LD	Linkage disequilibrium
LOD	Logarithm of odds
MAF	Minor allele frequency
Mb	Megabase
MICROS	Microisolates in South Tyrol Study
mRNA	Messanger RNA
MSY	Male-specific region of the Y chromosome
mtDNA	Mitochondrial DNA
OMIM	Online Mendelian inheritance in man
PC	Parent-child
PCA	Principal component analysis #correct remove s
RNA	Ribonucleic acid

S	Sibling
SNP	Single nucleotide polymorphisms
STR	Short tandem repeat
TDT	Transmission disequilibrium test
TSI	Toscani in Italia
CEU	Utah residents with Northern and Western European ancestry
WTCCC	Wellcome Trust Case Control Consortium

1 Introduction

1.1 Background and aim

Relationship and relatedness are two words which are often used to describe relatives and they have different meanings. The term ‘relationship’ tells us specifically how relatives are related to each other and typically specifies a pedigree connecting them. In this thesis, the terms ‘relationship’ and ‘pedigree’ are hence used interchangeably. The term ‘relatedness’, on the other hand, describes how closely the relatives are related to each other in terms of genetic sharing and is described by a single parameter such as a coefficient of kinship. A single degree of relatedness can correspond to many different relationships. For example, the three relationships: grandparent-grandchildren, half-sibling and uncle-nephew all have the same degree of relatedness and a common kinship coefficient.

Relationships and relatedness can be estimated from genetic data. This is relevant to many areas of application including genealogical research, forensic identification, linkage analysis and association analysis for various reasons. In genealogical research, lines of descent of a family could be traced for those who are interested in their family tree and for genetic counselling purposes. In forensic research, the identification of a victim might be possible after a disaster if the DNA of a relative is available (Gill et al., 1994, Olaisen et al., 1997). Relationships often need to be confirmed in inheritance claims and in immigration cases (Hansen and Morling, 1993). Pedigree data are used in traditional linkage analysis to locate putative genes for diseases. It is important to know that the relationships in the pedigree are correct. The result of a linkage analysis will be distorted if two founders in the pedigree are in fact related while they are assumed to be unrelated (Leutenegger et al., 2002). Estimating their true relationships could yield a larger and more informative pedigree for linkage analysis. Recently it has been claimed that pedigree-based kinship coefficients can be replaced with marker-based estimates in linkage analysis and pedigrees are not necessary any more (Day-Williams et al., 2011). Day-Williams et al. (2011) show that they have

obtained a LOD (logarithm of odds) score peak in a quantitative trait locus linkage analysis based on estimated kinship coefficients near the locus previously reported by a pedigree-based study. However, it is noted that their LOD score is much lower than the one reported in the pedigree-based study. So it is not evidence that pedigree is not needed any more in linkage analysis. Instead it just shows that it is possible to carry out a linkage analysis even if there is no pre-specified pedigree. Nowadays there are fewer pedigree-based linkage analysis studies. But this area is still controversial with many people arguing for the usefulness of large pedigrees in gene mapping (Wijsman, 2012, Pattaro and Saint-Pierre, 2013).

Association studies on unrelated individuals are commonly undertaken and they are good at finding common putative alleles contributing to disease. But disease alleles which are rare and have modest effects will be difficult to detect in population studies of unrelated individuals due to lack of power. One way to increase the power of finding those alleles is to use data from related individuals as they share longer haplotypes around the putative alleles and so are more informative than unrelated individuals (Boehnke and Cox, 1997, Goring and Ott, 1997, McPeek and Sun, 2000, Stankovich et al., 2005). This is because rare alleles usually would have arisen relatively more recently than common ones and so there is less recombination and mutation to break the haplotypes, while old alleles tend to become common or else disappear from the gene pool. In standard association studies, relationship estimation is also relevant. It can be used to detect unknown relatives in the sample as their existence could bias the results when all individuals are assumed to be unrelated (Choi et al., 2009, Thornton and McPeek, 2010). Relationship estimation is also widely used in research on animals and plants as well. The focus on this thesis will be on human data, but the methods are general and relevant to a broader range of applications.

My primary interest is to improve the accuracy of relationship estimation (using pedigrees) and improve our ability to detect relatives that cannot be easily detected using existing approaches, such as more distant ones, using genome-wide genetic data. I would also like to know whether we really need pedigrees for

finding good sets of relatives in population data. Moreover, there are many methods estimating relationships or overall relatedness, but it is not always clear what they are really estimating, what assumptions they are making and when they can be best used. It is of interest to investigate this as well.

When the exact relationship (pedigree) is of interest, one approach based on the pedigree likelihood is attractive in that it can easily deal with multiple individuals at the same time and prior information can, in principle, be incorporated (details in Chapter 4). Recent work indicates that we can exploit the ever-increasing availability of genome-wide single nucleotide polymorphism (SNP) data for estimating relationships between individuals using this likelihood approach (Skare et al., 2009). This approach will play an important role in this thesis.

1.2 Layout of the subsequent chapters

Chapter 2 introduces some background knowledge that is essential to the thesis. This includes human genetics and some terminologies and concepts such as identity by descent (IBD), linkage disequilibrium (LD) and pedigrees. Then different methods of calculating pedigree likelihood are described. The most important one for this thesis is the Lander-Green algorithm which is used to do all the pedigree likelihood calculation in this thesis.

In Chapter 3, the various software packages and datasets that have been used in this thesis are described. They are put together as one chapter for easy reference.

In Chapter 4, the pedigree-based likelihood method is introduced. It is an approach to distinguish one relationship from several alternative relationships rather than to estimate the unknown true relationship. Standard forensic problems, such as paternity testing, would typically look at two alternatives for example: the hypothesized relationship versus ‘unrelated’. This method is examined and ways to improve it are investigated. One question that I would like to answer is what will happen if the density of SNPs keeps increasing. One might presume that we would always get a better result when more markers are used. I also would like to know whether an extra genotyped individual will help estimate the relationship between two specific individuals and, if so, how to select this third individual.

Traditionally unlinked data (microsatellites) are used and there is hence no problem with LD. But when dense genome-wide genetic data are used as in this thesis, LD within the SNPs is unavoidable. In Chapter 5, the effect of LD on relationship estimation for the pedigree likelihood method is investigated and a possible way for dealing with LD is studied.

The most popular approach to deal with LD is to thin the data. But to what extent should we thin? There is no consensus in the literature. I would like to investigate how the thinning approach works. Since thinning dense SNP data will unavoidably lose some information, it is of interest to model LD rather than simply remove SNPs. It is investigated how LD modelling works for relationship estimation. In this chapter, the pedigree-based likelihood method is proposed to be used as a method to estimate the degree of an unknown relationship rather than just to distinguish between specific alternatives.

In Chapter 6, methods for estimating pairwise relatedness without using a pedigree are considered. They are used for the cases where the exact relationship is not of interest and just general relatedness is required. These methods include maximum likelihood methods, method of moments (MoM) and IBD segment detection methods. The aim is to understand how these methods work, their strength and weakness, and whether pedigrees are really necessary for detecting relatives when relatedness rather than exact relationship is all that is needed. Maximum likelihood methods are just briefly discussed. MoM methods and IBD segment detection methods are examined in some detail. These methods are compared with the pedigree likelihood approach that is proposed in Chapter 5. Work in Chapter 4, 5 and 6 has been written as a paper titled “On the Use of Dense SNP Marker Data for the Identification of Distant Relative Pairs” which has been accepted by the “Theoretical Population Biology”.

Chapter 7 is devoted to the usefulness of Y chromosome SNP data and mtDNA SNP data for relationship estimation. Such data are routinely collected in many SNP chips but, to my knowledge, they have not been used to supplement autosomal DNA for relationship or relatedness estimation before. The potential of

these additional data is investigated including how they can be incorporated with autosomal SNPs and in what situations they are most useful.

When data on more than two individuals are available, sometimes it is of interest to construct a pedigree for them. Chapter 8 looks at how to cluster a large number of individuals into groups of relatives and how the joint relationship of these relatives can be estimated all together and hence reconstruct a pedigree. One of the advantages of the pedigree likelihood method over pairwise estimation is that it can naturally consider the relationships between many individuals altogether.

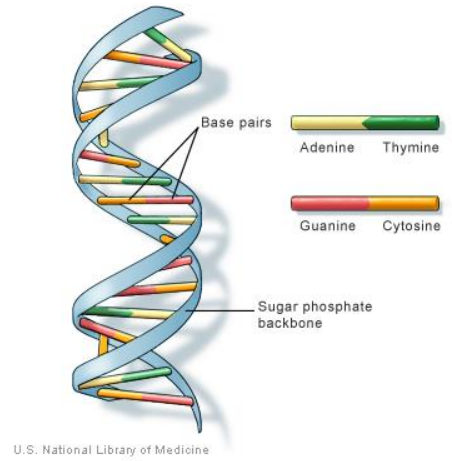
In Chapter 9, a summary is given for the findings in this thesis. I have also outlined some extensions for future work.

2 Background biology, terminology and concepts

2.1 Human genetics

Human genetic information is stored in DNA (deoxyribonucleic acid) molecules as a linear sequence of nucleotide bases and can be transmitted from parents to children. DNA is a two-stranded helix. Each strand consists of four different nucleotide bases: adenine (A), cytosine (C), guanine (G) and thymine (T). The bases on the two strands pair with each other following a rule of A pairing with T and G pairing with C (Figure 2.1). DNA is replicated by separating the strands and using each strand as a template to form two identical double-strands of DNA. The two strands are complementary, so we only need to know the sequence of one strand. A strand of DNA can also be used to produce RNA (ribonucleic acid) in a process called transcription, in which only one strand of DNA is used as a template to construct the complementary strand. RNA, unlike DNA, is usually single-stranded. There are two classes of RNA. One class is mRNA which carries the genetic information from the DNA to the synthesis of protein in a process called translation. Another class is noncoding RNA which is involved in the expression of other genes. The transmission of genetic information is through the activity of DNA and RNA regulating the formation of polypeptides, the basic component of proteins. A gene is a DNA segment which includes a sequence of bases which can be transcribed into mRNA and instructs the formation of protein. Protein is an important component of the human body and different types of protein can determine the different functions of cells. So changes in the DNA sequence can have an effect on our health by changing the functionality of proteins.

Figure 2.1 Picture of a standard double helix from U.S. national library of medicine.



The entire genetic material carried in humans is called the genome. The haploid human genome has a size of 3200 megabases (Mb). The genome is distributed among 23 pairs of chromosome inside the cell nucleus. 22 pairs are the autosomes and one pair consists of the sex chromosomes with XX in females and XY in males. Humans also carry mitochondrial DNA, a 16.5 kilobases (kb) molecule present in many copies in almost all cells, and housed outside the nucleus. Except for the sex chromosomes, nuclear chromosomes in each pair have the same basic structure and gene sequence with one originating from the father and one from the mother. The two chromosomes in a pair are called homologues. A locus is a particular position on a chromosome, and an allele is one of the sequence variants that could exist in a specific locus. For example, at the locus of the ABO blood group, there are three possible alleles: A, B and O.

Only 3% of our genome comprises coding sequences (Lander et al., 2001) and the number of genes that code for proteins is only about 24,000. 99.9% of genome is identical between individuals. But the remaining DNA sequence in chromosomes can vary in many ways. There are two widely studied types of DNA sequence variant - microsatellites and SNPs. A microsatellite, also referred to as a short tandem repeat (STR), is a repeated short sequence of about 2-6 base pairs (bp), e.g. ‘CACACA….’. Microsatellites are very variable in the numbers of repeats and hence informative since most people have different alleles at a given locus. A SNP is a variant of DNA where a single base is substituted for another, and

includes single-base insertion and deletion. Such variants usually have only two alleles.

Not all changes in DNA sequence change the structures of proteins. This is because 97% of the genome has no effect on proteins and also because there is redundancy in the genetic code, which means that several different codons (nucleotide triplets that can specify an amino acid or indicate the termination of translation) may produce the same amino acid. Those SNPs which can actually lead to the change of amino acids are called non-synonymous SNPs. We also need to note that changes outside of the protein-coding regions can affect the phenotype as well by altering the regulation of genes.

A genotype is the combination of the alleles at one locus on each of the two chromosomes. In a genotype, the order of the alleles is usually not important, e.g. Aa = aA by default. It is usually made explicit if they are ordered. Hardy-Weinberg Equilibrium (HWE) specifies a state that genotype frequencies are consistent with the two alleles being sampled independently from a population of alleles. For example, if the frequencies of allele A and a are p_A and p_a, the frequencies of the genotypes AA, Aa and aa under HWE are p_A^2, $2p_Ap_a$ and p_a^2 respectively. A haplotype is the sequence of the alleles at different loci along a chromosome. Obviously, a haplotype depends on the order of the alleles along the chromosome. Haplotypes also tell us the phase of the genotypes, in other words, which chromosome an allele belongs to. A phenotype, also called a trait, refers to an observable property of an individual. A phenotype could be a disease status e.g. presence or absence of a disease; it could also be a continuous characteristic, e.g. blood pressure, height, etc.

A person is homozygous at a locus if the two alleles at that locus on the pair of homologous chromosomes are the same and heterozygous if they are different. If a trait is expressed whenever one abnormal allele is present, then that abnormal allele is called dominant for this trait and the trait is called a dominant trait. If a trait only manifests itself when two copies of the abnormal alleles are present, the abnormal allele is said to be recessive for that trait and that trait is recessive. There is also an intermediate case, while both phenotypes are expressed and no

phenotype is dominant. The expressed phenotype is somewhere between two extreme states. This is called co-dominance.

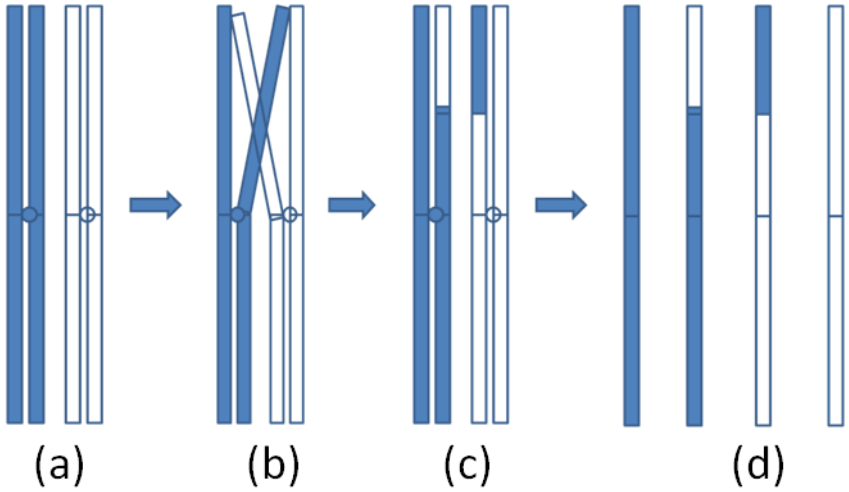
Cells can be categorized by the number of the chromosome sets contained in them. For human cells, each chromosome set consists of 23 chromosomes. But different cells in our body may contain different copy numbers of the chromosome set (ploidy). Most human cells contain two copies of the chromosome set and are called diploid. But sperm and egg cells only carry one copy of the chromosome set and are called haploid. Some cells, like red blood cells contain no chromosomes at all. In some species, the cell can be triploid or tetraploid.

There are two kinds of replication process for a cell, mitosis and meiosis. All the cells in our body are derived from one single zygote (the fertilized egg cell) which is formed by the sperm and egg cells. Both mitosis and meiosis involve the processes of replication of DNA and cell division. But mitosis produces identical chromosomes and the resulting cells remain diploid, while meiosis produces haploid cells which are all genetically different from each other. Meiosis only happens in the germ-line cells and produces gametes (sperm and eggs) while mitosis is the normal process for human cell division and it generates the cells required for an embryo to grow.

Meiosis starts from a diploid cell and goes through one replication process and two division processes leading to the formation of gametes. Each gamete randomly receives one member of the pair of homologous chromosomes which have experienced the crossover process (see below). So each gamete only has one copy of the chromosome set and contains 22 autosomes and one sex chromosome. In sperm cells, the sex chromosome could be either X or Y, but in egg cells, the sex chromosome is always X. When a sperm fertilizes an egg, a diploid zygote will be formed with chromosome constitution of either ‘46,XX’ or ‘46,XY’ (‘46,XX’ means there are 46 chromosomes in total and the sex chromosomes are XX).

Because each diploid cell contains two sets of chromosomes, it thus has two copies of each chromosome, one is paternally inherited and the other is maternal. During meiosis, when one cell divides into two daughter cells, the process that decides which homologue goes to which daughter cell is random. So there are 2^{23} possible combinations of parental chromosome in the gametes from just a single meiosis.

Figure 2.2 Crossover in meiosis. Two homologous chromosomes are replicated (a) -> crossover happens on two chromatids between two chromosomes (b)-> crossover is completed (c) -> four gametes are produced, each with one chromosome (d).



In the process of the gamete formation, the homologous chromosomes are not transmitted to the gametes completely. First, each single chromosome will be replicated and the replicated chromosome will be connected with the original chromosome, both of them are called chromatids (as shown in (a) of Figure 2.2). Then the cell will experience two division processes. A chromatid is one of the two identical copies of DNA making up a duplicated chromosome, which are joined at their centromeres, for the process of cell division. They are called sister chromatids. A process called crossover occurs between paternal and maternal chromatids leading to further genetic variation among the resulting gametes as shown in (b) of Figure 2.2. Crossover involves the breakage and the rejoining of chromosomes in paternal and maternal chromatids (as shown in (c) of Figure 2.2). The point at which a crossover happens is called a chiasma. Crossover can separate the alleles that originally appeared in one chromosome and could also

make two alleles with different sources appear on one chromosome. This process is called recombination. The random segregation of chromosomes in meiosis and the crossover between homologues ensures that one individual can produce a huge amount of genetic diversity in their gametes. Crossover is often used to assess the distance between two loci with 1cM (centiMorgan) denoting an average number of crossover of 0.01 in one generation. For two loci which are physically close on the same chromosome, the expected number of crossover between them is usually close to 0. Generally the greater the distance between two loci, the higher the probability of a crossover happening between them. The exception is in recombination hotspots (regions in a genome where crossover is more frequent).

It is the odd number of crossovers between two loci which causes the recombination of the genes on two loci because an even number of crossovers between two loci would take each allele back to their original chromosome. In each crossover process, there are always two sister chromatids remaining non-recombinant while another two recombine. So the maximum recombination rate, which can be assessed by the proportion of gametes that are recombinant, is 50% for a pair of well separated loci.

Mutations are variations in the usual DNA sequence of an organism which are the result of chemical or physical agents or DNA replication errors. Only those mutations in the coding regions of genes can affect the amino acid sequence of a protein. Regulatory mutations in the non-coding regions could affect the amount, location or timing of protein production. An organism could have original phenotypes or mutant phenotypes depending on whether a mutation occurs.

The Law of Segregation (first law of Mendel) states that every individual possesses a pair of alleles for any particular trait and that each parent passes a randomly selected allele to its offspring. The Law of Independent Assortment (second law of Mendel) states that separate genes for separate traits are passed independently from parents to offspring. That is, the selection of a particular gene in the gene pair for one trait to be passed to the offspring has nothing to do with the selection of the gene for any other trait. But we know now this is true only for

genes that are not *linked* i.e. are not located closely enough on a chromosome to be inherited together more often than not.

Mendelian characters are those characters whose expression is determined by a particular genotype at a single locus. There are six general inheritance patterns for Mendelian characters:

(1) Autosomal dominant. Under this pattern, an affected parent transmits the phenotype to both male and female children and affected males and females appear in each generation of the pedigree. One example of this type of disorder is Huntington's disease (OMIM number #143100).

(2) Autosomal recessive. Under this pattern, both sexes can be affected and two unaffected parents could have an affected child. One example of this type of disorder is cystic fibrosis (OMIM number #219700).

(3) X-linked dominant. Under this pattern, affected males pass the disorder to all daughters but to none of their sons, and affected heterozygous females married to unaffected males pass the disorder to half their sons and daughters. One example of this type of disorder is X-linked hypophosphataemic rickets (OMIM number #307800).

(4) X-linked recessive. Under this pattern, many more males than females show the disorder because males only have one X chromosome and the phenotype will expressed as long as that X chromosome contain the causal allele; all daughters of an affected male are ‘carriers’; none of the sons of an affected male show the disorder or are ‘carriers’. One example of this type of disorder is Haemophilia A (OMIM number #306700).

(5) Y-linked. A disorder with this pattern only affects males and all male offspring of an affected male are affected. One example of this type of disorder is Y-linked deafness (OMIM number #400043).

(6) Mitochondrial. A disorder with this pattern is passed by a mother to all of her offspring, and cannot be passed on by males, since females only pass

on mitochondrial DNA in the egg. One example of this type of disorder is Leber hereditary optic neuropathy (OMIM number #535000).

But most human characters are more complex than Mendelian characters. These are governed by genes on several loci and are called multifactorial.

The knowledge of human genetics in this thesis is mainly taken from the books of Strachan (2011) and Palmer (2011).

2.2 Identity by descent and identity by state

A set of genes at a locus are said to be identical by descent (IBD) if they have been inherited from a common ancestor. In contrast, two alleles are said to be identical by state (IBS) when they are of same type regardless of their origins. The sharing of genes IBD can be used to measure the relationship between individuals. Obviously, related individuals will have more IBD sharing than unrelated individuals and relatives are similar phenotypically because they share more genes IBD. These probabilities of IBD are important in identifying the relationship between people. The difference between IBD and IBS is thus important in linkage analysis where putative genes for disease are tracked through pedigrees but less important in association analysis where the focus is on the effects of the genes carried and not their origin.

Inbreeding refers to reproduction through the mating of two genetically related parents. It will result in increased homozygosity, which in turn increases the chances of offspring being affected by recessive traits. The inbreeding coefficient is the probability that two alleles at a locus in an individual are IBD. The kinship coefficient is the probability that the two alleles, randomly drawn from two individuals at a particular locus, are IBD. The inbreeding coefficient of an individual is hence the kinship coefficient between the father and the mother of the individual. Kinship coefficients and inbreeding coefficients are just probabilities, so they can only take values between 0 and 1. Denoting the kinship coefficient by θ we have

$$\theta = 1/2\times P(IBD=2)+1/4\times P(IBD=1)+0\times P(IBD=0) = 1/2k_2+1/4k_1+0k_0,$$

where P(IBD=2) = k_2 represents the probability that two individuals share both genes IBD, P(IBD=1) = k_1 represents the probability that two individuals share one IBD gene and P(IBD=0) =k_0 represents the probability that two individuals share zero IBD gene. The coefficient of ½ for k_2 arises from the fact that if two individuals share two IBD genes, then if we randomly take one gene from each individual, the chance that they are IBD is 50%. A similar argument applies for the coefficients of $k_1$and k_0.

There are many possible paths of genetic descent by which individuals could share IBD. For example, individuals A and B could share one gene IBD at a given locus which is from their common mother, or from their common grandmother. A genealogical relationship determines a probability for each path. One method of calculating the theoretical kinship or inbreeding coefficient, based on pedigrees, is the path-counting method which sums the kinship or inbreeding coefficients over all paths, originally introduced by Wright (1922).

2.3 Linkage disequilibrium

Linkage disequilibrium (LD) is the non-random association between alleles at two or more loci which results in the higher or lower frequency of some haplotypes in a population than what will be expected from a random formation of the haplotypes combining alleles on different loci. In contrary, linkage equilibrium (LE) describes the situation in which the haplotype frequencies in a population are same as the value that would be obtained if the genes at different loci are combined at random. LD is used to describe the association of alleles between loci within the population. There are many ways to measure LD. A simple one was proposed by Falconer and Mackay (1996). For diallelic loci,

$$D = P_{A1B1} - P_{A1}P_{B1} = P_{A2B2} - P_{A2}P_{B2} = P_{A1B1}P_{A2B2} - P_{A1B2}P_{A2B1},$$

where A and B are the two genes in question with alleles A1 and A2 for gene A and alleles B1 and B2 for gene B. P_{A1B1} is the frequency of the haplotype $A1B1$ in the population. P_{A1}, P_{B1} are the population allele frequencies, etc. So D measures the departure from linkage equilibrium in which case D = 0. This

measure depends on the allele frequencies and has a maximum and minimum value based on those frequencies. Another measure of LD is to scale D to D_{max}, the maximum possible value of D based on allele frequencies, $D' = D/D_{max}$. The range of values of D' is -1 to 1. D takes the maximum value whenever one haplotype frequency is 0. Another measure for diallelic markers is the squared correlation between pairs of loci (Hill and Robertson, 1968),

$$r^2 = D^2/[P_{A1}P_{A2}P_{B1}P_{B2}].$$

The value of r^2 ranges from 0 to 1. When r^2 is 1, the two markers provide exactly the same information. LD decays over time with recombination between the two loci and with mutation.

2.4 Pedigrees

A pedigree, which is also called a genealogy, can be formally defined as "a group of individuals together with a full specification of all the relationships among them" (Thompson, 1986). Pedigrees can be shown graphically. One example of a pedigree is shown in Figure 2.3. By convention a square is used to denote a male and a circle to denote a female. Horizontal lines below couples are used to represent marriages. Parents and their children are linked through vertical lines. The individuals who are in the same level in the pedigree are in the same generation, often denoted by Roman numerals. Each person in each generation is labelled by Arabic numbers. Individuals with data e.g. those who are affected by a disease, are shown by shading. By convention, it is often assumed that every individual either has no parent specified in a pedigree (called a founder), or both parents specified (called a nonfounder). Closed circuits formed by lines and individuals are called loops in a pedigree and are often the result of inbreeding, multiple mating or marriage exchanges. A pedigree with loops is called looped pedigree. A looped pedigree is shown in Figure 2.4. The loop is caused by the marriage between individuals 12 and 13 who have a relationship of first cousin once removed. Generations are often not distinct in the presence of loops.

Figure 2.3 An example of an unlooped pedigree with 2 affected members.

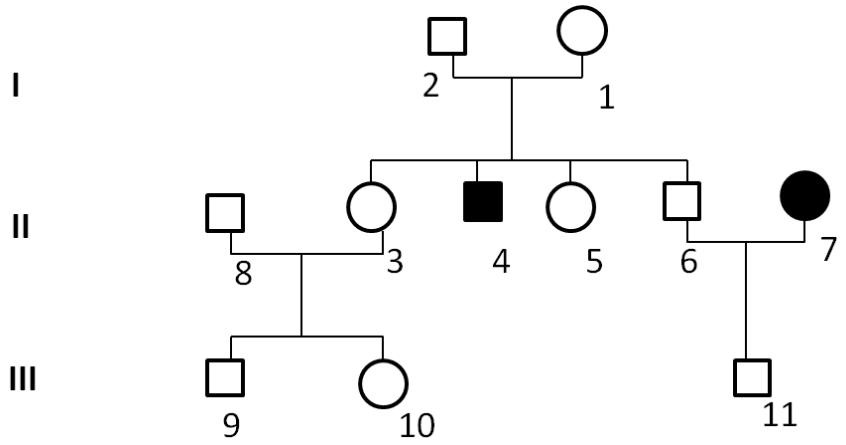
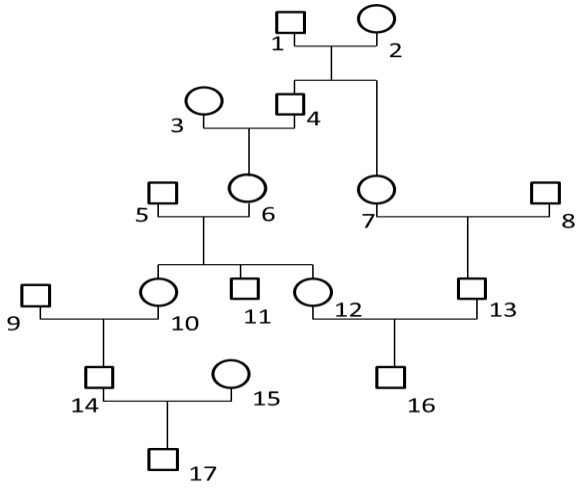


Figure 2.4 An example of looped pedigree.



Pedigrees are often used to determine the mode of inheritance (e.g. dominant, recessive, etc.) of genetic diseases. They are also essential to linkage analysis.

Linkage analysis is the analysis of the linkage in the inheritance between genes at different loci based on the observational phenotypes and the known pedigree structure (Palmer, 2011). Linkage between loci is the tendency for alleles of two or more loci close on the chromosome to be transmitted to the next generation together. So generally the closer two genes lie on a chromosome, the more likely they will show linkage. Genes located on different chromosomes, for example, do not show linkage. Genetic linkage studies aim to estimate the distance between a set of markers (polymorphic DNA sequences with known location) and a putative trait gene by estimating the recombination fractions. If a disease tends to be passed to offspring along with specific markers, then it can be concluded that the

gene(s) which are responsible for the disease are located close on the chromosome to these markers. The disease could be a Mendelian disease (caused by one gene) or a complex disease, which is caused by the action of many genes.

Linkage analysis is a method based on pedigree data. The information provided by pedigrees is determined by two factors. One factor is the structure and the size of the pedigree, e.g. number of generations, number of affected individuals. Another factor is the heterozygosity of the genetic markers. Higher heterozygosity is more informative because when a locus is homozygous we cannot see whether or not there is recombination between it and other loci.

Linkage analysis works well for Mendelian traits. But for complex diseases, the linkage analyses could only locate the genes to a large region (typically tens of cM). We could refine the locating to a smaller region by another method, association analysis, which can be a complementary to linkage studies.

Association analysis is used to establish the association between genotypes at one or more loci with a phenotype, which could be a quantitative character or a binary trait such as disease status. Association is a different concept from linkage. For linkage, there could be different alleles at the same locus linked with a same trait in different families. But for association, it is the same allele which is associated with the trait across the whole population. The availability of large numbers of SNPs and the reducing costs for genotyping have made association analysis an important study in genetic epidemiology. Importantly, association analyses are typically carried out on unrelated individuals.

A case-control study is a classical tool to carry out association analysis. Case-control studies use subjects who already have a disease or another trait and determine if these patients differs from those who do not have the disease or trait in any characteristics. In genetic case-control studies particularly, the frequencies of alleles or genotypes between the cases and controls are compared. The cases will have the disease or the trait under study; the controls will be unaffected and randomly selected from the population. A significant difference in the frequency of an allele or genotype of the genetic marker under consideration between these

two groups means that the marker could increase probability of having the disease or trait, or be in LD with a marker which does. Haplotypes can also show association with a disease or trait.

Pedigrees are not required for association studies but large population studies will undoubtedly contain relatives. These relationships, if ignored, could bias the results of an analysis. Methods for detecting relatives from genetic data would also be relevant to population association studies. Current genome-wide association studies provide a rich source of data with thousands of individuals typed for hundreds of thousands (even millions) of SNPs.

Confounding due to population stratification is another form of bias that can arise: misleading results can be produced when the individuals in the study are from different genetic backgrounds. Family-based designs, such as parent-case trios have been proposed to overcome such problems. In family-based designs, the alleles or genotypes transmitted to affected people are compared with those alleles or genotypes which are not transmitted. A family-based association test, transmission disequilibrium test (TDT), was proposed by Spielman et al. (1993).. All three individuals in each trio, two parents and one affected child, need to be genotyped. But the phenotypes of the parents do not need to be known. The TDT considers parents who are heterozygous for an allele and evaluates the frequency with which that allele is transmitted to affected offspring. For example, out of the total number of n parents with genotype A1A2, a parents have transmitted A1 to their children and b parents have transmitted A2 to their children ($a + b = n$). The departure of $\frac{a}{n}$ and $\frac{b}{n}$ from 50% are considered as a sign of association. A χ^2 test can be performed to test whether the association is significant.

2.5 Calculating likelihoods of pedigrees

The calculation of the likelihoods of pedigrees is important. It is required for traditional linkage analysis. The idea of using the computation of likelihoods to carry out linkage analysis was raised by Haldane (1934). Haldane and Smith (1947) developed the methods of using likelihood ratio and maximum likelihood

estimation to do linkage analysis, and calculating likelihood on the extended pedigrees.

Here I will introduce how the likelihood of a pedigree for genetic marker data can be calculated. I begin with unlinked genetic markers where the likelihood can be calculated one marker at a time. This includes the simple cases and more complex cases where some individuals are unobserved and a peeling method is needed (Thompson, 1986). Then the Lander-Green algorithm for likelihood calculation with linked markers is introduced.

2.5.1 Pedigree likelihood calculation when all individuals are observed

The likelihood that we want to calculate is the probability of the observed data under the hypothesized pedigree:

$$P(\text{observed data} \mid \text{hypothesized pedigree}). \quad (2.1).$$

When genotype data are available for all individuals on the pedigree, this probability can be written as

$$\prod_{\text{founders}} P(\text{genotype}) \prod_{\text{nonfounders}} P(\text{genotype}|\text{parental genotype}), \quad (2.2)$$

based on an assumption that the genotype of an individual at one locus is independent of the genotypes of all non-descendants in the pedigree given the genotypes of his/her parents. The likelihood of a pedigree can be easily calculated from Equation (2.2). $P(\text{genotype})$ is simply the genotype frequency in the population and $P(\text{genotype}|\text{parental genotype})$ can be calculated by Mendelian segregation.

I will describe the process of simulating genotype data on a pedigree. Consider the pedigree in Figure 2.5.

Suppose at one locus in this pedigree there are two alleles A1 and A2 with frequencies p(A1) and p(A2) respectively and p(A1) = p(A2) = 0.5. The founders are assumed to be unrelated and their genotypes are independent. The founder frequencies for different genotypes are 0.25 for A1A1, 0.5 for A1A2 and 0.25 for A2A2 assuming Hardy-Weinberg Equilibrium. Genotypes were simulated for the

founders 1, 2, 7 and 8 by randomly assigning genotypes A1A1, A1A2 and A2A2 to 1, 2, 7 and 8 according to their frequencies. The resulting genotypes are A2A2 for 1, A1A2 for 2, 7 and 8.

According to Mendelian segregation, the frequencies of offspring genotypes given the parental genotypes of (A1A2, A2A2) are 0.5 for A1A2 and 0.5 for A2A2. We can simulate the genotype of the 3, 4, 5 and 6 by randomly assigning genotypes A1A2 and A2A2 to them according to their frequencies. The result is A2A2 for 3, A2A2 for 4, A1A2 for 5 and A1A2 for 6. We can simulate the genotypes of individuals 9 and 10 based on the same frequencies. The resulting genotypes are A2A2 for 9, A1A2 for 10.

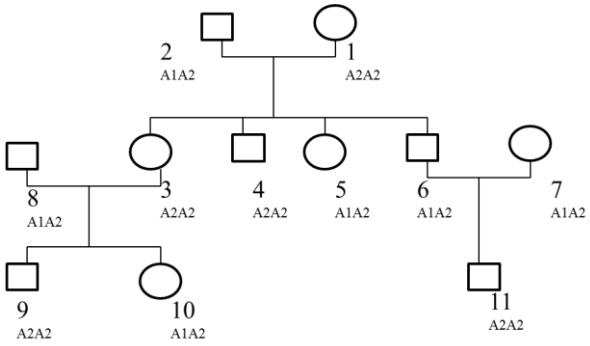
The genotypes of the parent of 11 are both A1A2 and the possibilities of 11's genotype and frequencies are 0.25 for A1A1, 0.5 for A1A2 and 0.25 for A2A2. The simulated genotype for 11 based on these frequencies is A2A2.

We can summarize the simulated genotypes in Table 2.1 and show them on the pedigree in Figure 2.5.

Table 2.1 The simulated genotypes for the pedigree in Figure 2.5.

Individual	1	2	3	4	5	6	7	8	9	10	11
Genotype	A2A2	A1A2	A2A2	A2A2	A1A2	A1A2	A1A2	A1A2	A2A2	A1A2	A2A2

Figure 2.5 Pedigree for genotype simulation and likelihood calculation.



For the pedigree with complete data in Figure 2.5, the likelihood can be simply computed by the formula

Likelihood = $\prod_{1,2,7,8} \mathrm{F(genotype)} \prod_{3,4,5,6,9,10,11} \mathrm{P(genotype|parental\ genotype)}$

$= (0.25\times0.5\times0.5\times0.5) \times (0.5\times0.5\times0.5\times0.5\times0.5\times0.5\times0.25)=0.5^{13}$.

2.5.2 Peeling method for likelihood calculation

When there are missing genotypes in the pedigree (some individuals are not genotyped), the likelihood calculation is more difficult. We need to sum up the probabilities of the observed genotypes for all the possible combinations of genotypes for those missing individual(s). This is called a numeration approach, which could be complex when the number of missing individuals is large and we need to calculate probabilities for the whole pedigree for every possible combination of the missing genotypes. Then the advantage of a peeling method is significant. In each step of it we only consider the possible genotypes of the missing individuals at that step.

Elston and Stewart (1971) first introduced the peeling method for unlooped pedigrees. Cannings et al. (1978) extended the method to more complex pedigrees of any form and size. I only use the unlooped pedigree to illustrate the peeling method for simplicity. For an unlooped pedigree, we can always find one member to partition the pedigree into two groups and such members are called pivots. One group connects the pivot through his parents and is called the 'above group' of the pivot. The other group connects the pivot through his offspring and is called the 'below group' of the pivot. The basis of the peeling method is that the probability of the genotypes above the pivot and the probability of the genotypes below the pivot are independent given the genotype of the pivot. We move the position of the pivot and sequentially each individual is peeled off and the information contained in those individuals is incorporated into a function on some remaining members of the pedigree.

The two most important formulae are as follows, where X is the pivot:

$$A_X(i) = P(\text{data above X \& X has genotype i}), \quad (2.3)$$

$$B_X(i) = P(\text{data below X \& X has genotype i}). \quad (2.4)$$

The formula (2.3) is used to peel down a pedigree from top to bottom and the formula (2.4) is used to peel up a pedigree from bottom to top.

For sequential computation, we need to define the following three components:

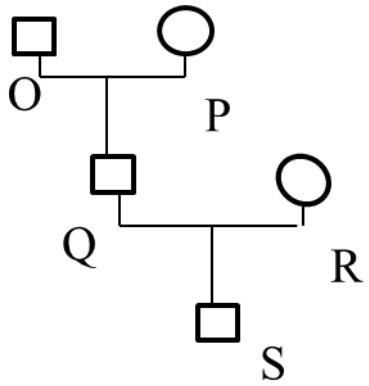
(1) The trait of interest is determined by a genotype at a locus. The frequency for genotype j in the population is F(j).

(2) The penetrance is defined as the probability that a person has a specified phenotype conditional on his genotype being j. S_A(j) = P(A has the phenotype | genotype of A is j). This item will always be 1 since only genotype data are used in this thesis.

(3) The transition probability is defined as the probability that a child has genotype i conditional on that his parents have genotypes j and k. T(i|j,k)=P(child has genotype i| parents having genotypes j, k). These would be 0, ¼, ½. 1 if inheritance was Mendelian, for example.

Suppose that in a pedigree we have 5 people in the order O, P, Q, R, S and Q is the pivot as shown in Figure 2.6 because it separates the pedigree into two unconnected parts.

Figure 2.6 A simple pedigree with pivot individual, Q.



Formula (2.3) in this case can be written as

A_Q(i) = P(data above Q & Q has genotype i)

= $\sum_j \sum_k$(P(genotype of O is k)P(genotype of P is j)P(O has the specific

phenotype | genotype of k)P(P has the specific phenotype | genotype of k)P(genotype of Q is i |O,P have genotype of j,k).

For simplicity, it can be written as

$$A_Q(i) = \sum_j \sum_k [F(j)F(k)S_O(k)S_P(j)T(i|j,k)]. \qquad (2.5)$$

Some rules need to be specified to make the peeling method work:

1) If there is no member above Q, then $A_q(i)$ = F(i) = P(Q has genotype i).
2) If there is no member below Q, then $B_Q(i) = 1$.
3) If a phenotype for Q is not specified, $S_Q(i) = 1$ for every i.

When working down the pedigree in this example, e.g. to a pivot Q, we use:

$$A_Q(i) = \sum_j \sum_k [A_M(j)A_F(k)S_M(j)S_F(k)T(i|j,k)], \quad (2.6)$$

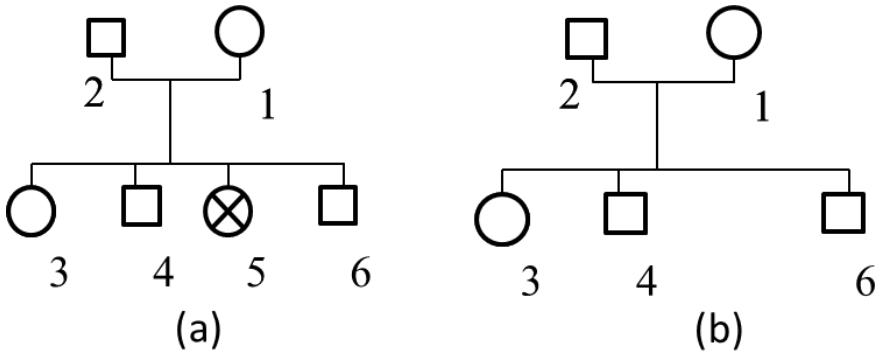
where M and F denote the mother and father of Q.

When working up a pedigree, Equation (2.7) is used to record information that each offspring C contributes to the combined parental genotypes (i,j) :

$$B^*_{M,F}(C: i,j) = \sum_k [S_C(k)B_C(k)T(k|i,j)]. \quad (2.7)$$

An unobserved individual can be ignored in calculating the likelihood if this individual does not have any offspring, illustrated with a simple example. This is important for simplifying calculations. If there are two parts of the pedigree linked by a couple without children, the likelihoods for the two parts can be calculated separately and multiplied together.

Figure 2.7 Pedigree with missing data for individual 5 (a) and Pedigree with individual 5 deleted (b).



If in the pedigree (a) of Figure 2.7 the only possible genotypes for 5 are i and j given the parental genotypes of 1 and 2, then P (i|1,2)+P (j|1,2) =1. The likelihood of observed data under this pedigree is:

$$P(1)P(2)P(3|1,2)P(4|1,2)P(5=i|1,2)P(6|1,2)+P(1)P(2)P(3|1,2)P(4|1,2)P(5=j|1,2)P(6|1,2)$$

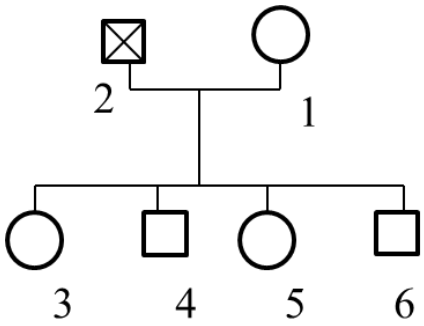
$$= P(1)P(2)P(3|1,2)P(4|1,\ 2)P(6|1,\ 2)[\ P(5=i|1,\ 2)+\ P(5=j|1,\ 2)]$$

$$= P(1)P(2)P(3|1,\ 2)P(4|1,\ 2)P(6|1,\ 2),$$

which is equal to the likelihood calculated for the pedigree (b) in Figure 2.7 with individual 5 deleted using Equation (2.2). The reason for this is that if an individual has no children and is also unobserved, then it does not provide any information to the probability of the genotypes of other people in the pedigree. This can be understood in another way that the relationship of all other individuals is unchanged by removing this individual. By the same reasoning, we can see that if two mating founders are not genotyped and they have only one child, then they can be removed and the child regarded as a founder; again this does not change the relationships between the other members of the pedigree.

Suppose that individual 2 is missing instead as shown in Figure 2.8.

Figure 2.8 Pedigree of Figure 2.7 (a) with individual 2 missing.



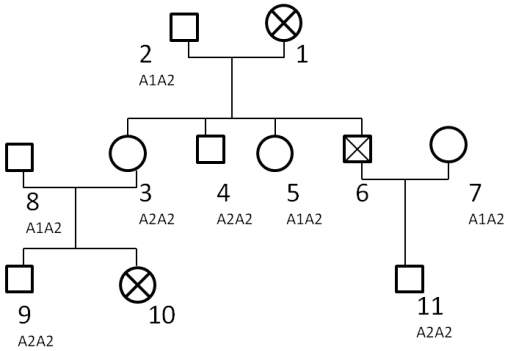
If the only possible genotypes for the individual 2 are i and j under this pedigree, then the likelihood is

P(1)P(i)P(3|1, i) P(4|1, i) P(5|1, i)P(6|1, i)+ P(1)P(j)P(3|1, j)P(4|1, j) P(5|1, j)P(6|1, j).

Individual 2 cannot be removed in calculating the likelihood for the observed genotypes as his/her possible genotypes are needed for calculating the probabilities for the genotypes of the children. And we can see that by removing individual 2, the relationships between other individuals are changed. For example, the relationship between 3 and 4 changes from certain full sibling to either full sibling or half sibling.

The peeling method will now be illustrated using the pedigree and simulated data of Figure 2.5 with individuals 1, 6 and 10 missing (Figure 2.9).

Figure 2.9 Pedigree of Figure 2.5 with multiple missing genotypes.



As shown above, individual 10 can be ignored as she has no data and no children in the pedigree. I will peel using the following sequence.

1) Peel 9 off and accumulate the information of 9 onto 8 and 3.
2) Peel 8 off and put all accumulated information onto 3.
3) Peel 3, 4 and 5 off and put all accumulated information onto 1 and 2.
4) Peel 1 and 2 off and put all accumulated information onto 6.
5) Peel 6 and 7 off and put all accumulated information onto 11.

The contribution of the data on 9 to the genotype probabilities of 8 and 3 is:
$\mathrm{B}^*_{3,8}(9{:}\,\mathrm{i},\mathrm{j}) = \sum_{\mathrm{k}}[\,B_{\mathrm{C}}(\mathrm{k})\mathrm{T}(\mathrm{k}|\mathrm{i},\mathrm{j})] = \mathrm{B}^*_{3,8}(9 : \mathrm{A2A2}, \mathrm{A1A2})$

$= B_9(\mathrm{A2A2})T(\mathrm{A2A2}|\mathrm{A2A2}, \mathrm{A1A2})]$

$=T(\mathrm{A2A2}|\mathrm{A2A2}, \mathrm{A1A2})]=0.5$

since $B_9(\mathrm{A2A2}) = 1$ because 9 does not have offspring.

Adding this contribution of 9 and information on 8 to 3, we get:
$B_3(\mathrm{A2A2})=A_8(\mathrm{A1A2})B^*_{3,8}(9{:}\,i,j)$

$=\mathrm{F}(\mathrm{A1A2}) \times 0.5 =0.25$ where $A_8(\mathrm{A1A2})=\mathrm{F}(\mathrm{A1A2})$ because there are no individuals above 8. Now we have accumulated all the information on 8, 9 to 3.

Next any information on 3, 4, 5 can be accumulated to their parents 1 and 2: The possible genotypes of 1 are A1A2 and A2A2 based on the genotypes of her children. P(A1A2)=0.5, P(A2A2)=0.25 according to the genotype frequencies of founders.

When genotype of 1 is A2A2,

$$B^*_{1,2}(3{:}\,\mathrm{i}, \mathrm{A1A2}) = B^*_{1,2}(3{:}\,\mathrm{A2A2}, \mathrm{A1A2}) = B_3(\mathrm{A2A2})T(\mathrm{A2A2}|\mathrm{A2A2}, \mathrm{A1A2})$$

$=0.25\times0.5=0.5^3$

$$B^*_{1,2}(4{:}\,\mathrm{i}, \mathrm{A1A2}) = B^*_{1,2}(4{:}\,\mathrm{A2A2}, \mathrm{A1A2}) = B_4(\mathrm{A2A2})T(\mathrm{A2A2}|\mathrm{A2A2}, \mathrm{A1A2})$$

$=1\times0.5 =0.5$ as 4 does not have descendent and $B_4(\mathrm{k})=1$ for any k.

$$B^*_{1,2}(5{:}\,\mathrm{i}, \mathrm{A1A2}) = B^*_{1,2}(5{:}\,\mathrm{A2A2}, \mathrm{A1A2}) = B_5(\mathrm{A1A2})T(\mathrm{A1A2}|\mathrm{A2A2}, \mathrm{A1A2})$$

=1×0.5=0.5 as B_5(k)=1 for any k.

If we accumulate all the information from all the offspring of 1 and 2 except 6 to 1 and 2 and specify it as $B_{1,2}^{**}$(i,j), we will get:

$B_{1,2}^{**}(\mathrm{i,A1A2}) = B_{1,2}^{**}(\mathrm{A2A2, A1A2})$

$= B_{1,2}^{*}(3\mathrm{:A2A2,A1A2})B_{1,2}^{*}(4\mathrm{:A2A2,A1A2})B_{1,2}^{*}(5\mathrm{:A2A2,A1A2})$

$=0.5^3\times0.5\times0.5=0.5^5$.

When the genotype of 1 is A1A2,

$B_{1,2}^{*}(3\mathrm{:i,A1A2}) = B_{1,2}^{*}(3\mathrm{:A1A2,A1A2}) = B_3(\mathrm{A2A2})T(\mathrm{A2A2|A1A2,A1A2})$

$=0.25\times0.25=0.5^4$

$B_{1,2}^{*}(4\mathrm{:i,A1A2}) = B_{1,2}^{*}(4\mathrm{:A1A2,A1A2}) = B_4(\mathrm{A2A2})T(\mathrm{A2A2|A1A2,A1A2})$

=1×0.25 =0.25 as B does not have descendent and B_4(k)=1 for any k.

$B_{1,2}^{*}(5\mathrm{:i,A1A2}) = B_{1,2}^{*}(5\mathrm{:A1A2,A1A2}) = B_5(\mathrm{A1A2})T(\mathrm{A1A2|A1A2,A1A2})$

=1×0.5=0.5 as B_5(k)=1 for any k

$B_{1,2}^{**}(\mathrm{i,A1A2}) = B_{1,2}^{**}(\mathrm{A1A2, A1A2})$

$= B_{1,2}^{*}(3\mathrm{:A1A2,A1A2})B_{1,2}^{*}(4\mathrm{:A1A2,A1A2})B_{1,2}^{*}(5\mathrm{:A1A2,A1A2})$

$=0.5^4\times0.25\times0.5=0.5^7$

This can be all accumulated onto 6 together with any information on 1 and 2 and produces A_6(i).

For i=A1A2,A_6(i)= $\sum_{\mathrm{j=A2A2\ or\ A1A2}}[\,A_1(\mathrm{j})A_2(\mathrm{A1A2})\ \mathrm{T(A1A2|j,\ A1A2)}\ B_{1,2}^{**}(\mathrm{j,\ A1A2})]$

= F(A2A2) × F(A1A2) × 0.5×0.5^5+ F(A1A2) ×F(A1A2) $\times0.5\times0.5^7$

$=0.25\times0.5\times0.5\times0.5^5+0.5\times0.5\times1\times1\times0.5\times0.5^7$

$=3\times0.5^{10}$

For i=A1A1,A_6(i)=$\sum_{\text{j=A2A2 or A1A2}}$[A_1(j)A_2(A1A2) T(A1A1|j, A1A2) $B^{**}_{1,2}$(j, A1A2)]

= F(A2A2) ×F(A1A2) ×0×0.5^5+ F(A1A2) ×F(A1A2) ×0.25×0.5^7

=0.5×0.5×0.25×0.5^7

=0.5^{11}

For i=A2A2, A_6(i)=$\sum_{\text{j=A2A2 or A1A2}}$[A_1(j)A_2(A1A2) T(A2A2 |j, A1A2) $B^{**}_{1,2}$(j,A1A2)]

= F(A2A2) F(A1A2) ×0.5×0.5^5+ F(A1A2) × F(A1A2) ×0.25×0.5^7

=0.25×0.5×0.5×0.5^5 +0.5×0.5×0.25×0.5^7

= $0.5^9 + 0.5^{11}$

=5× 0.5^{11}.

Then accumulate these information and any information of 7 onto 11, we get

A_{11}(A2A2)= $\sum_{\text{j=A1A2,A1A1 or A2A2}}$[A_6(j) A_7(A1A2) T(A2A2| A1A2, j)]

=3 × 0.5^{10} ×F(A1A2) ×0.25+0.5^{11}× F(A1A2) ×0+5×0.5^{11} × F(A1A2) ×0.5

=3×0.5^{13}+5× 0.5^{13}

=0.5^{10}

The total probability is A_{11}(A2A2) B_{11}(A2A2)) = 0.5^{10} ×1= 0.5^{10} and this is likelihood of the pedigree.

We can check this result by doing the calculation again with the enumeration method. Firstly, we enumerate all the possible genotype combinations for the genotypes of the missing individuals, and then calculate likelihood for each option. At the end, add the likelihoods up to provide the total likelihood. This is possible because there are only a few options in this case. Individual 10 is disregarded again and the possible combinations for the genotypes of individuals

(1, 6) which are consistent with the pedigree are (A2A2, A2A2), (A2A2, A1A2), (A1A2, A1A2), (A1A2, A2A2). We then calculate the likelihoods for the pedigree with each of these possibilities by the formula: P(1) ×P(2) ×P(7) ×P(8) ×P(3|1,2) ×P(4|1,2) ×P(5|1,2) ×P(6|1,2) ×P(9|3,8) ×P(11|7,6) and get $0.5^{11} + 0.5^{12} + 0.5^{13} + 0.5^{13} = 0.5^{10}$. This is the same as what we get using the peeling method.

There could be different sequences for peeling and some may be more efficient than others. In the above example, after we accumulated the information of individuals 8, 9 and 10 onto individual 3, we could have accumulated the information on individuals 7 and 11 onto individual 6 and then accumulated the information of all individuals onto individuals 1 and 2. This is easily seen for a simple example but finding a good peeling sequence can be computationally challenging for a large complex pedigree.

Only one locus is used in this example, but the peeling method can be extended to deal with multiple unlinked loci and looped pedigrees. For a looped pedigree, we cannot always partition the pedigree into two with a single pivot and may require a partition set of individuals instead. For complex pedigrees, these pivot sets can be quite large, e.g. two complex pedigrees mentioned by Sheehan (2000): the Polar Eskimo pedigree and the Pima Indian pedigree, need maximal cut sets of 50 and 75 respectively to peel (Sheehan, 2000). So the algorithm can easily run into storage problems. For similar reasons, linked loci, although manageable in principle, lead to computational problems.

2.5.3 The Lander-Green algorithm

The peeling method of the last section is based on the Elston-Stewart algorithm which sums the variables sequentially over a pedigree, but jointly over loci. Its complexity scales exponentially with the number of the loci and linearly with the number of people. Therefore it is suitable for pedigrees of arbitrary size, but only a few markers.

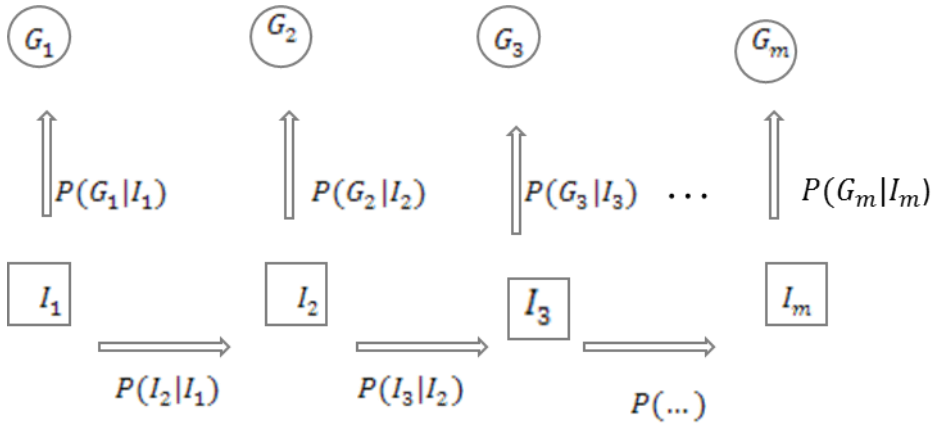
Here we consider another algorithm, Lander-Green algorithm, whose computation proceeds sequentially along the chromosome and jointly over all loci. Its complexity scales exponentially with the number of the size of the

pedigree, and linearly with the number of loci. So it is capable of dealing with a large number of loci, but the size of the pedigree that it can process is limited. Generally a pedigree with no more than 25 non-founders is practical (Lander and Green, 1987). Many types of computational software have been developed to implement the Lander-Green algorithm, such as ALLEGRO (Gudbjartsson et al., 2000) and Merlin (Abecasis et al., 2002).

To explain the Lander-Green algorithm, firstly we need to introduce the concept of the inheritance vector. Suppose there are f founders and n nonfounders in a pedigree. There will be 2n meioses with 2n gametes transmitted to the nonfounders. Each allele of one nonfounder could be from the paternal or maternal chromosome of his parent. For a locus M_j, we define an inheritance vector I_j as a vector with 2n coordinates corresponding to 2n gametes. Each coordinate takes the value of 0 if the gamete carries the DNA from the paternal chromosome of the parent, or 1 if the gamete carries the DNA from the maternal chromosome of the parent. Because each coordinate of the I_j has two possible values, there are potentially 2^{2n} different forms for I_j. But many of them are inconsistent with the observed data and can be excluded.

The probability that I_j on M_j differs from I_{j+1} on M_{j+1} is determined by the recombination fraction θ_j between M_j and M_{j+1}. It is assumed that, given the value of I_j at M_j, the probabilities of the value of I_{J+1} are independent of $I_1 \ldots . I_{j-1}$, which is the Markov property. So the inheritance vectors $I_1 \ldots . I_m$, where m is the number of loci form a Markov chain and are called a Hidden Markov Chain (HMM) because the state of I_j is not observable. But the genotype is observable and is influenced by the state of the inheritance vector. We can get some information of the state of I_j by $P(G_j|I_j)$, the probability of genotype G_j conditional on I_j. This HHM can be demonstrated by the following figure.

Figure 2.10 Demonstration of HMM in Lander-Green algorithm.



If we denote the genotype data of the whole pedigree over all the m loci as G, G_j as the genotype data of the whole pedigree on locus j, I_j as the inheritance vector for the locus j and I = $(I_1, I_2 \dots . I_m)$ is the set of inheritance vectors, the likelihood can be written as:

$$L = P(G) = \sum_I \{P(G|I)P(I)\} = \sum_{I_1} \dots \sum_{I_m} \{P(I_1) \prod_{j=2}^{m} P(I_j|I_{j-1}) \prod_{j=1}^{m} P(G_j|I_j)\} \quad (2.8)$$

where $P(I_1)$ is the prior probability of the state of the inheritance vector at locus M_1, $P(G_j|I_j)$ is the probability of the observed genotype given the inheritance vector state at locus j and $P(I_j|I_{j-1})$ is the transition probability of the heritance vector for state I_j at locus M_j given state I_{j-1} at locus M_{j-1}. This computation proceeds along the chromosomes jointly over all the meioses instead of proceeding sequentially over the pedigree and jointly over all loci.

From Equation (2.8) it can be seen that there are three ingredients for the likelihood calculation by Lander-Green algorithm:

1) Listing all possible inheritance vector states;
2) Calculating $P(I_j|I_{j-1})$ for each locus M_j;
3) Building the transition matrix for the HMM.

Then with all these ingredients ready, we can calculate the likelihood along a chromosome:

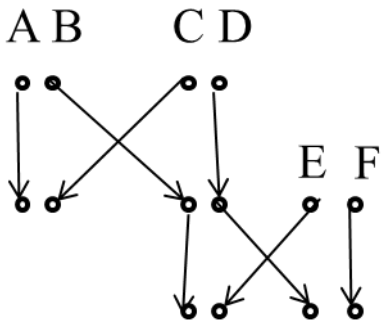
1) For a pedigree with n nonfounders, there are 2n meioses and the length of the inheritance vector is 2n with each coordinate corresponding to the result of one meiosis. There are 2^{2n} possible states for the inheritance vector by setting the value at each coordinate as 0 or 1.

   For example in a nuclear family with one father, one mother and one child, there are n=2 meioses. So the length of the inheritance vector at a specific locus is 2n=4 and there are 2^4=16 possible states for this inheritance vector, which can be listed as (0000,0001,0010,0011,0100,0101,0110,0111,1000,1001,1010,1011,1100, 1101,1110,1111).

   If one state is regarded as one element, all the possible states for I_j at locus M_j can be regarded as a vector and denoted as p_j.

2) At each locus M_j, calculate a probability vector q_j. Each element in this vector is set as $P(G_j|I_j)$, the probability of G_j conditional of each possible state of I_j. This $(G_j|I_j)$ can be calculated by studying the genetic descent graph (a graphical description of a specific inheritance vector) and the gene flow patterns. For each state of I_j list the possible sets of founder alleles which are compatible with the observed genotypes. The likelihood of the data, conditional on this state of I_j and one set of founder alleles, is just the product of the allele frequencies. Then sum them over all possible sets of founder alleles to get the probability $P(G_j|I_j)$ at that state of I_j. Figure 2.11 shows a genetic descent graph which displays how the founder alleles descend through the pedigree. This is a representation of an inheritance vector.

Figure 2.11 A genetic descent graph.



3) Build the transition matrix. Given the recombination fraction θ_j between the locus M_j and M_{j+1}, each coordinate of the inheritance vector I_j has the probabilities of θ_j to change value from 0 to 1 or from 1 to 0 and the probability of 1-θ_j of unchanged. So the transition probability from one state to another state is the product of the powers (exponents) of θ_j and $1-\theta_j$. The power of θ_j is the number of meioses where the value of the coordinate of the inheritance vector I_j changed and the power of $1-\theta_j$ is the number of meioses where the coordinate of the inheritance vector I_j unchanged. The transition matrix with one meiosis is

T= $\begin{bmatrix} 1-\theta_j & \theta_j \\ \theta_j & 1-\theta_j \end{bmatrix}$. The general form of the transition matrix T(θ_j) between locus M_j and M_{j+1} for a pedigree with n nonfounders is the Kronecker product of the above 2×2 matrix corresponding to transitions of each 2n coordinates in the inheritance vector. The Kronecker product, denoted by $\otimes$, is an operation on two matrices of arbitrary size resulting in a block matrix, like

$$A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix}.$$

$$T(\theta_j) = T^{\otimes 2n} = \begin{bmatrix} 1-\theta_j & \theta_j \\ \theta_j & 1-\theta_j \end{bmatrix}^{\otimes 2n} = T^{\otimes 2n-1} \times T$$

$$= \begin{bmatrix} (1-\theta_j)T^{\otimes 2n-1} & \theta_j T^{\otimes 2n-1} \\ \theta_j\, T^{\otimes 2n-1} & (1-\theta_j)T^{\otimes 2n-1} \end{bmatrix} .$$

The size of this matrix is $2^{2n} \times 2^{2n}$ for n nonfounders.

With these three components, we can run the Markov chain to calculate the likelihood:

1) Define a left conditional probability vector L_j=$L_{j-1}T_{j-1,j}\circ\ q_j$ where $T_{j-1,j}$ is the transition matrix from locus M_{j-1} to M_j and $\circ$ is the componentwise vector multiplication (derive a new vector by multiplying the corresponding components of two vectors with same length).

2) For j=1, set L_j=$I_1 q_1$ where $I_1$is the prior inheritance distribution, assumed to be a uniform distribution without extra information. Then iterate the process of L_j=$L_{j-1}T_{j-1,j}\circ\ q_j$ along the chromosome until locus M_m.

3) Then the overall likelihood is just the summation over all the elements of L_m.

Here I use an example to show the likelihood calculation by the Lander-Green algorithm. The pedigree is a nuclear family with the child genotyped over 5 unlinked loci and two parents untyped. The genotypes and alleles frequencies are shown in Table 2.2.

Table 2.2 The genotypes and alleles frequencies in an example for Lander-Green algorithm.

Locus	A1	A2	A3	A4	A5
Individual1	12	12	12	59	12

The number of alleles of loci and allele frequencies are in Table 2.3.

Table 2.3 The types of alleles of different loci and allele frequencies in the example for Lander-Green algorithm.

Locus	A1	A2	A3	A4	A5
No. of alleles	3	4	2	10	2
Allele 1	0.5	0.3	0.85	0.1 for each of	0.6
Allele 2	0.25	0.3	0.15		0.4
Allele 3	0.25	0.3		10 alleles	
Allele 4		0.1			

There are two meiosis in the pedigree and the possible states of the inheritance vector are (0, 0), (0, 1), (1, 0), (1, 1).

At locus A1, the probability for the inheritance state (0,0) is f(1)×f(2)=0.5×0.25=0.125 where f denote the allele frequency, for (0,1), (1,0) and (1,1) are all f(1) × f(2) =0.125. So q_1=(0.125,0.125,0.125,0.125). Similarly, q_2=(0.09,0.09,0.09,0.09), q_3=(0.1275,0.1275,0.1275,0.1275), q_4=(0.01,0.01,0.01,0.01), q_5=(0.24,0.24,0.24,0.24).

Because the 5 loci are unlinked so θ_j=0.5 for j=1,…,4 and the transition matrix is

$$\begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}^{\otimes 2}.$$

Run the Markov chain and get L_5= (8.60625e-07 8.60625e-07 8.60625e-07 8.60625e-07). Likelihood= 3.4425e-06. This result is checked and consistent with the likelihood calculated by P(A1) ×P(A2) ×P(A3) ×P(A4) ×P(A5) (R code in Appendix 10.1). We need to note that this example is just for demonstration purposes. For unlinked loci, the Lander-Green algorithm is not the most efficient algorithm.

Lander-Green algorithm is quite slow, but several refinements have been raised to improve the computing efficiency. Efforts have been made in two directions. One is to simplify the matrix computation. Another is to reduce to number of inheritance vectors. In the first direction, Idury and Elston (1997) proposed the divide and conquer algorithm, dividing the multiplication of a big matrix into the multiplication of smaller matrices. Kruglyak and Lander (1998) introduced the method of fast Fourier transforms. In another direction, Kruglyak et al. (1996)

used founder symmetry to reduce the number of inheritance vectors from 2^{2n} to 2^{2n-f}. Gudbjartsson et al. (2000) used founder couple symmetry to reduce the number of inheritance vector to 2^{2n-f-c} where c is the number of founder couples in the pedigree.

2.6 Summary

This chapter introduced the concepts and methods that will be used in the thesis. Firstly genetics was introduced as all data used in the thesis are genetic data, it is essential to know the basic concepts about genetics. Then pedigrees and different methods to calculate pedigree likelihood were described.

3 Datasets and software

3.1 Affymetrix 500K SNP allele frequency and map

Allele frequencies for the Affymetrix 500K Array Set derived from HapMap data have been used in simulations. They were downloaded from http://folk.uio.no/thoree/FEST. The physical map distances have been changed to genetic map distances (cM) in accordance with the Rutgers Combined Linkage-Physical map of the Human Genome. After quality control, there are 416,854 SNPs in total in this data set.

3.2 WTCCC Affymetrix SNP 6.0 dataset

I applied to the Wellcome Trust Case Control Consortium for Affymetrix SNP6.0 SNP genotype data of their control groups. This provides 2934 unrelated individuals, each with data on 893,634 autosomal SNPs. After data cleaning, I retain 2674 unrelated individuals and 711,020 SNPs. 1285 individuals are from the NBS (National Blood Service) control group and 1389 are from the 1958 Cohort control group.

3.3 SNP data from HapMap for CEU and TSI

I downloaded allele frequency data of two populations from the HapMap project. One population is denoted as ‘CEU’ on the HapMap website and the samples are from Utah residents with Northern and Western European ancestry. Another population is denoted as ‘TSI’ and the samples are from Tuscan Italians of Southern Europe. Because the numbers of SNPs in the allele frequency files for the two populations are slightly different, I carry out a filtering step when downloading the allele frequency files by only acquiring the SNPs common in both populations, with MAF ≥ 0.01 and for which I have a linkage map. This procedure yields 556,873 SNPs for each population.

3.4 The MICROS dataset

I have used real data from the MICROS study of Bolzano, Italy (Pattaro *et al*., 2007). Bolzano is the capital city of a province of Italy known as Southern Tyrol. The MICROS study is carried out in three villages of the Val Venosta around Bolzano which have been quite isolated due to geographical, historical and political reasons. The objective of this study was to screen the population for the presence of several traits by questionnaire and pedigree data together with clinical measurements and blood samples. Starting from the typed individuals, they have reconstructed pedigrees with 15 generations including 50,037 individuals, going back to the early 1600s. Most of the participants are connected by one large pedigree. The dataset that I obtained from Bolzano includes genotypes for 1286 individuals over 313,285 SNPs (including sex chromosomes). The platform used for genotyping is the Illumina 300K SNP chip. In this thesis, only markers on the autosomal chromosomes are used for these data. Allele frequencies are estimated from all the typed individuals. The linkage map for Illumina 550K is used, which covers all SNPs in Illumina 300K apart from several hundred. The number of autosomal SNPs that are kept is 303,783 in total. There are four pedigree files with 1, 2, 4, 8 and 12 generations above the sampled individuals respectively. The more generations there are, the more detailed our knowledge of the relationships between the individuals.

3.5 Merlin Software

The Merlin Software (Abecasis et al., 2002) was used to simulate the genotypes and calculates the pedigree likelihood using the Lander-Green algorithm as described in Chapter 1.

Four input files are needed by Merlin, each with a specified format. The first type of input file is a pedigree file with a suffix of ‘ped’ which describes the relationships between the individuals and contains the genetic data of the individuals. It includes five essential columns. These first five columns are: family identifier (a label which could be a name or a number), individual identifier, identifier of the father, identifier of the mother and the sex of the

individual (1 for male and 2 for female), respectively. There could be many families in one pedigree file. The columns after the first five columns are for any number of various types of genetic data, including phenotypes for discrete and quantitative traits and marker genotypes. The content in an example pedigree file hs1.ped is as following:

```
# hs1.ped

1 1 0 0 2 1   0 33  00

1 2 0 0 1 1   0 44 00

1 3 0 0 2 1   0 12 00

1 4 2 1 1 1   0 43 00

1 5 2 3 2 2   6 13 22
```

The pedigree in this file is a single family of half siblings with individuals 1, 2 and 3 as parents and individuals 4 and 5 as children. 0 means that the value is missing for an individual. Hence, the third and fourth columns indicate that 1,2 and 3 are all founders since they have no parents in the pedigree, whereas 4 and 5 have the same father (2) but different mothers (1 and 3). The sixth column is for disease status, where 1 means unaffected and 2 means affected. The seventh column is for a quantitative trait, where 6 is the value for the individual 5. The eighth and ninth columns are for one genetic marker with the two values coding the observed genotypes, each representing one allele. The number 1, 2, 3 or 4 ….represent different alleles, so Merlin allows multiple alleles. The genotypes are unordered, so '3 4' and '4 3' will be treated as the same genotype. The tenth and eleventh columns are for another genetic marker. 0 0 means the genotype is missing for that individual. The two alleles of a genotype can be separated by a slash sign as '4/4'. I only use genotype data in my work. So the columns for disease status and quantitative trait values are not needed. The first line in hs1.ped will be read like this: this individual belongs to family 1; he is the 1st individual in this family; the identifiers for his father and mother are both 0 because he is a founder and his parents are not included in the pedigree; he is a male; he is

unaffected; the data for his quantitative trait are missing; his genotype for the first marker is 3/3 and his genotype for the second marker is missing.

A pedigree file can include data of different types. So a second type of input file: *data file* is needed to describe the content of the pedigree file. A data file describes the content of the pedigree files apart from the first five essential columns. Data files have two columns. The first column is the type of data item and the second column provides a label (a name) for each data item. Each data item of the pedigree file corresponds to one row in the data file. The naming of the pedigree file and data file is somewhat confusing. In fact all the data are included in the pedigree file. Data files just explain the columns in the related pedigree files to show us what kind of data those columns represent. A data file corresponding to the pedigree file hs1.ped is hs1.dat and is shown below.

```
# hs1.dat

A  status of a disease

T  value of a trait

M  marker 1

M  marker 2
```

The third type of input file is a map file which specifies which chromosome each marker is on and its linkage mapping position (in cM) on it. Merlin will convert the linkage distance into a recombination fraction using the Haldane mapping function. An example of a map file is hs1.map as follows:

```
# hs1.map

CHROMOSOME   MARKER     POSITION

12           marker 1    112.3

12           marker 2    114.5
```

The fourth type of input file is a frequency file which specifies the number of alleles at each locus and their frequencies. The following is an example of the frequency file hs1.freq where marker 1 has four alleles and marker 2 has two alleles.

```
# hs1.freq

M marker 1

F 0.1 0.5 0.2 0.2

M marker 2

F 0.6 0.4
```

For Merlin to simulate genotype data, the pedigree file, data file and map file are all required. A frequency file is optional. Merlin's simulation function can generate random datasets that resemble the original data set in terms of marker positions, allele frequencies and missing patterns. If a genotype is missing in the original pedigree file, it will be missing in the simulated pedigree file as well. Merlin assigns random alleles to founders according to allele frequencies at each marker. If a frequency file is given to Merlin, the allele frequencies in this file will be used. Otherwise Merlin will estimate the allele frequencies from the genotype data in the pedigree file. Alleles are simulated independently at each marker for founders when we assume linkage equilibrium i.e. when we choose not to allow LD. These are then segregated through the pedigree using the relationships specified in the pedigree file and recombination fraction deduced from the map file. It replaces the original genotypes with these simulated genotypes, retaining the original pattern of missing data exactly (for example, if individual A is ungenotyped at a marker in the original pedigree file, individual A's genotype at that marker will be discarded in all simulation). Changing the random seed (with the -r command line option) will make Merlin generate a different set of founder chromosomes and segregation pattern, and consequently a different random simulation result.

Merlin can calculate the likelihoods of different hypothesized pedigrees for some individuals with genotype data. The calculation uses the Lander-Green algorithm (Lander and Green, 1987) because it involves large numbers of linked markers (for details see section 2.5.3 of this thesis).

While doing my study, I noticed that the official version of Merlin has a bug in calculating the likelihoods of distant pedigrees for 500K SNPs and the corrected version (from Øivind Skare) had to be used instead. The official release has not been debugged at the time of this thesis being written.

3.6 Mendel Software

The software package Mendel (Lange, et al., 2013) was used instead of Merlin in some simulation, because it enables tracking of simulated alleles at each locus of every non-founder to the founders. So we could obtain the exact kinship coefficient based on the realized IBD sharing between two individuals and compare this with the estimated kinship coefficient. To simulate data, the input files that Mendel requires are very different from Merlin. All the options are listed in a control file rather than in the command line, e.g. type ‘mendel control.in’ in Linux if the control file is named as ‘control.in’. The control file has the following format:

! Input Files

!

DEFINITION_FILE = ~/merlin/Mendel-130/affydef.in

MAP_FILE = ~/merlin/Mendel-130/affymap.in

PEDIGREE_FILE = ~/merlin/Mendel-130/affyped.in

!

! Output Files

!

NEW_PEDIGREE_FILE = affypedsourced.out

OUTPUT_FILE = Mendelaffy.out

SUMMARY_FILE = Summaryaffy.out

!

! Analysis Parameters

!

ANALYSIS_OPTION = Gene_dropping

REPETITIONS = 1

SEED = 1

MODEL = 1

KEEP_FOUNDER_GENOTYPES = False

MISSING_DATA_PATTERN = Existing_Data

MISSING_AT_RANDOM = 0.0

GENE_DROP_OUTPUT = Sourced

PEDIGREE_MAX_LINE_LEN = 2000000

PEDIGREE_LINKAGE_FORMAT = True

MAP_DISTANCE_UNITS= cM

Firstly the location and file names of the input files and output files need to be given. What come next are some options for Mendel. In the option of ANALYSIS_OPTION, ‘Gene dropping’ tells Mendel to simulate data. In the option of GENE_DROP_OUTPUT, ‘Unordered’ means that unordered genotypes be outputted. The label of the alleles in simulated data will be outputted if the value of this option is changed to ‘Sourced’. When ‘Sourced’ is chosen, every allele at all loci in the founders will be assigned a label and it is these labels that are passed down to the descendants. Then in the simulated data, we know the realized IBD status for every locus and between all individuals. But the simulation can be done only with either ‘Unordered’ or ‘Sourced’, and not both. So the

simulation needs to be done twice for every random seed to get both genotypes and labels.

3.7 R Statistical software

The statistical software R is an essential tool in this work and was used to prepare the input files for Merlin, to organize the output files of Merlin and to carry out analyses.

3.8 The high-performance computing facility (ALICE) of the University of Leicester

The ALICE, high-performance computing (HPC) cluster, of the University of Leicester was used in this study. ALICE can run multiple jobs at the same time, which can reduce the time carrying out the large numbers of simulations.

3.9 Eigensoft Software

Eigensoft (Patterson et al., 2006) is a package including many programs. Here I have only used one of its functions 'smartpca' which runs principal components analysis on genotype data and outputs principal components (eigenvectors) and eigenvalues. Then another of its functions was used to plot the top two principal components (or any specified pair of principal components).

3.10 Jenti Software

The software Jenti (Falchi and Fuchsberger, 2008) is a very useful tool for splitting large pedigrees into smaller ones to help visualize the pedigree and choose relative pairs with particular relationships. Jenti has been used to compute the expected kinship coefficients between individuals based on pedigrees. We can also specify the number of genotyped individuals in each sub-pedigree and the kinship coefficients between them when splitting pedigrees. When visualizing the sub-pedigrees we can specify how many generations above the genotyped individuals should be displayed.

3.11 ERSA and Germline Software

The software ERSA (Huff et al., 2011) was used to estimate the degree of relationship and it was compared with the method proposed in this thesis. The input file for ERSA is a list of IBD segments and their lengths which is outputted from software Germline (Gusev et al., 2009). The software Germline accepts a ped file and a map file in Plink format (Purcell et al., 2007) which is very similar to Merlin format. The only differences to Merlin format are that the ped file in Plink format has an extra column for phenotype (which can be set unknown) and the map file in Plink format has an extra column for physical map. Germline can detect IBD segments for the pairs of individuals included in the ped file.

4 Distinguishing relationships with pedigree-based likelihoods

This chapter is a more in-depth study of what was undertaken in Skare et al. (2009). That paper explored the potential of linked SNP markers and a pedigree-based likelihood method for estimating relationships. Other likelihood-based approaches and unlinked microsatellite markers have been used to estimate relationships traditionally (Milligan, 2003, Anderson and Weir, 2007) . With the availability of large numbers of linked SNP markers, it needs to be explored how these linked markers extend the range of identification problems that can be solved. Clearly the discriminatory power of one SNP marker will be less than that of one multi-allelic microsatellite marker and there will be a reduction of discriminatory power in linked markers compared to independently-inherited markers as there are correlations between linked markers. So a larger number of linked markers than unlinked markers will be needed. However, large numbers of linked markers are widely available in genome-scan data and the argument in Skare et al. (2009) is that these could be exploited to distinguish distant relationships. As in that paper, the approach taken here is that a pairwise relationship can be expressed by a pedigree which links the two relatives. There are many different existing software packages which can be used to calculate pedigree likelihoods.

The estimation problem in this chapter focuses on distinguishing specific alternative relationships rather than estimating unknown relationships. The paper of Skare et al. (2009) considered pairwise relationships and mainly focused on distinguishing a true relationship from an alternative relationship of ‘unrelated’. This is the classical forensic situation. For example, we only need to know whether an individual is a parent of another individual or not, whether an accused individual is the criminal or completely unrelated to the criminal, etc. Their results show that the information obtained from large sets of linked markers could increase substantially the number of problems that can be solved. With 500K autosomal SNPs, relationships up to the order of second cousins can be distinguished from ‘unrelated’ without ambiguity, which corresponds to m=6 meioses separating the two individuals of interest. Relationships up to third cousins can be distinguished from ‘unrelated’ with reasonable certainty. But any

relationship with m>8 separating meiosis seems to be beyond the scope of 500K markers.

4.1 Methods used and relevant notation

The ratio of the likelihoods of the two hypotheses of H_i and H_j is $\frac{L_i}{L_j}$, and is called the likelihood ratio. A likelihood ratio is typically expected for a forensic analysis. The Bayesian approach to inference is a method which updates the likelihood of a parameter by multiplying it with the prior distribution of that parameter to obtain a posterior distribution. Specifically, for a parameter θ and data vector x, the posterior probability distribution of θ is:

$$p(\theta|x) = \frac{p(x|\theta)}{p(x)} p(\theta),$$

leading to the proportionality:

$$\text{posterior probability} \propto \text{likelihood} \times \text{prior probability}.$$

The Bayesian approach will be used in this work, in addition to the simple likelihood. For cases with only two alternatives, the likelihood ratio is in fact more widely used and easier to interpret than the Bayesian posterior probability ratio. But the Bayesian method is more general: it allows us to consider cases with more than two alternative relationships and proper prior information can be accommodated if it is available. A flat prior is assumed throughout as all alternative relationships are assumed equally likely in the absence of any extra information.

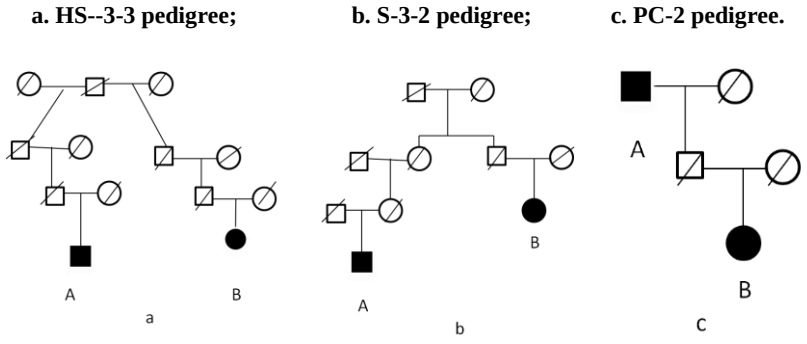
For n hypotheses concerning the relationship between two individuals, H_1, H_2,….. H_n, the prior probabilities are π_1, π_2,… π_n and $\pi_i = 1/n$ for a flat prior. The likelihood for H_i is the probability of the data under the hypothesis H_i and it is denoted as $L_i = P(data|H_i)$. Then the posterior probability of H_i will be:

$$\mathbf{P(H_i|data) = \frac{L_i \pi_i}{\sum_{i=1}^{n} L_i \pi_i} = \frac{L_i}{\sum_{i=1}^{n} L_i} \ as \ \pi_i = \frac{1}{n} \ for \ i = 1 \ldots n.} \qquad (4.1)$$

Note that $P(H_i|data) = \frac{L_i}{L_1+L_2}$ for $i = 1, 2$ when there are only two hypothesized relationships, so the posterior probability ratio of H_i and H_j is $\frac{P(H_i|data)}{P(H_j|data)} = \frac{L_i}{L_j}$ which is precisely the likelihood ratio.

Following Skare et al. (2009), all pedigrees connecting two individuals can be allocated into three types if looped pedigrees are excluded. The first type is $HS - n_1 - n_2$ where two individuals A and B share one common ancestor with n_1 being the number of generations between the common ancestor and A and n_2 being the number of generations between the common ancestor and B. For example, HS-1-1 represents the case where A and B are half-siblings. A HS-3-3, half second cousin, relationship is depicted in Figure 4.1a. The second type is $S - n_1 - n_2$ where two individuals A and B share two common ancestors and n_1, n_2 are the numbers of generations between the two common ancestors and A, B respectively. S-1-1 represents the case where A and B are full siblings. An S-3-2 (first cousin once removed) relationship is depicted in Figure 4.1b. The third type is PC-n where one individual A is an ancestor of another individual B and the number of generations between them is n. PC-1 is just a parent-child relationship. A PC-2, grandparent-grandchild, relationship is depicted in Figure 4.1c.

Figure 4.1 Examples of different types of unlooped pedigree.



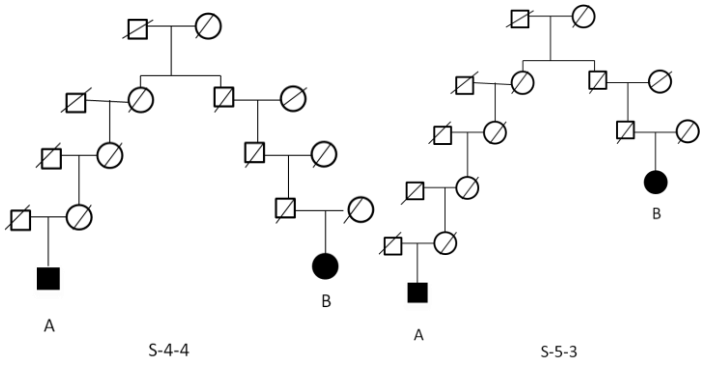
4.2 Verifying the results in Skare et al. (2009)

Firstly I verified the results reported from the R package FEST in the paper of Skare et al. (2009) using my own code. This provided a foundation for my extension work and also served as a code check. The software Merlin (Abecasis et al., 2002) was used to simulate the genotypes and calculate the likelihoods.

In this section, allele frequencies for the Affymetrix 500K Array Set derived from HapMap data are used. All simulations are based on real allele frequencies and genetic maps. Flat priors are used in the whole simulation study to report posterior probabilities. The Haldane map function was used to convert linkage distance to recombination fraction as defaulted in Merlin.

Before carrying out the major part of the work, the R code was checked using 400 simulations on a special case. The third cousin (S-4-4) and second cousin twice removed (S-5-3) relationships are not distinguishable by any amount of autosomal genetic information (Donnelly, 1983). This is because the likelihoods of both relationship types HS$-n_1-n_2$ and S$-n_1-n_2$ are determined by n_1, n_2 only through $n = n_1+n_2$. These two pedigrees are shown in Figure 4.2. The likelihoods of the two relationships were exactly the same for all 400 replicates of simulated data as expected. The same was checked to be true for relationships HS-6-6 and HS-5-7 as well.

Figure 4.2 Two pedigrees which cannot be distinguished by any number of autosomal markers.



4.2.1 Distinguish the true relationship from 'unrelated'

Genotypes were simulated for two individuals with different HS-n-n type true relationships and then it was tried to distinguish the true relationships from a single

alternative hypothesis of 'unrelated'. Likelihoods and posterior probabilities were calculated for both hypothesized pedigrees with a flat prior and the average posterior probabilities of the true relationship are shown in Equation (4.1). To see the effect of different numbers of markers, the simulations were performed with 22 (unlinked), 220, 2200, 22000 and 500K markers. 22 unlinked markers were generated by picking up one from each chromosome. 220, 2200, 22000 markers were generated by choosing 10, 100, 1000 markers evenly spaced from each chromosome respectively. 400 replicates were done when all 500K markers were used and 1000 replicates were done when less dense markers were used and the averages of the results were taken. Table 4.1 shows the average posterior probabilities of the true relationship when the only alternative relationship is that the two individuals are unrelated.

Table 4.1 Simulation results based on Affymetrix 500K SNP Array frequency and map data.

# of markers	HS-1-1	HS-2-2	HS-3-3	HS-4-4	HS-5-5	HS-6-6
22(unlinked)	0.608	0.512	0.500	0.500	0.500	0.500
220	0.923	0.580	0.507	0.501	0.500	0.500
2200	1	0.925	0.605	0.515	0.500	0.500
22000	1	1	0.947	**0.685**	0.550	0.547
500K	1	1	1	0.878	0.612	0.551

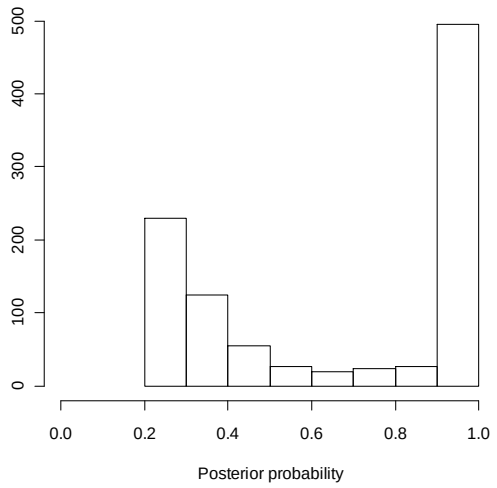
Average posterior probabilities for each true pedigree when the only alternative hypothesis is that the two individuals of interest are unrelated. Averages are taken over 400 replicates for 500K markers and over 1000 replicates otherwise.

From Table 4.1 we can see that 22 unlinked markers are not enough as the highest posterior probability is 0.608 for the HS-1-1 relationship. When the number of markers is increased to 220, this value becomes 0.923. So we have quite high power to distinguish half siblings from unrelated individuals with this few markers. With all 500K SNP markers we can solve this distinguishing problem with certainty for relationships up to HS-3-3 and with reasonable confidence up to HS-4-4. Note that HS-6-6 relatives are difficult to be distinguished from 'unrelated' with any number of SNPs used here.

These results are consistent with those reported in the paper of Skare et al. (2009). But the values are not exactly same of course because I used different random seeds in the simulations. Also I did not remove the markers with MAF (minor allele frequency) < 0.1 as they did.

It needs to be noted that the values in Table 4.1 are just the averages of the results of a large number of simulations. For example, 0.685 is the average of the 1000 posterior probabilities obtained from 1000 replicates with HS-4-4 as the true relationship using 22000 markers. The results of these replicates are quite varied as would be expected. A histogram of these 1000 posterior probabilities is shown in Figure 4.3. Due to this variability it is important to look at the results of individual replicates of the simulation, rather than just the average.

Figure 4.3 Histogram of posterior probabilities for 22,000 markers based on 1000 replicates, with HS-4-4 and 'unrelated' as the true and alternative relationships respectively.



4.2.2 Distinguish the true relationship from several alternative relationships

In this section, the true relationship is compared with several close alternative relationships, rather than just with 'unrelated'. 500K SNP genotype data were simulated for each of HS-1-1, HS-2-2……HS-5-5 and 'unrelated' relationships. Then for each simulated dataset, the true relationship and all other options were compared all together. Comparing several relationships all together is made possible by the Bayesian framework as a likelihood ratio can only be used to compare two relationships. Equation (4.1) was used for the calculation here. The results are shown in Table 4.2, from which we can see that as the relationship becomes more distant, it is more difficult to distinguish it from those close alternative relationships.

Table 4.2 Simulation results based on Affymetrix 500K frequency and map data with five alternative relationships for each true relationship.

True	HS-1-1	HS-2-2	HS-3-3	HS-4-4	HS-5-5	Unrelated
HS-1-1	1	0	0	0	0	0
HS-2-2	0	0.959	0.041	0	0	0
HS-3-3	0	0.034	0.748	0.189	0.028	0.001
HS-4-4	0	0	0.173	0.467	0.263	0.097
HS-5-5	0	0	0.023	0.275	0.388	0.313
Unrelated	0	0	0.002	0.089	0.326	0.583

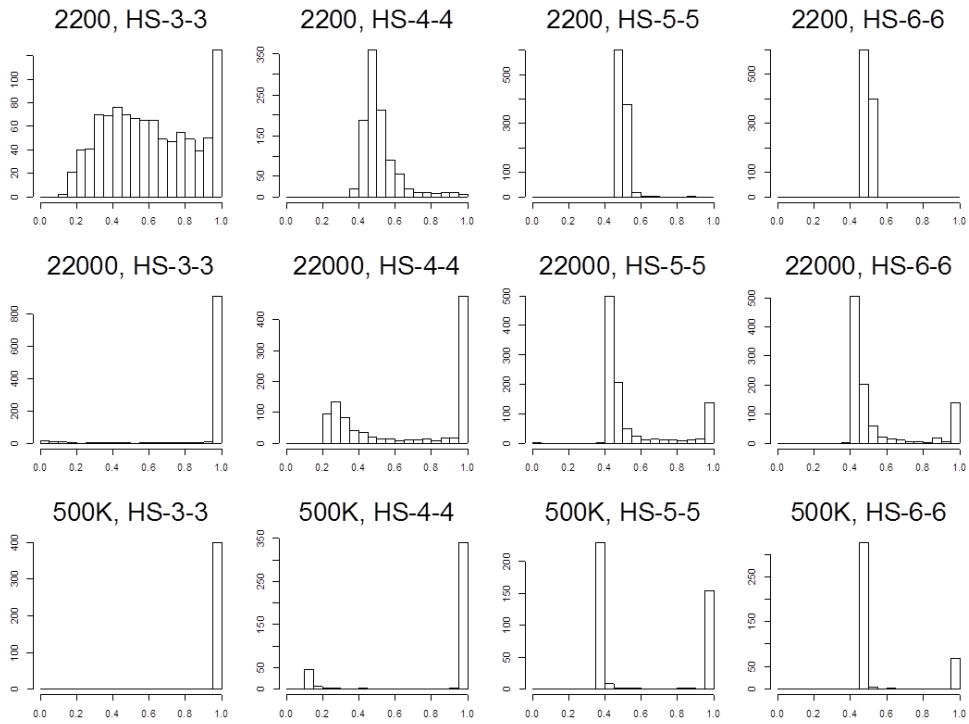
In the first column are the true relationships. Then for each true relationship, the posterior probabilities of the true relationship and five alternative relationships are shown. Each row may not sum to one due to rounding. Averages are taken over 400 replicates.

4.3 Some extensions to the paper of Skare et al. (2009)

4.3.1 Investigating the individual posterior probabilities in Table 4.1

Here I consider the complete posterior probabilities of all 400 or 1000 replicates for each simulation in 4.2.1 rather than just their averages. Histograms of the posterior probabilities corresponding to some of the averages reported in Table 4.1 are shown below in Figure 4.4. Complete histograms of the posterior probabilities for all entries in Table 4.1 are given in Appendix 10.2. Please note that I am going back to the case where there are only two options: a specific relationship versus ‘unrelated’.

Figure 4.4 Frequency histograms of the individual posterior probabilities of the true pedigrees contributing to each of the averages reported in Table 4.1 where there is only one alternative pedigree, 'unrelated', for each true pedigree. The three rows correspond to the different numbers of SNPs (2200, 22000, 500K) used in the simulation. The four columns correspond to the four different true pedigrees (HS-3-3, HS-4-4, HS-5-5, HS-6-6). The X-axis represents the posterior probability of the true pedigree.



From Figure 4.4 we can see two patterns in these posterior probability distributions. Firstly, when more markers are used, there is clearer separation in the posterior probabilities, some values are as high as 1, some are low values and there are very few values between the two extremes. Secondly, when dense markers are used, the proportion of values close to 1 gets smaller as the relationship becomes more distant. The explanation for these patterns is as follows. As noted by Donnelly (1983), the probability that there is no IBD sharing between two relatives increases when the number of meiosis between them increases, but if two relatives *do* share IBD, they will share quite substantial chromosomal segments. For example, the expected length of the shared IBD segments is 8.33cM for two relatives with 12 meioses (S-6-6) between them when they do have IBD sharing (Browning and Browning, 2012, Thompson, 2013). So there are two different situations in all these simulated genotype data sets. In some simulations, two related individuals shared some IBD chromosomal segments. In other simulations, there is no IBD sharing between them, therefore their genotypes look

like they are unrelated although they are related biologically. It suggests that those posterior probabilities close to 1 in my simulation correspond to the cases that there is IBD shared between the two relatives and those low values correspond to the cases that there is no IBD shared. When less dense markers are used, those two patterns in the histograms are not very clear because firstly, the less dense SNP markers cannot pick out all shared IBD segments in the relatives; secondly, the smaller number of markers do not give enough information to make the true pedigree distinguishable with certainty from 'unrelated' even when two relatives do share IBD.

Note also that the low values of the posterior probabilities for different true relationships in these histograms are different, being around 0.15 for HS-4-4, 0.35 for HS-5-5 and 0.45 for HS-6-6. The reason for these values lower than 0.5 is that when there is no IBD shared between two relatives, the likelihood of an 'unrelated' pedigree is higher and the likelihood of the true pedigree is lower. This is because the true pedigrees are not supported by the data and the observed genotypes are consistent with the hypothesis of 'unrelated'. The relationship HS-4-4 gave lower values than the relationship HS-6-6 because HS-6-6 is a more distant relationship than HS-4-4 and closer to 'unrelated' than HS-4-4. Therefore it is more 'wrong' to say HS-4-4 is 'unrelated' than to say HS-6-6 is 'unrelated'. So when there is no IBD shared by two relatives, the posterior probability of the true pedigree when HS-4-4 is the true pedigree is lower than when HS-6-6 is the true pedigree. This is further evidence that the low posterior probabilities correspond to the situations where there is no IBD shared between relatives.

I also found that these posterior probabilities have limits for both cases where the two relatives share IBD and where they do not share IBD. Therefore the average posterior probabilities have limits. It means that they cannot keep increasing no matter how informative the data are. Suppose the probability of no IBD sharing between two related individuals is β (note this is a fixed value for any specific relationship), then in the cases where there is no IBD shared by two relatives, the limit of the posterior probabilities of the true pedigree (relationship not supported by genotypes) is $\beta/(1+\beta)$. This can be seen as follows. The genotype data that were simulated are unrelated and there are only have two alternative pedigrees, the true pedigree and 'unrelated'. The probability that the 'unrelated' pedigree generates these unrelated genotypes is 1 and

the probability that the true pedigree generates these unrelated genotypes is β. With a flat prior probability distribution, the posterior probability of the true pedigree for these genotype data is β/(1+ β). This limit will be approached when dense markers are used. From the 500K SNP cases in Figure 4.4 we can estimate β for a HS-5-5 relationship at around 0.6 (240/400). It is estimated from the histograms in Figure 4.4 where 240 out of 400 simulated data sets having no IBD sharing when the true relationship is HS-5-5. Then the posterior probability of the true pedigree HS-5-5 when there is no IBD sharing is 0.6/1.6=0.375 which corresponds to the low values in the histogram. The expected posterior probability of the true pedigree is therefore $(1-\beta)\times 1+\beta\times\frac{\beta}{1+\beta}$ which is equal to 0.625 when $\beta = 0.6$. As another example, if we estimate $\beta = 0.8$ (320/400) for HS-6-6 relationship, the limit of the posterior probability for the true pedigree when no IBD shared is around 0.444 ($\frac{0.8}{1+0.8}$) and the expected posterior probability of the true pedigree is 0.556. Again it is estimated from the histograms in Figure 4.4 where 320 out of 400 simulated data sets having no IBD sharing when the true relationship is HS-6-6. These values are similar with the entries in Table 4.1 when 500K SNPs are used.

The fact that nearly all results lie close to the two limits of the posterior probability of 1 and 0.444 (see right bottom histogram in Figure 4.4) when 500K SNP markers are used suggests that we already have very high information from this number of markers for relationships as distant as HS-6-6. We can distinguish the true relationship from 'unrelated' with near certainty for any particular case where there is IBD sharing between the two relatives. This implies that we should not be able to make the average posterior probabilities in Table 4.1 increase significantly by simply increasing the number of markers. This is controlled by the fact that there is a high probability of no IBD shared between relatives when the relationship is distant. When 500K SNPs are used, it would seem that those average posterior probabilities are already very close to their limits.

Moreover, this is further evidence that for distant relationships we should not concentrate on the average posterior probability. For distant relationships, the average posterior probability of the true pedigree will always look small no matter how many SNP markers we have, but this does not mean that we can make no inference about

them. When two distant relatives do share IBD segments in their genome, there is potential to distinguish relationships even more distant than HS-6-6. Although the average looks quite small, in individual cases where two distant relatives do share IBD, we could have a very high likelihood ratio between the true pedigree and an alternative pedigree and thus be able to make reasonable inference. Simulation results show that we can distinguish the true pedigree for relatives as distant as HS-8-8 from 'unrelated' *provided* they share segments of chromosome IBD (results not shown). Unfortunately more distant relationships are beyond the calculating capability of Merlin. In summary, by taking a more detailed look into the results presented by Skare et al. (2009), we can significantly expand on the conclusions drawn in that paper.

4.3.2 After a certain point more SNPs do not help much

To check my hypothesis about the number of SNPs, I estimated allele frequencies for the Affymetrix SNP6.0 SNP data on the 2934 controls from the WTCCC and simulated genotype data for these 711,020 SNPs (I refer to this as 711K SNPs for simplicity) to repeat the simulation study in the last section. Recall (Chapter 3) that there are about 416,854 autosomal SNPs after cleaning in the previously used 500K SNP dataset. The first three rows in Table 4.3 are extracted from Table 4.1 and the last row is added for the results of 711K SNPs. We can see a clear improvement when the number of SNPs changes from 2200 to 500K SNPs, but there is no clear advantage, on average, in using 711K rather than 500K.

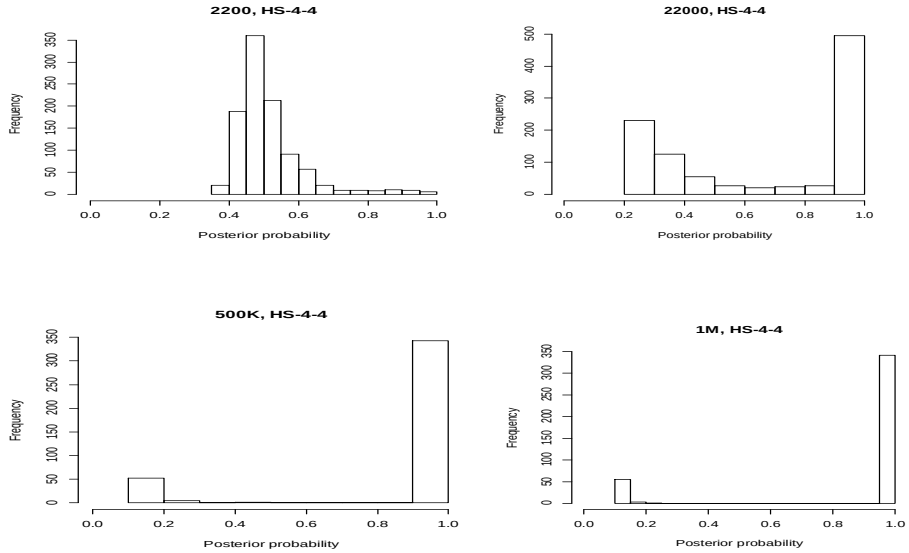
Table 4.3 Simulation results with 711K SNPs when the only alternative relationship is 'unrelated' and comparison with previous results. Averages are taken from 400 replicates.

# of markers	HS-1-1	HS-2-2	HS-3-3	HS-4-4	HS-5-5	HS-6-6
2200	1	0.925	0.605	0.515	0.500	0.500
22000	1	1	0.947	0.685	0.550	0.547
500K	1	1	1	0.878	0.612	0.551
711K	1	1	1	0.872	0.647	0.557

This point can also be demonstrated by looking at the histograms in Figure 4.5. The true pedigree HS-4-4 is used here for illustration. When the number of SNPs changes from 2200 to 500K SNPs, we get better clustering due to the increase in information from the additional SNPs. But there is almost no change at all when the number of

SNPs changes from 500K to 1 million. The reason for this has been discussed above in Section 4.3.1.

Figure 4.5 Histograms of posterior probabilities of the true pedigree when the true relationship is HS-4-4, for different numbers of SNPs.



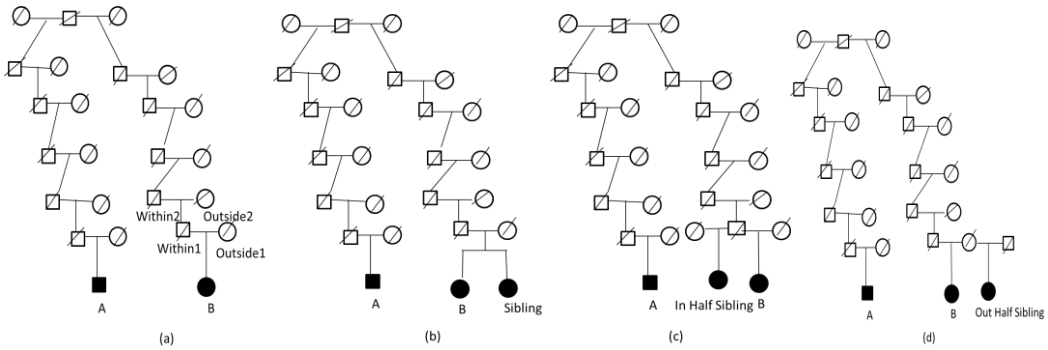
4.4 Distinguishing power when a third individual is genotyped for an unlooped pedigree

Sieberts et al. (2002) have shown that modelling three individuals together could help distinguish the true relationship between two individuals from a disputed relationship. They worked on a special case where the pairwise relationship between the individuals in a trio could only be MZ twins, full sibs, half sibs or 'unrelated'. They showed that there are 18 possible relationships for this trio and proposed a model to calculate the likelihood using an HMM (Hidden Markov Model). Linkage in the markers is incorporated in the calculation. Up to 285 genetic markers are used in their analysis. In one of their examples, the reported relationship of two individuals A and B is half-sibs, but its likelihood is very close to full-sibs. Therefore the two relationships are not distinguishable. They considered a third individual C whose relationship with B is full-sib which is confirmed by the available markers. Then the likelihood of the trio was calculated in their relationship space to find the most likely pedigree and the next-most-likely pedigree. The likelihoods of these two pedigrees were very different and in these two pedigrees the relationship between A and B are different, full-sibs in one

pedigree and half-sibs in another. Therefore the relationship between A and B can be found.

Pedigree-based likelihood approaches can easily incorporate genotype data on additional individuals which is advantageous over other pairwise estimation methods as will be discussed in Chapter 6. In the next simulation study I want to investigate whether one extra genotyped relative could improve the power to distinguish the true relationship between individuals A and B from alternative relationships in more general cases (i.e. not limited to those cases considered by Sieberts et al. (2002), and what the pattern of this improvement is when the extra individual features in different positions on the pedigree. There are many choices for the position of the third individual in the pedigree connecting the two individuals in question. I define some notation for the pedigree position of this extra typed individual and some of these positions are shown in Figure 4.6 (mainly those most practical). The third individual could be a parent, grand-parent, sibling, half-sibling, cousin, aunt/uncle or child of one of the two individuals of interest. By examining the different amounts of information associated with different positions of the third individual we can recommend which relatives would be best to genotype.

Figure 4.6 Denotation for different positions of the third genotyped individual.

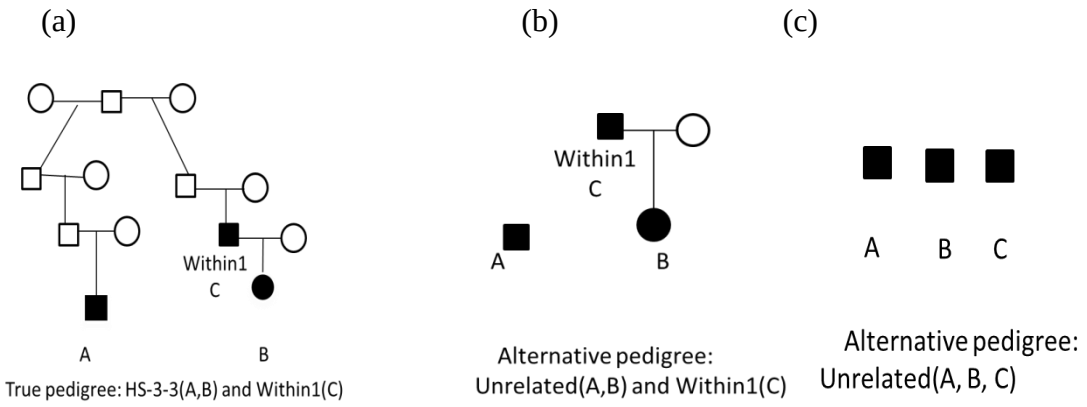


A third genotyped individual C, whose relationship with individual B is known, is said to be on a ‘within’ position if he/she is related to the other individual A. He/she is said to be on an ‘outside’ position if he/she is not related to another individual A. Several examples are as follows. ‘Within1’ denotes that the third genotyped individual is a parent of either of the two individuals of interest and is on the direct line of descent from the common ancestor(s). ‘Outside1’ denotes a parent either of the two individuals

of interest and who is a founder. 'Within2' denotes a grandparent of either of the two individuals of interest and who is on the direct line of descent from the common ancestor(s). 'Outside2' denotes a grandparent of either of the two individuals of interest and who is a founder (part a of Figure 4.6). The third individual is said to be at a 'sibling' position if he/she is a sibling of individual B (part b of Figure 4.6), an 'In Half Sibling' position if he/she is a half sibling of individual B and also related with individual A (part c of Figure 4.6) and an 'Out Half Sibling' position (part d of Figure 4.6) if he/she is a half sibling of individual B and unrelated to individual A.

It is assumed that we know the relationship between the third individual C and B because we are estimating the relationship of A and B based on the likelihood of the three individuals. This is only justifiable when the relationship between C and B is fixed throughout all hypothesized pedigrees and therefore there is only one variable in these hypothesized pedigrees which is the relationship between A and B. For example, when the true relationship between A and B is HS-3-3 and we have an individual C genotyped on 'within1' position as in (a) of Figure 4.7, the alternative pedigree that I use for the alternative hypothesis that A and B are unrelated is as in (b) of Figure 4.7. But it is not the only approach to look at the question. If we do not assume that the relationship between B and C is known, the alternative pedigree could be that all three individuals are unrelated as in (c) of Figure 4.7 .

Figure 4.7 Pedigrees illustrating that the relationship of the third individual with one individual in question should be kept same in alternative pedigrees.



If we do not keep the position of C fixed, then there are three relationships (AB, AC, BC) varying in the hypothesized pedigrees and changes in any of them affect the

likelihood. We cannot make any inference about the relationship of A and B by comparing likelihoods of the hypothesized pedigrees. If we use (c) as the alternative pedigree, when the true pedigree has higher likelihood than the alternative pedigree, all we can say is that the three individuals are not all unrelated, i.e. it is unlikely that all three individuals are unrelated with each other as shown in the alternative pedigree. But we cannot say specifically that individual A and B are unrelated or not. This can be easily illustrated by simulation.

Genotype data were simulated for the pedigree shown in (a) for 220 SNPs, which are evenly picked from Affymetrix 550K SNPs. 220 SNPs are used because the likelihood ratio will be smaller for a small number of SNPs and it is easier to see changes in likelihood. Allele frequencies from the HapMap data were used as before. If we only consider the data on A and B, the likelihood ratio of relationship HS-3-3 over 'unrelated' for individuals A and B is 1.406. We do not have good power to distinguish the true relationship of A and B from 'unrelated'. Next we want to see how a third 'within1' individual C could increase our distinguishing power. If we use the pedigree (c) in Figure 4.7 as the alternative pedigree, the likelihood ratio of the true pedigree and the alternative pedigree is extremely high: 1.135658e+12. But we cannot then conclude that the true relationship of A and B can be distinguished from 'unrelated'. This big difference in likelihoods of the two pedigrees (a) and (c) in Figure 4.7 could be caused by the change of the relationship between B and C (from PC-1 to 'unrelated'), rather than the change of the relationship between A and B (from HS-3-3 to 'unrelated'). If we compare the likelihood of the true pedigree with the pedigree in (b) in Figure 4.7 where A and B are unrelated, the likelihood ratio is just 2.367. In fact, when we use the correct alternative, we do not have such high distinguishing power to tell the true relationship of A and B from 'unrelated'.

For dense SNPs, likelihood ratios can be enormous e.g. of the order 10^{300}. However the basic pattern is the same: using the wrong alternative can lead to exaggerated evidence in favour of the relationship being tested. Consider simulated 500K SNP data for a more distant true relationship HS-4-4. When there is not a third individual, the likelihood ratio for the true relationship versus 'unrelated' is 10^{258}. When there is a third individual C available and its relationship with B is kept unchanged, the

likelihood ratio for the true pedigree and alternative pedigree is 10^{434}. But when there is a third individual available and the alternative pedigree of (c) in Figure 4.7 is used, the likelihood ratio for the true pedigree and alternative is 10^{54348}. Again using the wrong alternative (c) will give misleading results. Large numbers of replicates have been done for both examples and consistent results have been obtained but are not shown here.

When we keep the relationship between B and C as known and fixed in all pedigrees, the only thing that differs between the two alternatives is the relationship of interest between A and B (relationship of A and C will be known as long as that between A and B is known). Hence we can make inferences about the relationship of A and B based on likelihood.

Simulation was carried out with dense markers (500K SNPs) first, then with less dense markers and I started with third individuals who are ancestors of A or B. They are expected to provide higher information because those individuals on 'within' positions and on the direct line of descent from a common ancestor actually make the relationship that we are estimating closer. This is because C is in the middle of the meiosis chain that links A and B in this case and we know the relationship of A and B as long as we know the relationship between A and C. The simulation process is as follows: genotypes of the three individuals under a true pedigree were simulated by Merlin, then the probabilities of the genotype data were calculated both under the true pedigree and another pedigree on which A and B are unrelated and the extra individual C has the same relationship with B as that in the true pedigree. Posterior probabilities were then obtained as before. The true relationships used in this section are all HS-n-n type. Taking HS-3-3 relationship and position within1 as example, the true pedigree and corresponding alternative pedigree are shown in part (a) and part (b) of Figure 4.7. The cases for other true relationships and positions of the third genotyped individual will follow naturally.

The results based on 500K SNPs are shown in Table 4.4. The number of replicates is 400. The first column is for the different positions of the third individual in the pedigree. The first row is for different true relationships between A and B. The second row is for the corresponding values extracted from Table 4.1 when only two

individuals, A and B, are genotyped for the simulation and they are shown here for comparison. Each other value in the table is the average posterior probability of the true pedigree when a third individual C is available and the alternative pedigree is that A and B are unrelated with the relationship between B and C unchanged (Figure 4.7).

Table 4.4 Simulation results based on all 500K markers and a third individual being genotyped. The number of replicates is 400.

True relationship	HS-3-3	HS-4-4	HS-5-5	HS-6-6
Only two genotyped	**1**	**0.878**	**0.612**	**0.551**
Within1	1	0.978	0.770	0.587
Within2	1	0.998	0.861	0.638
Within3	1	1	0.968	0.763
Within4		1	1	0.865
Within5			1	0.961
Within6				0.993
Outside1	0.995	0.863	0.667	0.548
Outside2	0.995	0.891	0.637	0.545
Outside3	1	0.879	0.665	0.555
Outside4		0.849	0.647	0.544
Outside5			0.641	0.558
Outside6				0.558

Average posterior probabilities of the true pedigree are shown for each true relationship versus the single alternative of 'unrelated' and each position of the third individual.

From Table 4.4 two points seem to be suggested. Firstly, if the third individual is in a 'within' position and on the direct line of descent from the common ancestor in the pedigree, it will improve the power considerably (comparing posterior probabilities with those in the first row). The further away it is from the individual in question and hence closer to the common ancestor(s), the more improvement it gives. Secondly, there is no obvious increase in the average posterior probability if the third individual is in an 'outside' position. The first point is quite as expected because we have kept the relationship between C and B as known and C is on the direct line of descent from the common ancestor. Therefore the closer the position of C is to the common ancestor, the closer the relationship between A and C is. But it is unexpected to see that an extra genotyped individual in an 'outside' position does not necessarily help to distinguish

relationships. Especially the individual C on ‘outside1’ position should make some difference due to the parent-child relationship between B and C. For example, when C is homozygous for an allele (say 11) on a locus and B is heterozygous (12) for that same locus and A is homozygous (22) for the other allele, allele 1 of B must be from C, allele 2 of B must be from her father. Then this observation should favour the true pedigree over ‘unrelated’ because A and B are more likely to share the 2 alleles IBD under the true pedigree.

By repeating the simulation with less dense markers (2200 SNPs, 400 replicates), I found that when a third individual in an ‘outside1’ position is genotyped, clear increases in the posterior probabilities of the true pedigrees can be seen. If the true relationship between A and B is HS-3-3, the posterior probability of the true pedigree when a third individual in an ‘outside1’ position genotyped is 0.679 compared with 0.605 when there is no third individual genotyped. This means that when a third individual C at any ‘outside’ position is genotyped, the posterior probabilities can increase. The main reason that no clear increase in distinguishing power was seen when 500K SNPs were used is that likelihoods were converted into posterior probabilities. When likelihood ratios are very high, the increases in the posterior probabilities are less clear than the increases in likelihood ratios. For example, a likelihood ratio 2000:1 corresponds to a posterior probability of 0.9995002 and a likelihood ratio of 4000:1 corresponds to a posterior probability of 0.9997501. In this example the likelihood ratio doubles, but the posterior probability only increases by 0.00025. In cases where likelihood ratios are on a scale as large as 10^{100}, doubling the likelihood ratio yields an increase in the posterior probability of less than 10^{-100}, which is treated as 0 by computer. The second reason that we cannot see the increase of the posterior probability of the true pedigree when the third individual C is on ‘outside’ positions is that C only provides information when A and B share IBD in this case. I have done a calculation to illustrate that when A and B share no IBD, the likelihood ratio between the true pedigree and an alternative pedigree remains unchanged when a third individual on ‘outside1’ is observed (see Appendix 10.3).

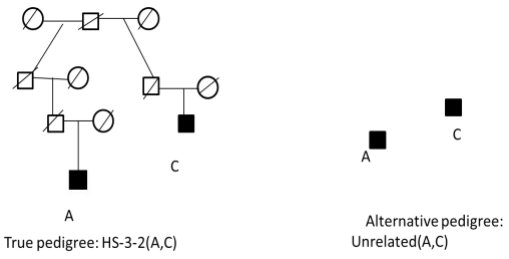
In cases where two individuals do share IBD we have great power to distinguish the two relatives from ‘unrelated’ with 500K dense markers. The likelihood of the true

pedigree is greatly higher than the likelihood of the alternative, so the posterior probabilities of the true pedigree are very close to 1 and are often rounded up to 1 by computer. Therefore we see no increase in the posterior probability of the true pedigree when a third individual is added although there is an increase in the likelihood ratio. The improvement in discriminating power due to a third individual can be seen more clearly when less dense markers are used because in those cases, the posterior probability of the true pedigree is less than 1, so there is scope for increase. But I still present posterior probabilities rather than likelihood ratios because I want to consider average performance over a large number of replicates. It is unreasonable to take averages of likelihood ratios because likelihood ratios could be on different scales for different replicates. But posterior probabilities are all on the same scale between 0 to1.

When the third individual is an offspring of one of the two individuals of question, it would seem intuitive that it will not increase our distinguishing power for the relationship of the two individuals of question. I find this holds when only unlinked markers are used, but is not the case when linked markers are used.

Let us consider the same pedigree in Figure 4.7, for a different problem. Suppose we want to estimate the relationship of A and C (HS-3-2 as the true relationship) and I want to see whether the third individual B as a child of C will increase our distinguishing power. When all three individuals are genotyped, the true pedigree and the alternative pedigree are as in Figure 4.7 (b). When only A and C are genotyped (B is not genotyped), the true pedigree and the alternative pedigree are as in Figure 4.8.

Figure 4.8 Pedigrees of HS-3-2 and 'unrelated'.



When only unlinked markers are used, the likelihood ratio between the true pedigree and the alternative can be easily calculated as it can be done for every marker separately. It is illustrated with one marker.

$$\mathrm{LR} = \frac{\mathrm{L(A,C,B|True)}}{\mathrm{L(A,C,B|Unrelated)}} = \frac{\mathrm{L(A,C|True)L(B|C)}}{\mathrm{L(A)L(B)L(B|C)}} = \frac{\mathrm{L(A,C|True)}}{\mathrm{L(A)L(B)}},$$

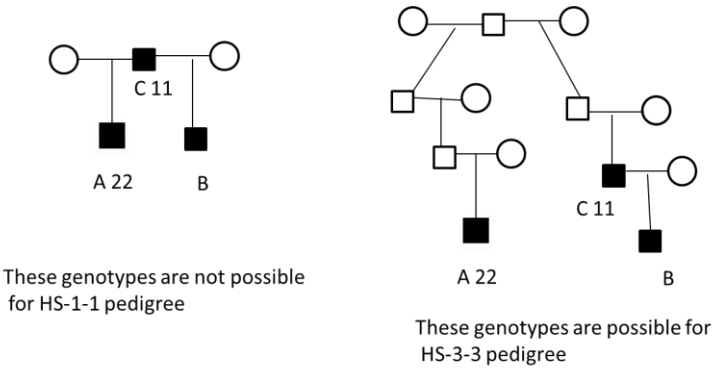
which is the same as the likelihood ratio when only A and C are observed. This means when markers are unlinked, the estimation of the relationship between A and C is independent of the genotypes of the child B. This explains the simulation results.

But if we use linked SNPs, the average posterior probability of the true pedigree based on 2200 SNPs and 400 replicates when all three individuals are observed is 0.8172621. The average posterior probabilities of the true pedigree based on the same SNPs and 400 replicates when the child B is not observed is 0.774, which is considerably less than 0.817. Variance of the posterior probability has been considered and the increase of the posterior probabilities is significant. Consistent results are observed when trying to distinguish the true pedigree from several close alternative pedigrees. This means that the genotype data of a third individual in an 'offspring' position does increase our information on the relationship of two individuals when linked markers are used. The reason could be that what happens at one marker is not independent of what happens at another marker when markers are linked and what segregates to the child in one marker can be informative for what segregates to another.

Next I show that a third genotyped individual helps to distinguish the true relationship from several alternative hypotheses as well as for just one alternative hypothesis. I take the positions of the third individual to be 'within1' and 'within2' for illustration.

When the third individual is in a 'within1' position, if the alternative relationships include HS-1-1, we could get bizarre results because the data generated from other distant relationships could be impossible for the HS-1-1 relationship. For example, genotype '1 1' at a locus of the third individual C who is a 'within1' parent of B, and genotype '2 2' at the same locus of the other individual in question, A, are not possible for a HS-1-1 pedigree because C is a common ancestor and will be a parent of A as well. But these genotypes could be generated by a HS-3-3 pedigree (Figure 4.9). Merlin will remove those markers which have inconsistent genotypes and the likelihoods we get will be for different numbers of markers and not comparable. So H-1-1 was excluded from the choice of the possible relationships.

Figure 4.9 One example of genotypes which can be generated by pedigree HS-3-3 but are not possible for pedigree HS-1-1.



Resulting posterior probabilities calculated as shown in Equation (4.1) are given in Table 4.5 and Table 4.6.

Table 4.5 Posterior probabilities of the true pedigree and several alternative pedigrees when a third individual is available in a 'within1' position (values in brackets are from Table 4.2 for only two individuals genotyped). 500K SNPs are used and averages are taken from 400 replicates.

True	HS-2	HS-3	HS-4	HS-5	Unrelated
HS-2	**0.987(0.959)**	0.013(0.041)	0/(0)	0/(0)	0/(0)
HS-3	0.010(0.034)	**0.894(0.748)**	0.092(0.189)	0.004(0.028)	0(0)
HS-4	0(0)	0.105(0.173)	**0.627(0.467)**	0.247(0.263)	0.021(0.097)
HS-5	0(0)	0.004(0.023)	0.274(0.275)	**0.473(0.388)**	0.249(0.313)
Unrelated	0(0)	0(0.002)	0.024(0.089)	0.249(0.326)	**0.727(0.583)**

Table 4.6 Posterior probabilities of the true pedigree and several alternative pedigrees when a third individual is available in a 'within2' position (values in brackets are for only two individuals genotyped). 500K SNPs are used and averages are taken from 400 replicates.

True	HS-2	HS-3	HS-4	HS-5	Unrelated
HS-2	**1(0.959)**	0(0.041)	0(0)	0(0)	0(0)
HS-3	0(0.034)	**0.961(0.748)**	0.039(0.189)	0(0.028)	0(0)
HS-4	0(0)	0.035(0.173)	**0.773(0.467)**	0.189(0.263)	0.002(0.097)
HS-5	0(0)	0(0.023)	0.178(0.275)	**0.692(0.388)**	0.131(0.313)
Unrelated	0(0)	0(0.002)	0.002(0.089)	0.132(0.326)	**0.865(0.583)**

From Table 4.5 and Table 4.6 we can see that when a third individual is genotyped, the true pedigrees have higher posterior probabilities than when only two individuals are genotyped, which means we can distinguish the true relationships from the alternative relationships with higher certainty. Note that we are only concerned with the values down the diagonals as they are the posterior probabilities of the true pedigrees.

For completeness, I also considered the effect of a third individual for the S-n-n relationships (results not shown). The findings are consistent with those for HS-n-n type relationships with respect to the patterns of influence on the distinguishing power associated with different positions of the third genotyped individual on the pedigree.

Finally, the effect of a third individual when its position is more general was investigated. I compared all positions of the third individuals which are most practical to be genotyped and then give a recommendation as to which relatives are potentially the most informative. In this simulation, the true relationship of A and B is HS-4-4 as shown in Figure 4.6 and the only alternative relationship is 'unrelated'. The positions considered for the third individual include 'within1', 'sibling', 'In Half Sibling', 'Out Half Sibling' (see Figure 4.6), 'In Avuncular' and 'In Cousin' (see Figure 4.10). 'In Avuncular' is the individual who is an uncle or aunt of B and is on the direct line of descent from the common ancestor. Likewise, 'In Cousin' is a cousin of B who is on a direct line descendent of the common ancestor. Results are shown in Table 4.7 and results for only two individuals are shown as well for comparison.

Figure 4.10 Positions of 'In Avuncular' and 'In Cousin' of the third individual on a HS-4-4 pedigree.

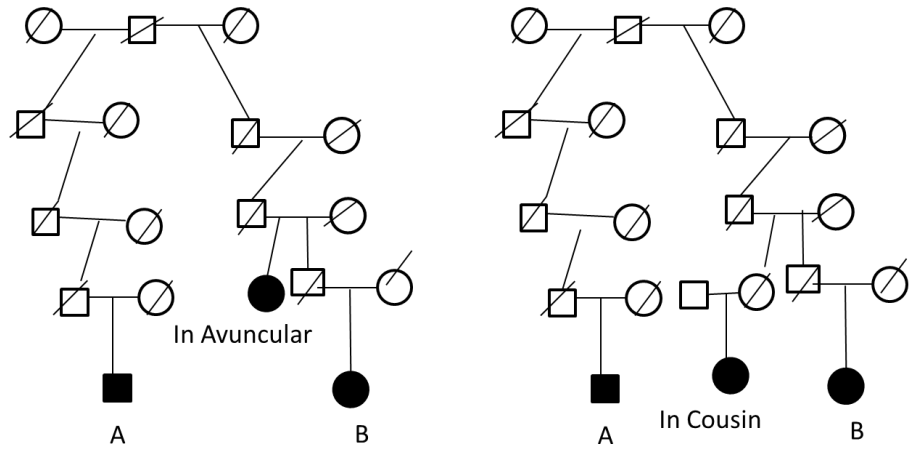


Table 4.7 The average posterior probabilities of the true pedigree when trying to distinguish the true relationship HS-4-4 from 'unrelated' based on 500K SNPs with a third individual in different positions. Averages are taken from 400 replicates.

Position of the third individual	No third individual	Within1	Sibling	In Half Sibling	Out Half Sibling	In Avunc-ular	In Cousin
Posterior probability of the true pedigree	**0.878**	0.978	0.941	0.942	0.874	0.978	0.970

It can be seen from Table 4.7 that the third individual in all the positions except 'Outside Half Sibling' increases the posterior probabilities of the true pedigrees. As noted previously, this is because any effect of the 'outside' individuals is unnoticed when very dense markers are used. These results show that 'sibling' and 'In Half Sibling' provide similar information and 'In Avuncular' and 'within1' provide similar information, which is higher than the information provided by 'sibling'.

My conclusion is that one extra genotyped individual always gives extra information in distinguishing the relationship of two individuals when linked markers are used. When dense SNPs are used, this increase of information relies on whether and how much the third individual C, whose relationship with B is known, shares IBD with A at loci besides those at which B and A already shares IBD. 'Outside' positions do not share

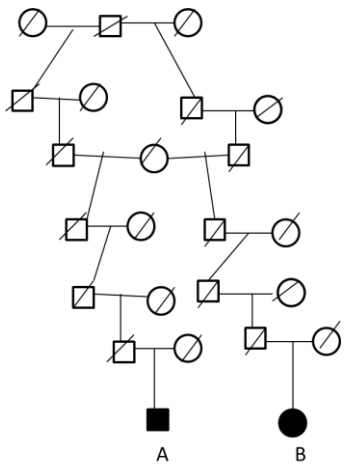
IBD with A at all and they increase the likelihood ratio only when A and B share IBD, in which case the likelihood will be already so high that the posterior of the true pedigree is 1 without the third individual. So there is no increase in the posterior probability. But the third individuals on 'within' positions could increase the posterior probability even when individuals A and B do not share IBD because they themselves could share IBD with A at extra loci. With the presence of the third individual at a 'within' position, the true pedigree will have posterior probability of 1 in more replicates than when there is no third individual. That is why we can see a big increase in the average posterior probability. In other words, the probability of detecting a relative pair, as estimated by the proportion of simulations in which the relationship is detected, is higher when there is a third individual in 'within' position. Some relative pairs who are otherwise not detectable (because there is no IBD shared between them) can be detected now. If the third individual is on a 'within' position, the closer it is to the common ancestor(s) of the two individuals of interest, the more IBD sharing it can have with A at additional loci beside those at which A and B already share IBD and the higher the increase in average posterior probability we can observe. For linked dense SNP data, 'outside' positions are not useful as they only provide information in the case where we already have plenty of information. The increase of information provided by the third individual in these cases is not very noticeable. For third individuals who are an ancestor of B, like parents or grandparents, the closer they are to the other individual A, the more information they provide if they are on a 'within' position. But at the same time, it is more likely that this third individual is actually on an 'outside' rather than 'within' position, because the more distant it is to B, the more ancestors B has on that generation who are symmetric on the pedigree and only one of them is on a 'within' position. As generally 'outside' positions are not very helpful as discussed previously, 'siblings' are safest to use as they will always help. If we can genotype both parents of A or B, or all four grandparents of A or B and try them all, then they would be better choice as they can provide more information.

4.5 Simulation with looped pedigrees

Realistically, human pedigrees generally will not have the simple unlooped structures considered so far. It is of interest to see what happens to our power to distinguish a true

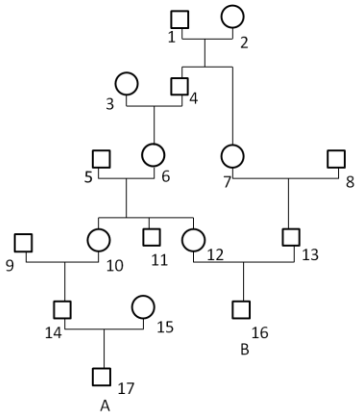
pedigree from an alternative if there are loops in the true pedigrees. Since a relationship between two individuals is defined when there is a path connecting them on a pedigree, looped pedigrees potentially represent multiple relationships between the two individuals. For example, on the pedigree shown in Figure 4.11, we can say that A and B have two relationships: HS-4-4 and HS-6-6. I carried out a simulation to see how well we can distinguish this special relationship (with a loop in the pedigree) from 'unrelated'. The posterior probability of the true pedigree is 0.901 when only the two individuals in question are genotyped and the only alternative relationship is that they are unrelated. This value is higher than that of 0.878 obtained when a true pedigree of HS-4-4 was assumed (Table 4.1). This is because the IBD sharing between A and B on this looped pedigree is higher than that of the normal HS-4-4 pedigree and our distinguishing power is based on the proportion of IBD sharing between two individuals. Recall that when calculating the theoretical IBD (pedigree-based) sharing between two individuals, different paths on the pedigree need to be summed over (Section 2.2).

Figure 4.11 Pedigree of HS-6-6 with a more recent HS-4-4 relationship nested within.



Another looped pedigree that I used as test structure shown in Figure 4.12 (Thompson, 1986), depicts A and B with a recent S-3-2 relationship via the common ancestors 5 and 6 nested within as S-5-3 relationship via 1 and 2. The posterior probability of the true pedigree as shown in Figure 4.12 is higher than that when the true pedigree is assumed to be S-3-2 when the only alternative relationship is 'unrelated'.

Figure 4.12 Pedigree taken from p17 of the book "Pedigree Analysis in Human Genetics" by Thompson (1986).



In this second pedigree (Figure 4.12), a third genotyped individual was assumed in different positions and posterior probabilities of the true pedigree versus the alternative pedigree, that A and B are unrelated, were calculated. The results are based on 1000 simulations with 220 SNP markers are shown in Table 4.8. This choice of SNP set is because we can distinguish the true relationship from unrelated with almost 100% certainty if denser markers are used, and so any increase of power provided by the third individual is not detectable. The results show similar effects of the third genotyped individual as seen for an unlooped pedigree. For example, 4, 6, 7 and 8 are positions for the third individual in the S-5-3 relationship between A and B via 1 and 2. Individual 8 provides very little information as it is an 'outside' position. Individual 4, 6 and 7 provide much more information as they are in 'within4', 'within3' and 'within2' positions respectively. As expected, individuals 4 and 6 provide more information than does 7. 10, 12, 13 and 14 are positions for third individual in the S-3-2 relationship between A and B via 5 and 6. Individual 13 is in an 'outside' position. Individual 10, 12 and 14 provide much more information as they are in 'within2', 'within1' and 'within1' positions respectively. Again, individual 10 provides more information than do individuals 12 and 14.

Table 4.8 The posterior probabilities of the true pedigree with a third individual genotyped for the pedigree in Figure 4.12 and the only alternative being that A and B are unrelated.

Extra's position	Only A and B genotyped	4	6	7	8	10	12	13	14
Posterior probabilities	0.609	0.780	0.795	0.693	0.635	0.957	0.754	0.640	0.773

4.6 Summary

In this chapter, I have presented a pedigree likelihood approach for estimating the relationship between two individuals and extended previous work using this method. Firstly the individual simulation results were examined more closely rather than just looking at the average results over a large number of simulations. By this we had a better understanding how this method performs. Together with the results from the simulation study with 1 million SNPs, it was shown that the distance of the relatives that we can detect from 'unrelated' is limited by the possibility of no IBD sharing between distant relatives rather than the number of markers. This is consistent with the theory of Donnelly (1983). Secondly it was shown that much denser SNPs data are not very helpful after a point. Therefore there is no need to keep increasing the number of SNPs used. We will return to this point when discussing the existence of LD within dense SNPs in the next chapter. Thirdly, how extra genotyped individuals help the relationship estimation of two individuals was investigated. One of the key advantages of this method over other approaches is that genotype data on additional individuals can easily be incorporated into the pedigree likelihood and thus considered jointly. Simulation results show that the best extra individual to genotype to help relationship estimation is the sibling of one of the two individuals in question.

5 Testing the likelihood approach on real data and accounting for LD

In previous studies with the pedigree likelihood method, some assumptions were made. For example, founders of a pedigree were assumed to be unrelated and different loci in the founder population were assumed to be in linkage equilibrium (genotype frequencies at one locus are independent of genotype frequencies at another locus). But the assumption of linkage equilibrium is not justifiable in general and especially when dense markers are used. It is desirable to explore the effect of this assumption.

5.1 Investigate the effect of ignoring LD with simulated data

As Merlin has the option of incorporating LD into genotype simulation and likelihood calculation, it is easy to take LD into account via this option, although it is not the only way to model LD. Merlin models LD by combining tightly linked markers into clusters and estimating haplotype frequencies within each cluster (Abecasis and Wigginton, 2005). The algorithm assumes that the markers can be grouped into non-overlapping clusters of consecutive markers such that (1) markers within a cluster can be in LD, (2) markers in different clusters have a very low level of LD, (3) the recombination rate within a cluster is extremely low. The LD within each cluster is described by the relevant haplotype frequencies. To make the method computationally tractable, two approximations are made: LD between clusters is ignored and the recombination rate within each cluster is assumed to be zero.

Users can provide Merlin a ‘clusters’ file describing a series of clusters within each of which there are several contiguous markers. Each cluster is described by a line which begins with the word ‘CLUSTER’ followed by a series of marker names. This line is followed by several lines which begin with the word ‘HAPLO’. Each of these lines specifies a haplotype and its frequency.

The first few lines of a ‘clusters’ file for three markers could be written as follows:

CLUSTER rs556920 rs553456 rs7989455

HAPLO 0.2500 3 2 1

HAPLO 0.3167 3 2 3

HAPLO 0.2000 1 4 1

HAPLO 0.2333 1 4 3

In this cluster, the first and second markers, rs556920 and rs553456, are in complete LD as allele 3 at the first marker always appears together with allele 2 at the second marker. Likewise allele 1 at the first marker is paired with allele 4 at the second marker.

If users provide their own 'Clusters', Merlin will check: firstly that the markers are contiguous within each cluster; secondly that the map position of the markers within each cluster are the same. Otherwise it will change those positions to be the same to ensure there is no recombination within each cluster.

If this 'clusters' file is not provided, Merlin can generate the file itself based on the input files of genotypes. There are different criteria for the formation of the clusters. Users can use the –distance option to specify a map distance. Then markers with a pairwise distance less than this value are included in one cluster. Or users can call the –rsq (standing for r^2 which is used to measure LD) option to specify a r^2 value and any two markers with pairwise r^2 greater than this value are included in one cluster. For both criteria, only contiguous markers will be included into clusters.

Then the haplotype frequencies within each cluster are estimated from available genotype data. This 'clusters' file can be saved for future use. Estimating maximum likelihood haplotype frequencies in pedigrees is complex and an E-M algorithm was proposed by Abecasis and Wigginton (2005). The observed genotypes of founders are used to estimate the haplotype frequencies. As each individual has two haplotypes, there are $2n$ haplotypes if the number of founders is n. However, the haplotypes of the founders are often not observed. In the E step, conditional on the starting estimate of the haplotype frequencies, the expected count of each haplotype is calculated by summing over all configurations of the founder haplotypes that are compatible with the

observed genotype data. The count of that haplotype in each configuration is weighted by the probability of that configuration, which is simply the product of the haplotype frequencies of the $2n$ haplotypes in that configuration. In the M step, the haplotype frequency estimates are updated by dividing the expected haplotype count of each haplotype by the number of the founder haplotypes in the sample. Then this process of iteration continues until convergence is reached.

In calculating the pedigree probability of the genotype data with LD, the Lander-Green algorithm is adapted: rather than iteration over markers, iteration over clusters of markers in LD is required (refer to Section 2.5.3). For each inheritance vector, the conditional probability of observed genotypes for all markers within the cluster is calculated based on estimated haplotype frequencies.

When Merlin simulates the data with the LD function chosen, the genotypes of the markers within a cluster will be simulated according to their haplotype frequencies, rather than their allele frequencies. Cluster haplotypes will be assigned to markers within a cluster for the founders and transmitted to descendants together as a unit because Merlin assumes no recombination between markers within a cluster. For markers outside the clusters, the process of simulation and likelihood calculation will be the same as when LD is not modelled.

In my first simulation study with LD, the real genotypes of 1285 unrelated individuals from the NBS sample (control group in WTCCC project selected from National Blood Service of UK) were used to estimate the haplotype frequencies. The total number of SNPs is 711,020. The threshold for combining SNPs into one cluster is 0.001cM, which means that the markers with pairwise genetic distance less than 0.001cM will be included in one cluster. This threshold of 0.001cM was chosen for practical reasons although it can be expected that with a greater threshold, LD will be better accounted for as the LD blocks will contain more SNPs. However, larger thresholds drastically increase the running time for the simulation process which makes large numbers of simulations, as are required here, impractical. These experiments are computationally expensive. The higher the threshold is set, the more time it takes to build the LD blocks. When 0.001cM is the threshold, the largest LD block contains up to 200 SNPs, which makes the computation very slow and it takes more than two days to generate

the haplotype frequencies file for Merlin. After the LD block file is ready, the genotype simulation and likelihood calculation for one pedigree could takes around 20 minutes when the whole set of 500K SNPs is used. For example, it took more than one week to generate the results for Table 5.1. Here I consider any experiment which takes much longer than a week as unpractical, but it does not mean that they cannot be done. By this simulation, the effect on relationship estimation of ignoring LD in the data is investigated. The results when several close alternative pedigrees are considered are shown in Table 5.1.

Table 5.1 The posterior probabilities of all alternative relationships when the true relationships are HS-4-4 and 'unrelated' respectively under different situations of LD. 'No LD simulated' means there is no LD in the simulated genotypes. 'LD simulated, not accounted for' means there is LD in the simulated genotypes, but the LD is not accounted for in the likelihood calculations when the relationships are estimated. 'LD simulated, accounted for' means there is LD in the simulated genotypes and it is accounted for correctly when the relationships are estimated.

True relation(below)	HS-1-1	HS-2-2	HS-3-3	HS-4-4	HS-5-5	Unrelated
HS-4-4 (No LD simulated)	0	0	0.173	0.467	0.263	0.097
HS-4-4 (LD simulated,not accounted for)	0	0.001	0.608	0.339	0.051	0.001
HS-4-4 (LD is simulated and accounted for)	0	0	0.164	0.479	0.279	0.077
Unrelated (No LD is simulated)	0	0	0.001	0.085	0.323	0.590
Unrelated (LD is simulated, not accounted for)	0	0	0.256	0.549	0.181	0.014
Unrelated (LD is simulated and accounted for)	0	0	0.001	0.085	0.324	0.589

From Table 5.1 we can see that when the true relationship is HS-4-4 and LD in the simulated data is accounted for perfectly, our inference will be the same as when there is no LD in the data. But when LD is not accounted for, HS-3-3 will have the highest average posterior probability among all alternative relationships considered. When the true relationship is 'unrelated' and the LD in the data is not accounted for, HS-4-4 will have the highest average posterior probability. This means that when LD is present and ignored, relationships will look closer than the true reltionships and, in particular, unrelated individuals will look related. Here we know the true LD model and can hence adjust appropriately. But in reality it is not that easy to model LD.

5.2 Estimating the degree of relationship using the pedigree likelihood method and showing the effect of ignoring LD with real data

The pedigree likelihood approach has only been used to distinguish the true relationship from some alternative relationships previously. This requires a sensible set of pedigrees to be selected, usually based on prior information. Here I propose to use pedigree likelihood to infer the degree of relationship or relatedness. In many situations, the precise relationship as defined by a pedigree may not be required, or relationship-specific alternatives may not be available. But it may be of interest to estimate the relationship of relatives up to a certain degree.

We know that the number of pedigrees connecting two individuals is infinite. Therefore two assumptions are made. The first assumption is that two individuals in question are outbred and under this assumption, we know that all possible relationships can be classified into three types, $S-n_1-n_2$, $HS-n_1+n_2$ and $PC-n$ (Section 4.1). The second assumption is that all relationships with more than n separating meioses are regarded as 'unrelated'. This means that we only consider relationships up to this level and not be able to estimate more distant relationships. With these two assumptions, the number of possible relationships for any pair of individuals is finite. Then the relationship which has the highest likelihood is the estimated relationship. Essentially we find the most likely relationship for the two individuals under the two assumptions. Since it is known that all $S-n_1-n_2$ relationships with a given $n=n_1+n_2$ have the same likelihood for all values of n_1 and n_2 which holds for $HS-n_1-n_2$ type of relationships as well, we can reduce the number of likelihoods that need to be calculated further. But at the same time, it means that we cannot estimate the precise relationship and can only estimate the degree of relationship. Therefore, when only considering relationships which are not more distant than $S-8-8$, we just need to calculate the likelihood of the pedigrees S-1-1, S-1-2, S-2-2, S-2-3, …, S-8-8, HS-1-1, HS-1-2, HS-2-2, …HS-7-8, PC-1, PC-2, …, PC-15 and 'unrelated'.

To reduce the number of likelihood calculations even further and speed up the process, when the requirement for accuracy is not very high, a small number of simple unlooped

pedigree structures, such as S-1-1, S-2-2, …, S-5-5, S-6-6 and 'unrelated' can be used as a rough template against which to test a given relative pair (so this method can also be phrased as 'Template' method). If any of the S – n – n structures has the highest likelihood the pair of individuals is regarded as related. Otherwise they are regarded as unrelated. This method does not give us the exact relationship between the two individuals, but could give us an estimate of the degree of relatedness in terms of the number of meioses separating them. When applying this method to estimate the degree of relationship in practice, it is not necessary to convert likelihoods into posterior probabilities. But in order to summarize the results of a large number of pairs of relatives when investigating the impact of LD on relationship and relatedness estimation, posterior probabilities are calculated assuming a flat distribution of the template pedigrees.

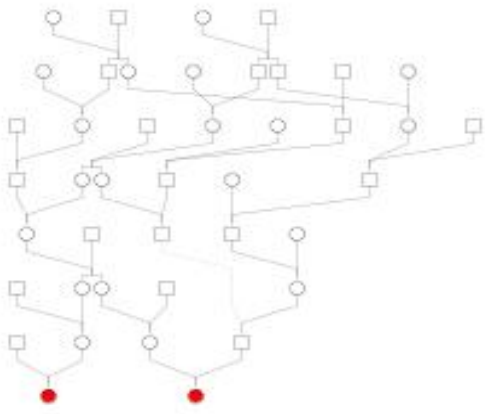
It needs to be noted that this is a method of maximum likelihood estimation of pairwise relatedness using pedigree as a parameter and it is different to the approach of using pedigree likelihood to distinguish the true relationship from a set of alternative relationships although they look similar. When distinguishing the true relationship from alternative relationships, it is assumed that the true relationship is included in the relationship set considered. But the true pedigrees could be complicated and there is no guarantee that the true relationship will be included in the set of alternatives. Hence, a set of alternative relationships with simple structures ranging from S-1-1 to S-6-6 etc. and 'unrelated' can be used as a template to estimate the general degree of relationship between the two individuals rather than the exact relationship itself. Later in this thesis, we can see that this method still works when the 'outbred' assumption is breached and it performs better than existing methods on estimating degree of relationship accurately.

The real data that I used are from the MICROS study (see description in Chapter 3) . Allele frequencies are estimated from all the individuals. Genotype data on 303,783 SNPs with a linkage map are used for every individual. We have four pedigree files of the MICROS study covering the same genotyped individuals, but with different numbers of generations in each of them. The full 12-generation pedigree is very big and unwieldy to work with. The 8-generation pedigree is used in this whole thesis. The software Jenti (Falchi and Fuchsberger, 2008) is used to split large pedigrees into

smaller ones to help visualize the pedigree and also to choose pairs of individuals with desired relationships. Jenti was also used to compute the expected kinship coefficient between individuals based on the pedigree data provided. Another possible choice to visualize pedigrees is the R package 'kinship'.

Firstly several pairs of relatives with most recent relationship ranging from S-3-3 to S-5-5 were chosen to get an idea of how pedigree likelihood estimation performs for such data., A larger number of relative pairs (101) whose most recent relationship is S-3-3 were then considered for average performance. Second cousins (S-3-3) could be distinguished from 'unrelated' easily (Section 4.2), but they could not be so easily distinguished from close alternatives, which is what I am looking at now. The first pair of relatives (230 and 1193) is connected via the sub-pedigree in Figure 5.1, and has a true most recent relationship of S-3-3: the kinship coefficient between the two individuals is 0.01833677 based on the large 8-generation pedigree and 0.0177 based on the sub-pedigree while the expected kinship coefficient of an S-3-3 relationship is 0.015625.

Figure 5.1 An inbred pedigree connecting individuals 230 and 1193 (shaded) of the MICROS Study.



The likelihood of the 'true relationship' was compared with several very close alternatives and posterior probabilities of all relationships were calculated. The posterior probabilities of all the hypothesized relationships are shown in Table 5.2.

Table 5.2 Posterior probabilities of several close alternative relationships for individuals 230 and 1193 with the true relationship S-3-3.

Relationships	S-3-3	HS-2-2	HS-3-2	S-4-3	HS-3-3	Unrelated
Posterior probabilities	0.9946751	9.662598e-35	3.710342e-30	0.005324923	1.825236e-28	0

There are two surprising phenomena. Firstly, the posterior probability of the true pedigree is very high even though the alternative pedigrees are very close to it; but we do not usually have such high distinguishing ability for simulated S-3-3 pedigree data even when there is no LD. Secondly, the S-3-3 pedigree can easily be distinguished from the alternative HS-3-2, even though these two pedigrees have the same expected kinship coefficient. In simulated data, these two pedigrees give very close posterior probabilities. I found this is partly due to the fact that I did not include any closer alternative relationship, and partly due to that all $\mathrm{HS}-\mathrm{n}_1-\mathrm{n}_2$ type of hypothesized relationships tends to have extremely lower posterior probabilities compared to $\mathrm{S}-\mathrm{n}_1-\mathrm{n}_2$ type of relationships when the true pedigree is inbred. This point has been verified with simulated inbred pedigree data (results not shown). The reason for this phenomenon could be that the IBD sharing between two individuals with a complex relationship is more consistent with two common ancestors and we know that the variances of the realized sharing for extending siblings and extending half-siblings are different even though the expected value may be the same (Hill and Weir, 2011). This could imply that when we use the 'Template' method to estimate relationships in real data where inbreeding is likely, only extended siblings relationships should be used as alternative relationships. For the next four pairs of relatives, closer alternative relationships are added, and in each case it is the closest alternative relationship that has the highest likelihood.

We now consider the 101 pairs of relatives whose most recent relationship is S-3-3. The results show that in all 101 relative pairs, it is S-1-1 that has the highest likelihood. This means that the estimated degree of relationship is always closer than the true relationship which is consistent with the findings from simulated data.

This could be due to two reasons: the true relationship of the two individuals is actually closer than the most recent relationship because of inbreeding and the presence of LD in dense SNP markers makes individuals look more closely related than they really are. The second reason is more likely as the true relationships must be more distant than S-1-1 however much inbreeding is present in these relatives.

5.3 Is there a way to solve the problem of LD

It has been shown that the pedigree-based likelihood method is biased in the presence of LD. Here in this section, I consider how to deal with it. One simple way to remove the effect of LD, which is commonly done in practice, is to reduce the density of the SNP markers to be used (Berkovic et al., 2008, Pemberton et al., 2010). Berkovic et al. (2008) and Pemberton et al. (2010) simply used 10,000 markers without giving reason for this number. Kling et al. (2012) did relationship estimation for different levels of true relationship using different numbers of SNPs. There are large numbers of relative pairs for each true relationship. A pedigree likelihood approach is used in their work as well. They consider the number of SNPs giving the highest proportion of correct estimates. It is shown that this number is different for different true relationships. They suggested using no more than 20,000 markers to obtain reliable result. I will test whether thinning works with both simulated data and real data bearing in mind that it discards a lot of the available information. Then I will attempt to model LD in real data.

Merlin was used throughout to model LD. As described before, there are two criteria to build LD blocks, distance between markers and pairwise r^2. The Plink Software was also looked at, but it is very similar to Merlin in the way it models LD and only uses r^2 as the criterion to build LD blocks. The real data that were used are the SNP data from the MICROS dataset for 101 selected pairs of individuals whose most recent relationship is S-3-3. 303,783 autosomal SNPs are available for every individual. It should be noted that the true relatedness of these relatives is varied due to inbreeding although their most recent relationships are same. Based on the 8-generation pedigree, the expected kinship coefficient for some pairs is closer to that of an S-3-2 relationship. Hence we do not expect to get an estimate of S-3-3 for every pair of relatives.

5.3.1 Is thinning SNPs a valid method to get rid of LD?

I began by considering whether the effect of LD can be removed by simply thinning the SNPs. So SNP data with LD were simulated for different true relationships and then different subsets of SNPs were selected for the relationship estimation. If N is the total number of available SNPs, a thinned subset of size N/n is obtained by taking every nth SNP in sequence.

The first simulation I did is an extension of that summarized in Table 5.1 where 711,020 SNP data with LD were simulated for the true relationships of HS-4-4 and 'unrelated' respectively and the posterior probabilities of alternative pedigrees HS-1-1,…HS-5-5 and 'unrelated' were calculated. In that simulation, LD was modelled by setting that only SNPs with pairwise distances less than 0.001cM are in LD. I have shown that LD will make the estimated relationship closer than the true relationships if it is not accounted for, but will not cause a problem if it is accounted for correctly. Now for the same simulated data, posterior probabilities of the alternative pedigrees were calculated with thinned SNPs. The results are shown below in Table 5.3 along with those from Table 5.1 for comparison.

Table 5.3 Averaged posterior probabilities over 400 replicates of several alternative pedigrees when 711,020 SNPs are simulated with and without LD, but only thinned SNPs are selected (italic).

True relation(below)	HS-1-1	HS-2-2	HS-3-3	HS-4-4	HS-5-5	Unrelated
HS-4-4 (No LD simulated)	0	0	0.173	0.467	0.263	0.097
HS-4-4 (LD simulated, not accounted for)	0	0.001	0.608	0.339	0.051	0.001
HS-4-4 (LD is simulated and accounted for)	0	0	0.164	0.479	0.279	0.077
HS-4-4 (only use 20K SNPs)	*0*	*0.001*	*0.215*	*0.367*	*0.246*	*0.171*
HS-4-4 (only use 30K SNPs)	*0*	*0.0003*	*0.218*	*0.406*	*0.240*	*0.136*
HS-4-4 (only use 50K SNPs)	*0*	*0.0014*	*0.207*	*0.435*	*0.242*	*0.154*
Unrelated (No LD is simulated)	0	0	0.001	0.085	0.323	0.590
Unrelated (LD is simulated, not accounted for)	0	0	0.256	0.549	0.181	0.014
Unrelated (LD is simulated and accounted for)	0	0	0.001	0.085	0.324	0.589
Unrelated(only 20K SNPs are used)	*0*	*0*	*0.021*	*0.197*	*0.352*	*0.429*
Unrelated(only 30K SNPs are used)	*0*	*0*	*0.013*	*0.168*	*0.353*	*0.466*
Unrelated(only 50K SNPs are used)	*0*	*0*	*0.006*	*0.140*	*0.352*	*0.502*

It can be seen that when thinned SNPs (20K, 30K or 50K) are used, the estimates are not as biased as when all SNPs are used (row 3 of Table 5.3). In all three cases, it is the true pedigree which has the highest average posterior probability. For both true pedigrees of HS-4-4 and 'unrelated', when 50K SNPs are used, the true pedigree has greater posterior probability than when 20K SNPs are used, but the average performance is still not as good as when all SNPs are used and LD modelled appropriately (row 4 of Table 5.3). We cannot find an ideal number of SNPs which gives the best estimate in the presence of LD with this simulation as this number will be dependent on the true relationship (Kling et al., 2012). In this simulation, genotypes with LD are simulated and LD blocks estimated by Merlin, which are unrealistically short (see Table 5.7) and this LD can be removed easily with thinning. This is why the 50K SNP set works well here. The simulation study was repeated with LD blocks built using $r^2 > 0.2$ as a threshold and similar patterns of the results were observed when thinning was applied. It is intuitive that less dense markers reduce the LD but they also provide less information than denser markers. So there is a balance to maintain when

thinning the density of SNPs to reduce the effects of LD as reduced numbers also lead to reduced distinguishing power and flatter posterior probability distributions.

Determining whether SNPs are in LD simply by their pairwise distances is very arbitrary and the LD between SNPs located more distantly apart is ignored. Therefore in the next simulation, I modelled LD from the MICROS dataset using r^2 as the criterion for building LD blocks and estimated the haplotype frequencies from all 1285 genotyped individuals. These samples in this dataset are related which is not ideal, but we do not have enough genotyped founders to do anything else. Estimating haplotype frequency from the sample will induce bias. But this is what everyone does when population frequencies are not available. The criterion of putting SNPs into LD blocks is that their pairwise r^2 is greater than 0.2. The true pedigree is S-3-3. Likelihoods and posterior probabilities for several closer alternative pedigrees were calculated when there was no LD in the simulated data, when there was LD and the LD was not dealt with and when there was LD but a subset of SNPs was used to do the estimation. The results are shown in Table 5.4.

Table 5.4 Averaged posterior probabilities of several alternative pedigrees when 300,000 SNPs are simulated with and without LD. All and thinned SNPs are used in calculation.

True relation(below)	S-1-1	S-2-2	S-3-3	S-4-4	S-5-5	Unrelated
S-3-3 (No LD simulated)	0	0.0002	0.906	0.093	0.0006	0
S-3-3 (LD simulated, not accounted for)	0.235	0.764	0.0005	0	0	0
S-3-3 (only use 20K SNPs)	0	0.001	0.871	0.116	0.011	0.001
S-3-3 (only use 50K SNPs)	0	0.001	0.923	0.074	0.001	0

Again we can see that when thinned SNPs are used, the effect of LD is greatly reduced as we saw in the previous simulation with HS-4-4 as the true pedigree. Thinned SNPs give results which are not very different to the results given by dense SNPs in linkage equilibrium for this level of relatedness. If we only consider two hypothesized pedigrees, the true pedigree and 'unrelated', then the posterior probability of the true pedigree changed from 1 to 0.994 when the number of SNPs changed from 300K to 20K. So thinning SNPs to around 20K does keep most of the distinguishing power while removing LD. But if the true pedigree is more distant like HS-4-4 (Table 5.3),

the distinguishing power is reduced as well while removing the effect of LD when thinned SNPs are used. These results imply that for relationships up to S-3-3, thinning SNP data to around 20K is feasible in that we do not lose too much distinguishing power while bias caused by LD is removed.

5.3.2 Modelling LD in the real MICROS data

From simulation results it seems that thinning SNPs is an acceptable method to deal with LD. But it will waste information that we have on dense SNP data. So in this section I try to see whether LD in dense real SNP data can be modelled well. Here I want to consider all 101 pairs of relatives that were selected from MICROS study whose most recent relationship is S-3-3. By examining the results for a large number of relatives, we can have a better understanding of the performance of my modelling of LD on average.

Firstly I used pairwise distance as the criterion for building LD blocks and then used pairwise r^2. For every pair of relatives, the likelihood was calculated for hypothesized pedigrees S-1-1 to S-6-6 and 'unrelated', and the posterior probability is calculated under the assumption of a flat prior distribution. Results are presented as average posterior probabilities of the different hypothesized relationships for the 101 pairs of relatives.

Firstly let us see how the likelihood method performs when LD is not considered. When all 300K SNP are used in the estimation, the average posterior probability is 1 for S-1-1 and 0 for all other alternative pedigrees. Pedigree S-1-1 as the closest considered pedigree has highest likelihood in all 101 pairs, which is very biased.

When LD was modelled by Merlin with a distance of 0.001cM as the threshold to build LD blocks, the estimation did not improve. Results show that in all 101 pairs of relatives, the posterior probability of the pedigree S-1-1 is 1 in all 101 pairs of relatives. When LD was modelled with r^2 as the blocking criterion, a threshold of 0.6 was first considered and Merlin performed the analysis in a window (or grid) of 5 cM along the chromosome. The results are shown in Table 5.5.

Table 5.5 Averaged posterior probabilities of several hypothesized pedigrees and the number of times that each hypothesized pedigree has the highest likelihood, over 101 pairs of relatives whose most recent relationship is S-3-3 when LD is modelled and the threshold of r^2 is set at 0.6.

Hypothesized pedigree	S-1-1	S-2-2	S-3-3	S-4-4	S-5-5	S-6-6	Unrelated
Average posterior probability	0.620	0.380	0	0	0	0	0
Number of times giving the highest likelihood	63	38					

It can be seen that in some pairs of relatives the estimated relationships have changed from S-1-1 to S-2-2, but these estimated relationships are still closer than the suspected 'real' relationship.

Setting r^2 at 0.2 and a grid of 10 cM allows SNPs in weaker LD to be included and larger LD blocks. Hopefully LD will be better accounted for. For every SNP, increasing the grid will increase the number of SNPs to be checked for LD with that SNP. The results for these 101 pairs with real data are as follows in Table 5.6.

Table 5.6 Averaged posterior probabilities of several hypothesized pedigrees and the number of times that each hypothesized pedigree has the highest likelihood, over 101 pairs of relatives whose most recent relationship is S-3-3 when LD is modelled and the threshold of r^2 is set at 0.2.

Hypothesized pedigree	S-1-1	S-2-2	S-3-3	S-4-4	S-5-5	S-6-6	Unrelated
Average posterior probability	0.278	0.613	0.109	0	0	0	0
Number of times giving the highest likelihood	28	63	10				

The estimated relationship is still biased towards closer relationship, but the bias is less than when LD is not modelled. On average it is S-2-2 which has the highest posterior probability rather than S-1-1. These results show that when the grid is increased and the clustering criterion is lowered, the bias caused by LD is greatly reduced. By lowering the criterion of clustering SNPs, the number of SNPs that can be considered for an LD block is increased.

The LD blocks formed with different thresholds are examined and part of the reason why the distance threshold performed so badly compared with the r^2 threshold become evident.

Table 5.7 The number of LD blocks formed, the number of SNPs included in LD blocks in total, the number of SNPs in the largest LD block and the median of the number of the SNPs in all LD blocks when different thresholds are used to model LD in MICROS dataset. The total number of SNPs is 303783.

Threshold for LD block	# of LD blocks	# of SNPs included in LD blocks	# of SNPs in the largest LD block	Median of # of SNPs in all LD blocks
Distance of 0.0001cM	7,601	15,764	7	3
Distance of 0.001 cM	23,078	51,785	192	2
Distance of 0.01 cM	59,616	181,507	587	2
r^2 0.6	41,180	195,526	21	4
r^2 of 0.2	38,671	283,658	21	6
r^2 of 0	14,475	303,780	21	21
Standard D'	69,280	245341	48	3

From Table 5.7, it can be seen that when an r^2 threshold of 0.2 is used, nearly all SNPs are included into LD blocks and the sizes of LD blocks are limited to 21 SNPs by Merlin, which results in a flatter distribution of the number of SNPs in LD blocks. However, when the distance threshold of 0.001cM is used, the proportion of SNPs that are included in LD blocks is very small, which means that most SNPs are not included in LD blocks and the LD between them is therefore not accounted for. With the distance threshold, there is no upper limit on the size of the LD blocks in Merlin, which results in some extremely large LD blocks (the number of these huge blocks is small, only 3 blocks have more than 50 SNPs when 0.001cM is the threshold) although it is stated that Merlin can only handle LD blocks with fewer than 20 SNPs (Abecasis and Wigginton, 2005). 0.001cM threshold and r^2 0.2 are compared because the software's running time for building LD blocks is similar for the two cases. The conclusion is that modelling LD with distance as the threshold may be acceptable in simulations to gain an insight into the effect of LD, but it is not good for real data. It is very arbitrary to set

LD blocks simply based on the distance between markers. Another reason is that SNPs are not evenly spaced on the chromosomes physically. If a distance threshold is used, some very large LD blocks could be formed at a higher threshold resulting in slow running of computers, so a much lower threshold has to be used in practice and then most of the LD in the data will not be accounted for. Another reason that distance threshold performs badly in modelling LD is that distance between markers does not necessarily reflect the level of LD between them. In later sections, I will model LD using r^2 as the criterion for building LD blocks which will include more SNPs, on average, in each block. But due to an artificially created limit of 21 built into Merlin when the r^2 criterion is used, which is very restrictive, neither method is ideal.

Kling et al. (2012) reported that Merlin is unable to handle extended pedigrees such as third cousins or more than 5000 clusters. But my experience shows that Merlin is fine with pedigrees as distant as sixth cousin and as many as 50,000 clusters. The only problem is that its speed will be very low. It could take a couple of days to build LD blocks and estimate frequencies of the haplotypes for all clusters and half hour to calculate the likelihood of a HS-6-6 pedigree with around 300,000 SNPs. The problem encountered by Kling et al. (2012) could be down to hardware capacity.

However, the estimate that we get by modelling LD is still not perfect even the r^2 threshold of 0.2 is used. This could be for several reasons: the LD blocks will certainly not be long enough to cover all SNPs that are in LD; Haplotype frequencies were estimated for these LD blocks from the sample itself; the presence of high inbreeding in the population. But it shows that it is possible for LD to be modelled in this pedigree likelihood method and the estimate can be greatly improved.

Another approach for LD blocks construction (Gabriel et al., 2002) which is implemented in the R package 'LDexplorer' (Taliun et al., 2014) is also considered. Instead of r^2, it uses D' (see Section 2.3) to measure LD between markers. Rather than arbitrarily choosing a threshold as we have done for r^2, they have proposed standard criteria for LD blocks.

1) A pair of SNPs is in strong LD if the lower bound of D' 90% confidence interval is > 0.7 and the upper bound of the D' 90% confidence interval is >0.98.

2) A pair of SNPs shows evidence of recombination if the upper bound of D' 90% confidence interval is <0.9.

3) SNP pairs which do not fall into either 1) or 2) are called non-informative. Informative pairs are those satisfying conditions 1) or 2).

4) Then, a LD block is defined as a region of adjacent SNPs where the outer-most SNPs are in strong LD and 95% of all informative SNP pairs are in strong LD.

With this approach applied to the real MICROS data, there are 69280 clusters formed which put 245341 SNPs into LD blocks. The biggest block has 48 SNPs and the median of the block sizes is 3 (Table 5.7). The same estimation process is carried out and the results seem inferior to what we obtained using Merlin with threshold $r^2 > 0.6$ (Table 5.8).

Table 5.8 Average posterior probability of several hypothesized pedigrees and the number of times that each hypothesized pedigree has the highest likelihood, over 101 pairs of relatives whose most recent relationship is S-3-3 when standard D' threshold is used to form LD blocks.

Hypothesized pedigree	S-1-1	S-2-2	S-3-3	S-4-4	S-5-5	S-6-6	Unrelated
Average posterior probability	0.876	0.124	0	0	0	0	0
Number of times giving the highest likelihood	88	13					

It seems that modelling LD in dense SNPs with strong LD will still give biased results because none of the models will necessarily capture the true LD patterns. Modelling LD with LD blocks might reduce the bias that arises from ignoring LD while maintaining the information provided by dense markers. But it should be used with caution and we need to be vigilant of the possible bias in the estimation. It could be useful to model LD in dense SNPs if the aim is just to find relatives rather than the

exact relationship. However, if unbiased estimates of relationship are required, we need either to reduce the number of markers to an essentially LD-free set or to model LD so that all SNP pairs with even minimal LD are put into LD blocks and there is no association between these blocks. It is the association between LD blocks in the current models that is most likely to be causing a problem.

5.3.3 Estimating relationships in real data with thinned SNPs

The modelling of LD in the last session is very computationally intensive and the results are not very satisfactory. Now let us look at how thinning performs on these 101 pairs of relatives whose most recent relationship is S-3-3. The estimation results are shown in Table 5.9, Table 5.10, Table 5.11, and Table 5.12 respectively for 50K, 30K, 20K and 10K SNPs.

Table 5.9 Average posterior probability of several hypothesized pedigrees and the number of times each hypothesized pedigree has the highest likelihood, over 101 pairs of relatives whose most recent relationship is S-3-3 when 300K SNPs in real data are thinned to 50K SNPs.

Hypothesized pedigree	S-1-1	S-2-2	S-3-3	S-4-4	S-5-5	S-6-6	Unrelated
Average posterior probability	0.064	0.567	0.368	0	0	0	0
Number of times giving the highest likelihood	7	59	35				

Table 5.10 Average posterior probabilities of several hypothesized pedigrees and the number of times that each hypothesized pedigree has the highest likelihood, over 101 pairs of relatives whose most recent relationship is S-3-3 when 300K SNPs in real data are thinned to 30K SNPs.

Hypothesized pedigree	S-1-1	S-2-2	S-3-3	S-4-4	S-5-5	S-6-6	Unrelated
Average posterior probability	0.010	0.349	0.632	0.008	0	0	0
Number of times giving the highest likelihood	1	34	65	1			

Table 5.11 Average posterior probabilities of several hypothesized pedigrees and the number of times that each hypothesized pedigree has the highest likelihood, over 101 pairs of relatives whose most recent relationship is S-3-3 when 300K SNPs in real data are thinned to 20K SNPs.

Hypothesized pedigree	S-1-1	S-2-2	S-3-3	S-4-4	S-5-5	S-6-6	Unrelated
Average posterior probability	0	0.251	0.740	0.009	0	0	0
Number of times giving the highest likelihood		23	78				

Table 5.12 Average posterior probabilities of several hypothesized pedigrees and the number of times that each hypothesized pedigree has the highest likelihood, over 101 pairs of relatives whose most recent relationship is S-3-3 when 300K SNPs in real data are thinned to 10K SNPs.

Hypothesized pedigree	S-1-1	S-2-2	S-3-3	S-4-4	S-5-5	S-6-6	Unrelated
Average posterior probability	0	0.112	0.851	0.030	0.005	0.001	0.001
Proportion of giving the highest likelihood		10	87	3			

We can see that thinning SNPs is an acceptable way to adjust for LD, at least for relationships up to the level of second cousins (S-3-3). As the number of SNPs decreases, the bias towards closer relationship caused by LD reduces while at the same time, the distinguishing power is still high.

It seems that when the number of SNPs is reduced to around 20K, the effect of LD in the data is almost negligible. This is consistent with other findings (Kling et al., 2012). It needs to be noted that we do not expect that the S-3-3 is the only correct estimate because MICROS pedigree is highly inbred and the true relatedness between some S-3-3 relatives could be closer than what is suggested by their most recent relationships (Figure 6.9). S-2-2 can be regarded as correct for some pairs of relatives when the relatedness, rather than the exact relationship, is of interest. When the number of SNPs is 10K, the likelihoods of more distant relationships like S-4-4, S-5-5 start to increase. This is due to the decrease in distinguishing power of the reduced number of SNPs. So

there is a tradeoff between distinguishing power and LD: fewer SNPs reduce bias caused by LD, but the distinguishing power is reduced as well.

Since genome-wide STR data are also available for the MICROS study, the applicability of STRs for relationship estimation was investigated as well. All 1285 individuals were genotyped at 1067 autosomal STRs. Usually STRs are considered unlinked when the pedigree likelihood is calculated. But as it would seem unreasonable to ignore linkage with this many markers, the Lander-Green algorithm was used to calculate the pedigree likelihood with a linkage map, as was done for SNP data. A simulation study (results not shown) showed that one unlinked STR has distinguishing power analogous to that of about12 SNPs (will depend on allele frequency). But the distinguishing power provided by 1067 linked STRs is slightly less than 10,000 SNPs with one STR equal to around 8 SNPs. This is because linked markers have less information than the same number of unlinked markers. The estimation results using 1067 STRs on real data are shown in Table 5.13.

Table 5.13 Average posterior probabilities of several hypothesized pedigrees and the number of times that each hypothesized pedigree has the highest likelihood, over 101 pairs of relatives whose most recent relationship is S-3-3 when 1,067 STRs in real data are used.

Hypothesized pedigree	S-1-1	S-2-2	S-3-3	S-4-4	S-5-5	S-6-6	Unrelated
Average posterior probability	0.0014	0.0684	0.7924	0.0769	0.0254	0.0186	0.0169
Proportion of giving the highest likelihood	1	8	84	7	1		

It can be seen that these results are close to the results obtained from 10,000 SNPs (Table 5.12). This suggests that 1,067 STRs are adequate for estimating relationships up to S-3-3. At the same time, they are less likely to be affected by LD compared with SNPs. Therefore when the targeted relationship is not very distant, they are a good choice for relationship estimation if available. But for very distant relationships, it is expected that they may not provide enough coverage over the shared chromosomal segments due to their small number, in which case the dense SNPs have advantage.

5.4 Adding the genotype of a third individual in real data

Now that LD has been considered, I return to investigate the effect of including a third individual in real data. Only relatives more distant than second cousins are considered here because when dense markers are used, close relationships are easy to estimate and an extra individual is not needed. Examples of real sub-pedigrees were selected from the MICROS dataset. Because the MICROS pedigree is very inbred and outbred pedigrees are preferred for my purpose, the software Jenti was used to search for relatives whose expected kinship coefficient based on the overall 12-generation pedigree is exactly the same as their expected kinship coefficient based on their most recent relationship. Thus for a pair of individuals whose most recent relationship is defined by 10 separating meioses, a (full pedigree) kinship coefficient of 0.0009765625 (i.e. 0.5^{10}) would give us confidence that there is no hidden inbreeding. They also need to have some relatives genotyped for us to see the effect of including a third individual in estimating their relationships. Relatives of S-5-5 were sought, but other extended sibling relatives of the same degree like S-6-4 were also accepted. If we use a new notation, S-n, to denote the extended sibling relationships separated by n meioses, they could all be denoted as S-10. For S-5-5 relatives, individuals 1173 and 79, there are additional genotyped relatives at 'sibling', 'within1' and 'outside1' positions. For another pair of S-5-5 relatives, individuals 1021 and 2, there are typed relatives at 'sibling', 'within1' and 'In-avuncular' positions. For 2 pairs of S-4-6 relatives, individuals 183 and 1048, 183 and 1076, there are potential third individuals at 'sibling', 'within1', 'In-avuncular' and 'outside1' positions. MICROS samples were genotyped for 303,783 autosomal SNPs. We know there is LD in dense real SNP data and it can make the estimated relationship much closer than the true relationship in pedigree likelihood methods. Therefore the density of the SNPs was decreased here to alleviate the effect of LD. As seen above, 20K, 30K and 50K SNPs, evenly picked out from the 303,783 SNPs, are all acceptable subsets. As the true relationships in question are distant, the 50K SNPs set was preferred to have better power for distinguishing the hypothesized relationship from the single alternative of 'unrelated'.

In earlier simulation studies, average of the posterior probabilities of the hypothesized relationships was taken to show the effect of a third individual. Here, when considering individual cases, it is more appropriate to report the likelihood ratio of the true relationship and alternative relationship as we have shown previously that the posterior probability of the true pedigree will nearly always be approximated to 1 when the two individuals in question actually do share IBD and so the increase of information obtained from the third individuals cannot be seen. The likelihood ratio always reflects the additional information.

Table 5.14 Likelihood ratios between the true relationship and the alternative relationship 'unrelated' when only two individuals of interest are observed and when a third individual at different positions is observed.

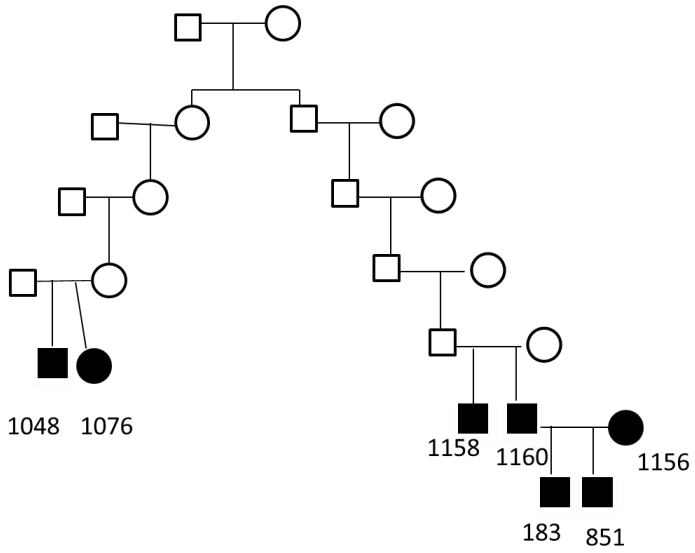
Individuals of interest	True relationship	Only two individual	Sibling	Within1	In-avuncular	Outside1
79 and 1073	S-5-5	6.1e+20	3.9e+23	6.1e+49		6.5e+45
2 and 1021	S-5-5	1.3e+21	4.0e+29	1.2e+28	6.9e+15	
1048 and 183	S-4-6	2.5e +5	6.5e+20	2.8e+14	8.7e+6	4.1e+13
1076 and 183	S-4-6	2.15	4.87e+14	412.47	1.20	0.97

It can be seen from the first three examples (Table 5.14, note that these are ridiculously large numbers) that generally the third individual always increases the likelihood of the true relationship versus the alternative relationship of 'unrelated'. But due to the high variability in the true IBD sharing between distant relatives, the contribution of different third individuals may not be exactly consistent with averaged results for simulated data. In fact, due to the limited number of real relatives that were found here, the cases where the third individual is potentially most useful are not seen i.e. when two relatives that are not detectable become detectable with the genotype of the third individual.

Note that for individuals 1076 and 183, the likelihood ratio between the true relationship S-4-6 and 'unrelated' is only about 2 without a third individual considered, which could imply that they share little or no IBD although they are related. But the

increase of information obtained from a ‘sibling’ is much higher than that obtained from a ‘within1’ relative, which is unexpected. Extensive check on pairwise relationships among all the genotyped individuals in Figure 5.2 (shaded) indicate that 1076 and 183 are possibly related through another parent 1156 of individual 183, which is not represented in the MICROS genealogy. Results imply that both individuals 1048 and 1076 have an S-8 relationship with individual 1156 which is unrevealed by the constructed genealogy in MICROS study. It could hence be the case that 1076 shares little IBD with 1160, and less or no IBD with 183, but shares a lot more IBD with the individual 851 through 1156 by an unrecorded relationship explaining why the inclusion of 851 as a third individual allows us to clearly distinguish the true relationship between 1076 and 183 from ‘unrelated’. These unrecorded relationships are plausible as it is confirmed (personal communication) that individual 1156 is an immigrant from a nearby village. Therefore it is possible for the relationship between her and other individuals missing from the genealogy.

Figure 5.2 A pedigree extracted from MICROS real pedigree.



5.5 Summary

In this chapter, the pedigree-based likelihood method was applied to real MICROS data and the issue of LD, which is inevitable in real data, was investigated. A method of

estimating relatedness with pedigrees, when there is no prior information on the true relationship, was presented. In this method a series of alternative pedigrees such as S-1-1, S-2-2, … etc. and 'unrelated' were considered and their likelihoods calculated to find the pedigree with the highest likelihood (or posterior probability). It would seem that when the 'Template' method is used to estimate relatedness in real data where inbreeding is likely, extended sibling relationships are the sensible choice over extended half-sibling relationships for alternative relationships.

It was seen that LD will bias the estimate of relationship towards closer relationships in both simulated and real data. To simulate data with LD, two different datasets, WTCCC with 1 million SNPs and MICROS with 300K SNPs, indicated that the results are not particular to a given dataset. Then two ways to deal with the problem of LD were considered. One such way is to thin the SNPs, another is to model LD with LD blocks implemented in Merlin (Abecasis and Wigginton, 2005). As there are two approaches to building LD blocks in Merlin, they were compared and it is found that the distance threshold performs quite badly. When the r^2 threshold was used, estimation is much better than when LD was not modelled, but the effect of LD still could not be removed completely. If an unbiased estimate of relationship is desired, it is best to thin the SNPs. Thinning SNPs seems to work reasonably well and is easy to implement. However, the extent to which we should thin depends on the level of LD in the data, the number of SNPs available and, to some extent, the genotyping platform. This suggests that although we have high distinguishing power with dense SNP markers, in real cases when LD is present, dense SNPs cannot be used without causing bias.

6 Review of methods of pairwise relatedness and relationship estimation without a pedigree

In previous chapters, relationships were described by a well-defined pedigree, where the true relationship between two individuals was clearly specified. But this is not always what is estimated in practice. There are many estimators which only estimate the relatedness (how closely or distantly related) between two individuals, but not the exact relationship between them. In this case, only certain parameters need to be estimated. The most common parameter used to specify relatedness is the kinship coefficient. If we know the pedigree we will know the relatedness, but the reverse is not true. All the methods considered in this chapter involve pairwise estimation without recourse to a pedigree.

6.1 Introduction

Identification of relationship utilizes the concept of IBD. Two alleles are IBD if they are both copies of the same ancestral allele. Two alleles are IBS (identical by state) if they are of the same type. Obviously IBD alleles must be IBS and non-IBD alleles could be IBS as well, just by chance. So IBD will cause an excess of IBS sharing than would be expected by chance. Relatives are different from unrelated individuals because they share chromosomal segments IBD. The amount of IBD sharing can be used to estimate the closeness of relationships. We could not consider all ancestors because they can go back infinitely and have to cut a line somewhere, e.g. fifty generations back and all individuals earlier than that time is regarded as unrelated and call the population at that time as reference population. Astle and Balding (2009) reviewed that 15 IBD probabilities are needed to fully describe the relatedness between two diploid individuals and this number reduces to 9 if the pair of alleles within each individual is regarded as unordered (Jacquard, 1970). Only eight probabilities need to be estimated as these probabilities sum to one. These parameters give the probability that each subset of the four alleles at an arbitrary locus, two from each of two

individuals, is IBD. If no inbreeding is allowed, the number of coefficients can be reduced to just three (Cotterman, 1940). These three coefficients specify the probabilities that the two individuals share none, one and two alleles IBD, respectively. Again only two coefficients need to be estimated as the three coefficients sum to 1. The more coefficients we estimate, the more accurately we can estimate the relationship. But it is too difficult to estimate all these IBD probabilities accurately, especially when the relationship between the two individuals is distant. Hence interest tends to focus on the overall degree of relatedness. What is usually estimated for relatedness is either the kinship coefficient θ, or the coefficient of relationship r which is twice θ. The kinship coefficient is defined as the probability that two alleles, one drawn randomly from each of the two individuals at the same locus, are IBD. The coefficient of relationship has a natural interpretation as the expected fraction of genome shared IBD by two individuals.

The kinship coefficient is easy to compute for pedigrees by tracing allelic lineages back to common ancestors. This is called the *expected kinship coefficient*. Thompson (1986) provides a recursive algorithm to evaluate the kinship coefficient between two individuals in a complex pedigree. There are many software packages available to calculate pairwise kinship coefficients based on pedigrees, such as Merlin (Abecasis et al., 2002) and Jenti (Falchi and Fuchsberger, 2008). The relatedness estimated from genetic marker data is the *realized relatedness* and it could be different from the expected relatedness evaluated from the underlying pedigree. There is great variation in the biological process by which relatives share IBD genes due to the stochastic nature of inheritance and this variation will affect every estimation method. There will also be variability in the estimated relatedness for the same degree of realized relatedness. This is generated by the process of estimation and can be controlled.

6.2 Method of Moments (MoM) estimators

One important approach to estimating pairwise relatedness is the method of moment estimation (MoM). There have been several MoM methods proposed (Queller and Goodnight, 1989, Li et al., 1993, Ritland, 1996, Lynch and Ritland, 1999, Wang, 2002). Generally, they consider some statistic, T, of the genotype data. The expected value of this statistic is expressed as a function of the kinship coefficient or other

parameter specifying the relatedness. Then the expected value of T is replaced by its observed value and the function is solved to get an estimate of the kinship coefficient or other relatedness parameter of interest. All MoM estimators essentially estimate the relevant relatedness coefficient from excess IBS between two individuals. They are all defined for the population frequencies but these allele frequencies typically have to be estimated in practice. The theory of four MoM estimators is described below. Some of them estimate coefficient of relationship instead of kinship coefficient. But the kinship coefficient can be obtained easily as just half of the coefficient of relationship and these estimators will be used to estimate both.

6.2.1 Estimator LI

The first MoM estimator to be considered will be denoted as LI here. Li et al. (1993) gave a measure of similarity and used this measure to calculate the expected DNA similarity of two unrelated individuals due to chance only, and the expected similarity between two related individuals due to both chance and relatedness. Based on these results, the degree of similarity due to relatedness only can be calculated. If the four allelic types at one locus are a, b, c and d, traditional methods measure the similarity, S, of the two individuals as 1 when the two individuals share two alleles in common (e.g. aa/aa or ab/ab), 0 when they share no allele in common (e.g. aa/bb or ab/cd), and 0.5 when they share one allele in common (e.g. aa/ac or ab/ac). But Li et al. (1993) suggested that S should be 0.75 in the case of aa/ac and the value of S is unchanged from traditional measurement for other cases. In their method, the similarity is thus defined as the probability that there is an allele IBS in the second individual to a randomly chosen allele from the first individual. For example, when the genotypes of the two individuals are aa/ac, if any a is chosen, there is an allele IBS in the other individual; if the allele c is chosen, there is no allele IBS in the other individual. Because each of these four alleles has the same probability of being the chosen allele of the first individual, the probability that there is an allele in the second individual IBS to a randomly chosen allele from the first individual is 0.75.

Firstly, the similarity U of unrelated individuals in a random-mating population is given as a benchmark. Then the similarity of relatives of different relatedness can be calculated based on U. U can be obtained by a method of enumeration, but this is quite

complex and has some unusual notation. Another simpler method of calculating U was also provided by Li et al. (1993) which is described as follows. Suppose n is the number of allelic types at the locus and p_i is the frequency of i^{th} allele, x is the first individual and y is the second individual. The allele chosen from x could be of any of the n allelic types with a probability equal to its allele frequency. So allele i has a probability p_i of being picked out and the probability that individual y also has an allele of type i is $p_i^2 + 2p_i(1 - p_i) = 2p_i - p_i^2$, where p_i^2 is the probability that individual y is homozygous for type i and $2p_i(1 - p_i)$ is the heterozygous probability where only one allele of individual y is of type i. The similarity that we want for one locus is just the sum of $p_i \times (2p_i - p_i^2)$ over all n allelic types of that locus, which is

$$S_{(unrelated)} = U = \sum_{i=1}^{n} p_i^2 (2 - p_i). \qquad (6.1)$$

In the case of a SNP marker, $U = p_1^2(2 - p_1) + (1 - p_1)^2(1 + p_1) = p_1^2 - p_1 + 1$.

The similarity, S, of several other relationships such as parent-child, siblings, grandparent-grandchild are also calculated. The formula for these specific relationships is then used to derive the formula for general relationships. The general form of similarity, S, between relatives with degree of relatedness r $(r = 2\theta)$ can be written as

$$S = r + (1 - r)U. \qquad (6.2)\ .$$

Replace S with the observed similarity, which is either 1, 0.75, 0.5 or 0 for bi-allelic loci where more than two types of alleles are possible, and solve for r to get an estimate of r:

$$r = \frac{S-U}{1-U}. \qquad (6.3)$$

The multi-locus estimate of r is the non-weighted average of the single locus estimates and the estimate of kinship coefficient can be obtained by taking half of the estimate of r.

6.2.2 Estimator DW

A second MoM estimator is proposed by Day-Williams et al. (2011) and will be denoted as DW in this thesis. The authors considered estimation of both global kinship

coefficient and local kinship coefficient. They defined the global kinship coefficient between two individuals in the same way as the kinship coefficient that is normally used. The local kinship coefficient between two individuals measures their relatedness at a specific locus conditional on all observed genotypes. Here we only focus on their method for estimating the global kinship coefficient. They write the expected number of IBS matches between individuals x and y as

$$e_{xy} = \sum_{i=1}^{m}[\theta_{xy} + (1 - \theta_{xy})(p_i^2 + q_i^2)] , \qquad (6.4)$$

where m is the number of the SNPs, θ_{xy} is the kinship coefficient for x and y, p_i is the major allele frequency in the population at SNP i and $q_i = 1 - p_i$ is the minor allele frequency. The first term in the summation accounts for the matches that are IBD and the second term accounts for the matches that are not IBD but are IBS. Solve for θ_{xy}, to get

$$\theta_{xy} = \frac{e_{xy} - \sum_{i=1}^{m}(p_i^2 + q_i^2)}{m - \sum_{i=1}^{m}(p_i^2 + q_i^2)} . \qquad (6.5)$$

To get an estimate of θ_{xy}, we can replace e_{xy} with the observed number of IBS matches over all SNPs. For each SNP, the number of matches is defined as the conditional expectation of number of matches for two alleles, one drawn randomly from each individual, given the observed genotypes of individual x and y. Therefore the observed number of IBS matches at SNP i is

$$o_{xy}^{i} = \frac{1}{4}[1_{\{I_i = K_i\}} + 1_{\{I_i = L_i\}} + 1_{\{J_i = K_i\}} + 1_{\{J_i = L_i\}}], \qquad (6.6)$$

where I_i and J_i represent the alleles of x at SNP i, K_i and L_i represent the alleles of y at SNP i and 1 is the usual indicator function taking the value 1 when the condition is met and the value 0, otherwise. The observed number of matches is obtained by summing o_{xy}^{i} over all SNPs.

In DW, the similarity is measured as the probability that two alleles are IBS if one allele is picked from each individual randomly. This is different from the similarity in the LI estimator, e.g. the genotype pair ab/ab is measured as ½ in DW but 1 in LI and the genotype pair aa/ab is measured as ½ in DW and ¾ in LI. Another feature of the DW estimator is that the formula

$e_{xy} = \sum_{i=1}^{m}[\theta_{xy} + (1 - \theta_{xy})(p_i^2 + q_i^2)]$ is written over all loci, rather than for a single locus unlike what most other estimators do.

6.2.3 Estimator RI

The third MoM estimator considered was proposed by Ritland (1996) and is denoted as RI here. RI estimates the kinship coefficient locus by locus, and then combines the estimates over all the loci. Suppose there are i types of alleles at a locus (i will be 2 for SNP markers).

Firstly an indicator variable is defined as S_i which is the proportion of matches between two individuals for allele type i. There are four ways of sampling two alleles, one from each of two individuals. S_i takes the value of the number of times that a match is obtained for allele type i averaged by the four ways of sampling. For example if the genotypes of the two individuals are $a_i a_i$ and $a_i a_j$, the observed S_i is 1/2, if the genotypes are $a_i a_j$ and $a_i a_j$, the observed S_i is ¼. The expectation of S_i is

$$s_i = \theta p_i + (1 - \theta)p_i^2, \qquad (6.7)$$

where θ is the kinship coefficient of the two individuals and p_i is the population frequency of allele type i. The first term is the similarity due to IBD and the second term is the similarity due to chance.

Secondly, by equating the observed quantities to their expectations in (6.7), the estimate of kinship coefficient for each allele i at a single locus is

$$\theta_i = \frac{S_i - p_i^2}{p_i(1-p_i)}, i = 1, \dots, n,$$

where p_i is the estimated frequency of allele i. To get the full locus estimate, a weighting is required to combine the estimates for different alleles. The optimal weights are those minimising the variance of θ and the process is complicated. Unless $\theta = 0$ or $\theta = 1$ is assumed, the weights have to be obtained numerically. A simpler estimator of kinship coefficient is given by Ritland (1996) by assuming $\theta = 0$ in the calculation of weights. The weighting is $w_i = q_i/(n-1)$ for allele i, where $q_i = 1 - p_i$, when there are n alleles. Therefore the sinlge locus estimator of kinship coefficient is

$\theta = \sum_{i=1}^{n} \frac{S_i - P_i^2}{(n-1)P_i}$, which is equal to $\frac{S_1 - P_1^2}{P_1} + \frac{S_2 - P_2^2}{P_2}$ in the case of SNP data.

Finally the estimates over all loci need to be combined. Ritland (1996) showed that the variance of a single locus estimator is proportional to 1/(n-1) where n is the number of allelic types at that locus and the inverse of this value is used as weight for that locus in combining the estimates over all loci. For SNP markers, n takes a fixed value of 2.

This simplified multi-locus estimator of relatedness is

$$\theta = \sum_{l=1}^{L} \sum_{i=1}^{n_l} \frac{S_{il} - P_{il}^2}{P_{il}} / \sum_{l=1}^{L} (n_l - 1), \qquad (6.8)$$

where L is the number of loci and n_l is the number of types of allele at locus l. When the markers are SNPs, $n_l = 2$, so the multi-locus estimator θ is just an arithmetic average of the θ of all loci.

6.2.4 Estimator LR

Another MoM estimator was proposed by Lynch and Ritland (1999) which is also called the 'Regression' estimator. I denote this estimator as LR. It is called a regression method because one individual is regarded as a reference and the probability of the genotype of the second individual is conditional on the genotype of the first individual; however, basically it is still a MoM estimator. The parameter for relatedness used in this method is the coefficient of relationship, r, which is twice the kinship coefficient, θ, the parameter that is used in most other MoM estimators. This method not only estimates the coefficient of relationship, but also estimates the Cotterman coefficients of relatedness (k_0, k_1, k_2) simultaneously (Section 2.2). Here we denote k_1 as the probability that individuals x and y only have one allele IBD and k_2 as the probability that two alleles of x are IBD with both alleles of y at a locus,

$$r_{xy} = \frac{k_1}{2} + k_2. \qquad (6.9)$$

LR estimates r_{xy} and k_2 simultaneously. Previous estimators model the genotypes by considering the joint probability of two individuals whereas in the LR approach, the probability of the second individual is considered conditional on the genotype of the first individual at the same locus. The two alleles of the reference individual x are

denoted as a and b, and the two alleles of the individual y are labelled c and d. Then the conditional probability of the genotype y is expressed as a function of k_1 and k_2 and allele frequencies,

$$\mathrm{P(y = cd|x = ab) = P_0(cd) \times (1 - k_1 - k_2) + P_1(cd|ab) \times k_1 + P_2(cd|ab) \times k_2,} \quad (6.10)$$

where $P_1(cd|ab)$ and $P_2(cd|ab)$ denote the probabilities of the genotype cd in y given the genotype ab in x when y and x share one and two IBD genes, and $P_0(cd)$ denotes the probability of the genotype cd in y when y and x share no IBD gene, which is just the product of the frequencies of alleles c and d.

Firstly I will show how the LR estimator is obtained for a special case where the reference individual is homozygous. Then the general form of this estimator will be given.

$P(aa|aa)$ and $P(a.|aa)$ denote the conditional probabilities that individual y has two and one alleles IBS with the reference individual x, respectively, given that the genotype of x is aa, where . stands for any allele other than a. We have that ,

$$\mathrm{P(aa|aa) = p_a^2 + p_a(1 - p_a)k_1 + (1 - p_a^2)k_2}$$

$$\mathrm{P(a.|aa) = 2p_a(1 - p_a) + (1 - p_a)(1 - 2p_a)k_1 + 2p_a(1 - p_a)k_2} \text{ ,}$$

where p_a is the allele frequency of a and is assumed to be known. By replacing $P(aa|aa)$ and $P(a.|aa)$ with their estimates and solving the two equations we get

$$\mathrm{k_1 = \frac{(1 + p_a)P(a.|aa) + 2p_aP(aa|aa) \ - 2p_a}{(1 - p_a)^2}}$$

$$\mathrm{k_2 = \frac{p_a^2 - p_aP(a.|aa) + (1 - 2p_a)P(aa|aa)}{(1 - p_a)^2}}$$

and from Equation (6.9), $\mathrm{r_{xy} = \frac{P(a.|aa)+2P(aa|aa)-2p_a}{2(1-p_a)}}$. (6.11)

$P(aa|aa)$ and $P(a.|a)$ take the value 1 if the genotype pair $aa|aa$ or $a.|aa$ are observed, otherwise 0.

A general form of the LR estimator was given by introducing 'indicator variables' for the sharing of alleles. They are $(S_{ab}, S_{ac}, S_{ad}, S_{bc}, S_{bd}, S_{cd})$. If the two indexed alleles are of the same type, these indicator variables take the value of 1, otherwise 0.

If the allele frequencies of a and b are denoted as p_a and p_b, the general expression for the coefficient of relationship for one locus is finally obtained via

$$r_{xy} = \frac{p_a(S_{bc}+S_{bd})+p_b(S_{ac}+S_{ad})-4p_ap_b}{(1+S_{ab})(p_a+p_b)-4p_ap_b}. \qquad (6.12)$$

Because there is no reason to use one of the two individuals as the reference, in practice this estimator is computed twice with each individual taken as the reference and the average taken as the final estimate. Again note that r_{xy} is twice the kinship coefficient that was discussed previously. LR is undefined for diallelic loci when the reference individual is heterozygous and $p_a = p_b$ (Lynch and Ritland, 1999). Such loci will be deleted from the calculation when implementing this method.

To combine the estimates over *L* loci, a weighted average is taken under the assumption of unlinked, or statistically independent, loci by which the weights that minimize the sampling variance of the overall estimate of r_{xy} are just the inverse of the sampling variance of r_{xy} at each locus. But the variance of r_{xy} is a function of the parameters that we are trying to estimate, so an approximation of this variance is obtained by assuming the two individuals are unrelated, which is a similar approach as in the RI case for assigning weights. The locus-specific weight is given by the inverse of the sampling variance of the estimate of r_{xy} of unrelated individuals x and y. Its general expression for locus l is

$$w_{r,x}(l) = \frac{1}{Var[r_{xy}(l)]} = \frac{(1+S_{ab})(p_a+p_b)-4p_ap_b}{2p_ap_b}$$

and the overall estimate of r_{xy} is

$$r_{xy} = \frac{1}{\sum_{l=1}^{L} w_{r,x}(l)=} \sum_{l=1}^{L} w_{r,x}(l) r_{xy}(l) .$$

6.2.5 Discussion of MoM estimators

Although the concepts of IBD and kinship coefficient are used in MoM methods, they are not well defined. By definition, the kinship coefficient is the probability that the two alleles randomly picked out from two individuals are IBD while being IBD means that two alleles are from the same ancestor. Both IBD and ancestor are only well defined on pedigrees. But MoM methods consider two individuals without pedigrees. One confusing issue is that MoM methods can generate negative value for the kinship coefficient which, by definition, is a probability, and therefore should be non-negative. The kinship coefficient that is used in MoM really measures the correlation coefficients between the allele types in the two individuals, rather than a probability. These methods usually have low bias - perhaps because they can assume negative values.

Like the likelihood method, MoM methods make strong assumptions. Firstly they assume markers are unlinked when they average the estimates from all loci to make the final estimate and make another assumption about the true value about the kinship coefficient when calculating the weights for the estimates at different loci because the weights depend on the true value.

Another problem for MoM methods is that the statistics used in those methods are not necessarily sufficient. Hence, some information in the data is not used, potentially making the estimates less reliable.

6.3 Implementing and comparing different MoM estimators

The four chosen MoM estimators will first be compared with each other on a simple outbred population. Effects of inbreeding, linkage and LD will be investigated at a later point.

6.3.1 Comparisons of different MoM estimators in an outbred population

Firstly different estimators wre compared with simulated data on outbred pedigrees. The software package Mendel (Lange et al., 2013) was used in the simulation, because it enables tracking of simulated alleles at each locus from founders through the pedigree and hence the simulated alleles of every non-founder at every locus can be

traced back to the founders. The observed, or realized, kinship coefficient can be calculated straightforwardly according to the definition of kinship coefficient as now the IBD status of the alleles between the two individuals is known. For every locus, there are four possible ways of picking out one allele from each individual. The kinship coefficient is the number of ways in which the two alleles being picked out are IBD, divided by 4. Therefore different estimators can be compared by their accuracy in estimating the realized kinship coefficient.

110,000 SNPs (evenly spaced 5000 SNPs on each autosomal chromosome) selected from an Affymetrix 500K dataset (Section 3.1) and their allele frequencies were used to simulate genotypes for relative pairs with three levels of relatedness: S-2-2, S-4-4 and S-6-6. 400 replicates were done for each relationship. 110,000 rather than 500K SNPs were used here in order to increase the calculation speed. Because my purpose in this section is to compare the performance of the four estimators, the number of SNPs is not important as long as the same number of SNPs is used for each estimator. The same SNPs and allele frequencies were used and the same number of replicates were done in all the simulation work of this section unless otherwise specified.

When comparing estimators, other authors tend to choose between two approaches. In a more popular method, a large number of replicates of simulation are carried out for a specified relationship and the variances of different estimators are compared. The estimator with the smallest variance is deemed the 'best' (Bink et al., 2008, Day-Williams et al., 2011). The second method considers a large number of simulations for several relationships with different levels of relatedness at the same time and then calculates the correlation coefficients between the true pedigree-based kinship coefficients with the estimated kinship coefficients for each estimator respectively (Santure et al., 2010). The estimator with the highest correlation coefficients is deemed to be desirable.

Here, genotypes were simulated 400 times for each true relationship using Mendel. Then correlation coefficients between the realized kinship coefficient and each of the estimated kinship coefficients from the four MoM estimators were calculated respectively for each replicate. In this way we can know how each estimator performs in estimating the actual IBD sharing between the relatives, rather than just estimating

the expected IBD sharing. These correlations were plotted in R to show the results graphically (Figure 6.1, Figure 6.2, Figure 6.3) and the corresponding correlation coefficients are shown in Table 6.1. The plot of a good estimator should be close to a straight line. The variance of the estimation is mainly caused by the stochastic nature of inheritance which yields a large variation in IBD sharing even for relatives with the same relationship. From these plots, increasing variability can be seen when the relationship becomes more distant.

Table 6.1 Correlation coefficients between the realized kinship coefficients and the estimated kinship coefficients for different estimators and different true relationships (the number of replicates is 400).

True pedigree	DW	LI(1993)	RI(1996)	LR(1999)
S-2-2	0.9771734	0.9736674	0.9870069	0.9880383
S-4-4	0.7394734	0.6693761	0.8711707	0.8710923
S-6-6	0.2350226	0.2026086	0.3585096	0.3587111

Figure 6.1 Plots of the realized kinship coefficients (x axis) and the estimated kinship coefficients (y axis) from the four MoM estimators when the true relationship is S-2-2.

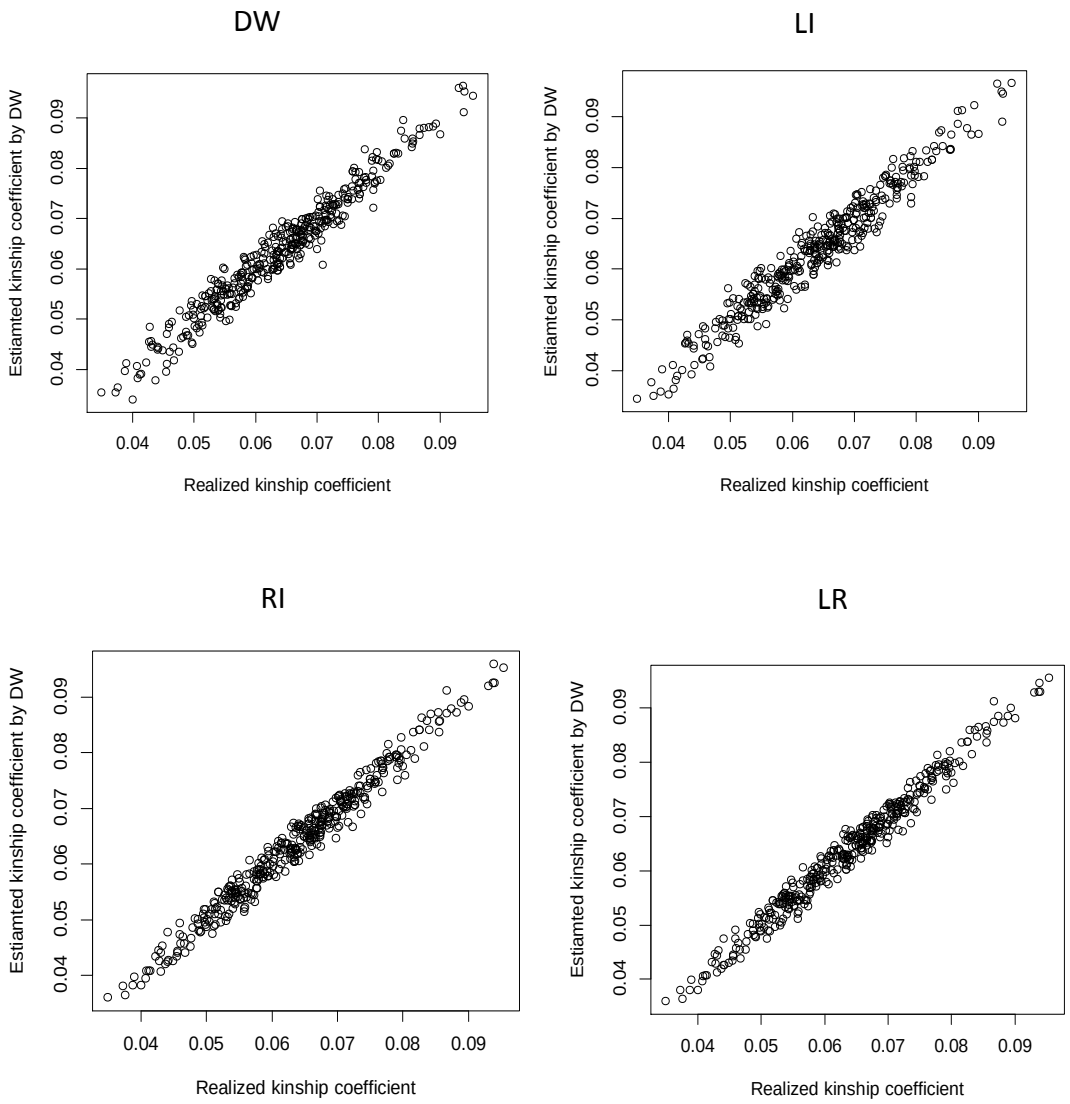


Figure 6.2 Plots of the realized kinship coefficients (x axis) and the estimated kinship coefficients (y axis) from the four MoM estimators when the true relationship is S-4-4.

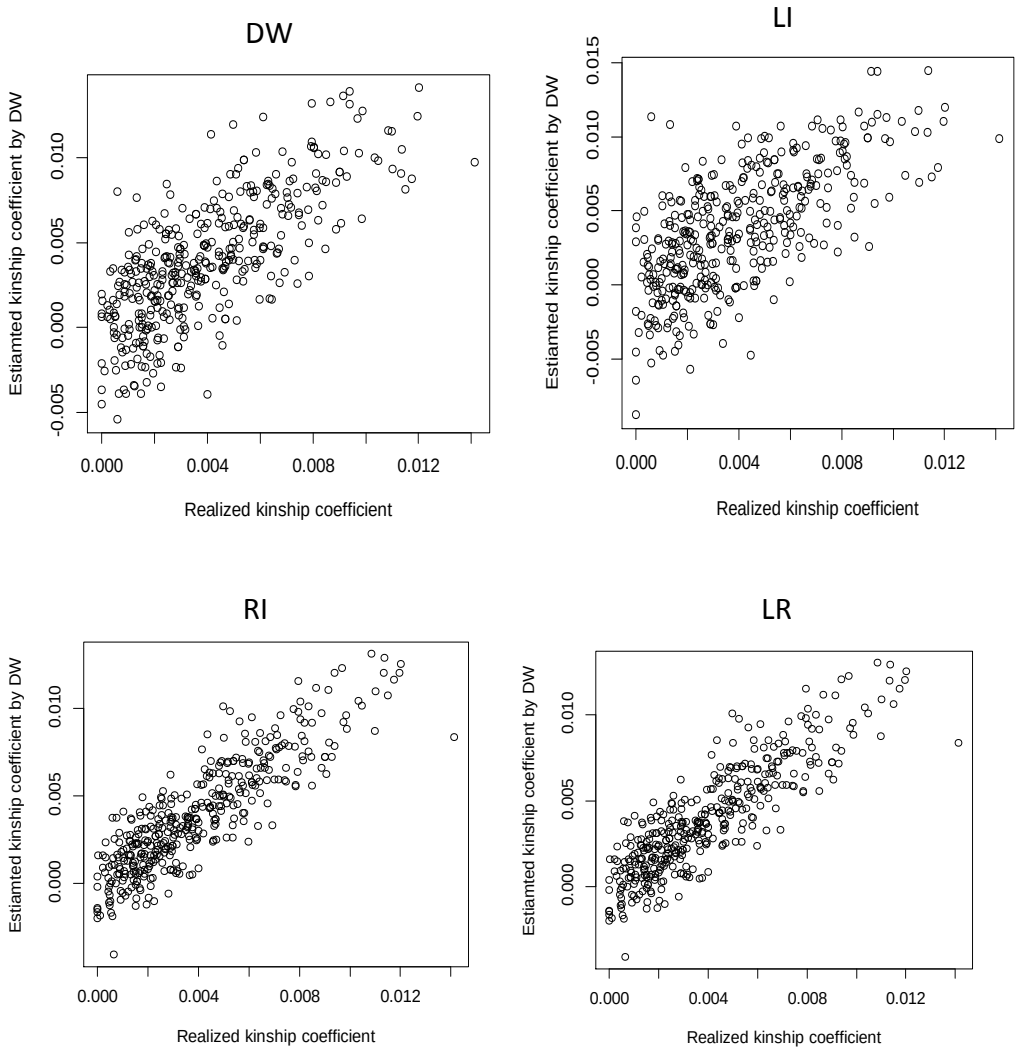
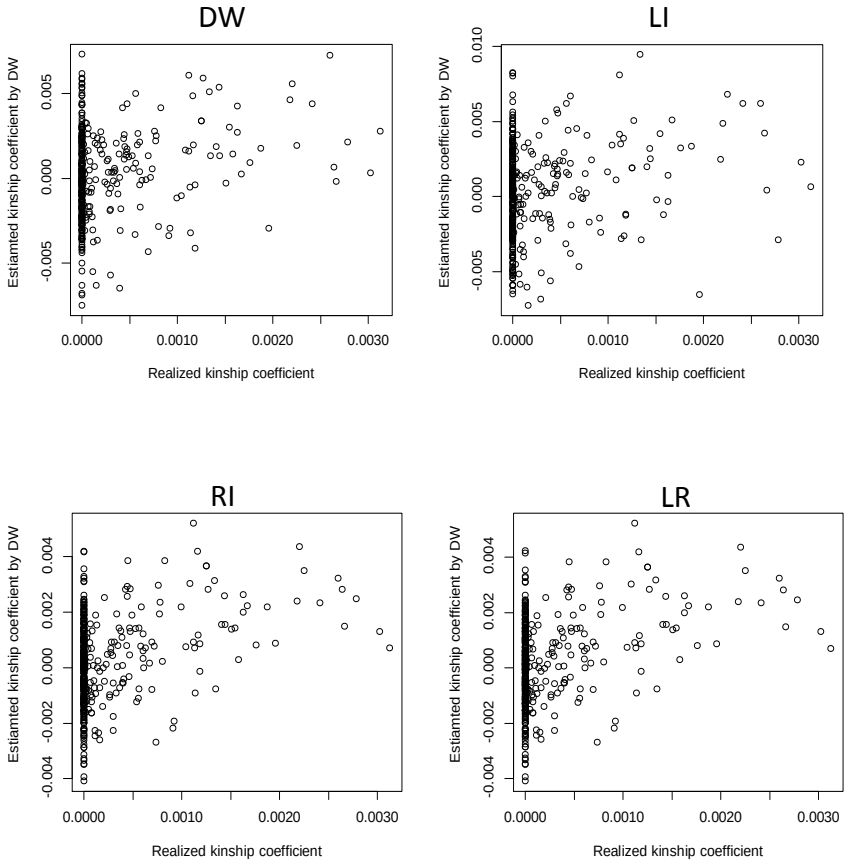


Figure 6.3 Plots of the realized kinship coefficients (x axis) and the estimated kinship coefficients (y axis) from the four MoM estimators when the true relationship is S-6-6.



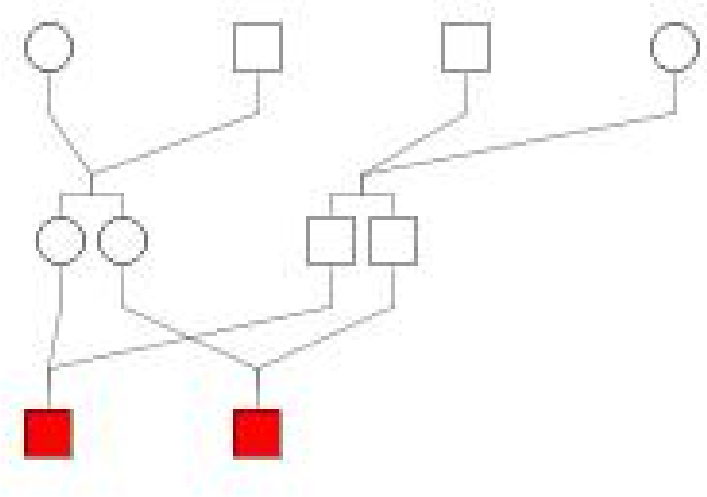
It can be seen from these results that the performance of the RI and LR estimators are better than the performance of DW and LI in this outbred situation. When the true relationship is close, like S-2-2, all estimators are very good at estimating the realized relatedness. But this does not necessarily imply that we can estimate the expected pedigree-based relatedness accurately because of the variance in the realized relatedness. When the true relationship is as distant as S-6-6, none of the four MoM estimators can accurately estimate the realized relatedness, let alone expected relatedness.

6.3.2 MoM estimators and inbreeding

After considering these estimators for unlooped extended sibling pedigrees, I now consider them on a small looped pedigree taken from the MICROS study. In this pedigree two brothers in one family married two sisters of another family and we are

interested in detecting the relationship between two double first cousins resulting from this marriage exchange. The pedigree is shown in Figure 6.4.

Figure 6.4 Pedigree for double first cousin relationship resulting from a marriage exchange.

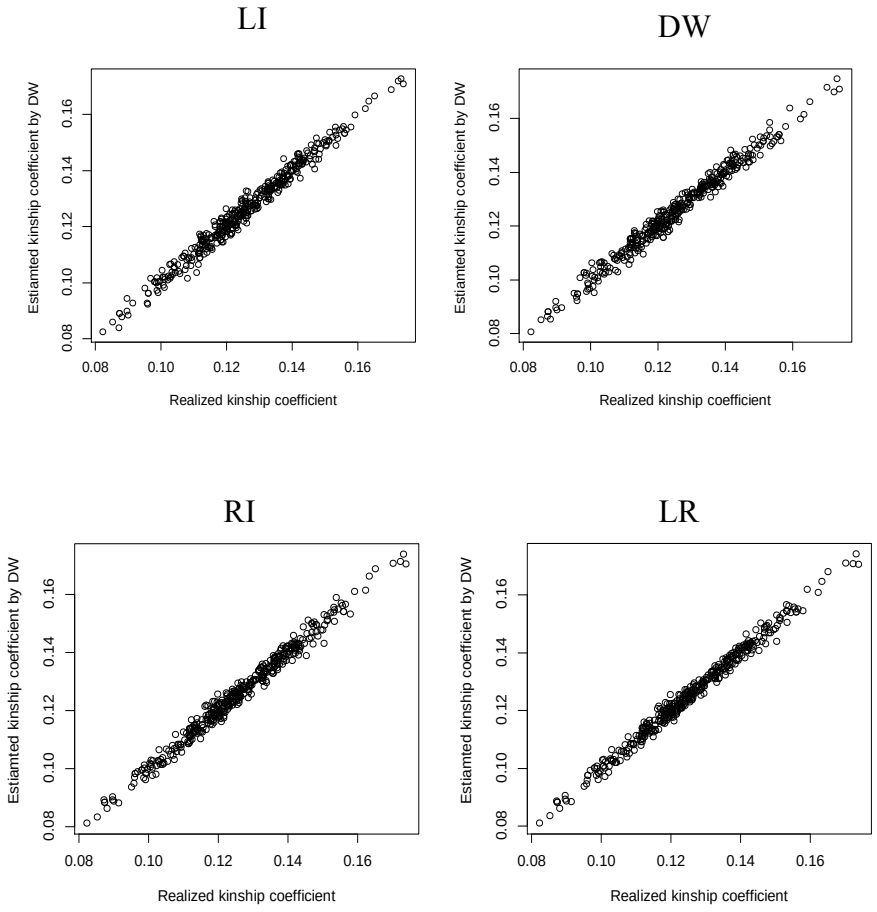


Again 110,000 SNPs from the Affymetrix 500K SNP dataset were used to simulate genotypes and 400 replicates of the simulation were done. The results are shown in Table 6.2 and Figure 6.5.

Table 6.2 Correlation coefficients between the estimated kinship coefficients and the realized kinship coefficients from four different MoM estimators based on 400 replicates when the true relationship is as shown in Figure 6.4.

True relationship	DW	LI	RI	LR
Double cousin	0.989636	0.9894192	0.9920499	0.9938748

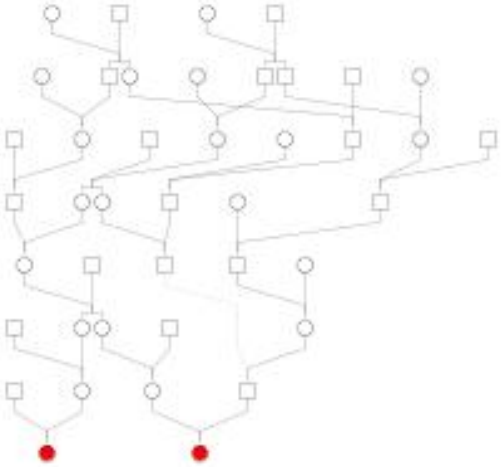
Figure 6.5 Plots of the estimated kinship coefficients from four MoM estimators against the realized kinship coefficients when the true pedigree is as shown in Figure 6.4.



It can be seen that the correlations between the realized and estimated kinship coefficients are very high. So for this simple looped pedigree, the MoM estimators work well, maybe due to the closeness of the relationship (expected kinship coefficient is 0.125).

Next I am going to use the four MoM estimators on an inbred pedigree in Figure 5.1 which has been extracted from the large 8-generation pedigree in the MICROS dataset. The expected kinship coefficient between the individuals 230 and 1193 is 0.0177 based on the true pedigree. The distance of their relationship is somewhere between that of an S-3-3 and S-3-2 relationship.110,000 SNPs were used in simulation and 400 simulations were performed.

Figure 6.6 Pedigree from Figure 5.1.



The results are shown in Table 6.3.

Table 6.3 Means and variances of the realized and the estimated kinship coefficients.

Estimator	Realized	DW	LI	RI	LR
mean	0.01756877	0.01735078	0. 01516173	0.01746704	0.01738742
variance	5.412875e-05	6.090953e-05	6.64343e-05	5.642762e-05	5.579036e-05
Correlation to realized kinship coefficient	1	0. 9530898	0.9016092	0.9760077	0.9763112

Comparing the realized and estimated kinship coefficients, we can see that even for a complex pedigree, the MoM estimators are still accurate in detecting the realized kinship coefficient between two close relatives. The four estimators all perform similarly except the estimator LI in this inbred pedigree. However, if the most recent relationship is of interest we can look at this example from a different perspective: when the most recent relationship of the two individual is S-3-3 and their common ancestor is inbred, both realized kinship coefficient and the estimated kinship coefficient are greater than the expected kinship coefficient of an outbred S-3-3 relationship which is 0.015625.

6.3.3 The issue of linkage and LD for MoM

In most papers discussing these MoM estimators, only a limited number of markers (mainly microsatellites) were used. Linkage and LD were not accounted for in all these methods. Here I want to investigate whether linkage and LD will cause problems for MoM estimators because I am using dense genome-wide SNPs to estimate relationship.

All current methods ignore the linkage between markers (except the pedigree-based likelihood method). They assume the markers are unlinked, which means the data are independent on those loci. Linkage is a concept inherent to meiosis and can be taken into account in pedigrees easily, but is difficult to take into account in pairwise relatedness estimation without a pedigree. In this case, the effect of linkage is to cause association between genotypes at different markers. The strength of the association between the markers with a specific distance (therefore specific linkage) for different degrees of relatedness will be different. If the two individuals are truly unrelated, then the two markers are independent because there is no association between them at population level. But if the relationship between two individuals is close (e.g. siblings), it is wrong to say the genotypes at the two markers are independent. Then here the effect of linkage is similar to that of LD: both of them cause association between markers. The effect of ignoring linkage and LD in the MoM estimates is not known for dense markers because they were mostly used on small numbers of markers by authors and linkage was not discussed.

All pairwise relatedness estimators sum the information over individual loci to estimate the relatedness. Assuming no linkage is equivalent to assuming independent sampling, but linkage makes these samples dependent when individuals of interest are related. Bias due to linkage was not reported in the original papers of the estimators LI, LR and RI. That could be because a small number of SNPs were used. The method of Day-Williams et al. (2011) used dense SNPs and they did not report any problems with linkage, but this issue was not really discussed. Milligan (2003) mentioned that the performance of the MoM estimators, especially RI, has a strong dependency on the actual degree of the relatedness, the unknown quantity that is being estimated. This could be caused by ignoring linkage.

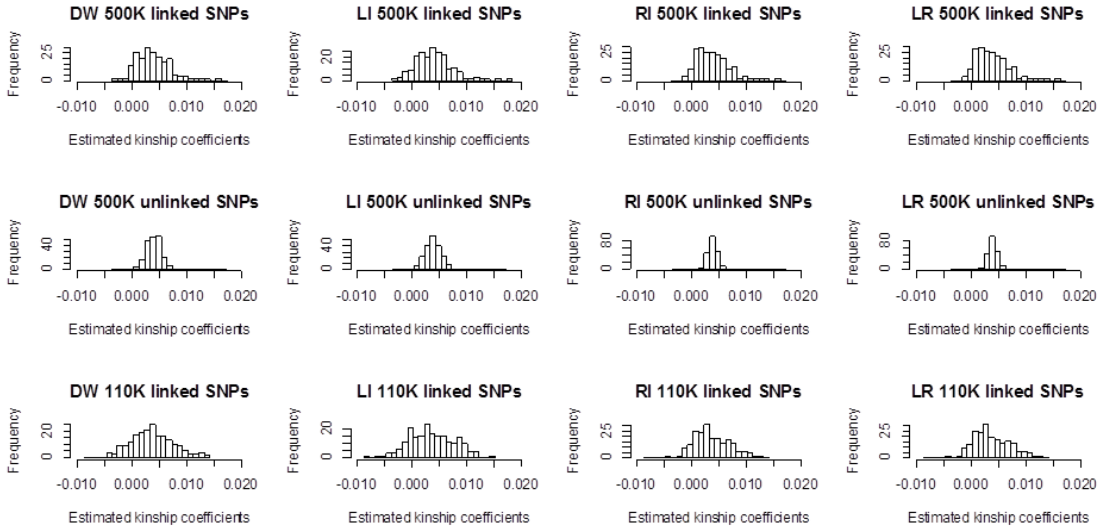
Simulations to assess the consequences of ignoring linkage were carried out using Merlin to simulate data for relationships S-2-2 and S-4-4 and then using all four MoM estimators to make estimation. The simulation was repeated 200 times for every case. Firstly, I compared the same number of markers with different linkage between the markers. Genotypes were simulated for all Affymetrix 500K SNPs. But in the first instance the real linkage map was used and in the second instance the linkage map distance between the SNPs was scaled up 1000-fold, which should reduce the linkage to a minimal level. Secondly, I compared results based on 110,000 linked SNPs, which were spaced out from 500K SNP data with the results based on all 500K linked SNPs with the real linkage map, so that we increase the number of SNPs, and simultaneously strengthen the linkage between SNPs.

The results are shown in Table 6.4 and Figure 6.7. Looking at the means and variances of the estimated kinship coefficients, it seems that ignoring the linkage in dense SNP markers does not make MoM estimators more biased. The effect of the linkage between markers is just to increase the variance of the estimates. When the same number of linked SNPs are used instead of unlinked SNPs (both 500K), the MoM estimators behave similarly on average, but the variances of the estimators increase. So it seems that disregarding linkage is not causing a big problem for MoM estimators on average apart from decreasing the reliability slightly. The variance decreases when the number of SNPs increases from 110,000 to 500K although the linkage between SNPs increases due to the increased density. This is because the information obtained from the larger number of SNPs offsets some of the disadvantages caused by linkage.

Table 6.4 Means and variances (in brackets) of the estimated kinship coefficients using linked and unlinked markers when the true pedigree is S-4-4 and the expected kinship coefficient is 0.00390625 (the number of replicates is 200).

Estimator	DW	LI	RI	LR
500K Linked markers	0.003842702 (1.014113e-05)	0.003979949 (1.12022e-05)	0.003931101 (9.898561e-06)	0.003931211 (9.894794e-06)
500K Unlinked markers	0.003945009 (1.581985e-06)	0.003931338 (2.112009e-06)	0.003905399 (6.391839e-07)	0.003905372 (6.36612e-07)
110K Linked markers	0.003707552 (1.509773e-05)	0.003565808 (1.648612e-05)	0.003703687 (1.020428e-05)	0.00370324 (1.019471e-05)

Figure 6.7 Histograms of the estimated kinship coefficients using 500K linked SNPs, 500K unlinked SNPs and 110,000 linked SNPs respectively when the true pedigree is S-4-4 and the expected kinship coefficient is 0.00390625 (the number of replicates is 200).

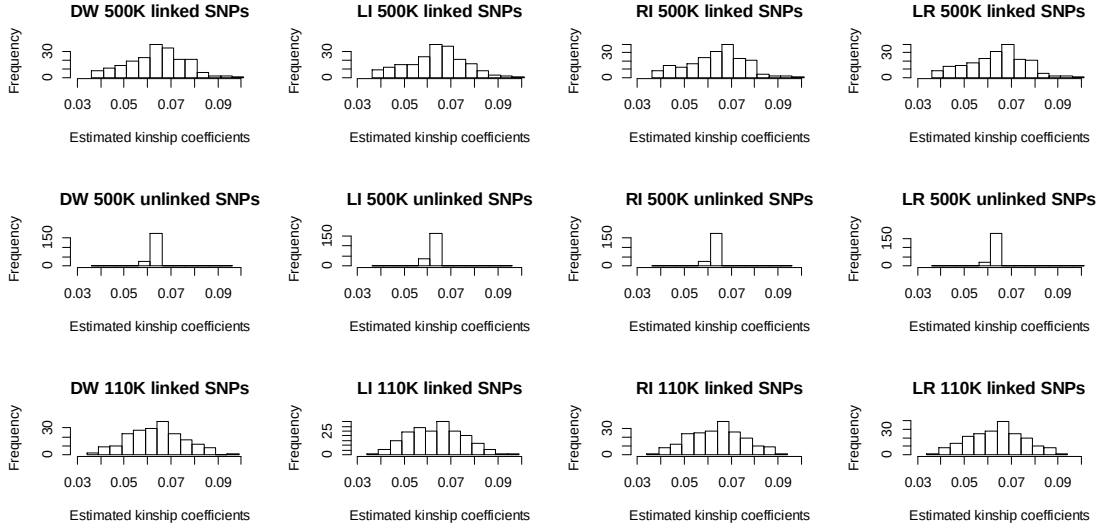


To see whether it will be problematic to ignore linkage for closer relationships in MoM estimators, I repeated this process for a true relationship of S-2-2 and the results are shown below.

Table 6.5 Means and variances (in brackets) of the estimated kinship coefficients using linked and unlinked markers when the true pedigree is S-2-2 and the expected kinship coefficient is 0.0625 (the number of replicates is 200).

Estimator	DW	LI	RI	LR
500K Linked markers	0.0636278 (1.465987e-04)	0.06371204 (1.495493e-04)	0.06361142 (1.468686e-04)	0.06362577 (1.470662e-04)
500K Unlinked markers	0.06256129 (1.750727e-06)	0.06248735 (1.910057e-06)	0.06252397 (1.29441e-06)	0.06251913 (1.179018e-06)
100K Linked markers	0.06382754 (1.302056e-04)	0. 06354161 (1.359421e-04)	0. 06359901 (1.379809e-04)	0. 0638072 (1.379809e-04)

Figure 6.8 Histograms of the estimated kinship coefficients using 500K linked SNPs, 500K unlinked SNPs and 110,000 linked SNPs respectively when the true pedigree is S-2-2 and the expected kinship coefficient is 0.0625 (the number of replicates is 200).



With the true relationship is S-2-2, the variance of the estimated kinship coefficient is increased slightly when the number of SNPs is increased from 110K to 500K. That means the estimation becomes less precise by increasing the number of markers. The reason could be that when the true underlying relationship is close, the association between markers caused by linkage is stronger than when the true relationship is more distant and cause higher variance. This is consistent with the statement of Milligan

(2003) above that the performance of MoM estimator is dependent on the actual relationship.

My conclusion is that MoM estimators can be used with linked markers. We can get generally unbiased estimates and the accuracy will generally increase when the number of the linked markers increases.

I next investigated whether LD in the genotype data will cause problems for these MoM estimators in the cases where LD was artificially generated and when haplotype frequencies were estimated from real data. LD was modelled by Merlin as described in the Chapter 5. To create LD in the simulated genotypes, I arranged that SNPs with contiguous pairwise distance less than 0.001cM were put in one cluster and there is stronger or weaker LD within a cluster. SNPs from different clusters, or not in any cluster, are not in LD. The stronger LD is represented as: within one cluster, allele 1 on one SNP is always on the same haplotype with allele 1 of other SNPs and allele 2 on one SNP is always on the same haplotype with allele 2 of the other SNPs. The weaker LD is represented as: within one cluster, allele 1 on one SNP is three times more likely to be on the same haplotype with allele 1 of the other SNPs than to be on the same haplotype with allele 2, and allele 2 on one SNP is three times more likely to be on the same haplotype with allele 2 of the other SNPs than to be on the same haplotype with allele 1.

Firstly I simulated genotypes for two S-4-4 relatives with artificially created LD and compared the MoM estimates with the results when there is no LD. 400 replicates were done for each case. Affymetrix 500K SNP allele frequency data and map (Section 3.1) were used in the simulation. I went back to the all 500K SNPs because in this simulation we want to study LD, which is stronger in denser data. No obvious bias can be seen in the results as shown in Table 6.6 (the expected kinship coefficient of a relationship of S-4-4 is 0.00390625). The variance remains in the same scale generally.

Table 6.6 Means and variances (in brackets) of the estimated kinship coefficients by four MoM estimators for the true relationship of S-4-4 when there is artificially created LD in the genotype data over 400 replicates.

Estimator	DW	LI	RI	LR
Stronger LD	0.003912881 (1.272177e-05)	0.003908717 (1.23142e-05)	0.003889089 (1.063776e-05)	0.003888941 (1.063574e-05)
Weaker LD	0.003827033 (9.076264e-06)	0.003891193 (9.416982e-06)	0.003791742 (8.605419e-06)	0.003792179 (8.602354e-06)
No LD	0.003842702 (1.014113e-05)	0.003979949 (1.12022e-05)	0.003931101 (9.898561e-06)	0.003931211 (9.894794e-06)

The process was repeated with true relationship S-4-4 using the haplotype frequencies and allele frequencies modelled from the WTCCC dataset and MICROS dataset respectively to simulate data with LD. There are 1M SNPs in the WTCCC dataset and 300K SNPs in the MICROS dataset. The results (Table 6.7) show quite different estimates for the two datasets. This could be due to the fact that different SNPs are included in these two datasets and different numbers and sizes of LD blocks are formed. We cannot conclude whether there is bias due to the limited number of replicates. But it seems that the bias, if there is, will be very minor and not strong enough to make the estimated relatedness move to another degree, unlike what we saw when the pedigree likelihood method was used.

Table 6.7 Means and variances of the estimated kinship coefficients by four MoM estimators for the true relationship of S-4-4 when there is LD in the genotype data. The LD is simulated with LD blocks and haplotype frequencies that are modelled from WTCCC and MICROS real data respectively. The number of replicates is 400.

Estimator	DW	LI	RI	LR
Mean (WTCCC LD)	0.003650581	0.003691031	0.003858229	0.003865781
Variance (WTCCC LD)	1.027867e-05	1.022799e-05	9.382549e-06	9.372978e-06
Mean (MICROS LD)	0.004257487	0.00385722	0.004376937	0.004372208
Variance (MICROS LD)	9.239893e-06	9.64238e-06	8.58513e-06	8.566145e-06

I then considered the complex inbred pedigree in Figure 5.1 and used the haplotype frequencies modelled from the MICROS dataset to simulate data with LD. The expected kinship coefficient of the two individuals based on the pedigree is 0.0177. The results are shown in Table 6.8. Most estimators look unbiased except LI. It is also noticed that the variances of the estimates here are higher than what are observed in Table 6.7 with the same SNP data. Inbreeding seems to increase the variance of the estimate.

Table 6.8 Means and variances of the estimated kinship coefficients by four MoM estimators when the true pedigree is the one in Figure 5.1 and there is LD in the genotype data and the LD is simulated with LD blocks and haplotype frequencies that are modelled from MICROS real data. The number of replicates is 400.

Estimator	DW	LI	RI	LR
Mean	0.01772558	0.01501998	0.0178681	0.01776715
Variance	4.025724e-05	4.29283e-05	3.931681e-05	3.882128e-05

All results show that the LD in the simulated data does not have a fixed pattern in its effect on these MoM estimators, unlike in the likelihood method, where it makes the estimated relationship closer than the true relationship. Overall, for MoM estimators LD does not cause big problems and does not bias the result greatly. So these methods can be used for GWAS data. But they are not good to be used for distant relationships since their variances are too high when the relationships are distant.

The reason why the pedigree likelihood method makes pairwise relationships look closer than they really are when LD is ignored could be as follows. When the likelihood is calculated with the Lander-Green algorithm, markers are assumed to recombine with a rate determined by their linkage map positions. However, when there is LD in the genotype data, markers in LD do not recombine, or recombine with extremely low rate, and they will have the same effect as IBD markers. This perceived extra 'IBD' sharing causes the estimated relationship to be closer than it actually is. But MoM methods are not affected by LD in this way, because they measure the overall similarity between the two individuals without considering the positions of the markers. On average, the similarity between two individuals should be equal to what is expected for their relationship, no matter which markers we use and what their positions are on the chromosomes.

6.3.4 Using MoM on real data

Next I applied these MoM estimators on the real MICROS data. 303,783 autosomal SNPs were used. The allele frequencies were estimated from the sampled individuals by Merlin because allele frequency data from a big independent population is lacking. This could cause bias in the estimation result because firstly, we are estimating allele frequencies from the samples and secondly, the samples are related. For example, if two relatives share a rare allele, it is possible that this allele has been kept within the family for several generations and its estimated allele frequency will be higher than the true allele frequency. Then the estimated relatedness based on the higher frequency of these alleles will not be as close as their true relatedness because it is not unusual for unrelated individuals to share a common allele.

Five pairs of relatives whose most recent relationship are between S-3-3, S-4-3 or S-4-4 were first selected and their kinship coefficients estimated using the four MoM estimators to show how the MoM estimators perform on real data before a large number of pairs of relatives is considered. The results for these first five pairs are shown in Table 6.9.

Table 6.9 Estimated kinship coefficients for several pairs of relatives by MoM estimators.

IDs of Relatives	Most recent relationship	Kinship based on pedigree	DW	LI	RI	LR
230 and 1193	S-3-3	0.01833677	-0.005395225	-0.008325708	0.002548815	0.004410028
415 and 742	S-4-3	0.009594	0.0169984	0.01436857	0.009948151	0.01038311
1224 and 634	S-3-3	0.0160	0.010727434	0.009113624	0.017961001	0.018094706
1280 and 607	S-4-4	0.0040779	- 0.006582	-0.0124967	-0.0002911	0.0012308849
879 and 203	S-4-4	0.004258	-0.011242	-0.032621	0.0001753959	0.0044873295

What can be seen is that: 1) for the inbred MICROS pedigrees, the RI and LR estimators perform much better than DW and LI on real data; 2) the estimated kinship coefficients are less than the expected kinship coefficients calculated from pedigrees which means that the estimated relationships are more distant than the real relationships.

To examine these findings in more detail, 101 pairs of individuals from MICROS dataset (the same that are used in Chapter 5) whose most recent relationships are S-3-3 were used. The pedigree connecting each pair is complex and each pair can be tracked to their ancestors back for 12 generations. It needs to be noted that they could have common ancestors above the known pedigree which are ignored, but that ignored relatedness should be trivial. The expected kinship coefficients between every pair based on the large pedigree were calculated. Then MoM estimators were used to estimate the kinship coefficients for these relative pairs using their real genotype data.

Figure 6.9 The histogram of the expected kinship coefficients based on the large pedigree for 101 pairs of relatives whose most recent relationship is S-3-3.

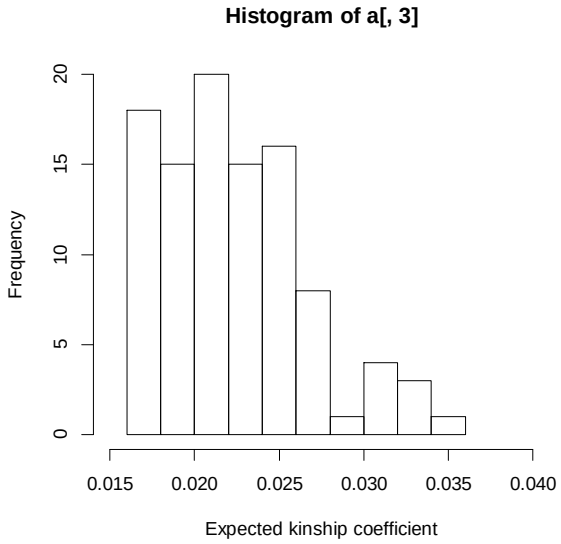
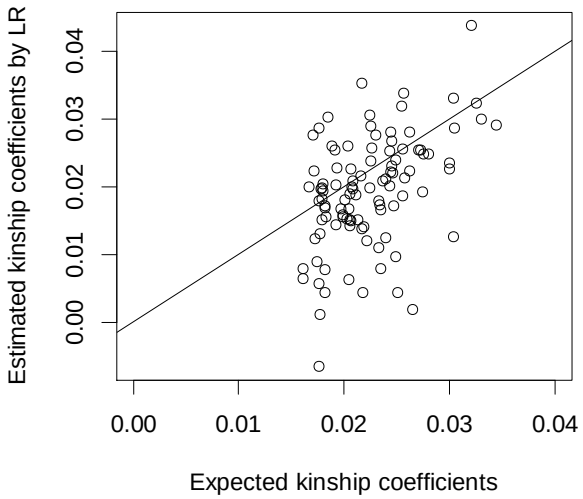


Figure 6.9 shows how varied the expected kinship coefficients between individuals could be in an inbred population even though all are 0.015625 if the most recent relationships are considered. Unknown background relatedness can change the picture greatly. We cannot know the realized kinship coefficients for these relatives because these are real data, therefore we cannot see how well the MoM estimators predict the realized kinship coefficient. Instead we can see how well the MoM estimators can predict the expected kinship coefficient. Although all these relatives have the same recent relationship, their expected kinship coefficients based on the large pedigree are different. So an average of their expected kinship coefficients was taken and compared with the average of the estimated kinship coefficients from MoM estimators in Table 6.10. The estimated kinship coefficients using the LR estimator are plotted against the expected kinship coefficients for all 101 pairs of relatives to have a closer look at the estimate of each individual in Figure 6.10.

Table 6.10 The average of the expected kinship coefficients of the 101 pairs and the averages of the kinship coefficients estimated by MoM estimators.

Expected	Estimated (DW)	Estimated (LI)	Estimated (RI)	Estimated (LR)
0.02234791	0.01916018	0.01360039	0.01788149	0. 01930725
Variance	1.347381e-04	2.21402e-04	8.889837e-05	6.858069e-05

Figure 6.10 Scatterplot of the estimated kinship coefficients using the LR estimator against the expected kinship coefficients for all 101 pairs of relatives.



From Table 6.10, it can be seen that the estimated kinship coefficients tend to be less than the expected kinship. These results are consistent with the findings of Gazal et al. (2014). This negative bias could be due to several possible reasons: lack of independent estimates of population allele frequencies, population stratification, etc. From Figure 6.10, it can be seen that the real relatedness can be predicted quite poorly by the estimated kinship coefficient. For several pairs of relatives, the estimated kinship coefficient is so close to 0, they can hardly be distinguished from 'unrelated'. This is very different from the results obtained in Section 5.3.3 with the pedigree-based likelihood method, where all 101 pairs can be clearly distinguished from 'unrelated'.

6.3.5 Comparing MoM with 'Template' method

The pedigree-based likelihood method ('Template' in Section 5.2) and Method of Moments are the two main approaches that have been discussed in this thesis. In previous sections, it has been shown that the two approaches are different in that: the pedigree likelihood method gives an estimate of the number of meioses separating two individuals while MoM gives an estimate of kinship coefficient; the pedigree likelihood method can consider many individuals at the same time while MoM can only do pairwise estimation; the likelihood method is biased by LD, but MoM is not. Now I will further investigate whether the pedigree-based likelihood method performs better than MoM estimators at detecting distant relatives.

In this part of the simulation study, again the software Mendel was used which labels founder alleles so the number of IBD alleles in the simulated genotypes of any two relatives is known in each replicate and therefore the realized kinship coefficient can be calculated.

Firstly, 110,000 SNPs selected from the Affymetrix 500K SNP dataset were used to simulate genotypes for 400 pairs of S-6-6 relatives. For each pair of relatives, the MoM estimator LR, which seems to be the best among the four estimators considered in this dissertation, was used to estimate the kinship coefficients. Using the pedigree-based method, the likelihoods of the genotypes for eight extended sibling relationships of S-1-1, S-2-2,…, S-8-8 and 'unrelated' were calculated. The relationship with highest likelihood is our estimate of the relationship of the two relatives.

Based on the labels of alleles which track the IBD status between individuals on all SNPs, out of those 400 pairs of relatives, 141 pairs share at least one IBD allele, 259 pairs share no IBD allele. Out of the 259 pairs of relatives with no IBD allele shared between them, in 134 pairs (51.7%) the estimated kinship coefficients using the LR estimator are higher than 0. Out of 141 pairs of relatives with IBD allele shared between them, in 91 pairs the estimated kinship coefficients are higher than 0, so the rate of correct detection of S-6-6 relatives when there is IBD shared between them is 91/141=64.5%. This indicates that the MoM estimator is useless at detecting relatives as distant as S-6-6. Even if two individuals share no IBD, LR has about 50% probability of giving a positive kinship coefficient.

Now consider the performance of the pedigree method on the same data. In 249 pairs out of the 259 pairs of relatives with no IBD sharing, 'unrelated' gives the highest likelihood, with a rate of 'correct inference' of 249/259=96.1%. In 40 pairs out of 141 pairs of relatives with IBD sharing, 'unrelated' gives the highest likelihood. In other 101 pairs of relatives, one of the S-1-1,…, S-8-8 relationships has the highest likelihood. That means we can detect that the two individuals are related with a probability of 101/141=71.6%. It may not be the true pedigree which has the highest likelihood, but we can detect that the two individuals are related. This is very different from what we saw for the MoM estimators where the distant relationship S-6-6 was detected nearly randomly. So the pedigree-based likelihood method works better in detecting distant relatives than the MoM estimators. In particular, the pedigree-based method is much better than the MoM method when two relatives share no IBD as it nearly always infers them to be unrelated while the MoM method has a 50% probability of regarding them as related.

Secondly, genotype data were simulated 400 times, again using labelled founder alleles, for 500K Affymetrix SNPs with allele frequencies estimated from the WTCCC data for the true pedigree of S-6-6. The likelihoods of the true pedigree and the only alternative pedigree 'unrelated' were calculated. Here I would like to know what happens to the 'Template' method and how well it performs when there is little IBD sharing of IBD.

The expected kinship coefficient of S-6-6 is 0.0002441406 which corresponds to 407 IBD alleles shared by the two relatives. This is because the kinship coefficient can be estimated by the number of shared IBD alleles, either 1 or 0, divided by 4 at each locus and the total number of SNPs is 416,854 here i.e. $0.0002441406 \times 4 \times 416{,}854 \approx 407$. The results show that out of 400 replicates of the simulation, there are 220 replicates where the two S-6-6 relatives share no IBD allele at all and out of these 220 replicates, the alternative 'unrelated' has higher likelihood than the true pedigree in 218 replicates. There are 180 replicates where the two relatives do share some alleles IBD, the number of the alleles shared IBD ranges from 12 to 4816 with corresponding kinship coefficient of 7.196764e-06 and 0.002888301. When the realized number of alleles IBD exceeds the expected value, the true pedigree has posterior probability of 1

(see Figure 6.11 in which the x axis represents the realized kinship coefficients and the y axis represents the posterior probabilities of the true pedigree). For the MoM estimator, when the realized kinship coefficient is around the expected kinship coefficient 0.5^{12}, the estimate is almost as likely to be positive as negative (Figure 6.12).

Figure 6.11 Posterior probabilities of the true pedigree S-6-6 when the only alternative pedigree is 'unrelated' against the corresponding realized kinship coefficients (the vertical line corresponds to the expected kinship coefficient for S-6-6 and the horizontal line corresponds to 0.5).

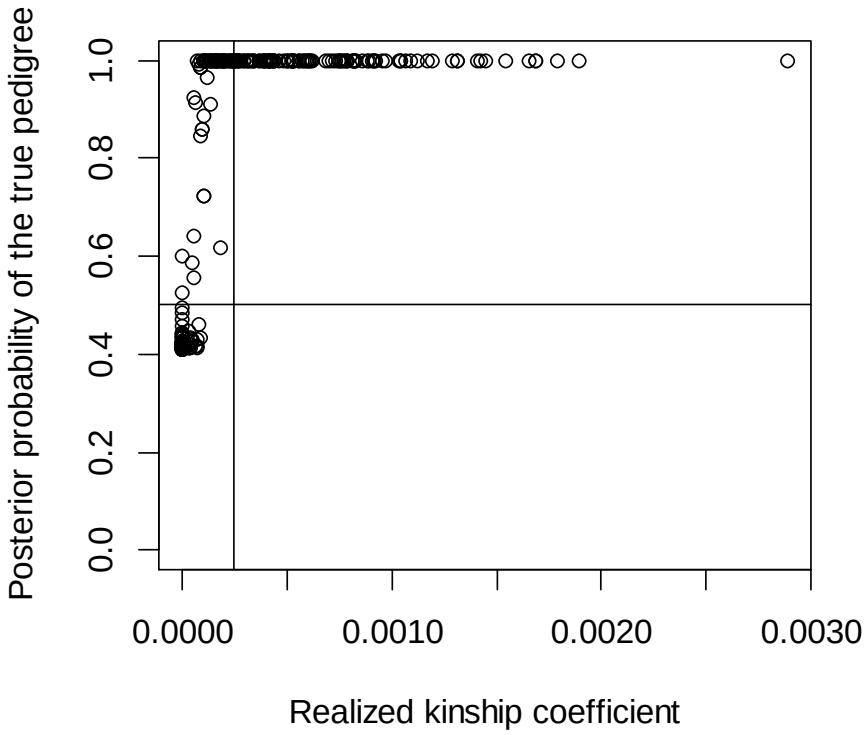
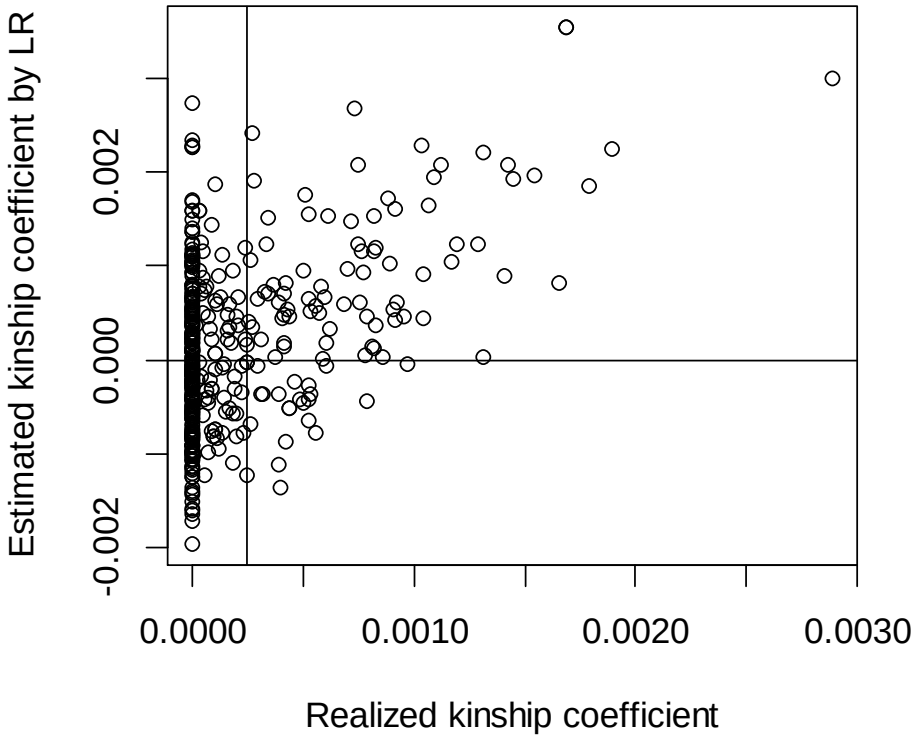


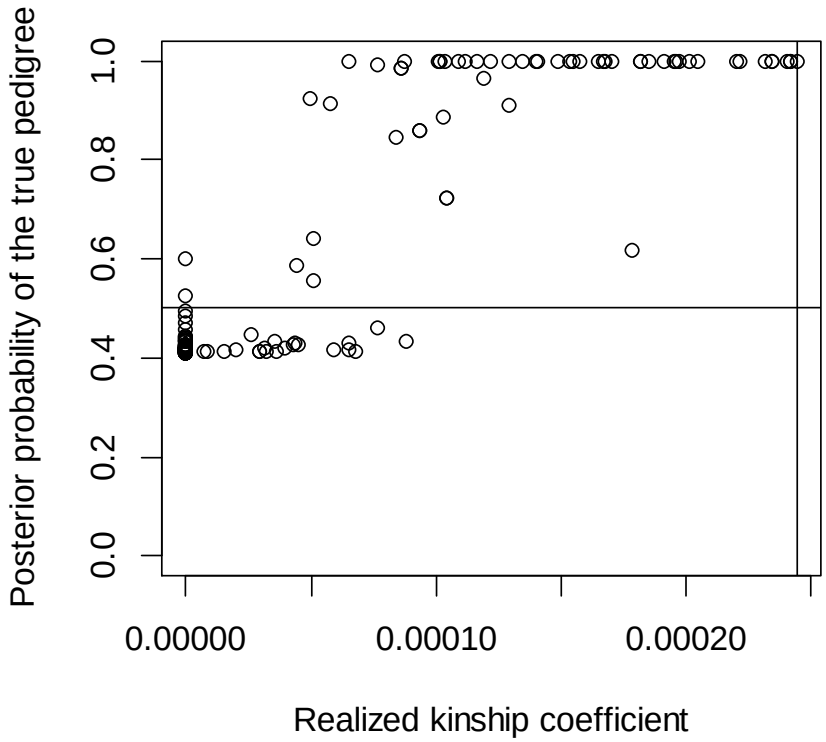
Figure 6.12 Plot of the estimated kinship coefficients by MoM LR with the realized kinship coefficients when 500K SNP are simulated and the true pedigree is S-6-6. The vertical line shows the expected kinship coefficient of S-6-6.



In order to see clearly what happens when the number of alleles shared IBD is small, Figure 6.13 only shows the cases where the realized kinship coefficients are less than the expected value.

Figure 6.13 Posterior probabilities of the true pedigree S-6-6 when the only alternative pedigree is 'unrelated' for cases in which the realized kinship coefficients are between 0 and the expected kinship coefficient (the

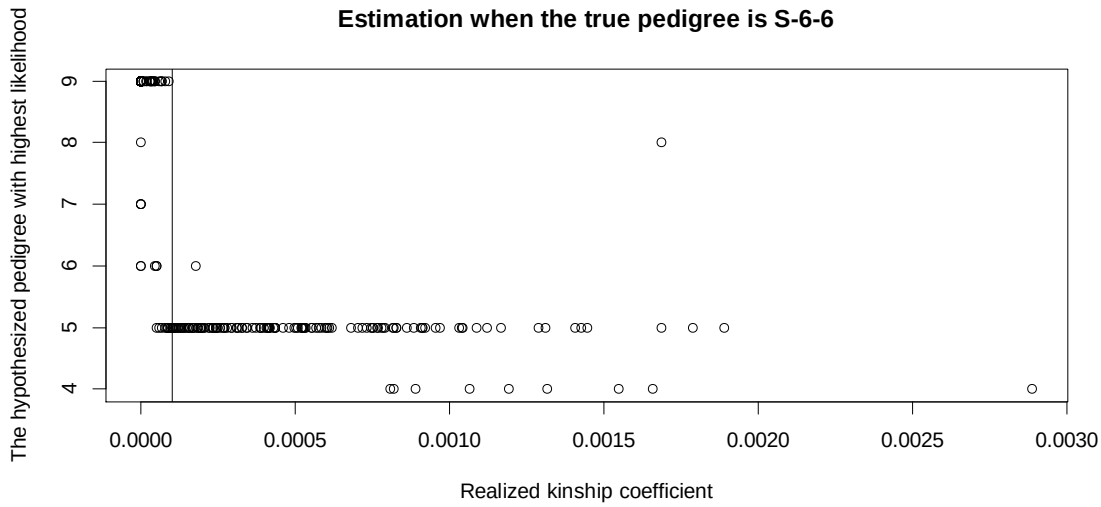
vertical line corresponds to the expected kinship coefficient for S-6-6 and the horizontal line corresponds to 0.5).



It can be seen from the above plot, that even when the realized kinship coefficient is less than the expected value, the posterior probability of the true pedigree is still 1 in most cases. In all cases where the realized kinship coefficient is higher than 0.0001, which corresponds to 167 alleles and less than half of the expected value for pedigree S-6-6, the likelihood of the true pedigree is higher than 0.5 which means we could successfully distinguish the relatives from 'unrelated'. But when the realized kinship coefficient is less than 0.00005, which corresponds to 88 alleles, the likelihood of 'unrelated' is higher than the true pedigree.

Realized kinship coefficients are plotted against the pedigrees which have the highest likelihood among the hypothesized pedigrees from S-1-1…to S-8-8 and 'unrelated' in all replicates of simulation (Figure 6.14) using the same simulated 500K SNP data.

Figure 6.14 Plot of the estimated pedigrees against the realized kinship coefficients when the true pedigree is S-6-6 and the number of SNPs is 500K. The number 4 in the y axis represents S-4-4 and so on, except that the number 9 represents 'unrelated'. The vertical line corresponds to 0.0001.



This graph shows that when the realized kinship coefficient is greater than 0.0001 (166 shared IBD alleles), the pair will be estimated as related. Generally the more IBD alleles they share, the closer the estimated relationship could be. Also, the estimated relationship for the true relationship S-6-6 ranges from S-4-4 to S-8-8. This graph shows information which is not available when looking at the average posterior probabilities like in Table 4.2. The hypothesized pedigree which has the highest average posterior probability does not necessarily have the highest likelihood in all replicates of the simulation. In fact the true pedigree S-6-6 hardly ever has the highest likelihood (as shown in Figure 6.14) although it has the highest average posterior probability. One interesting point is that when the two relatives do share IBD, their IBD sharing is very likely higher than expected for their true relationship. That is why the estimated relationship is closer than the true relationship. This implies that those relatives who are detectable will tend to be inferred to be closer than their true relationship.

6.4 Other relatedness estimators

In addition to MoM estimators, there are two other approaches for pairwise relatedness and relationship estimation. One is maximum likelihood estimation which will be

briefly outlined here as it can only deal with unlinked markers and the focus of this thesis is on using genome-wide data. It needs to be noted that this method is the 'likelihood method' that is usually referred to whereas the likelihood approach that I have discussed in this thesis is a pedigree-based likelihood method. The second approach is based on shared IBD segment detection and is more recent than MoM approaches.

6.4.1 Maximum likelihood estimation of pairwise relatedness

The maximum likelihood approach to estimating pairwise relatedness from unlinked genetic marker data was first proposed by Thompson (1976). Milligan (2003) and Anderson and Weir (2007) had extended the method. Under the assumption of no inbreeding, maximum likelihood (ML) methods find the three Cotterman coefficients, also referred as the k-coefficients ($k_{0,}k_{1,}\ k_{2,}$) which maximize the likelihood of the genotype data. Hepler (2005) had extended MLE to inbred populations and the estimation of the 9 Jacquard coefficients. However, the maximum likelihood approach is limited in its applicability if the number of markers is small (Lynch and Ritland, 1999, Bink et al., 2008). In particular, the MLE method can be biased for small numbers of markers (Ritland, 1996, Milligan, 2003).

The probabilities for the genotypes of two outbred individuals at one locus can be calculated conditional on there being 0, 1 or 2 IBD alleles. These probabilities for all possible genotypes are shown in Table 6.11 where p_i denotes the frequency of the i^{th} type of allele. Note that there could be more than two types of alleles at one locus for STR markers. For example, when the genotypes of the two individuals are both (i,i) at this locus, all the four alleles are independent conditional on 0 IBD allele, therefore the probability of observing these genotypes are the product of observing each of them and should be p_{i}^4. Conditional on 1 IBD allele, there are only three independent alleles because two among the four must be of same type, therefore the probability becomes p_{i}^3. Similarly, conditional on 2 IBD alleles, the probability becomes p_{i}^2.

Table 6.11 The probabilities for the genotypes of two outbred individuals at one locus conditional on the different number of IBD alleles.

	(i,i)(i,i)	(i,i)(i,j)	(i,j)(i,j)	(i,i)(j,k)	(i,i) (j,j)	(i,j)(i,k)	(i,j)(k,l)
0 IBD	p_i^4	$2p_i^3p_j$	$4p_i^2p_j^2$	$2p_i^2p_jp_k$	$p_i^2p_j^2$	$4p_i^2p_jp_k$	$4p_ip_jp_kp_l$
1 IBD	p_i^3	$p_i^2p_j$	$p_ip_j(p_i+p_j)$	0	0	$p_ip_jp_k$	0
2 IBD	p_i^2	0	$2p_ip_j$	0	0	0	0

For pairwise relationships, the algorithm is as follows. The genotypes of the two individuals at locus j are denoted as G_j^1 and G_j^2 , then the probability of the genotypes (G_j^1, G_j^2) can be written as P(G_j^1, G_j^2), which can then be written as

$k_0 \times$ P$(G_j^1, G_j^2$ |0 common IBD)+ $k_1 \times$ P$(G_j^1, G_j^2$|1 common IBD)+ $k_2 \times$ P$(G_j^1, G_j^2$|2 common IBD),

where k_0, k_1, k_2 are the probabilities that there are 0, 1 or 2 IBD genes between the two individuals with relationship of R_i. If we have J unlinked genetic markers, the probability for all the genotypes for J markers is the product of P(G_j^1, G_j^2), in which the set K of (k_0, k_1, k_2) contains the parameters which need to be estimated. Then the relationship which maximizes the above probability is our maximized likelihood estimate for the relationship between the two individuals. If the degree of relatedness is all that is required, kinship coefficient can be calculated easily from the estimated K.

If the exact relationship is of interest, more detailed inference can be obtained by the fact that each relationship corresponds to a specific set K (Thompson, 1986), although some relationships may have same set of K. In practice, we could try several common relationships whose K values are known already rather than try to find an arbitrary relationship with the maximized likelihood (Thompson, 1986).

Below is an example for pairwise relationship estimation when several common relationships whose K values are known. This is a simple case for illustration purpose. In Table 6.12 are the simulated genotypes over 5 unlinked markers for two individuals, whose true relationship is sibling (Thompson, 1986). Then the likelihoods for the different hypothesized relationships were compared.

Table 6.12 The simulated genotypes over 5 unlinked markers for two individuals.

Locus	L1	L2	L3	L4	L5
Individual1	12	12	12	59	12
Individual2	12	12	12	22	22

The number of alleles at these loci and allele frequencies are shown in the Table 6.13.

Table 6.13 The number of alleles at 5 loci and allele frequencies.

Locus	A1	A2	A3	A4	A5
No. of alleles	3	4	2	10	2
Allele 1	0.5	0.3	0.85	0.1 for each of 10 alleles	0.6
Allele 2	0.25	0.3	0.15		0.4
Allele 3	0.25	0.3			
Allele 4		0.1			

The following table shows the IBD probabilities for common non-inbred relationships (Weir et al., 2006) and the likelihood of the above simulated genotypes under the hypotheses of these relationships.

Table 6.14 The IBD probabilities for common non-inbred relationships and the likelihood of the above simulated genotypes.

Relationship	k_0	k_1	k_2	Likelihood for the simulated genotypes
Full siblings	1/4	1/2	1/4	1.934500e-08
Parent-child	0	1	0	0
Identical twins	0	0	1	0
Double first cousins	9/16	3/8	1/16	1.527958e-08
Half siblings	½	½	0	1.122806e-08
First cousins	¾	¼	0	1.049389e-08
unrelated	1			8.09015e-09

From the above calculation with these five markers, the likelihood for the true relationship (siblings) is the highest among all the hypothesized relationships.

This maximum-likelihood estimator only considers unlinked markers as it uses the fact that the likelihood for all loci is the product of the likelihood at every locus. Existing software packages for maximum likelihood estimation, such as ML-Relate (Kalinowski S.T., 2006), for example, only deal with microsatellite markers (unlinked).

6.4.2 IBD segment detection

Another approach to detect relatedness is to look for chromosomal segments (haplotypes) that are identical by descent (IBD) between individuals. Firstly let's look at the process of the IBD segment sharing. At any specific locus in the genome, two individuals who share a common ancestor from n generations back, have a probability $\beta = 2^{-(2n-1)}$ of sharing genome IBD because there are $2n$ meiosis between them and they could share either of the two alleles of the common ancestor. This probability decreases quickly when the relationship between the two individuals get more and more distant. However, DNA is inherited between generations in large segments and if the two relatives do share IBD segments, they can be of substantial length (Donnelly, 1983, Thompson, 2013). The segments shared from the same ancestors will be broken only by recombination with a rate of approximately one recombination per Morgan per meiosis. The lengths of IBD segments, that result from a common ancestor n generations in the past, are approximately exponentially distributed with mean 1/(2n) Morgans (Thompson, 2013, Browning and Browning, 2012). For example, two relatives with 12 meioses (S-6-6) between them more often do not share any IBD segment, but the expected length of the shared IBD segment is 8.33cM when they do have IBD sharing.

The principle behind the detection of IBD sharing is that, if the haplotypes shared by individuals are so long or rare that they are unlikely to be observed more than once in independently sampled individuals, they must be IBD. Current methods for detecting shared IBD segments can be divided into two types. The first type is rule-based, which uses length threshold to do quick searches for shared haplotypes. The second type is model-based, which uses a probabilistic model for the IBD status.

6.4.2.1 Rule-based methods

There are different ways to apply the rule-based method for genotype data and haplotype data. For genotype data, a simple way to search for IBD segments shared by

two individuals is to find long chromosomal segments that are consistent with being IBD, which means there is at least one allele common between the two individuals (Miyazawa et al., 2007). Loci where both individuals are homozygous for different alleles will thus be breakpoints in potential IBD segments. The length threshold for consistent segments to be inferred as IBD segments will be affected by several factors. The first one is the allele frequency. For SNPs with a higher frequency for the major allele, a higher number of IBD-consistent SNPs is needed for the threshold. The second factor is LD. When there is LD within SNPs, a higher number of IBD-consistent SNPs is needed to infer the segments to be IBD. If the threshold is set too high, the power to detect IBD segments will be low, but if the threshold is too low, the false positive rate will be too high. So there is a trade-off between false positives and false negatives to choose the right threshold. Miyazawa et al. (2007) use a threshold of 3.0cM as they argue that it gives a small false positive rate while the false negative rate is acceptable.

If phased haplotype data are available, the haplotypes can be directly compared to find shared haplotypes. If two chromosomal segments with the same haplotype are long enough, we have evidence that they could be IBD. Germline (Gusev et al., 2009) and Beagle fastIBD (Browning and Browning, 2011) are the two most popular rule-based IBD segment methods for haplotype data. Germline uses the length of the shared haplotype as the criterion of inferring IBD segment while Beagle fastIBD uses the haplotype frequency of the shared haplotypes as the criterion for inferring IBD segments. The shared haplotypes are estimated to be IBD if the estimated haplotype frequency of those haplotypes is below a predetermined threshold. When phased data are not available, unphased genotype data can be phased before applying these methods. This will add more uncertainty to the estimation process.

6.4.2.2 Model-based methods

Model-based methods for IBD segment detection require modelling IBD status with a hidden Markov Model (HMM). The latent IBD status is regarded as a Markov chain along the chromosome. Leutenegger et al. (2003) first proposed a HMM method for two chromosomes to detect inbreeding within a single individual. Only two IBD states (IBD or non-IBD) are considered. Purcell et al. (2007) extend this model to four chromosomes (two individuals) and the IBD states considered become 0, 1 and 2 IBD alleles shared between two individuals (implemented in the Plink software). The model

of Browning and Browning (2010), an older version of BEAGLE before fastIBD, applies to two individuals too, but only considers two IBD status: whether there is IBD or not between them. These models mainly comprise two components: the probability of the observed genotype data conditional on the IBD status and the transition of IBD status between markers. LD is not dealt with by Plink, therefore SNP data have to be to be thinned before using Plink. Beagle models LD in its newest version. But according to Gazal et al. (2014), the HMMs modelling LD did not give better results than HMMs with simply pruned, or thinned, SNPs and the relatedness estimation results were still highly biased although the bias caused by LD was significantly reduced. This is similar to what we found when modelling LD with the pedigree likelihood period in Chapter 5 that the bias can be reduced but not completely removed by modelling LD and the estimation does not necessarily become better when LD is modelled rather than simply thinning the SNPs.

6.4.2.3 Estimating relatedness and relationship from detected IBD segments

A simple way to infer relatedness from the detected IBD segments is to divide the total length of detected IBD segment by the total length of all chromosomes. Browning and Browning (2010) claimed that this method is more accurate than MoM for estimating the proportion of shared IBD between individuals.

When only used to calculate the proportion of the genome that is estimated to be IBD, IBD segment sharing methods only provide average IBD sharing between individuals, and are hence similar to MOM methods. In a very recent paper, Gazal et al. (2014) compared the performance of MoM estimators (single-point estimators) and different IBD segment estimation methods, both rule-based and model-based, by simulation. They found that MoM generally performs similarly to segment sharing methods for the above application although rule-based segment sharing methods are best. They also found that the performance of segment sharing methods relies heavily on the choice of the right threshold for segments of chromosome to be classified as IBD. LD is an issue for segment detection methods and there is no perfect way to get around of it. They concluded that for IBD segment sharing approaches, better estimation was achieved using a sparse set of markers rather than modelling LD for dense markers.

There is a theory that all variation of IBD sharing between individuals is represented in the variation of the number, length and position of the shared IBD segments (Hill and Weir, 2011). If these statistics of IBD segments are known, relatedness can be inferred. Both Huff et al. (2011) and Hill and White (2013) have developed likelihood methods to estimate relatedness based on detected IBD segment. The software ERSA (Estimation of Recent Shared Ancestry) of Huff et al. (2011) uses the numbers and lengths of the estimated IBD segments to do likelihood ratio tests. The null hypothesis is that two individuals have no recent common co-ancestor and the alternative hypothesis is that the two individuals share a recent ancestor. They assume the number of shared IBD segments, under the null hypothesis, follows a Poisson distribution with mean equal to the sample mean of the number of segments shared in the population that they belong to. Moreover, the length of the shared IBD segments is assumed to follow an exponential distribution. They also assume the number of the IBD segments and the length of the IBD segments are independent of each other. A threshold length t for IBD segments needs to be specified. They then calculate the likelihood under the null hypothesis, of the number n and the lengths of the shared IBD segments (included in the IBD segment set s). The alternative hypothesis is constructed by introducing new parameters: d, representing the number of generations separating two individuals, and n_a, which is the number of shared IBD segments inherited from the recent ancestors (in contrast to the number of IBD segments from the background population). A maximum likelihood function under the alternative hypothesis is then written with these new parameters. The ratio of the two likelihoods is said to follow a χ^2 distribution approximately and a likelihood ratio test can be done. If the result of the test is that the two individuals do share recent common ancestors, a relationship out of all possible relationships which maximize the likelihood function in the alternative hypothesis is obtained. While they refer to 'all possible relationships', what they really mean is 'all possible value for d', the number of meioses separating two individuals rather than literally all relationships. This is similar to what I did in Section 5.2, where my estimate outcomes will be relationships of different degrees.

Hill and White (2013) have shown that many of the assumptions made in ERSA do not hold. For example, according to Huff et al. (2011), the expected length of shared IBD segments is $\frac{1}{d}$, and the expected number of shared IBD segments for one chromosome

is $E(n) = (dl + 1)(\frac{1}{2})^{(d-1)}$ for extended half–sibling under the assumption of independence between the number and length of IBD segments. But if we divide the expected proportion of chromosome IBD, which is $(\frac{1}{2})^{(d-1)}$ for two relatives separated by d meiosis, by this $E(n)$, the expected length of the IBD segment should be $\frac{1}{d+\frac{1}{l}}$ rather than $\frac{1}{d}$. By simulation, Hill and White (2013) also showed departures of the expectation $E(n)$ from variance $Var(n)$ under the Poisson assumption of Huff et al. (2011) which should be the same. Hill and White (2013) proposed to calculate the likelihood based on the empirical distributions of number, total length of IBD segments (not the length for each single segment) and position of the IBD segments that are obtained from simulation. Then the likelihood of different pedigrees can be compared and to find the one with the highest likelihood. Two relationships can be distinguished by their likelihood ratio. This theory is very similar to the 'Template' method except that it calculates the likelihood for the number and length of IBD segments while 'Template' calculates the pedigree likelihood for the original genotype data. Hill and White (2013) also found that in their method, most information is provided by the total number of the shared IBD segments without the length and the positions of the segments. It seems that segment position hardly gives any extra information unless the relationship is very close. Both methods of Huff et al. (2011) and Hill and White (2013) have to assume that the IBD segments are detected accurately.

Huff et al. (2011) reported that ERSA's estimates are accurate to within one degree of relationship for 80% of sixth-degree and seventh-degree relatives, which correspond to S-3-4 and S-4-4 relatives in terminology of this thesis. The degree of relative is usually defined by the proportion of genes shared by two relatives which is determined by the number of separating meiosis, e.g. first-degree relatives shares about half of their genes and second-degree relative shares about one-quarter of their genes, etc. It is of great interest to compare 'Template' method (Section 5.2) with approach of Huff et al. (2011) in estimating relationships up to a certain degree. Implementation of ERSA is much more complicated than 'Template' method. It requires accurate estimates of IBD segments from software Germline and if there is missing data for some loci, which is very common for real data, software Beagle is needed to do imputation as Germline

does not accept input files with missing data. In this simulation study, 416854 SNPs genotype were simulated for relatives of sixth-degree (such as S-4-3) and seventh-degree (such as S-4-4) using allele frequencies of Affymetrix 500K the same as in Section 4.2. Then ERSA and 'Template' method were used to estimate the degree of the relationships and this process were replicated 400 times for each degree of relationship. When all 416854 SNPs were used, the estimate results of ERSA are accurate to within one degree of the true relationship for 83% of the sixth-degree relatives and 75% of the seventh-degree relatives while the corresponding figures for 'Template' method are 94.5% and 89%. From the results in Chapter 5 we know that 'Template' method is affected by LD seriously. So I reran 'Template' method with only 22,000 evenly selected SNPs although there is no LD in these specific simulated data, to make the comparison fairer to ERSA which was claimed not to be affected much by LD (confirmed by another simulation, results not shown). The corresponding figures from 'Template' method are 85.75% and 81.5% respectively. Therefore even with just 22,000 SNPs (with this density of SNPs, we can be quite confident that it will not be affected by LD), the 'Template' method gives better results than ERSA. Due to the variance of the estimate and the limited number of replicates, we cannot say the 'Template' method is better, but it at least has accuracy of the same order with ERSA with much less SNPs and it is much easier in implementation.

Then the real data for sixth-degree (e.g. S-4-3) and seventh-degree (e.g. S-4-4) relatives were searched for in the MICROS study. The expected kinship coefficient is $\frac{1}{2^7}$ for sixth-degree relatives and $\frac{1}{2^8}$ for seventh-degree relatives. Because the pedigrees in MICROS study are inbred, the number of relatives with an exact kinship coefficients of $\frac{1}{2^7}$ or $\frac{1}{2^8}$ is very small. Relatives with the kinship coefficient within the range of $(\frac{1}{2^7}, \frac{1.01}{2^7})$ and $(\frac{1}{2^8}, \frac{1.01}{2^8})$ were searched for respectively. In total, 116 pairs of sixth-degree relatives and 148 pairs of seventh-degree are found. Then 20,000 SNPs evenly selected from the original 300K SNPs were used to estimate the degree of relationship. The estimation result is accurate to within one degree of relationship for 94% of sixth-degree relatives and 77% of seventh-degree relatives. This is even better than the results in the simulated data. Unfortunately, I did not run ERSA successfully for these real data as no IBD segment was returned by Germline. My simulations results of

ERSA are consistent with the reported results of ERSA in the original paper of Huff et al. (2011). Therefore we should have made a reasonable comparison.

Hill and White (2013) reported that with their method, they can achieve a 75% correct assignment of second cousin once removed (S-3-2) and third cousin (S-3-3). I did not implement their method, but my simulation results show that 'Template' method has the rate of correct assignment of 88% for the same problem.

It may be true that the number, length and positions of IBD segments contain all the information that we need for estimating pairwise relatedness. But this information is contained in the original genotype data together with the linkage map as well. The number, length and positions of IBD segments are, at most, sufficient statistics of the original genotype data, which is exploited by 'Template' method. Moreover, IBD segments have to be estimated before ERSA can be used. Some IBD segment detection methods require phased data and phasing need to be done from the original genotype data. Calculating the likelihood for the genotypes directly is much simpler and more straightforward than going through a phasing process, then detecting IBD segments and finally calculating the likelihood. My simulation also shows that 'Template' method seems to give more accurate estimate than ERSA. However it is slower than ERSA (including the process IBD segment detection). So which method is to be used really depends on the applications. On the other hand, if the requirement on accuracy is not very high or the targeted relationships are not very distant (e.g. only close relative like first cousins are needed), MoM methods are even faster than both of 'Template' method and ERSA, and they are simple both in the theory and implementation.

6.5 Population stratification and allele frequencies

Population stratification is the presence of a systematic difference in allele frequencies between subpopulations in a population possibly due to different ancestry, which is also referred to as population structure. Such stratification is important in association analysis because observed association could be caused by underlying population structure rather than by disease-associated loci. It will cause problems for relationship estimation as well which depends on what is assumed about allele frequencies and

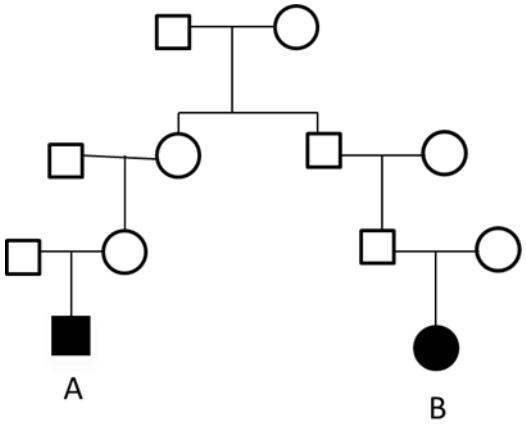
these are affected by population stratification. When estimating relationships in a large sample, it is desirable to check whether there is population stratification.

6.5.1 Population stratification causes bias in estimating relatedness

Population stratification can be considered in two different scenarios. The first scenario is that the ancestors are from different subpopulations, but there has been admixture for some generations. The second scenario is that the population is composed of different subpopulations without admixture. In this section, I will consider the effect of population stratification on relatedness estimation by reviewing recent literature on the first scenario and doing a simulation study on the second scenario of population stratification.

For these simulations I used the allele frequency data for two populations, CEU (Utah residents with Northern and Western European ancestry) and TSI (Toscani in Italia), from the HapMap project. After quality control, there are 556,873 SNPs for each population. The level of stratification in a population is usually measured by the fixation index (F_{ST}). A popular definition of it is based on the variance of allele frequencies between subpopulations. F_{ST} is defined as $F_{ST} = \frac{\sigma_S}{\sigma_T}$ (Holsinger and Weir, 2009) if the variance in the frequency of an allele between different subpopulations is σ_S and the variance of the allelic state of this allele in the total population is σ_T. This definition, which is due to Wright, illustrates that F_{ST} measures the amount of genetic variance that can be explained by population structure. It is frequently estimated from genetic polymorphism data, such as SNPs or microsatellites. The estimated value for F_{ST} between the two populations of CEU and TSI using the allele frequency data that I obtained is 0.007920339, which is consistent with what is commonly known of the population difference between different European countries as the typical range of the value between Germans and Italians is 0.0029-0.0080 (Nelis et al., 2009).

Figure 6.15 Pedigree of S-3-3 used in population stratification study



Suppose there are two subpopulations and each of them includes 15 S-3-3 pedigrees. There are 12 individuals for each S-3-3 pedigree (Figure 6.15), 6 founders and 6 descendants. So there are 180 individuals, 90 founders and 90 descendants, in each subpopulation. Data are simulated for the two subpopulations with different allele frequencies. The S-3-3 relatives in the first subpopulation only are chosen for relatedness estimation and the MoM estimator RI is used. This process is repeated 20 times with a different random seed every time and 300 pairs of S-3-3 relatives are obtained in total. For every simulation of the genotypes, the kinship coefficients are estimated with six different sets of allele frequencies respectively. These allele frequencies are: true allele frequencies (i.e. frequencies used to simulate the data) of the first subpopulation (AF1), average of the true allele frequencies of two subpopulations (AF2), estimated allele frequencies using founders in the first subpopulation (AF3), estimated allele frequencies using all individuals in the first subpopulation (AF4), estimated allele frequencies using founders in the whole population (AF5), estimated allele frequencies using all individuals in the whole population (AF6). By comparing the estimation results with these different allele frequencies, I can see not only the effect of ignoring the population structure, but also the effect of estimating allele frequencies from the samples themselves. We know that estimates of relatedness based on allele frequencies estimated from samples, which include relatives that we want to infer, will be biased (Wang, 2002).

Firstly, I want to see the effect of ignoring the population structure. This can be done by comparing the estimation results using allele frequencies for the whole population

and the estimation results using allele frequencies for the homogeneous first subpopulation. Secondly, for the effect of estimating allele frequencies from samples, estimation results using estimated frequencies can be compared with results using the true frequencies from which the data were simulated. Due to the small number of samples, we will reasonably expect the estimated allele frequencies to be different from the true allele frequencies. When estimating allele frequencies from samples, I apply two options: estimating from founders only and estimating from all individuals. The relationship estimation results of these two options are also compared to see which option gives better results. The results are shown in Table 6.15.

Table 6.15 Means and variances of the estimated kinship coefficients by MoM estimator RI for 300 pairs of S-3-3 relatives from the first subpopulation (expected kinship coefficient is 0.015625) based on, respectively, AF1 (true allele frequencies of homogeneous first subpopulation only), AF2 (average of the true allele frequencies of two subpopulations), AF3 (estimated allele frequencies using founders in the first subpopulation), AF4 (estimated allele frequencies using all individuals in the first subpopulation), AF5 (estimated allele frequencies using founders in the whole population) and AF6 (estimated allele frequencies using all individuals in the whole population).

	AF1	AF2	AF3	AF4	AF5	AF6
Mean	0.01553728	0.01974073	0.01074966	0.00735127	0.0161448	0.01144355
Variance	4.168311e-05	4.122482e-05	4.788052e-05	3.469359e-05	4.187162e-05	4.195883e-05

The results are consistent with the findings of Wang (2002) that MoM estimators are sensitive to the allele frequency. From Table 6.15, we can see that the variances of the estimated kinship coefficients are similar when different allele frequencies are used, but the means of the estimates are very different. If we compare the estimates based on the correct subpopulation with estimates based on the whole population (between AF1 and AF2, or between AF3 and AF5, or between AF4 and AF6) we can see the effect of ignoring population structure is that the estimate of kinship coefficient is greater than expected. That means the estimate of kinship coefficient is biased when population structure is ignored and relatives will tend to appear to be more closely related than they really are. If we compare the results between AF3 and AF4 or between AF5 and AF6, it can be seen that it is better to use founders only than to use all samples to estimate allele frequencies when we use estimated allele frequencies. If we compare

the results between AF1 and AF3 or between AF2 and AF5, it can be seen that the effect of using estimated allele frequencies instead of true allele frequencies is to make the estimated kinship coefficient less than the expected value and relatives will appear to be more distantly related than they really are. Interestingly, the effects of ignoring population structure and using allele frequencies estimated from small number of samples nearly cancelled each other in these simulations as they cause biases to different directions.

These results go some way towards explaining what we observed when estimating 101 relative pairs in the real MICROS data (Section 5.3.3). We know that the samples in the MICROS study are from three different isolated villages although they are of same origin anciently. Therefore there is a possibility of population stratification. In previous work we have ignored population structure and used estimated allele frequencies from the samples. The average expected kinship coefficient is 0.02234791 and the average kinship coefficient estimated by RI is 0.01788149, which is 0.00446642 less than the expected value. If we look at the results in the above simulation study, the expected kinship coefficient is 0.015625. The estimated kinship coefficient using true allele frequency is 0.01553728, which is close to the expected value. When the pooled true allele frequencies for the whole population (including two subpopulations with equal numbers of individuals) are used, the estimated kinship coefficient increases to 0.01974073. When the allele frequencies estimated from all samples in the whole population are used, the estimated kinship coefficient decreases to 0.01144355 from 0.01974703. Overall, the estimated kinship coefficient 0.01144355 is a decrease of 0.00418145 from the expected kinship coefficient 0.015625. This is the same pattern seen in the results for the real data. So the reason that we observe a lower estimated kinship coefficient than is expected in the MICROS data could be a combination of two effects: ignoring population structure increases the estimate of the kinship coefficient, and using estimated allele frequencies decreases the estimated kinship coefficient even more.

6.5.2 Dealing with population stratification

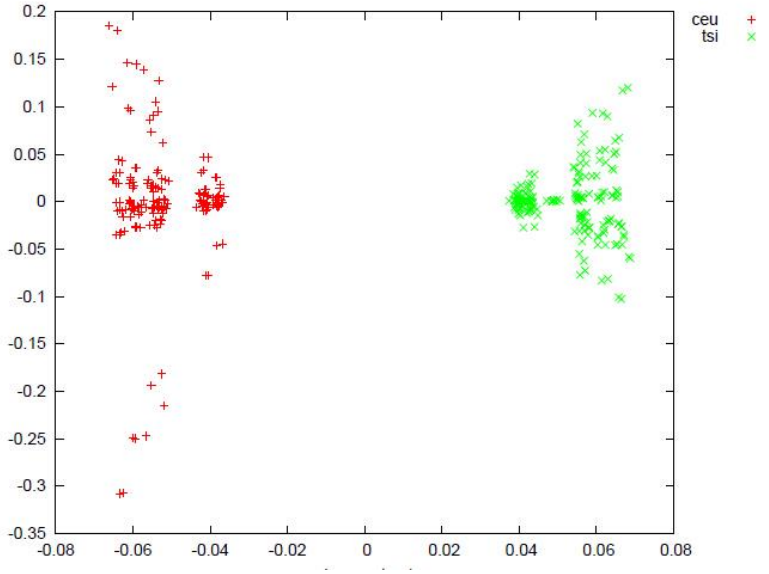
Efforts have been made by several authors to obtain accurate estimates of kinship coefficients in the first scenario of structured populations (admixed) (Thornton et al.,

2012, Morrison, 2013). Morrison (2013) proposed a SNP pruning approach to deal with population stratification, which is called AIM-based SNP pruning. AIM stands for ancestry informative marker and it is based on results from a principal components analysis. The idea is to remove those SNPs whose allele frequencies are very different in the distinct subpopulations. A principal components analysis is carried out first, then a threshold is set on the P-value for the association between SNPs and top principal components to find AIMs. Only markers that are supposed to be non-AIMs are used in the relationship estimation. Theoretically this is a good approach. Although the population is structured generally, it is still possible that there are many markers in which there is no difference between subpopulations and the population structure problem is avoided if we only these markers are used.

In Morrison's simulation, about 600,000 SNPs were reduced to about 110,000 SNPs after pruning, which is a big drop in the density of the SNPs. It was shown that AIM-based pruning reduces the variance and the bias of the estimate. However the effect of simply reducing the number of the markers by thinning instead of applying the AIM-based pruning was not investigated. It is plausible that the reduction in variance after AIM-pruning is simply caused by the reduction of the density of the markers. To test this, I evenly thinned the number of SNPs from 556,873 to about 110,000 for the simulation in the beginning of Section 6.5.1, and found that simply reducing the number of SNPs did not achieve the same effect as reported for the AIM method. There was no noticeable reduction in the variance of the estimated kinship coefficient. So it would seem that it is the AIM pruning rather than simply thinning of SNPs which led to the results reported by Morrison (2013) .

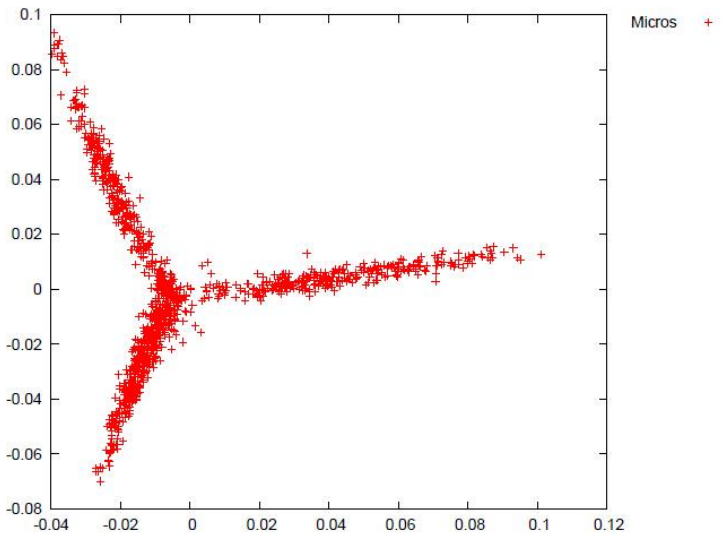
As an alternative, I now consider whether it is possible to cluster samples from different subpopulations by simply doing principal component analysis (PCA). Then relationship estimation can be carried out within the resulting subpopulations without stratification. Simulation results in Section 6.5.1 indicate that we have a better estimate of relatedness when it is done in a subpopulation without stratification. The principal component analysis was done using the software Eigensoft (Patterson et al., 2006). I first tried PCA on the simulated data of Section 6.5.1. A plot of the first two principal components is shown in Figure 6.16.

Figure 6.16 Plot of the first two principal components of 360 simulated individuals, half with allele frequencies of CEU and half with allele frequencies of TSI.



The simulated individuals from the two populations can be clustered successfully in this case. However, the simulated data could be too simplistic as the two simulated subpopulations have very different sets of allele frequencies. I then tried PCA on the real data of MICROS which includes 1285 individuals from three villages and the plot of the first two principal components is shown in Figure 6.17. Based on the knowledge of the source of the samples in MICROS study, these three arms in the plot most likely represent the three villages. Eigensoft has been recommended for unrelated individuals while a small number of relatives should not cause problems. In my application on simulated and real data, both of which include large number of relatives, but it seems that the individuals from different subpopulations can be successfully clustered.

Figure 6.17 Plot of the first two principal components of 1285 individual from MICROS study.



This result suggests that PCA can be used to cluster samples from a mixed population. The estimate of relatedness will be better in each homogenous subpopulation and we have seen that even a small difference between the subpopulations can cause serious bias of the estimated kinship coefficient. But this approach will work best in the case where samples are collected from different homogenous sub-populations. For truly admixed cases where individuals from different subpopulations have been mixed for generations, the AIM method of Morrison (2013) should be used. In this section, population stratification problem is only illustrated by MoM, but it should also affect other methods as they all depend on allele frequencies.

6.6 Summary

In this chapter, some other methods for estimating relationship and relatedness were discussed. Firstly, four MoM estimators for pairwise relatedness were introduced and implemented, and their performance in different situations compared. It seems that the overall performance of the LR estimator is the best of them and it is chosen to be used for clustering in Chapter 8 when we try to reconstruct a pedigree. One MoM estimator was chosen to be compared with the pedigree likelihood method. Results show that the pedigree likelihood method is better than MoM estimators in detecting distant relatives. Linkage and LD will not necessarily bias the estimates of pairwise relatedness, but they

may increase the variances of the estimates which can offset the information provided by increased density of markers. Genome-wide genetic data can be used for these MoM estimators although most of these estimators were proposed when STR data were prevalent with a quantity of less than hundreds. Generally if the only aim is to detect close relatives, less dense markers can be used which can both reduce the calculation time and increase performance. For example, I have shown for relationships as close as S-2-2, 110,000 SNPs perform better than 500K SNPs. We could expect that even fewer SNPs will perform better for closer relationships like S-1-1.

Traditional maximum likelihood estimators of relatedness (which estimate the degree of relatedness rather than a pedigree relationship) were also discussed along with a more recent approach, IBD segment detection and the likelihood methods based on the detected IBD segments. It was found that the estimation results from these likelihood methods based on the detected IBD segments (such as ERSA) are very similar to what can be achieved with the pedigree likelihood approach based on the original genotype data. It seems that estimation of IBD segments is perhaps not always necessary for relationship estimation purpose. Of course, detecting IBD segments between individuals can have other applications, which are not the main focus of this thesis.

Finally, it was shown that even small levels of population stratification can cause serious problems for relationship estimation. Population stratification affects the estimators through its effect on allele frequency. In other words, we should use allele frequencies that are as accurate as possible when using MoM estimators, i.e. try to do the estimation in a homogeneous population and either use independent allele frequencies or estimate them from a large sample.

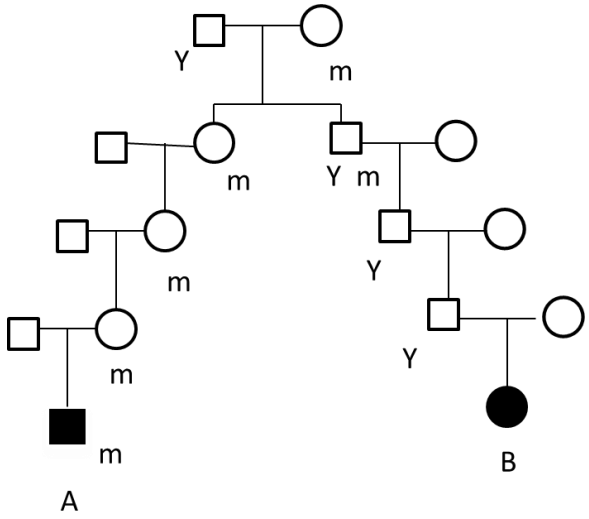
7 The use of the Y chromosome and mtDNA in pairwise relationship estimation

7.1 Introduction

SNP data on the male-specific region of the Y chromosome (MSY) and mitochondrial DNA (mtDNA) are routinely collected by many widely-used SNP chips. But they are not used in mainstream pairwise relationship estimation. One reason could be that these data are difficult to process. Also, it is hard to combine them with autosomal DNA in current relationship estimation methods due to haploidy arising from their uniparental inheritance. In this chapter I investigate the ways that they can be used in relationship estimation.

The MSY and mtDNA are special in that they pass down the generations unchanged except by mutation. MSY only exists in males and it only transmits from father to son, since it carries the male-determining gene *SRY*. By contrast, both males and females have mtDNA, but fathers do not pass their mtDNA to their descendants. This is due to the relatively low number of copies carried by sperm cells together with a mechanism that appears to eliminate paternal mitochondria following fertilization. Since the oocyte carries many mitochondria, mothers pass their mtDNA to both male and female descendants. In the following Figure 7.1, a genealogy connecting third cousins A and B is shown to illustrate the transmission of MSY and mtDNA.

Figure 7.1 Illustration of the transmission of MSY and mtDNA. Y represents the MSY haplotype of the common male ancestor and m represents the mtDNA haplotype of the common female ancestor. In this particular example, A and B do not share either Y or mtDNA IBD.



Two classes of markers on the MSY are short tandem repeats (Y-STRs – also known as microsatellites) and single nucleotide polymorphisms (Y-SNPs). STRs on the MSY have received more attention than SNPs in forensic application. Y-STRs have been used in male-specific identification, paternity testing and estimation of the geographic region of origin of a male (Kayser et al., 1997, Jobling et al., 1997). The diversity of Y-STR haplotypes on the global level makes them suitable for this purpose. One famous application of the MSY STRs in relationship inference is the Thomas Jefferson paternity case (Foster et al., 1998), in which a shared haplotype between a putative male-line descendant and male-line descendants of Jefferson's paternal uncle supported his paternity of Eston Hemings Jefferson, son of one of his slaves. Broader applications include making links over many generations between men who carry the same surname (King and Jobling, 2009). At present, standard kits contain up to 23 Y-STRs (Purps et al., 2014) and up to 186 Y-STRs have been analysed (Ballantyne et al., 2010).

However, Y-STRs have high mutation rates – typically ~2 x 10^{-3} per STR per generation, but can be as high as 7.44×10^{-2} (Ballantyne et al., 2010). This gives rise to very high haplotype diversity and is an advantage in individual identification, but it complicates pairwise relationship estimation as the mutation rate has to be incorporated. It also means that sometimes even close relatives may not share the same Y-STR haplotype. By contrast, Y-SNPs are less discriminating than Y-STRs. The

haplotypes defined by Y-SNPs are usually called ‘haplogroups’, some of which are common in particular populations, but their mutation rates are so low (~3 x 10^{-8} per nucleotide per generation (Xue et al., 2009)) that mutation among SNPs can be ignored in relationship estimation. Haplogroups formed by Y-SNPs have been used to find ancestral origins of populations and their potential in forensic identification has been discussed (Geppert et al., 2011, Jobling, 2001). However, a concrete way to apply them in general relationship estimation problems has not been proposed.

mtDNA SNPs have been used in forensic identification as well, and one famous case in which they have been applied in relationship inference is in the identification of the Romanov royal family (Gill et al., 1994), and later of their two missing children (Coble et al., 2009). mtDNA is much shorter than MSY (16.5 kb compared to ~24 Mb of usable Y sequence) with almost all variation being due to SNPs. Most forensic analysis focuses on the ~350-bp Hypervariable Segment I (HVSI), which has particularly high sequence diversity. SNPs in this region have mutation rates about 100 times that of SNPs in the nuclear genome including MSY (Soares et al., 2009), but this rate is still low enough to be neglected in most pedigree applications. SNPs outside HVSI (in the so-called coding region) have mutation rates about 10 times the nuclear genome rate. Overall, SNPs on mtDNA can be treated without considering mutation, though it should be borne in mind that particular SNPs in HVSI may show recurrent mutation. Recurrent mutation is a specific nucleotide change that has occurred more than once. It is also worth noting that SNPs can be more reliably typed from degraded DNA than STRs and hence have the potential to be of practical use in identification cases.

7.2 Defining SNP haplotypes for MSY and mtDNA from Affymetrix 6.0 SNP chip

The raw data that I have are the SNP genotypes from the Wellcome Trust Case Control Consortium (WTCCC2, 2008) which include 2987 unrelated individuals (1483 males and 1504 females) from the two control groups in the study: the 1958 birth cohort and the National Blood Service sample (NBS). The platform for genotyping is the Affymetrix 6.0 SNP chip. This chip includes 901 SNPs for the MSY and 445 SNPs for

mtDNA. SNP haplotypes (referred to as haplogroups here) for both MSY and mtDNA were defined from these raw data.

The majority of MSY SNPs shows no useful variation and thus can be discarded. These include some SNPs in which only one type of allele is called for some individuals but a proportion of individuals have failed to give an allele call. Genotypes at the remaining SNP loci with meaningful variation are then compared and fitted to the standard Y haplogroups (the nomenclature system of Y haplogroups in Karafet et al. (2008) was used, supplemented with information from: http://www.isogg.org/tree/). 26 distinct haplogroups are inferred for the MSY for these data. Recently software including YTool (Peng et al., 2014) has been developed to retrieve Y-chromosomal haplogroups from GWAS data.

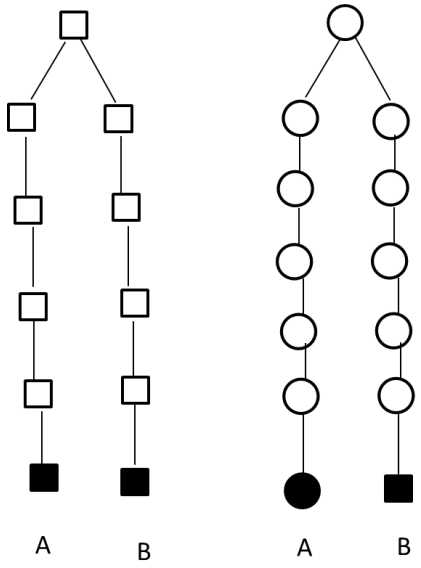
The software Haplogrep (Kloss-Brandstatter et al., 2011) was used to call mtDNA haplogroups. It compares mtSNP calls at specified sites with a database of previously analysed whole mtDNA sequences and returns the best fit. 167 haplogroups for mtDNA are inferred. The inferred MSY and mtDNA haplogroups and their frequencies are shown in Appendix 10.3.

7.3 Using MSY and mtDNA information as a complement to autosomal data

It could be expected that in most cases, these markers will not help in relationship estimation due to their uniparental inheritance, e.g. for the pedigree shown in Figure 7.1. However, if two individuals happen to be on a patrilineal or matrilineal line of descent from a common ancestor, they would share a haplotype of MSY or mtDNA even if they were very distantly related (Figure 7.2). This could provide evidence for relatedness, but the strength of the evidence would depend on the haplotype frequency in the general population from which the individuals derive. Some haplotypes may be common, in which case evidence for a particular relationship is weak. However, if the shared haplotype is rare, the evidence could be very strong. Donnelly (1983) showed that the probability that two relatives share autosomal IBD alleles decays exponentially with increasing numbers of separating meioses. When the relationship between two individuals is distant, they often share no IBD at all. In that case, based on autosomal

data only, their relationship could be undetectable. Provided there are patrilineal or matrilineal connections between the individuals, MSY or mtDNA haplotypes are expected to be shared IBD no matter how many meioses separate the individuals in question.

Figure 7.2 Pedigrees showing patrilineal and matrilineal lines of descent.



7.3.1 Method

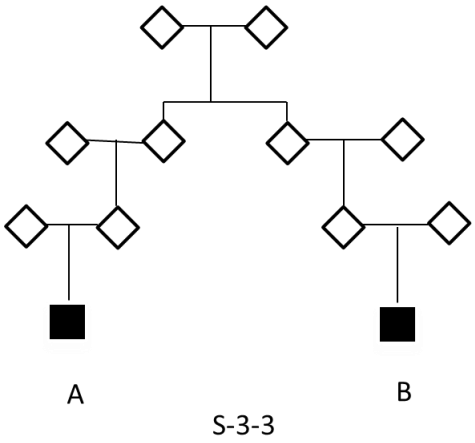
Here I will present a general way to combine SNPs data of MSY and mtDNA with autosomal SNPs in likelihood methods of relationship estimation. Because the inheritance of MSY and mtDNA is independent of autosomal DNA, the probability of observing autosomal DNA data can be multiplied by the probability of the observed MSY and mtDNA data, to get the probability of observing both. This makes it convenient to combine the data from MSY and mtDNA with that from autosomal chromosomes. Furthermore, the likelihood ratios between hypothesized relationships of two individuals obtained from autosomal data, MSY and mtDNA data, respectively, can be multiplied as well. The overall likelihood ratio of the two hypotheses of H_i and H_j is hence

$$\frac{L_i}{L_j} = \frac{L_i(Autosome)}{L_j(Autosome)} \cdot \frac{L_i(Y)}{L_j(Y)} \cdot \frac{L_i(mtDNA)}{L_j(mtDNA)}. \quad (7.1)$$

Here L_i represents the likelihood of the hypothesis H_i based on all the data; $L_i(Autosome)$, $L_i(MSY)$ and $L_i(mtDNA)$ represent the likelihoods of the hypothesis H_i using autosomal DNA, MSY and mtDNA data respectively.

The next issue is to calculate the probability of the MSY and mtDNA data. Here I consider only the problem of pairwise relationship estimation and assume that the sex of the unobserved individuals in the pedigree linking the two individuals in question is unknown. This is because all we are concerned with is the distance of the relationship between the two individuals and the sex of the ancestors is generally not available. Although in previous chapters pedigrees were depicted with sex, this is irrelevant for the calculations using autosomal SNP data, whereas the sex of an individual crucially matters in the case of MSY- and mtDNA-based estimation. For example, a pedigree where only two individuals of interest are observed is shown in Figure 7.3. As is convention, individuals with unknown sex are represented by diamonds. Note that, even when the sex of an observed individual is unrecorded, males and females can be differentiated from SNP-chip data simply by the presence or absence of data for the MSY SNPs.

Figure 7.3 Pedigree of S-3-3 where sex is unknown except those of observed individuals.



Firstly I show how the probability of observed MSY haplotypes for the hypothesized pairwise relationship of extended siblings ($S - n_1 - n_2$) can be calculated when the sex of other individuals in the pedigree is unknown.

$$P(Y) = P(Y|IBD) * P(IBD) + P(Y|NonIBD) * (1 - P(IBD), \quad (7.2)$$

where Y represents observed haplogroup data for the two individuals of interest and P(IBD) is the probability that the two individuals (both males) inherit their MSY from the same male ancestor.

If the number of meioses between the two relatives is n ($n = n_1 + n_2$ for a relationship $S - n_1 - n_2$), P(IBD) =$1/2^{(n-2)}$ because in every generation there is only a probability of 0.5 that the connecting individual is a male (assuming a 50:50 sex ratio) and therefore inherits a Y chromosome from the common male ancestor. Since we already know that the two observed individuals are males, these individuals inherit their MSY from their fathers with probability of 1. Values for $P(Y|IBD$ and $P(Y|NonIBD)$, the probability of observing specific Y haplogroups in two males when they are IBD and not IBD, are shown in Table 7.1.

Table 7.1 Probabilities of observing the Y haplotypes of two male individuals when they are and not IBD.

Observed Y haplogroup	IBD	Non-IBD
i i	p_i	p_i^2
i j	0	$p_i p_j$

As an example, consider two males both having the same MSY haplogroup i with frequency 0.01. The probability of these observed haplogroup data under pedigree S-2-2 is

$$p_i * P(IBD) + p_i^2 * (1 - P(IBD) = 0.01 * 1/4 + 0.01^2 * (1 - 1/4) = 0.002575$$

where p_i represents the frequency of the haplogroup i. The probability of these MSY haplogroups under pedigree 'unrelated' is $0.01^2 \times 1 = 0.0001$. The likelihood ratio of the two pedigrees is 25.75, which gives moderate support for pedigree S-2-2. But if this MSY haplogroup is a common one with frequency, say 0.2, the above likelihood ratio will become just 2, which gives limited support for pedigree S-2-2. Evett and Weir (1998) suggested a convention of interpreting likelihood ratios into evidence to support one hypothesis against another (Table 7.2).

Table 7.2 Likelihood ratios and their verbal equivalent of support (excerpted from (Evett and Weir, 1998)).

Likelihood ratio	Verbal equivalent
1 to 10	Limited support
10 to 100	Moderate support
100 to 1000	Strong support
more than 1000	Very strong support

The calculation of the probability with observed mtDNA haplogroups is exactly the same as that for MSY, except that it can be done when the two observed individuals are of either sex, since all individuals carry mtDNA. The common female ancestor passes her mtDNA to all her children with probability 1. After that generation, every generation of ancestors of the two individuals in question has a probability of 0.5 to pass the common ancestor's mtDNA to the next generation, because males do not transmit mtDNA. So the two observed individuals (of either sex) have a probability of $1/2^{(n-2)}$ to share a mtDNA haplogroup IBD.

7.3.2 Simulation Study

To show whether we can obtain useful extra information from the MSY and mtDNA data, a simulation study was carried out along the lines of what was done for autosomal markers. We know that MSY data are only available for males and mtDNA data are available for both males and females so, in order to consider both kinds of data together, I focused on cases where the two individuals in question are both male. 11,000 evenly spaced autosomal SNPs, MSY haplogroups and mtDNA haplogroups were simulated for different true relationships ranging from S-1-1 to S-5-5. The sex of the individuals on the pedigree were supposed unknown (with equal probability of being either sex) except for the two observed individuals. Then for each true relationship, the likelihoods for several hypothesized relationships - S-1-1, S-2-2, S-3-3, S-4-4, S-5-5 and 'unrelated'- were calculated. The purpose was to see how well we can distinguish the true relationship from those specific alternative relationships. 10,000 replicates were carried out for each true relationship. Then the likelihoods were converted into posterior probabilities $P_i = \frac{L_i}{\sum_{i=1}^{n} L_i}$ assuming a flat prior, where P_i and L are the posterior probability and likelihood of the i^{th} hypothesized relationship

respectively (Section 4.1). The reason that 11,000 autosomal SNPs were used is that with this number we do not need to consider the issue of LD (Berkovic et al., 2008, Pemberton et al., 2010, Kling et al., 2012). 11,000 SNPs usually are not sufficient to distinguish distant relatives, but here I just want to see whether the distinguishing power is increased when MSY and mtDNA data are combined with these autosomal SNPs. The results when only autosomal SNPs were used in the calculation are shown in Table 7.3 to be used as a baseline.

Table 7.3 Average posterior probabilities of several close alternative relationships when 11,000 SNPs are used for each true relationship.

True	S-1-1	S-2-2	S-3-3	S-4-4	S-5-5	Unrelated
S-1-1	**1**	0	0	0	0	0
S-2-2	0	**0.9941**	0.0059	0	0	0
S-3-3	0	0.0059	**0.8194**	0.1465	0.0222	0.0061
S-4-4	0	0	0.1457	**0.4275**	0.2591	0.1677
S-5-5	0	0	0.0222	0.2589	**0.3545**	0.3644

In the first column are the true relationships. Then the posterior probabilities are shown for each true and alternative relationship. Each row may not sum to one due to rounding.

The results of incorporating MSY and mtDNA haplogroup information in the likelihood and corresponding posterior probability calculations are shown in Table 7.4.

Table 7.4 Average posterior probabilities of several close alternative relationships when 11,000 autosomal SNPs, Y chromosome and mtDNA are used for each true relationship.

True	S-1-1	S-2-2	S-3-3	S-4-4	S-5-5	Unrelated
S-1-1	**1**	0	0	0	0	0
S-2-2	0	**0.9941**	0.0059	0	0	0
S-3-3	0	0.0055	**0.8212**	0.1455	0.0219	0.0060
S-4-4	0	0	0.1449	**0.4288**	0.2592	0.1670
S-5-5	0	0	0.02188	0.2588	**0.3548**	0.3645

In the first column are the true relationships. Then the posterior probabilities are shown for each true and alternative relationship. Each row may not sum to one due to rounding.

Comparing the values in bold in Table 7.4 with that in Table 7.3, we can see that when the true relationship is more distant than S-2-2 the values for the average posterior probability of the true pedigree are slightly increased with the addition of MSY and

mtDNA data, although the increase is negligible. Because MSY and mtDNA help only in a small number of the replicates due to their uniparental inheritance and lack of sex information, the information provided by them appears trivial by averaging over a large number of the replicates. But in fact, in those cases where two individuals do share an MSY or mtDNA haplotype IBD, the increased information could be dramatic. In Table 7.5, I show that when MSY and mtDNA are included, the number of relatives that can be detected increases. This is especially apparent when the true relationship is distant and the relatives share fewer autosomal SNPs IBD.

Table 7.5 The number of replicates where relationship 'unrelated' has the highest likelihood among the six hypothesized relationships considered out of the 10,000 replicates for each true relationship. 'Unrelated' has the highest likelihood means that the two individuals of interest are mistakenly estimated as unrelated, otherwise they will be estimated somehow related although it may not be the true relationship which has the highest likelihood.

True relationship	Only autosomal SNPs	MSY and mtDNA	Reduced
S-1-1	0	0	0
S-2-2	0	0	0
S-3-3	47	44	3
S-4-4	2416	2389	27
S-5-5	6228	6185	43

When the true relationship is S-3-3, the number of replicates in which 'unrelated' has the highest likelihood decreases from 47 to 44, which mean 3 more pairs of relatives are estimated as related when MSY and mtDNA are included. Similarly, when the true relationship is S-4-4, 27 more pairs of relatives are detected as related when MSY and mtDNA are used. And when the true relationship is S-5-5, 43 more pairs of relatives are detected as related. When the true relationship is S-1-1 and S-2-2, there are no replicates in which 'unrelated' has the highest likelihood, which means that the two relatives are estimated as related in all replicates.

The reason that MSY and mtDNA help more for cases of distant relatives is that when the true relationship is more distant, there is a higher probability that they share few or even no autosomal markers IBD, while for close relationships, there is so much

information in the autosomal data that the effect of MSY and mtDNA data is negligible .

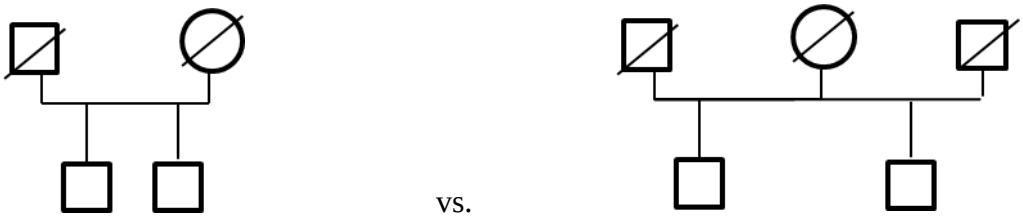
7.4 Special cases

Above I have shown how MSY and mtDNA SNP haplotypes can be used in a general pairwise relationship estimation problem where sex is unknown except for the two individuals of interest, and presented a method for systematically combining them with autosomal SNPs. In this section I consider their uses in cases where the sex of more pedigree members is known. A very simple use of MSY and mtDNA markers is to verify pedigrees for sex consistency for any applications requiring pedigrees e.g. linkage analysis.

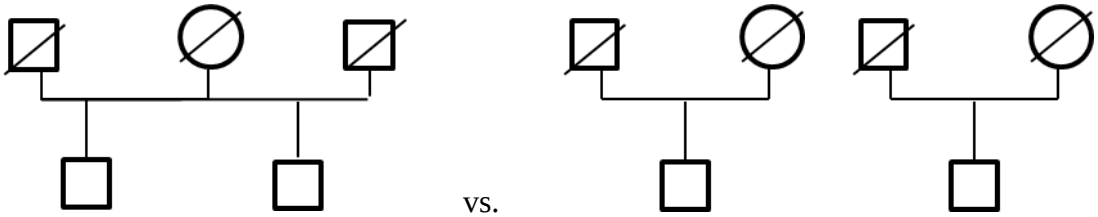
Next I give examples of four special cases where MSY and mtDNA can be very useful. In case 1, I want to know whether two males are siblings or maternal half-siblings when no parent is observed. In case 2, I want to know whether the two males are maternal half-siblings or unrelated when no parent is observed. In case 1, the question that I want to answer is whether the two individuals share a common father, while in case 2, the question is whether the two individuals share a common mother.

Figure 7.4 Graphs of case 1 and case 2 (individuals with a line crossing are not observed).

Case 1

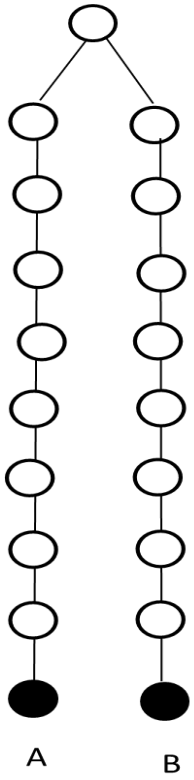


Case 2



In case 1, if the Y haplogroups of the two individuals in question are different, we can exclude the first pedigree and choose the second pedigree immediately. In this case it becomes a problem of excluding a relationship for inconsistency of MSY haplogroups. If they are of the same type and the frequency of this shared haplogroup is denoted as p, then the likelihood of the first pedigree is p and the likelihood of the second pedigree is p^2. Therefore the likelihood ratio of the two pedigrees is $1/p$. If the shared MSY haplogroup is a rare one with a low p, we will obtain stronger evidence supporting the first pedigree. The idea is the same for case 2, but here I consider their mtDNA haplotype. The above constitutes the evidence that we can get from MSY data or mtDNA alone. These likelihood ratios could be multiplied by those obtained from autosomal data if they are available.

Figure 7.5 Graph of case 3.

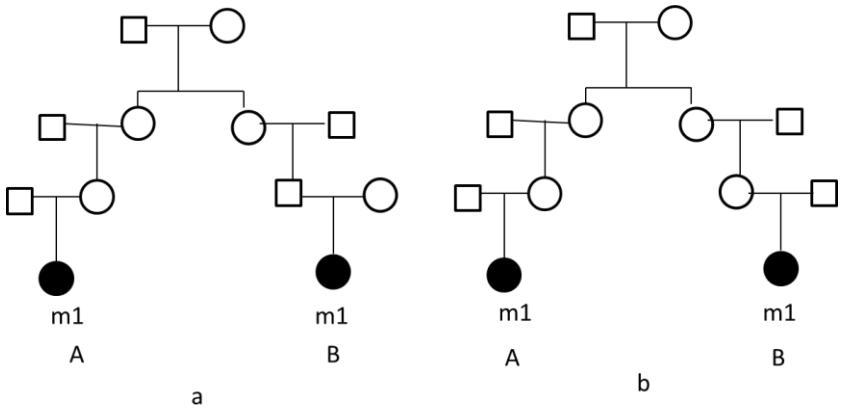


In case 3, a reported pedigree is as shown in Figure 7.5 in which two individuals share a common female ancestor nine generations ago via matrilineal descent. The proposed pedigree is known to be credible and A is known. We want to know whether an observed individual is the individual labelled B in this pedigree, or just some random person, i.e. unrelated to A. In this case, the use of mtDNA is more important because for relationships this distant, the autosomal data are unlikely to be very helpful since such relatives have a very high probability of sharing no autosomal allele IBD. Sharing a rare mtDNA haplogroup with frequency of 0.001 will give a likelihood ratio of 1000/1 between the reported relationship and the proposition that they are unrelated. This situation is not unlike the case of the DNA identification of the remains of King Richard III, in which a living individual known to be a 16-generation matrilineal descendent of Richard's sister, Anne of York, has been shown to share a rare mtDNA haplotype with the skeletal remains (King et al., 2014). In this case, the whole mtDNA genomes were sequenced.

Autosomal data could tell us how two individuals are related, such as cousins, second cousins etc. They can never tell us the sex of the individuals in the pedigree linking the

two individuals, which we may want to know in some applications. But MSY and mtDNA data can give us some hints. If the two relatives share a rare Y chromosome or mtDNA haplotype, we can suspect that they are likely to be linked by a pedigree as shown in Figure 7.2 because if there is a break in the patrilineal or matrilineal line of descent, the probability that the two relatives share the same rare haplogroup is very low. For example, two females A and B are second cousins (Figure 7.6) and we want to know whether B is related to A through her father or mother. If m1 is a rare haplogroup, the pedigree in part (a) of Figure 7.6 is very unlikely and we have strong evidence to support the pedigree in part (b) of Figure 7.6, which means that B is related to A through her mother. In another scenario, if we know the mtDNA haplogroup of the mother of B is a different type m2, we have evidence to support that B is related to A through her father.

Figure 7.6 Two examples where mtDNA haplotype is used to exclude hypothesized pedigrees (black colour represents observed individuals, m1 and m2 are different mtDNA haplotypes)



Clearly, these data (Y and mtDNA) are most useful when sex information is available. Such special cases could be generalized to include pedigree verification for any applications requiring pedigrees e.g. linkage analysis.

7.5 Discussion

It is worth noting that the way of calculating likelihood for MSY and mtDNA data in Section 7.3.1, is specially proposed for situations where the sex of the pedigree member are all unknown except the two individuals of interest. For situations where the sex of the hypothesized pedigree is known (based on prior information perhaps),

the probability that the two individuals share IBD MSY or IBD mtDNA should not be calculated by the formula P(IBD) $=1/2^{(n-2)}$ where n is the number of meioses between the two individuals of interest. The probability of the IBD status of the MSY or mtDNA between two individuals is either 1 or 0 as long as the pedigree (together with sex) in specified.

A case in which MSY and mtDNA data are very useful, is that the two individuals, A and B, are distantly related and share little or no autosomal DNA IBD. If we compare the true relationship that A and B are distantly related with another two hypotheses that A and B are closely related, and that A and B are unrelated, the hypothesis that A and B are closely related could be easily rejected by the fact that A and B share little autosomal DNA IBD. It may be appealing to increase the weight for the information provided by MSY or mtDNA because it could increase our capability of rejecting the wrong hypothesis that A and B are unrelated. But doing that will unavoidably increase the probability that unrelated individuals, who share the same MSY or mtDNA haplogroup simply by chance, are concluded as distantly related because the likelihoods obtained from autosomal DNA are similar for the two hypotheses of 'distantly related' and 'unrelated', and all that makes difference are their MSY and mtDNA data.

However, when prior information is available which enables us to test a specific hypothesis where the sex of all the pedigree member are known such as in the special case 3 in this chapter, against a hypothesis of 'unrelated', MSY and mtDNA give more information than when we test a hypothesis where the sex on the pedigree is unknown. Use the special case 3 as an example: the likelihood ratio of this S-9-9 relationship over 'unrelated', based on them sharing a mtDNA haplogroup with frequency 0.001, is 1/0.001=1000 in this case. If we do not know the sex of the individuals on this S-9-9 pedigree and simply compare the two hypotheses that the two individuals of interest are S-9-9 related and that they are 'unrelated', the likelihood ratio is 1.015244 (based on method in Section 7.3.1). This example illustrates that we get greater information from MSY and mtDNA data in special cases where prior information designates the sex of the individuals on the pedigrees that are considered. But when this prior information is not available and the sex of the linking individuals between the two

individuals of interest is unknown, it is not justified to up-weight the likelihood ratio obtained from MSY or mtDNA. We have no way to know whether two individuals share a rare mtDNA simply by chance or because they share a distant ancestor. Note that this kind of uniparental pedigrees shown in Figure 7.2 are so rare that the probability that two individuals share this kind of relationships could even be less than the probability they share a rare MSY or mtDNA haplogroup by chance.

7.6 Summary

In this chapter, it was studied that how MSY and mtDNA haplogroups can be inferred from SNP data and a general way was shown to incorporate MSY and mtDNA SNPs in the likelihood method for relationship estimation as a supplement to autosomal SNP data. This study showed that they can help to detect distant relatives who may not be detectable using only autosomal data. Some other scenarios in which MSY and mtDNA SNPs can be useful are discussed as well. It can been seen that in some cases, MSY and mtDNA SNPs just provide extra information to autosomal DNA data, but in some cases they can solve problems which cannot be solved by autosomal DNA data. Since there is no extra cost to obtaining these data for many widely-used SNP chips, it makes sense to harness the information.

8 Reconstructing pedigrees from genetic data

So far only the relationships between two individuals have been considered although it is known that the pedigree-based method can easily incorporate more individuals. When data on more than two relatives are available, it is sometimes of interest to estimate their relationship altogether and reconstruct a pedigree for them. In traditional linkage analysis, pedigrees are needed to ascertain the mode of the trait of interest segregating in the family. The co-segregation of the trait and the putative causal genes in the pedigree is the foundation of model-based linkage analysis. Pedigree reconstruction is also needed in the situation where the remains of many individuals are found such as from a grave or a disaster scene, and DNA is the only information available to identify their relationships, e.g. the finding of the Romanov family (Gill et al., 1994, Coble et al., 2009). Another benefit of considering the relationships of many individuals jointly is that we may find relationships that cannot be found by pairwise estimation. Imagine a case where all individuals on an S-5-5 pedigree are genotyped (this may not be realistic for humans, but is for other species). The relationship of the two individuals at the bottom of the pedigree may be hard to find as there is a high chance that they do not share any DNA IBD. But when the whole pedigree is reconstructed, their relationship is known. In this chapter I consider how to find the relationships among a group of individuals by reconstructing a pedigree.

8.1 Maximum likelihood method

There have been many approaches to pedigree reconstruction. Most of them are based on maximum likelihood pedigree reconstruction approach in which a pedigree with the maximum likelihood for observed genetic data is sought. To date, most methods using this approach only work well on complete samples. A complete sample is a sample of individuals where either both parents of each individual are included in the sample, or this individual is a pedigree founder. A maximum likelihood method was first developed by Thompson (1976). Here I give a simple description of Thompson's method.

When everyone is observed, likelihoods for different hypothesized pedigrees can be easily calculated for unlinked markers by decomposing the pedigrees into sets of nuclear families which only include parents and children. The following formulae are presented by Thompson (1986). The likelihood of the 'basis pedigree', where all individuals are unrelated, is simply the product of the probabilities of all genetic data over individuals and loci:

$$\mathrm{P(data|all\ unrelated)} = \prod_{\mathrm{h}} \prod_{\mathrm{j}=1}^{\mathrm{s}} \mathrm{P(data\ on\ individual\ h\ at\ locus\ j)}. \quad (8.1)$$

This is not a plausible pedigree, but is just as a base point for reconstruction.

When all individuals are observed, the likelihood of a hypothesized pedigree is the product of the founder genotypes probabilities multiplied by the probabilities of genotypes of offspring given those of their parents:

$$\prod_{\mathrm{j}=1}^{\mathrm{s}} [\prod_{\mathrm{founders}} \mathrm{P(data\ on\ founder\ at\ locus\ j)} \prod_{\mathrm{nonfounders}} \mathrm{P(data\ on\ individual\ at\ locus\ j|parent\ data)}] \quad (8.2).$$

If we denote the mother and father of nuclear family i as M_i and F_i, and the children in the family as $\mathrm{C_{ik}}$ ($\mathrm{k} = 1, \ldots, \mathrm{i_k}$ is the $\mathrm{k^{th}}$ child in the family i), in the log-likelihood difference of the two equations (8.1) and (8.2), the terms corresponding to founders cancel and leave

$$\mathrm{Log[(2)]} - \mathrm{log[(1)]} = \sum_{\mathrm{j}=1}^{\mathrm{s}} \sum_{\mathrm{i}} \sum_{\mathrm{k}=1}^{\mathrm{i_k}} \mathrm{log\{P(G_j(C_{ik})|G_j(M_i), G_j(F_i))/P(\ G_j(C_{ik}))\}}, \quad (8.3)$$

where $\mathrm{G_j(C_{ik})}$ represents the genotype of individual $\mathrm{C_{ik}}$ at locus j.

Every hypothesized pedigree has a value corresponding to Equation (8.3). Thompson (1976) regarded pedigrees as a collection of sib-ships and aimed to reconstruct the pedigree by finding all compatible sib-ships in the pedigree. She started the pedigree reconstruction process from the 'basis pedigree' and sequentially updated it. For every new sib-ship, the putative parents which gave highest increase in value for Equation (8.3) would be selected. With her likelihood approach, the process for the pedigree reconstruction is a sequential acceptance of a set of sib-ships which gives the highest value for Equation (8.3). In order to guarantee the sib-ships are compatible, some extra information is needed. For example, one child cannot have two mothers or two fathers, so sex information of the individuals are needed. Similarly age information is needed to

exclude the possibility that an individual is his own grandfather, which is possible due to the symmetry of the parent-child relationship. A restriction on the size of sib-ship is also important as big sib-ship size tends to generate the largest value of Equation (8.3). There are some limitations to this method. Firstly, conditional on the non-exclusion of the sibling as a parent, the sibling often gives a higher likelihood than the true parent. This has been confirmed by my simulations later on in this chapter. Siblings usually will not be mistaken as parent-child as parents and children must share at least one allele in common. But when siblings happen to share at least one allele in all loci, although this becomes less likely when more markers are considered, they are often estimated as parent-child relationship. Secondly, a relationship falsely assigned at an earlier stage in the sequential procedure may result in the true pedigree never being found.

Almudevar (2003) extended the sequential algorithm of Thompson and developed an annealing algorithm for maximum likelihood pedigree reconstruction. He enumerated parent-offspring triplets as the first step, then assembled these triplets into a pedigree based on the maximization principle. This optimization problem was then reformulated to divide the set of all admissible pedigrees into subsets on which the likelihood is easily maximized. In his method, age and sex information are not always needed, but a complete sample is still required. Both sequential algorithm and annealing algorithm can only find a pedigree with high likelihood, but cannot guarantee to find the pedigree with maximal likelihood (Cussens et al., 2013). Riester et al. (2009) took an MCMC sampling approach to improve the accuracy of the pedigree reconstruction. Cowell (2009) adapted the Bayesian network learning algorithm and developed a dynamic programming exhaustive search algorithm. This algorithm is guaranteed to return a pedigree of highest likelihood, but it is computationally intensive and is feasible only for up to around 30 individuals. Cussens et al. (2013), Barlett and Cussens (2013) proposed a constraint-based integer programming approach to maximum likelihood pedigree reconstruction and it could handle much larger samples. In the next section, the method of Cussens et al. (2013) will be introduced in more detail as it seems to work best with a complete sample in terms of both ‘guaranteeing a highest likelihood estimate’ and ‘handling large numbers of individuals’.

Methods have also been proposed to construct sib-ships from a sample of individuals which includes siblings or half-siblings and unrelated individuals (Blouin, 2003, Jones and Wang, 2010a) without attempting to construct a pedigree. What these methods do is just partitioning a sample into sib-ships and they are not pedigree reconstruction strictly as they do not do multiple-generation pedigree reconstruction. Instead they can be regarded as a special case of clustering (Section 8.2). The fact that they only look for sib-ships, which is a very close relationship, makes their use in clustering limited although they could have applications in evolution and conservation of wild populations (Jones and Wang, 2010b, Carpenter et al., 2005, Gottelli et al., 2007).

8.1.1 Constraint-based integer linear programming approach in maximum likelihood pedigree reconstruction

The method of Cussens et al. (2013) works under the same assumptions as the previous algorithms: complete sample, complete genetic data, no mutation, no linkage and Mendelian inheritance and linkage equilibrium. The method is implemented in the software GOBNILP (www.cs.york.ac.uk/aig/sw/gobnilp/), which is free to be downloaded. The integer linear programming optimization problem is to find the values to some variables, which maximizes some objective functions while respecting all constraints. In order to use integer linear programming to do optimization, both the function to be optimized and the constraints need to be expressed as linear functions of a set of variables, some of which are integers, then off-the-shelf solvers can be used to find the solution. An important task is to formulate the pedigree reconstruction problem in the correct form.

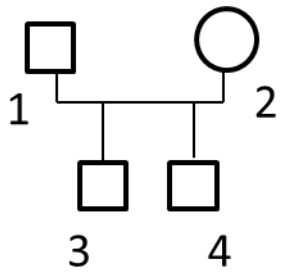
The probability that individual v has the observed genotype g_v on a pedigree G conditional of the genotype of his parent can be denoted as $\tau(v, Pa(v, G))$ where $Pa(v, G)$ denotes the parent set of v in pedigree G, which could include 2, 1, or 0 elements. When $Pa(v, G)$ is an empty set, $\tau(v, \emptyset)$ is equal to the marginal probability of individual v having the genotypes of g_v. Because a pedigree is completely specified by parent-child relationships, under the assumption of a complete sample and Mendelian inheritance, the probability of any single configuration of genotypes on a pedigree, and hence the likelihood, decomposes into a product of conditional probabilities and can be written as $L(G) = \prod_{v \in V} \tau(v, Pa(v, G))$, which is equivalent to

Equation (8.2). As it is often more convenient to work with log-likelihood, the optimisation problem is to find G which maximize the log-likelihood

$$\mathrm{l}(\mathrm{G}) = \log \mathrm{L}(\mathrm{G}) = \sum_{\mathrm{v}\in\mathrm{V}} \tau(\mathrm{v}, \mathrm{Pa}(\mathrm{v}, \mathrm{G})). \qquad (8.4)$$

In order to reformulate the above log-likelihood as an integer linear programming problem, binary indicator variables $\mathrm{I}(\mathrm{W} \to \mathrm{v})$ are created for each possible parent set W of individual v. Set $\mathrm{I}(\mathrm{W} \to \mathrm{v}) = 1$ if and only if W is the correct parent set for individual v and $\mathrm{I}(\mathrm{W} \to \mathrm{v}) = 0$ otherwise, while$\lceil \mathrm{W} \rceil \leq 2$ for diploid individuals who have at most two parents in the pedigree. For example, for a simple pedigree as shown in Figure 8.1, $\mathrm{I}(\emptyset \to 1) = 1$ because both parents of individual 1 are missing; $\mathrm{I}\{1,2\} \to 4 = 1$ because $\{1,2\}$ is the full set of the parents of individual 4 on the pedigree; $\mathrm{I}\{1\} \to 4 = 0$ because $\{1\}$ does not include all parent of individual 4 on the pedigree; $\mathrm{I}\{1,3\} \to 4 = 0$ because $\{1,3\}$ is not the correct parent set of individual 4 on the pedigree.

Figure 8.1 A pedigree used to illustrate how GOBNILP defines variables.



The log-likelihood shown in Equation (8.4) can be rewritten in terms of these binary variables as

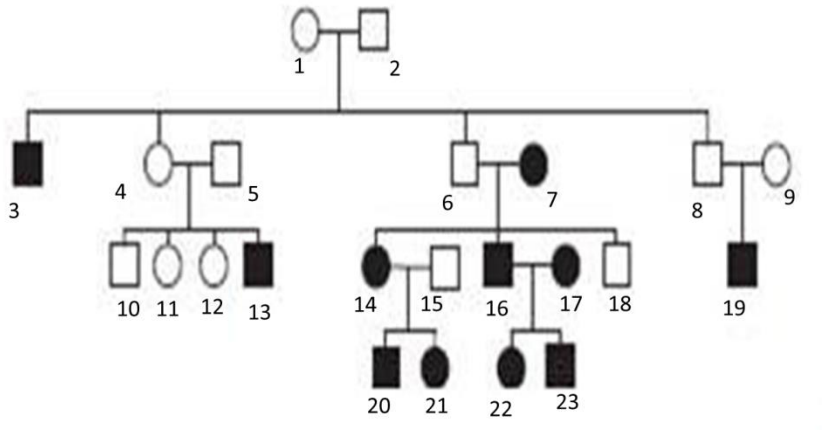
$$\mathrm{l}(\mathrm{G}) = \log \mathrm{L}(\mathrm{G}) = \sum_{\mathrm{v},\mathrm{W}} \tau(\mathrm{v}, \mathrm{W}) \mathrm{I}[(\mathrm{W} \to \mathrm{v})(\mathrm{G})] \,, \qquad (8.5)$$

where $\mathrm{I}[(\mathrm{W} \to \mathrm{v})(\mathrm{G})] = 1$ only when $\mathrm{W} = \mathrm{Pa}(\mathrm{v}, \mathrm{G})$. Now the maximum likelihood pedigree reconstruction problem becomes finding a set of $\mathrm{I}(\mathrm{W} \to \mathrm{v})$ to maximize the quantity in Equation (8.5), which is a linear function, subject to the constraint that the pedigree represented by such an instantiation is valid. A pedigree is valid if and only if every individual is included in the pedigree once; no-one is their own ancestor; every

individual only has one father and one mother. These constraints and some prior information, if available, can be simply expressed as linear functions.

In order to see the performance of the software GOBNILP and provide advice to the developers on how it behaves when the sample is not complete, I have done extensive testing on it for different situations. It is currently designed only for complete sample problems and has been shown to work well for that scenario (Cussens et al., 2013, Sheehan et al., 2014). I mainly tested it on pedigrees with missing individuals. When all individuals on the pedigree shown in Figure 8.2 (Day-Williams et al., 2011) are observed (both shaded and unshaded), this pedigree was reconstructed by GOBNILP correctly in all 100 replicates with just 30 STRs. But when there were missing individuals on a pedigree, some parts of the pedigree were estimated as unrelated to other parts of the pedigree if there was no data on the linking individuals. When some individuals had only one single parent observed, the direction of the parent-child relation could be wrong sometimes. Again for the pedigree shown in Figure 8.2, when only the shaded individuals were observed, individuals 7, 14, 16, 17, 20, 21, 22, 23 often form one pedigree while individuals 3, 13, and 19 are estimate unrelated to them in the results.

Figure 8.2 A pedigree used to test GOBNILP.



8.2 Clustering

Rather than reconstruct a pedigree for all individuals altogether, Cowell and Mostad (2003) proposed a divide-and–conquer approach for pedigree reconstruction. Firstly

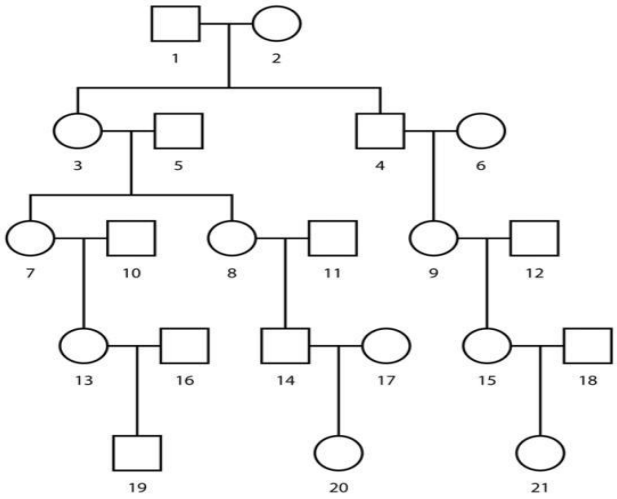
people are clustered into small groups of closely related individuals based on the distance of their pairwise relatedness. The distance criterion for clustering individuals can be adjusted as desired. The pedigree for each group is easier to reconstruct as the pedigrees are smaller. Then they suggested that these small pedigrees can be constructed together. However they did not propose any new method of reconstructing a pedigree either within each cluster or for connecting all clusters. So what they did is just clustering, rather than pedigree reconstruction although they used the term 'pedigree reconstruction'. Day-Williams et al. (2011) discussed clustering as well while they talked about doing linkage analysis with just clusters of relatives rather than pedigrees. The criterion used by Day-Williams et al. (2011) to cluster is the kinship coefficient, which is a mainstream pairwise relatedness measure while the distance measure of Cowell and Mostad (2003) is not quite the same thing and is not commonly used. I will concentrate on the work of Day-Williams et al. (2011) when discussing clustering individuals.

In Section 8.1.1, it was mentioned that when there are missing individuals in the pedigree, the order of the parent-child relationship reconstructed by GOBNILP can be wrong and two groups of individuals with a gap between them will be treated as two unlinked pedigrees. Therefore in this situation, GOBNILP can cluster individuals into groups that are all connected by parent-child links. I applied GOBNILP on an example presented in the paper of Day-Williams et al. (2011) and compared with their results.

Genotype data were simulated for two independent replicates of the pedigree shown in Figure 8.3. What Day-Williams et al. did is to cluster the two pedigrees using their method without constructing the pedigree. They calculated the pairwise kinship coefficients for all individuals and grouped them in one cluster if their pairwise kinship coefficient is greater than a threshold. Three different thresholds were used to do the clustering and their results are shown in Table 47 of the supplementary material of their paper (Day-Williams et al., 2011). They firstly considered the case where all individuals are observed, then moved to the case where individuals on positions of 7, 8, 9, 10, 11, 12 are unobserved. Basically this means the whole third generation is unobserved. Their results show that when all individuals are observed, they can cluster the two pedigrees correctly with all three different thresholds. When there are missing individuals and a low threshold is used, they can cluster individuals into the two

pedigrees correctly as well. When there are missing individuals and a high threshold is used, many small clusters are formed and distant relatives are grouped into different clusters.

Figure 8.3 A pedigree used to compare GOBNILP and clustering based on pairwise relatedness (Day-Williams et al., 2011).



I want to see whether GOBNILP could do the same or better. It is known that GOBNILP will do the clustering and reconstructing at the same time. As GOBNILP currently does not accept the pedigree IDs when there is more than one pedigree in the dataset and every individual needs to have a unique ID, I added number '100' before the original ID of the first pedigree and added '200' before the original ID of the second pedigree to make every individual have a unique ID number while their positions on the original pedigree are kept. The simulation was done for the case that all individuals are observed first, then for the case that the individuals on positions of 7, 8, 9, 10, 11 and 12 unobserved. 100 runs were carried out for both situations and for 15, 30 and 150 STR markers respectively.

The results can be summarized as follows. When all individuals are observed, the two pedigrees are correctly clustered and both are constructed correctly with just 15 markers. When the middle generation is missing, generally the top part and bottom part of each pedigree will form separate clusters and are not linked together. Therefore the clustering method of Day-Williams et al. (2011) has an advantage in that they can cluster distant relatives together even if there are missing individuals (with a lower threshold for clustering) which GOBNILP cannot do at present since it always seeks

parent-child edges. The advantage of GOBNILP is that when doing clustering, the pedigree within each cluster are reconstructed which the method of Day-Williams et al. (2011) cannot do.

8.3 Reconstructing pedigrees using pairwise relatedness

Most methods in literature on pedigree reconstruction are likelihood based, even for sib-ship construction. Pairwise relatedness has been utilized to find sib-ships only (Bentzen et al., 2001, Rodriguez-Ramilo et al., 2007), but not for multi-generation pedigree reconstruction, although the term 'pedigree reconstruction' has been used. With the clustering method in mind, I propose a method of reconstructing pedigrees based on pairwise relatedness. The reason that this method could work is based on the fact that with dense SNP markers, first-degree relatives (relatives who are one meiosis away from each other and share about 50 percent of their genes, such as parent, children and siblings) and second-degree relatives can be separated with minimal uncertainty. This method works on a large number of individuals which are not necessarily a complete sample and includes two clustering steps and one construction step.

The LR MoM estimator was used to calculate pairwise kinship coefficients between every pair of individuals in a sample and build a relatedness matrix. Remember that a pedigree determines a relatedness matrix, but a relatedness matrix does not determine a pedigree, e.g. a pedigree where 3 individuals all are siblings and a pedigree where 2 individuals are sibling and 1 individual is their parent, have the same relatedness matrix.

The purpose of the first clustering is to divide the whole sample into groups within which all individuals are connected together without a gap. The criterion of the first clustering is: put a new individual into an existing cluster as long as he has a first-degree relationship with at least one individual who is already in the cluster. A plausible threshold value for the kinship coefficient is 0.1875 which is the midpoint of 0.25 for a first-degree relationship and 0.125 for a second-degree relationship.

In order to construct the pedigrees for each cluster obtained from the first clustering, we need another clustering. The idea is to divide a cluster into 'units'. A unit is a special cluster in that pairwise relationships between any pair of its members are first-degree. So the criterion for the second clustering is: put a new individual into an existing 'unit' only if he has a first-degree relationship with **all** individuals who are already in the 'unit'. The same kinship coefficient of 0.1875 as in first clustering can be used. Each unit includes only one parent and all of his children because the pairwise relatedness between the two parents are generally very low in outbred population and do not meet the criterion of the second clustering. Then these units can be linked up to form a pedigree in the construction step by finding those individuals who are parents in one unit and children in another unit. The whole pedigree will be known as long as all 'units' are known.

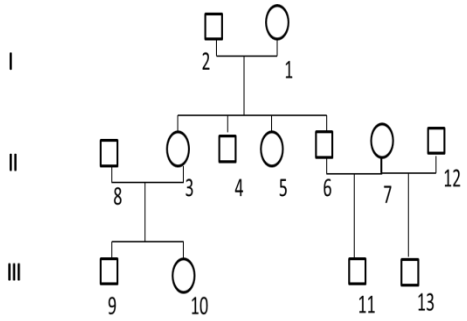
One important issue that has to be considered in the second clustering is whether we could distinguish first-degree relationships from second-degree relationships (where relatives share about a quarter of their genes) reliably. Blouin, *et al.* (1996) showed that with 20 unlinked microsatellite markers, they could distinguish half-sib pairs from full-sib pairs with accuracy of 80%. Bentzen, *et al*. (2003) reconfirmed the results. Some other papers have used pairwise relatedness to distinguish full-siblings from half siblings as well (Fernandez and Toro, 2006, Rodriguez-Ramilo et al., 2007), but they also used less than 100 STR markers. The rate of correct relationship assignment that they reported is not high enough as we need a much higher accuracy to be able to construct 'units' confidently. Only recently has it been shown that by using dense SNP markers the first-degree relationships can be distinguished from second-degree relationships. Day-Williams et al. (2011) have shown that with 500K SNPs, there is no single overlap in the estimated kinship coefficients for a S-1-1 relationship and a S-1-2 relationship out of 500 replicates. To confirm this, I did my own simulation with relationships S-1-1 and HS-1-1. I found that with 110,000 SNPs selected from Affymetrix 500K Array Set, first-degree relatives (represented by S-1-1) can be distinguished from second-degree of relatives (represented by HS-1-1) with accuracy of higher than 99%.

There are two ways to achieve this second clustering using the R software. 1) Set one individual up as the first 'unit'. Then check all other individuals respectively. For every

individual, if the kinship coefficients between him and **all** individuals existing in the current 'unit' are greater than our threshold, add him to the 'unit'; if he does not meet the criterion for any current 'unit', build a new cluster for him. 2) Set one cluster for each individual first. The number of 'units' will be the same as the number of individuals. Then for each cluster, search from all other individuals for those who meet the criterion and add them into that 'unit'. There will be repeated 'units' after this step because the exact same 'unit' will be constructed starting from every individual included in this 'unit'. The redundant clusters should be removed. This clustering will generate 'units' which include full siblings and one of their parents.

Next what needs to be done is to find who is the parent in the 'units' in the construction step. We could find who is the parent within this 'unit' with the help of pairwise k_2 (the probability that two individuals share two alleles IBD) and age information. We know that parent-child relationship has an expected k_2 of 0 while siblings have an expected k_2 of 0.25 although the expected k_1 (the probability that two individuals share one allele IBD) is the same for the two relationship. I did a simulation with 110K SNPs to see whether the parent-child relationship can be distinguished from the sib-sips with the estimated k_2. There is only 1 overlap for the k_2 estimate of the two relationships, PC-1 and S-1-1, out of 400 replicates. So PC1 and S-1-1 can be distinguished from each other with very high accuracy. Another way to distinguish relationship PC-1 from relationship S-1-1 is by the fact that parents and children should share at least one allele IBS at all loci. We could assign a relationship to be S-1-1 if the number of loci with genotypes inconsistent for a PC-1 relationship is greater than a threshold. Each 'unit' only includes one parent, but the number of children varies. If there is more than one child within a 'unit', the parent is the individual that has a parent-child relationship with everyone else in the 'unit'. But if there is only one child in a 'unit', then age information is needed to tell who is parent by symmetry of the parent-child relationship. Once the parent in each 'unit' is found, 'units' can be linked by individuals who are a parent in one 'unit' and a child in another 'unit', hence the pedigree can be reconstructed by linking 'units' together.

Figure 8.4 Example: A complete pedigree, no missing individual, no inbreeding.



8.3.1 Testing my pedigree reconstructing method based on pairwise relatedness

I have tested this method on the small pedigree shown in Figure 8.4 with simulated SNP and STR data and it performed well. 110, 000 SNPs instead of 500K were used because it has been shown that MoM estimators perform better when using 110, 000 than using 500K SNPs for close relationships and the speed of calculation is faster when less markers are used. 1000 STR markers were tested later to check whether we can use fewer STR makers to replace dense SNP markers because there are $n(n-1)/2$ kinship coefficients to calculate for a sample with n individuals and the speed of calculation is very important. The results shows that this method works very accurately on this small pedigree and 1000 STRs work almost as well as 110,000 SNPs.

Unlike most pedigree reconstruction methods that apply a likelihood method, this method is based on pairwise relatedness estimates only and does not need a complex program or software. It is convenient to be used for small pedigrees. Another difference with likelihood methods is the mechanism how they work: this method looks for parent and children altogether while likelihood methods tend to look for parent sets for each individual. It is quite similar to the sib-ship construction methods with pairwise relatedness. The difference is that I apply the pairwise relatedness to construct multi-generation pedigrees, rather than just construct sib-ships. Multi-generation pedigrees could be reconstructed only if there are parents included in the sample. If there is no parent in the sample, the resulting 'pedigrees' will just be sib-ships.

8.4 Incomplete sample

This pedigree reconstruction method based on pairwise relatedness may not be superior to a likelihood approach for 'complete sample' cases. But it provides another perspective to the pedigree reconstruction problem and has the potential to link clusters together when the 'complete sample' assumption is not satisfied. GOBNILP can do very well for 'complete sample' cases, even without the help of age or sex information. It can reconstruct a pedigree with 1614 individuals and 8 generations with high accuracy (Sheehan et al., 2014). But when there are missing individuals in a pedigree (incomplete sample), what we get is a set of sub-pedigrees. It is a difficult task to link these sub-pedigrees to a whole pedigree. Cowell and Mostad (2003) suggested that a likelihood method could be used to do this, but this has not been done so far. I am currently involved in updating GOBNILP to incorporate missing individuals and a paper based on it is expected soon. On an ad hoc basis, these sub-pedigrees can be linked by the pairwise relatedness estimates between the members of different sub-pedigrees, conditional on all pairwise relationships in the formed pedigree being consistent with the estimated degree of relatedness. We may not get a completely correct pedigree, especially when the sub-pedigrees are distantly related. But we can get an idea of the positions of the sub-pedigrees on the overall pedigree.

A similar approach using the software PRIMUS (Staples et al., 2014) was published at the same time as the work in this chapter was being developed. The idea is similar to what is mentioned in Section 8.3 in that it attempts to construct the pedigree based on pairwise IBD sharing and can handle incomplete samples. Instead of the ad hoc approach proposed here, it uses a program to search extensively for ways of linking individuals. Up to third-degree relatives (e.g. first cousin) are searched for every individual in the sample. Pedigrees in which the pairwise relationships between all individuals are consistent with their pairwise relatedness (k_0, k_1, k_2) are sought. By searching for relatives up to third-degree rather than just parents for every individual, PRIMUS could allow some missing individuals in the pedigree which cannot be done by maximum likelihood approaches currently. However, PRIMUS cannot distinguish between relationships which share the same (k_0, k_1, k_2), such as half-sibling and avuncular, first cousin and great avuncular. This is perhaps an issue inherent to all

approaches on reconstructing pedigrees by estimated pairwise relationships. Potentially this issue can be solved by a likelihood approach where the likelihood of all individuals is considered altogether using linked markers, although missing individuals have to be allowed to outperform PRIMUS.

8.5 Summary

In this chapter, firstly the maximum likelihood methods of reconstructing pedigrees were introduced. Next, the software GOBNILP, which can do a complete search over pedigree space and guarantee the return of a maximum likelihood pedigree, was introduced and evaluated. GOBNILP works well with very few genetic markers. But no method based on a likelihood approach can allow missing individuals in the pedigree. Then another approach to find clusters of relatives based on pairwise relatedness estimation was discussed. I found that it is possible to reconstruct pedigrees using simple pairwise IBD sharing estimates and indeed the recent approach of Staples et al 2014 confirmed this in a more formal way. With this approach, the problem of missing individuals can be handled to some extent. Currently work is ongoing, in which I am participating, to deal with the problem of missing individuals in pedigree reconstruction with likelihood methods. Substantial progress has been made and a relevant paper should be available soon.

9 Discussion

9.1 Summary of the thesis

This thesis has considered estimation of relatedness and relationships, especially detection of distant relatives, using genome-wide SNP data. Different approaches to relatedness estimation were examined and their performances compared. An approach based on the pedigree likelihood was given more attention due to some of its obvious advantages: it is straightforward to incorporate extra information and it can consider extra individuals. For instance, SNP data such as MSY and mtDNA which are usually available from the same chips used to genotype autosomal chromosomes can be easily incorporated to increase our chance of detecting unknown relatives. When additional individuals are genotyped, we can often estimate the relationship between the two individuals of interest more accurately.

This approach, which is originally used to distinguish one relationship from a specified set of alternatives relationships, can also be used as a method for estimating pairwise relatedness, or the degree of relationship. The exact relationship for many relative pairs can only be given by a pedigree. Non-pedigree-based approaches cannot do this as they only do pairwise estimation. The pedigree likelihood approach also yielded satisfactory results when used simply as a way of estimating the degree of relatedness. Indeed, using this approach, we can detect distant relatives more accurately than current published methods based on MoM estimators and IBD segment detection approaches. Although a pairwise relatedness estimate, such as by MoM, does not tell us the exact relationship between two individuals, lots of pairwise estimates over a group of relatives can sometimes suffice to reconstruct a pedigree.

LD is an issue which cannot be ignored when genome-wide SNP data are used to estimate relationship or relatedness. MoM methods are not badly affected by LD and generally perform well for close relatives, but they are not accurate in predicting distant relatives. The pedigree-based likelihood approach is seriously affected by LD, whether we use it to distinguish an exact relationship from putative relationships or to estimate

relatedness. The stronger the LD, the more biased the estimate is. However, LD can be modelled although it is hard to remove LD completely.

In Chapter 4, the use of the pedigree likelihood approach for distinguishing a true pedigree from alternative pedigrees was studied based on simulated data. It is well known that with as many as 500K SNPs we can distinguish relationships as distant as second cousins from 'unrelated' nearly with certainty but when the relationship is more distant we cannot make assertive inference. However this conclusion is based on the average posterior probability of the true relationship when the only alternative relationship is 'unrelated'. I found that this average posterior probability is misleading. After investigating the distribution of the posterior probabilities of the true relationships, I found that even if the relationship between the two relatives is much more distant than second cousin, we still have good power to distinguish them from 'unrelated' in some cases. That means that our discriminating power for specific distant relatives is, in fact, higher than implied by published findings based on average performances. Whether we can distinguish a relationship from 'unrelated' is largely dependent on whether the two individuals share IBD segments or not. Continuing to increase the number of SNPs from SNP data which are already quite dense will not necessarily improve the results. But it should be noted that although it may be easy to distinguish distant relationships from 'unrelated' when they do share IBD, it is far harder to distinguish them from a close alternative relationship.

I found that an extra individual which has a known relationship with one of the two individuals of interest always helps to estimate the relationship between the two individuals. The information provided by the extra individuals on different positions of the pedigree is different. It was recommended to use a sibling as the extra individual because siblings provide quite high information and are generally easier to find. More importantly, they always help, unlike parents and grandparents who help only if they are on the direct line of descent from the common ancestors of the two individuals of interest. But I admit that this finding applies more to cases where specific alternative pedigrees are hypothesized such as in a forensic or genetic counselling setting where the relationship of the third individuals can be known.

In Chapter 5, the effect of LD on the pedigree likelihood method was investigated. It was shown that when LD is not dealt with, the relationship will look much closer than the true relationship. When LD was modelled with a blocking approach, which is a standard way of modelling LD, the estimate was greatly improved in general, but the effect of LD could not be removed completely. Another simpler approach is to thin the density of the SNPs. It reduces the information unavoidably, but we can achieve a nearly unbiased estimate of the relationship, although we should be cautious about how many SNPs that should be used as the ideal number will vary for different situations. In this chapter, I also proposed using a template of pedigrees as a way of estimating the degree of relationship, or relatedness rather than distinguishing exact relationship. The performance of this method was compared with other pairwise relatedness estimating approaches in Chapter 6. The pedigree likelihood approach was applied to the real data of MICROS study and the results are generally consistent with what were seen for simulated data.

In Chapter 6, other methods of estimating relatedness without using pedigrees were introduced. I have investigated how MoM methods perform on dense genome-wide SNPs data as most of them were proposed in the time when only STRs were available. Basically, they can perform well for dense SNPs data and it seems the LD will not cause too much bias. For close relatives they can give good estimates and predict the realized kinship coefficients well. But the accuracy of MoM methods decreases quickly when the true relationship gets more and more distant. They will not give reliable results for relationships more distant than first cousins. My ‘Template’ method based on the pedigree likelihood works much more accurately in estimating distant relatives than MoM. Another approach for estimating relatedness is to detect IBD segments and then infer degree of relationship based on the number and length of the detected IBD segments, such as software ERSA. ERSA is quite accurate in estimating the degree of relationship, but it is more complicated to implement than the ‘Template’ method and it requires one more step of estimating IBD segments. For example, I did not run it on real inbred MICROS data successfully as I could not get any output of IBD segments from software Germline. On the other hand, ‘Template’ method seems more accurate than ERSA, but it could be slower than ERSA due to the large number of pedigree likelihood calculations. So it is dependent on the applications to decide which approach

should be used in reality. Population stratification can cause problems for all relationship and relatedness estimation methods which use allele frequencies. Care needs to be taken to check possible population structure before estimating relationships. Estimation should either be done in homogeneous populations or based on those non-ancestor-informative markers.

Chapter 7 was devoted to the incorporation of MSY and mtDNA SNP data into the relationship and relatedness estimation. I showed here how to infer haplogroups of MSY and mtDNA from SNPs obtained from common genotyping chips. It is straightforward to combine MSY and mtDNA SNP data with autosomal SNP data in the pedigree likelihood approach. Although information from MSY and mtDNA is limited in that each of them can be regarded as a single polymorphic genetic marker, I showed that we can increase the number of distant relatives that are detected with these extra data and usually there is no extra cost to obtain them. Sometimes they could be more useful when there is no or little autosomal DNA shared by two relatives but a rare MSY or mtDNA haplogroup is shared by them.

In Chapter 8, it was discussed how to estimate the relationships of more than two individuals jointly and reconstruct a pedigree. When all individuals of the pedigrees are genotyped, it is an easy problem. There are many methods proposed and they work very well for this scenario. However, when there are missing individuals it is more complicated. A large pedigree could be constructed as several small sub-pedigrees when there are missing individuals in it. Pairwise relatedness could be used to link these sub-pedigrees although the problem becomes harder when the number of missing individuals is large. There is software available based on pairwise IBD sharing to reconstruct pedigrees with a limited number of missing individuals (Staples et al., 2014).

In summary, relatedness and relationships can be inferred from genetic data. It was shown that with genome-wide SNP data we can solve more problems than with smaller numbers of markers (either SNPs or microsatellites). The approach based on the pedigree likelihood seems to be quite accurate for estimating the degree of pairwise relationship and is easy to implement. But when really dense SNPs are used, a good model for LD has to be found. It was found that with additional information such as an

extra genotyped individual, or MSY and mtDNA data, we can increase the accuracy of the relationship estimation. Other non-genetic prior information, such as age and sex, could be easily combined with genetic data in the pedigree likelihood method, although it was not investigated in this thesis. However, all methods are limited by the number of alleles shared IBD by individuals, which could vary substantially for the same relationship. Having the right allele frequencies or haplotype frequencies is also important.

One limitation to this study is that genotyping error has been ignored whose existence is possible in real data. Sieberts et al. (2002) have proposed a model dealing with genotyping error. However it seems that genotyping error, if there is any, in the real data that were used, MICROS, has not caused any noticeable problem to the work carried out in this study.

9.2 Future work

9.2.1 Develop software

The most immediate and relevant further work is to develop software which will implement the idea of using pedigrees to estimate relatedness as is proposed in the thesis. This software will take the genotypes data as input. Output could include the most likely degree of relationship and several other degrees of relationship with high likelihood together with the likelihood ratio of the most likely relationship over the two individuals being unrelated. Next, more pedigrees with real data can be used to test it head-to-head with ERSA, Beagle and other methods based on IBD segments. This approach has been shown to work well on simulated data and on the MICROS data. The pedigree in MICROS study violates the ‘outbred’ assumption in this approach, but it still works well. We could imagine it will work only better if more suitable outbred pedigrees can be obtained to test it.

9.2.2 Modelling LD

It is noticed that the current model for LD cannot remove the LD completely which makes the estimate of ‘Template’ method biased when really dense SNPs are used. As a compromise, SNP data have to be thinned. The problem with current LD models based on LD blocks is that they are too simplistic. They assume no recombination for

SNPs within LD blocks and no LD between LD blocks. If the LD block is too short, only part of the LD is modelled and there will still be LD between those markers which are not clustered. However if the LD block is too large, the assumption of no recombination within blocks is seriously breached. I think there are other possible approaches that we can take. One approach could be to allow large LD blocks and at the same time, allow recombination within these blocks with a reduced rate, although this is computationally intensive.

9.2.3 Applying methods of MSY and mtDNA on real data

All work in Chapter 7 is based on simulated data although some real relationships are considered. I am still waiting for some real data from my collaborators. Once these data are available, extra work can be done on them and a paper is planned to be written based on these methods.

9.2.4 Application in linkage analysis

One application of estimating relationship is in linkage analysis. Currently one problem faced by many genetic association studies is that genes that are found to be associated with complex traits often cannot be replicated. This is due to the fact that the effect of each single allele is so small in line with the hypothesis of ‘common disease, common allele’. Potentially there could be rare alleles with relatively large effects and there could be different alleles causing the same disease in different families. Such rare alleles will be hard to find in population based association studies of unrelated individuals. It is well accepted that linkage studies based on relatives have higher power than association analyses in finding rare alleles. Current association studies could contain more than ten thousand individuals. It is very likely that they will contain some relatives. It is of interest to apply the methods discussed in this thesis to these GWAS studies to find some relatives and construct pedigrees. Then linkage analyses can be conducted to investigate whether we can improve the power for mapping causal genes with current data without incurring extra cost.

9.2.5 Dealing with the problem of missing individuals in pedigree reconstruction combining pairwise estimates with GOBNILP

We know that GOBNILP could reconstruct a large pedigree with missing individuals into many sub-pedigrees. Each sub-pedigree generally can be reconstructed very accurately. It will be of interest to develop a method to link these sub-pedigrees systematically by pairwise relatedness estimates from methods, such as MoM.

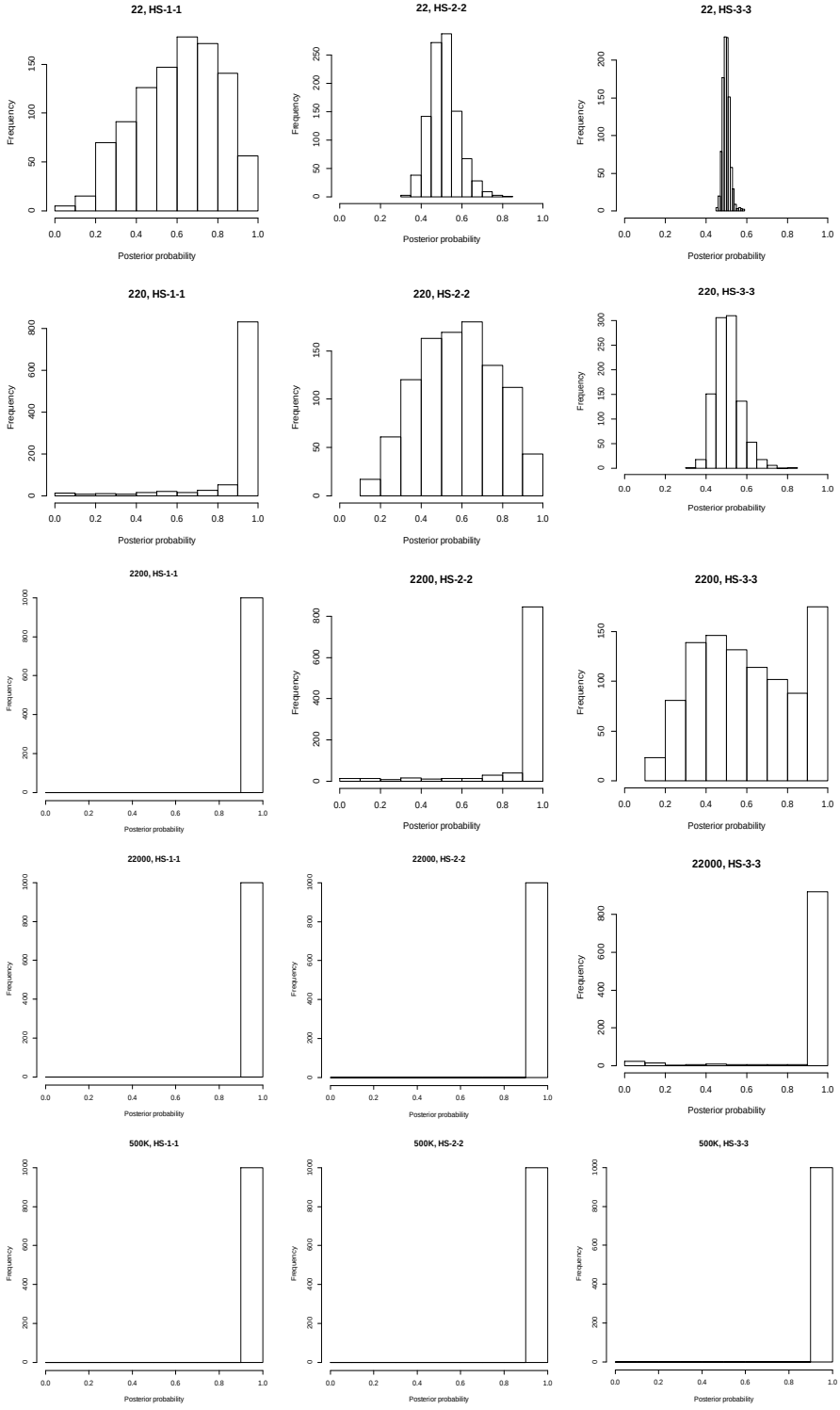
10 Appendix

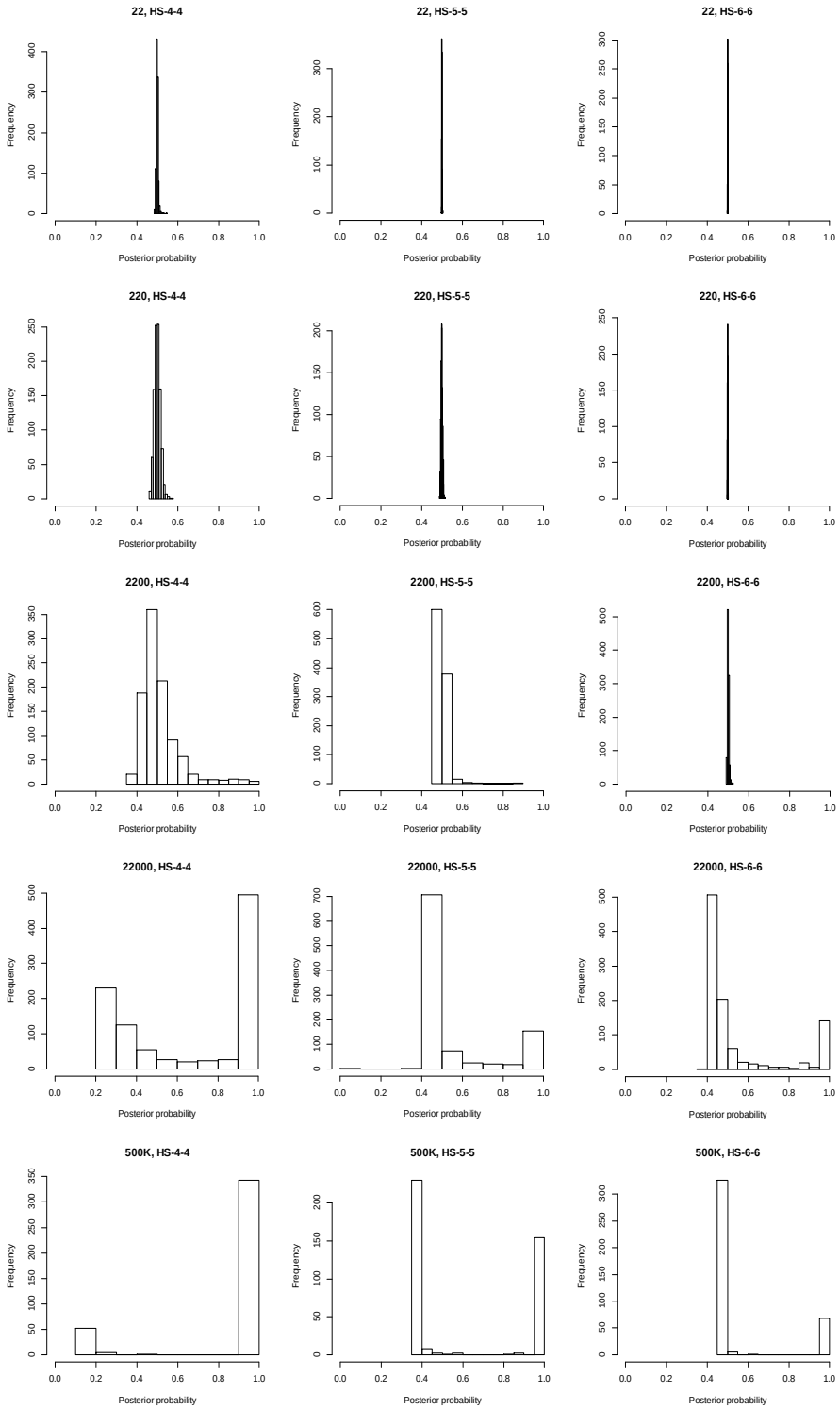
10.1 R code to calculate pedigree likelihood for an example by Lander-Green algorithm

```
q1=c(0.125,0.125,0.125,0.125)

q2 =c(0.09,0.09,0.09,0.09)

q3=c(0.1275,0.1275,0.1275,0.1275)

 q4=c(0.01,0.01,0.01,0.01); q5=c(0.24,0.24,0.24,0.24)

T=matrix(rep(0.25,16),nrow=4)

sum(rep(1/4,4)*q1%*%T*q2%*%T*q3%*%T*q4%*%T*q5)
```

10.2 A complete version of Figure 4.4

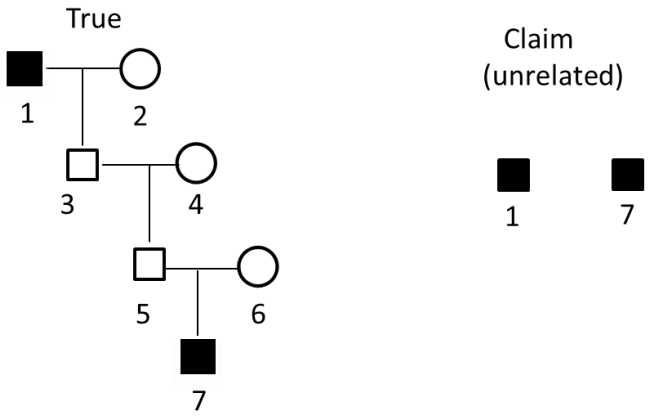
Figure 10.1 Frequency histograms of the individual posterior probabilities of the true pedigrees contributing to each of the averages reported in Table 4.1 where there is only one alternative pedigree: 'unrelated' for each true pedigree. The five rows correspond to the numbers of SNPs (22, 220, 2200, 22000, 500K) used in the simulation. The six columns (over two pages) correspond to the six different true pedigrees (HS-1-1, HS-2-2, HS-3-3, HS-4-4, HS-5-5, HS-6-6). The X-axis represents the posterior probability of the true pedigree.




22, HS-4-4
22, HS-5-5
22, HS-6-6
220, HS-4-4
220, HS-5-5
220, HS-6-6
2200, HS-4-4
2200, HS-5-5
2200, HS-6-6
22000, HS-4-4
22000, HS-5-5
22000, HS-6-6
500K, HS-4-4
500K, HS-5-5
500K, HS-6-6
Frequency
Posterior probability

10.3 Likelihood calculation when the extra individual is on ‘within1’ or ‘outside1’ position

Figure 10.2 Example pedigree for likelihood calculation when an extra individual is on ‘within1’ or ‘outside1’ position.



LR_U represents the likelihood ratio unconditional on the third individual.

LR_C represents the likelihood ratio conditional on the third individual.

$\mathrm{P(I=0)}$ represents the probability that two individuals share 0 IBD allele.

$\mathrm{P(I=1)}$ represents the probability that two individuals share 1 IBD allele.

1. When individual 1 is genotyped as 1/1 and individual 7 is genotyped as 2/2.

$LR_U=\frac{\mathrm{P(1,7|True)}}{\mathrm{P(1,7|Claim)}}=\frac{\mathrm{P(1,7|I=0)\times P(I=0)+P(1,7|I=1)P(I=1)}}{\mathrm{P(1)P(7)}}=\frac{p_1^2p_2^2(1-0.5^2)}{p_1^2p_2^2}=\frac{3}{4}$.

1) Assume individual 5, whose relationship with 7 is undisputed and on a ‘within1’ position, is genotyped as 1/2.

P(1,5,7|True)=P(1|True)P(5|1,True)P(7|5,True)= P(1,5|True)P(7|5,True)=$0.5p_2\times$ P(1,5|True)

$=0.5p_2\times(2p_1^3p_2\times P(I=0)+p_1^2p_2\times P(I=1))$

$=0.5p_2\times(2p_1^3p_2\times 0.5+p_1^2p_2\times 0.5)$

$=0.5p_1^2p_2^2(p_1+0.5)$

P(1,5,7|Claim)=P(1) ×P(5) ×P(7|5)= $p_1^2 \times 2p_1p_2 \times 0.5p_2 = p_1^3p_2^2$.

$LR_C = \frac{\text{P(1,5,7|True)}}{\text{P(1,5,7|Claim)}} = \frac{0.5p_1^2p_2^2(p_1+0.5)}{p_1^3p_2^2} = 0.5+\frac{0.25}{p_1} > \frac{3}{4} = LR_U$ as $0 < p_1 < 1$.

So when conditional on the genotype of the third individual 5, there is stronger evidence for the true pedigree.

2) Assume individual 6, whose relationship with 7 is undisputed and on an 'outside1' position, is genotyped as 1/2.

Because individual 5 is a parent of individual 7, his possible genotypes are 1/2 or 2/2 and we can calculate the probability P(1,6,7|True) conditional on the genotypes of individual 5 using the law of total probability.

P(1,6,7|True)= P(1,6,7|5= $'1/2'$,True) ×P(5= $'1/2'$)+ P(1,6,7|5= $'2/2'$,True) ×P(5= $'2/2'$)

= P(1,6,7,5= $'1/2'$,True)+ P(1,6,7,5= $'2/2'$,True)

P(1,6,7,5= $'1/2'$|True)=P(1,5= $'1/2'$|True) ×P(6|True) ×P(7|5= $'1/2'$,6,True)

= [$P(1,5 = '1/2'|I = 0$,True)× $P(I = 0)$+ $P(1,5 = '1/2'|I = 1$,True)× $P(I = 1)$] × P(6|True) ×P(7|5,6,True)

=(2×$p_1^3p_2 \times \frac{1}{2}$+$p_1^2p_2 \times \frac{1}{2}$) ×$2p_1p_2 \times \frac{1}{4}$

=$\frac{1}{4}p_1^3p_2^2(2p_1+1)$

P(1,6,7,5=$'2/2'$|True)=P(1,5=$'2/2'$|True)P(6|True)P(7|5=$'2/2'$,6,True)

= [$P(1,5 = '2/2'|I = 0$,True)× $P(I = 0)$+ $P(1,5 = '2/2'|I = 1$,True)× $P(I = 1)$] × P(6|True) ×P(7|5=$'2/2'$,6,True)

= ($p_1^2p_2^2 \times \frac{1}{2}$+0×$\frac{1}{2}$) ×$2p_1p_2 \times \frac{1}{2}$

= $\frac{1}{2}p_1^3p_2^3$.

Therefore P(1,6,7|True) = $\frac{1}{4}p_1^3p_2^2\times(2p_1+1)+\frac{1}{2}p_1^3p_2^3=\frac{1}{4}p_1^3p_2^2\times(2p_1+1+2p_2)=\frac{3}{4}p_1^3p_2^2$

P(1,6,7|Claim)=P(1) ×P(6) ×P(7|6)= $=p_1^2\times 2p_1p_2\times\frac{1}{2}p_2=p_1^3p_2^2$

$LR_C=\frac{\mathrm{P(1,6,7|True)}}{\mathrm{P(1,6,7|Claim)}}=\frac{3}{4}=LR_U.$

In this case, the individual 6 does not increase the information.

3) Assume individual 5, whose relationship with 7 is undisputed and on a 'within1' position, is genotyped as 2/2.

P(1,5,7|True)=P(1|True)P(5|1,True)P(7|5,True)= P(1,5|True)P(7|5,True)= $p_2\times$ P(1,5|True)

$=p_2\times(p_1^2p_2^2\times P(I=0)+0\times P(I=1))$

$=p_2\times(p_1^2p_2^2\times 0.5)$

$=0.5p_1^2p_2^3$

P(1,5,7|Claim)=P(1) ×P(5) ×P(7|5)= $p_1^2\times p_2^2\times p_2=p_1^2p_2^3$

$LR_C=\frac{\mathrm{P(1,5,7|True)}}{\mathrm{P(1,5,7|Claim)}}=\frac{0.5p_1^2p_2^3}{p_1^2p_2^3}=0.5<\frac{3}{4}=LR_U.$

In this case the data of individual 5 decreases the likelihood ratio of the true pedigree and the alternative pedigree. But we need to note that it is more likely to have the genotype of 1 / 2 than 2 / 2 for individual 5.

4) Assume individual 6, whose relationship with 7 is undisputed and on a 'outside1' position, is genotyped as 2/2.

P(1,6,7|True)=P(1|True)P(6|True)P(7|1,6,True)

$=p_1^2\times p_2^2\times(p_1\times p_2\times\frac{1}{2}+p_2\times\frac{1}{2}\times\frac{1}{2}+p_2\times\frac{p_2}{2}\times 1)$

$=p_1^2\times p_2^2\times\frac{3}{4}p_2$

$=\frac{3}{4}p_1^2p_2^3.$

P(1,6,7|Claim) = P(1) ×P(6) ×P(7|6) = $p_1^2 \times p_2^2 \times p_2 = p_1^2 p_2^3$

$LR_C = \frac{\mathrm{P(1,6,7|True)}}{\mathrm{P(1,6,7|Claim)}} = \frac{3}{4} = LR_U.$

In this case, the individual 6 does not increase the information as well.

It can be understood that when the two individuals do not share IBD (genotype of individual 1 is '1/1', genotype of individual 7 is '2/2' in this case), the third individual on 'outside' positions does not increase the likelihood ratio of true pedigree and 'unrelated'.

2. When individual 1 is genotyped as 1/1 and individual 7 is genotyped as 1/2.

$$LR_U = \frac{\mathrm{P(1,7|True)}}{\mathrm{P(1,7|Claim)}} = \frac{\mathrm{P(1,7|I=0)\times P(I=0)+P(1,7|I=1)P(I=1)}}{\mathrm{P(1)P(7)}} = \frac{2p_1^3 p_2 \times (1-0.5^2)+p_1^2 p_2 \times 0.5^2}{p_1^2 \times 2p_1 p_2}$$

$$= \frac{\frac{3}{2}p_1 + \frac{1}{4}}{2p_1} = \frac{3}{4} + \frac{1}{8p_1}$$

As this is a decreasing function of p_1, it can be seen that sharing an uncommon allele gives high likelihood ratio in favour of the true pedigree.

Assume individual 6, whose relationship with 7 is undisputed and on an 'outside1' position, is genotyped as 2/2.

P(1,6,7|True)=P(1|True)P(6|True)P(7|1,6,True)

$= p_1^2 \times p_2^2 \times (p_1 \times p_1 \times 1 + p_1 \times p_2 \times \frac{1}{2} + p_2 \times \frac{p_1}{2} \times 1 + p_2 \times \frac{1}{2} \times \frac{1}{2})$

$= p_1^2 \times p_2^2 \times (p_1^2 + \frac{1}{2}p_1 p_2 + \frac{1}{2}p_1 p_2 + \frac{p_2}{4})$

$= p_1^2 p_2^2 (p_1 + \frac{1}{4}p_2)$

P(1,6,7|Claim)=P(1) ×P(6) ×P(7|6)= $= p_1^2 \times p_2^2 \times p_1 = p_1^3 p_2^2$

$LR_C = \frac{\mathrm{P(1,6,7|True)}}{\mathrm{P(1,6,7|Claim)}} = \frac{p_1^2 p_2^2 (p_1 + \frac{1}{4}p_2)}{p_1^3 p_2^2} = 1 + \frac{p_2}{4p_1} = \frac{3}{4} + \frac{1}{4p_1} > \frac{3}{4} + \frac{1}{8p_1} = LR_U.$

In this case, the data of individual 6 increases the likelihood ratio between the true pedigree and the alternative pedigree.

10.4 MSY and mtDNA haplogroups inferred from 2987 unrelated controls of WTCCC2 dataset and their frequencies

Table 10.1 MSY haplogroups inferred from controls of WTCCC2 dataset and their frequencies.

E1b1_P181;	0.019125683
E1b1_P181;(rs9785815)	0.003415301
F(xG,IJ)	0.00068306
G_M201.	0.023907104
G_M201.*	0.00068306
I1(rs34626372)	0.020491803
I1.	0.090163934
I2*	0.00068306
I2.	0.028688525
I2a2_P217;S23.	0.010928962
I2a2a_P223;S117.	0.038251366
I2a2a1a_L126;S165.	0.005464481
J	0.011612022
J(rs1011954)	0.018442623
P	0.00204918
R*	0.00136612
R1_P236;	0.00068306
R1a	0.032103825
R1a1a_M512;	0.00068306
R1b1_(xR1b1a_L320)	0.00068306
R1b1a2a	0.019125683
R1b1a2a1a1_ L52;	0.665300546
R1b1a2a1a1_ P311;S128	0.00068306
R1b1a2a1a1_(rs1469371) Z278.	0.00136612
R1b1a2a1a1_(rs4044090)	0.00204918

YxK	0.00136612

Table 10.2 mtDNA haplogroups inferred from controls of WTCCC2 dataset and their frequencies.

A12	0.001339136
B4j	0.000669568
H	0.161365919
H1	0.145965852
H1+16311	0.002678273
H11a2a	0.002678273
H11a4	0.000334784
H13a2b2	0.002343488
H13a2c	0.000334784
H13c	0.000334784
H14b	0.000334784
H15a1b	0.001339136
H1ak1	0.000669568
H1an2	0.000334784
H1au	0.000669568
H1b1e	0.000669568
H1ba	0.001339136
H1bt1	0.000669568
H1bu	0.000334784
H1f+16093	0.00167392
H1h1	0.001004352
H1j7	0.000334784

H1n1b	0.002008704
H26b	0.000669568
H27+16093	0.00167392
H2a	0.005021761
H2a+152	0.001339136
H2a1a	0.001339136
H2a2	0.007030465
H2a2a	0.00167392
H2a2a1	0.006695681
H2a2a1d	0.000669568
H2a2b5a	0.000334784
H3+16311	0.006695681
H3ag1	0.004686977
H3ap	0.000669568
H3b1b1	0.000334784
H3h5	0.000334784
H3s	0.00167392
H3u1	0.001339136
H3v	0.001004352
H3v+16093	0.003013057
H3x1	0.000334784
H3y	0.000669568
H4	0.017073987
H5a1+16093	0.014395715
H5a1b	0.004352193

H5a1g	0.003347841
H5a1g1	0.000334784
H5a6	0.000669568
H5a6a	0.000334784
H5g	0.001004352
H5j	0.000334784
H61	0.001004352
H6a1b	0.008034818
H6a1b3a	0.001339136
H8a	0.001339136
H8c2	0.002008704
H96	0.001004352
HV	0.039169736
HV17	0.001004352
HV6	0.002008704
I1a1	0.005021761
I1d	0.020756612
I5a1b	0.000334784
I5a4	0.000334784
J	0.019417476
J1	0.067626381
J1b1a	0.00167392
J1b1a1	0.012721794
J1c1b1a	0.001339136
J1c2a3	0.00167392

J1c2b5	0.002343488
J1c2m	0.001004352
J1c3c	0.001004352
J1c3d	0.00167392
J1c3j	0.000334784
J1c8	0.000334784
J1c8a	0.006695681
J2a1a1a1	0.000669568
J2b1a6	0.000334784
J2b1c1	0.000334784
JT	0.000334784
K	0.024104453
K1	0.016739203
K1a1b1d	0.000334784
K1a4a1	0.01975226
K1a4a1a2b	0.001339136
K1b1+(16093)	0.008034818
K1b1a	0.003347841
K1b1c	0.007700033
K1b2b	0.00167392
K2a2a1	0.000334784
K2b1a1	0.001339136
L0a2	0.000334784
L1b1a	0.000334784
L3b	0.000334784

L3e	0.001339136
M	0.001004352
M12a1a2	0.000334784
M1a1e2	0.000669568
M30	0.000334784
M30b	0.000334784
M30d1	0.000334784
M34	0.001339136
M4\67"	0.000334784
M7	0.000334784
N1	0.000334784
N1a1	0.000669568
N1a1a	0.001339136
N1a3a2	0.000334784
N1a3a3	0.000334784
N1b	0.00167392
N2	0.000334784
N9a1	0.000334784
N9a3	0.000334784
R	0.004017409
R0a1a2	0.000669568
R0a1b	0.002008704
R0a2b	0.001004352
R12'21	0.000334784
R2b	0.002343488

R8a	0.000334784
T	0.039169736
T1a1c	0.000669568
T2	0.041848008
T2b4a	0.002678273
T2b6a	0.00167392
T2c1a1	0.000669568
T2e	0.003013057
T2f1a	0.001004352
U	0.001004352
U1b	0.000334784
U2'3'4'7'8'9	0.030800134
U2e1b1	0.000669568
U2e1f	0.000334784
U3a1	0.010043522
U4a2b	0.000334784
U4b1a2	0.000334784
U4b1b1a	0.000334784
U4b1b2	0.000334784
U5a2b2	0.057917643
U5a'b	0.027117509
U5b1c1a1	0.002343488
U5b1c2a	0.000669568
U5b1i	0.000669568
U5b2a1a	0.002343488

U5b2a1a+16311	0.000334784
U5b2a1a2	0.001004352
U5b2a1b	0.002343488
U5b2a2a2	0.000334784
U6	0.000334784
U6c	0.000334784
U6d1	0.000334784
U8	0.000669568
U8a	0.002343488
U8b1b	0.000334784
V10a	0.002343488
V2	0.00167392
V21	0.001004352
V3	0.001339136
W	0.013056579
W1c1	0.000334784
W1e1a	0.000334784
W3a	0.003013057
W6	0.000669568
X	0.011717442

11 Bibliography

ABECASIS, G. R., CHERNY, S. S., COOKSON, W. O., et al. 2002. Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet,* 30**,** 97-101.

ABECASIS, G. R. & WIGGINTON, J. E. 2005. Handling marker-marker linkage disequilibrium: Pedigree analysis with clustered markers. *American Journal of Human Genetics,* 77**,** 754-767.

ALMUDEVAR, A. 2003. A simulated annealing algorithm for maximum likelihood pedigree reconstruction. *Theor Popul Biol,* 63**,** 63-75.

ANDERSON, A. D. & WEIR, B. S. 2007. A maximum-likelihood method for the estimation of pairwise relatedness in structured populations. *Genetics,* 176**,** 421-40.

ASTLE, W. & BALDING, D. J. 2009. Population Structure and Cryptic Relatedness in Genetic Association Studies. *Statistical Science,* 24**,** 451-471.

BALLANTYNE, K. N., GOEDBLOED, M., FANG, R., et al. 2010. Mutability of Y-chromosomal microsatellites: rates, characteristics, molecular bases, and forensic implications. *American Journal of Human Genetics,* 87**,** 341-53.

BARLETT, M. & CUSSENS, J. 2013. Advances in Bayesian Network Learning using Integer Programming. *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence (UAI 2013)***,** 182-191.

BENTZEN, P., OLSEN, J. B., MCLEAN, J. E., et al. 2001. Kinship analysis of Pacific salmon: insights into mating, homing, and timing of reproduction. *J Hered,* 92**,** 127-36.

BERKOVIC, S. F., DIBBENS, L. M., OSHLACK, A., et al. 2008. Array-based gene discovery with three unrelated subjects shows SCARB2/LIMP-2 deficiency causes myoclonus epilepsy and glomerulosclerosis. *Am J Hum Genet,* 82**,** 673-84.

BINK, M. C., ANDERSON, A. D., VAN DE WEG, W. E., et al. 2008. Comparison of marker-based pairwise relatedness estimators on a pedigreed plant population. *Theor Appl Genet,* 117**,** 843-55.

BLOUIN, M. S. 2003. DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. *TRENDS in Ecology and Evolution,* 18**,** 503-511.

BOEHNKE, M. & COX, N. J. 1997. Accurate inference of relationships in sib-pair linkage studies. *American Journal of Human Genetics,* 61**,** 423-429.

BROWNING, B. L. & BROWNING, S. R. 2011. A Fast, Powerful Method for Detecting Identity by Descent. *American Journal of Human Genetics,* 88**,** 173-182.

BROWNING, S. R. & BROWNING, B. L. 2010. High-Resolution Detection of Identity by Descent in Unrelated Individuals. *American Journal of Human Genetics,* 86**,** 526-539.

BROWNING, S. R. & BROWNING, B. L. 2012. Identity by Descent Between Distant Relatives: Detection and Applications. *Annual Review of Genetics, Vol 46,* 46**,** 617-633.

CANNINGS, C., THOMPSON, E. A. & SKOLNICK, M. H. 1978. Probability Functions on Complex Pedigrees. *Advances in Applied Probability,* 10**,** 26-61.

CARPENTER, P. J., POPE, L. C., GREIG, C., et al. 2005. Mating system of the Eurasian badger, Meles meles, in a high density population. *Molecular Ecology,* 14**,** 273-284.

CHOI, Y., WIJSMAN, E. M. & WEIR, B. S. 2009. Case-Control Association Testing in the Presence of Unknown Relationships. *Genetic Epidemiology,* 33**,** 668-678.

COBLE, M. D., LOREILLE, O. M., WADHAMS, M. J., et al. 2009. Mystery solved: the identification of the two missing Romanov children using DNA analysis. *PLoS One,* 4**,** e4838.

COTTERMAN, C. W. 1940. A calculus for statistico-genetics. *Dissertation, Ohio State University*.

COWELL, R. G. 2009. Efficient maximum likelihood pedigree reconstruction. *Theor Popul Biol,* 76**,** 285-91.

COWELL, R. G. & MOSTAD, P. 2003. A clustering algorithm using DNA marker information for sub-pedigree reconstruction. *J Forensic Sci,* 48**,** 1239-48.

CUSSENS, J., BARTLETT, M., JONES, E. M., et al. 2013. Maximum likelihood pedigree reconstruction using integer linear programming. *Genet Epidemiol,* 37**,** 69-83.

DAY-WILLIAMS, A. G., BLANGERO, J., DYER, T. D., et al. 2011. Linkage Analysis Without Defined Pedigrees. *Genetic Epidemiology,* 35**,** 360-370.

DONNELLY, K. P. 1983. The Probability That Related Individuals Share Some Section of Genome Identical by Descent. *Theoretical Population Biology,* 23**,** 34-63.

ELSTON, R. C. & STEWART, J. 1971. A general model for the genetic analysis of pedigree data. *Human heredity,* 21**,** 523-42.

EVETT, I. & WEIR, B. S. 1998. *Interpreting DNA evidence : statistical genetics for forensic scientists,* Sunderland, Mass., Sinauer Associates.

FALCHI, M. & FUCHSBERGER, C. 2008. Jenti: an efficient tool for mining complex inbred genealogies. *Bioinformatics,* 24**,** 724-6.

FERNANDEZ, J. & TORO, M. A. 2006. A new method to estimate relatedness from molecular markers. *Mol Ecol,* 15**,** 1657-67.

FOSTER, E. A., JOBLING, M. A., TAYLOR, P. G., et al. 1998. Jefferson fathered slave's last child. *Nature,* 396**,** 27-8.

GABRIEL, S. B., SCHAFFNER, S. F., NGUYEN, H., et al. 2002. The structure of haplotype blocks in the human genome. *Science,* 296**,** 2225-2229.

GAZAL, S., SAHBATOU, M., PERDRY, H., et al. 2014. Inbreeding coefficient estimation with dense SNP data: comparison of strategies and application to HapMap III. *Hum Hered,* 77**,** 49-62.

GEPPERT, M., BAETA, M., NUNEZ, C., et al. 2011. Hierarchical Y-SNP assay to study the hidden diversity and phylogenetic relationship of native populations in South America. *Forensic Sci Int Genet,* 5**,** 100-4.

GILL, P., IVANOV, P. L., KIMPTON, C., et al. 1994. Identification of the Remains of the Romanov Family by DNA Analysis. *Nature Genetics,* 6**,** 130-135.

GORING, H. H. H. & OTT, J. 1997. Relationship estimation in affected rib pair analysis of late-onset diseases. *European Journal of Human Genetics,* 5**,** 69-77.

GOTTELLI, D., WANG, J. L., BASHIR, S., et al. 2007. Genetic analysis reveals promiscuity among female cheetahs. *Proceedings of the Royal Society B-Biological Sciences,* 274**,** 1993-2001.

GUDBJARTSSON, D. F., JONASSON, K., FRIGGE, M. L., et al. 2000. Allegro, a new computer program for multipoint linkage analysis. *Nature Genetics,* 25**,** 12-13.

GUSEV, A., LOWE, J. K., STOFFEL, M., et al. 2009. Whole population, genome-wide mapping of hidden relatedness. *Genome Res,* 19**,** 318-26.

HALDANE, J. B. S. 1934. Methods for the detection of autosomal linkage in man. *Annals of Eugenics,* 6**,** 26-65.

HALDANE, J. B. S. & SMITH, C. A. B. 1947. A new estimate of the linkage between the genes for colour-blindness and haemophilia in man. *Annals of Eugenics,* 14**,** 10-31.

HANSEN, H. E. & MORLING, N. 1993. Genetic Investigations in Immigration Cases and Frequencies of DNA Fragments of the Vntr Systems D2s44, D5s43, D7s21, D7s22, and D12s11 in Turks. *Forensic Science International,* 60**,** 23-35.

HEPLER, A. B. 2005. Improving forensic identification using Bayesian networks and relatedness estimation. *Dissertation, North Carolina State University, Raleigh*.

HILL, W. G. & ROBERTSON, A. 1968. Linkage disequilibrium in finite populations. *TAG Theoretical and applied genetics Theoretische und angewandte Genetik,* 38**,** 226-31.

HILL, W. G. & WEIR, B. S. 2011. Variation in actual relationship as a consequence of Mendelian sampling and linkage. *Genetics Research,* 93**,** 47-64.

HILL, W. G. & WHITE, I. M. 2013. Identification of pedigree relationship from genome sharing. *G3 (Bethesda),* 3**,** 1553-71.

HOLSINGER, K. E. & WEIR, B. S. 2009. Genetics in geographically structured populations: defining, estimating and interpreting F(ST). *Nat Rev Genet,* 10**,** 639-50.

HUFF, C. D., WITHERSPOON, D. J., SIMONSON, T. S., et al. 2011. Maximum-likelihood estimation of recent shared ancestry (ERSA). *Genome Res,* 21**,** 768-74.

IDURY, R. M. & ELSTON, R. C. 1997. A faster and more general hidden Markov model algorithm for multipoint likelihood calculations. *Human Heredity,* 47**,** 197-202.

JACQUARD, A. 1970. Structures G´en´etiques des Populations. *Masson & Cie, Paris.*

JOBLING, M. A. 2001. Y-chromosomal SNP haplotype diversity in forensic analysis. *Forensic Sci Int,* 118**,** 158-62.

JOBLING, M. A., PANDYA, A. & TYLER-SMITH, C. 1997. The Y chromosome in forensic analysis and paternity testing. *Int J Legal Med,* 110**,** 118-24.

JONES, O. R. & WANG, J. 2010a. COLONY: a program for parentage and sibship inference from multilocus genotype data. *Mol Ecol Resour,* 10**,** 551-5.

JONES, O. R. & WANG, J. 2010b. Molecular marker-based pedigrees for animal conservation biologists. *Animal Conservation,* 13**,** 26-34.

KALINOWSKI S.T., W. A. P., TAPER M.L. 2006. ML-Relate: a computer program for maximum likelihood estimation of relatedness and relationship. *Molecular Ecology Notes* 6**,** 576-579.

KARAFET, T. M., MENDEZ, F. L., MEILERMAN, M. B., et al. 2008. New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. *Genome Res,* 18**,** 830-8.

KAYSER, M., CAGLIA, A., CORACH, D., et al. 1997. Evaluation of Y-chromosomal STRs: a multicenter study. *Int J Legal Med,* 110**,** 125-33, 141-9.

KING, T. E., FORTES, G. G., BALARESQUE, P., et al. 2014. Identification of the remains of King Richard III. *Nature Communications,* 5.

KING, T. E. & JOBLING, M. A. 2009. What's in a name? Y chromosomes, surnames and the genetic genealogy revolution. *Trends Genet,* 25**,** 351-60.

KLING, D., WELANDER, J., TILLMAR, A., et al. 2012. DNA microarray as a tool in establishing genetic relatedness--Current status and future prospects. *Forensic Sci Int Genet,* 6**,** 322-9.

KLOSS-BRANDSTATTER, A., PACHER, D., SCHONHERR, S., et al. 2011. HaploGrep: A Fast and Reliable Algorithm for Automatic Classification of Mitochondrial DNA Haplogroups. *Human Mutation,* 32**,** 25-32.

KRUGLYAK, L., DALY, M. J., REEVEDALY, M. P., et al. 1996. Parametric and nonparametric linkage analysis: A unified multipoint approach. *American Journal of Human Genetics,* 58**,** 1347-1363.

KRUGLYAK, L. & LANDER, E. S. 1998. Faster multipoint linkage analysis using Fourier transforms. *Journal of Computational Biology,* 5**,** 1-7.

LANDER, E. S., CONSORTIUM, I. H. G. S., LINTON, L. M., et al. 2001. Initial sequencing and analysis of the human genome. *Nature,* 409**,** 860-921.

LANDER, E. S. & GREEN, P. 1987. Construction of Multilocus Genetic-Linkage Maps in Humans. *Proceedings of the National Academy of Sciences of the United States of America,* 84**,** 2363-2367.

LANGE, K., PAPP, J. C., SINSHEIMER, J. S., et al. 2013. Mendel: the Swiss army knife of genetic analysis programs. *Bioinformatics,* 29**,** 1568-70.

LEUTENEGGER, A. L., GENIN, E., THOMPSON, E. A., et al. 2002. Impact of parental relationships in maximum lod score affected sib-pair method. *Genetic Epidemiology,* 23**,** 413-425.

LEUTENEGGER, A. L., PRUM, B., GENIN, E., et al. 2003. Estimation of the inbreeding coefficient through use of genomic data. *Am J Hum Genet,* 73**,** 516-23.

LI, C. C., WEEKS, D. E. & CHAKRAVARTI, A. 1993. Similarity of DNA fingerprints due to chance and relatedness. *Hum Hered,* 43**,** 45-52.

LYNCH, M. & RITLAND, K. 1999. Estimation of pairwise relatedness with molecular markers. *Genetics,* 152**,** 1753-66.

MCPEEK, M. S. & SUN, L. 2000. Statistical tests for detection of misspecified relationships by use of genome-screen data. *American Journal of Human Genetics,* 66**,** 1076-1094.

MILLIGAN, B. G. 2003. Maximum-likelihood estimation of relatedness. *Genetics,* 163**,** 1153-67.

MIYAZAWA, H., KATO, M., AWATA, T., et al. 2007. Homozygosity haplotype allows a genomewide search for the autosomal segments shared among patients. *Am J Hum Genet,* 80**,** 1090-102.

MORRISON, J. 2013. Characterization and Correction of Error in Genome-Wide IBD Estimation for Samples with Population Structure. *Genetic Epidemiology,* 37**,** 635-641.

NELIS, M., ESKO, T., MAGI, R., et al. 2009. Genetic structure of Europeans: a view from the North-East. *PLoS One,* 4**,** e5472.

OLAISEN, B., STENERSEN, M. & MEVAG, B. 1997. Identification by DNA analysis of the victims of the August 1996 Spitsbergen civil aircraft disaster. *Nature Genetics,* 15**,** 402-405.

PALMER, L., BURTON, P. AND SMITH, G. 2011. *An introduction to genetic epidemiology,* Bristol, The Policy Press.

PATTARO, C. & SAINT-PIERRE, A. 2013. Family-based studies to the rescue of genome-wide association studies in renal function. *Kidney International,* 83**,** 196-198.

PATTERSON, N., PRICE, A. L. & REICH, D. 2006. Population structure and eigenanalysis. *PLoS Genet,* 2**,** e190.

PEMBERTON, T. J., WANG, C., LI, J. Z., et al. 2010. Inference of unexpected genetic relatedness among individuals in HapMap Phase III. *Am J Hum Genet,* 87**,** 457-64.

PENG, M.-S., HE, J.-D., FAN, L., et al. 2014. Retrieving Y chromosomal haplogroup trees using GWAS data. *Eur J Hum Genet,* 22**,** 1046-1050.

PURCELL, S., NEALE, B., TODD-BROWN, K., et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet,* 81**,** 559-75.

PURPS, J., SIEGERT, S., WILLUWEIT, S., et al. 2014. A global analysis of Y-chromosomal haplotype diversity for 23 STR loci. *Forensic Sci Int Genet,* 12**,** 12-23.

QUELLER, D. C. & GOODNIGHT, K. F. 1989. Estimating Relatedness Using Genetic Markers *Evolution,* 43**,** 258-275.

RIESTER, M., STADLER, P. F. & KLEMM, K. 2009. FRANz: reconstruction of wild multi-generation pedigrees. *Bioinformatics,* 25**,** 2134-9.

RITLAND, K. 1996. Estimators for pairwise relatedness and individual inbreeding coefficients *Genetics Research,* 67**,** 175-185.

RODRIGUEZ-RAMILO, S. T., TORO, M. A., MARTINEZ, P., et al. 2007. Accuracy of pairwise methods in the reconstruction of family relationships, using

molecular information from turbot (Scophthalmus maximus). *Aquaculture,* 273**,** 434-442.

SANTURE, A. W., STAPLEY, J., BALL, A. D., et al. 2010. On the use of large marker panels to estimate inbreeding and relatedness: empirical and simulation studies of a pedigreed zebra finch population typed at 771 SNPs. *Molecular Ecology,* 19**,** 1439-1451.

SHEEHAN, N. A. 2000. On the application of Markov chain Monte Carlo methods to genetic analyses on complex pedigrees. *International Statistical Review,* 68**,** 83-110.

SHEEHAN, N. A., BARTLETT, M. & CUSSENS, J. 2014. Improved maximum likelihood reconstruction of complex multi-generational pedigrees. *Theor Popul Biol,* 97**,** 11-9.

SIEBERTS, S. K., WIJSMAN, E. M. & THOMPSON, E. A. 2002. Relationship inference from trios of individuals, in the presence of typing error. *Am J Hum Genet,* 70**,** 170-80.

SKARE, O., SHEEHAN, N. & EGELAND, T. 2009. Identification of distant family relationships. *Bioinformatics,* 25**,** 2376-2382.

SOARES, P., ERMINI, L., THOMSON, N., et al. 2009. Correcting for purifying selection: an improved human mitochondrial molecular clock. *Am J Hum Genet,* 84**,** 740-59.

SPIELMAN, R. S., MCGINNIS, R. E. & EWENS, W. J. 1993. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *American journal of human genetics,* 52**,** 506-16.

STANKOVICH, J., BAHLO, M., RUBIO, J. P., et al. 2005. Identifying nineteenth century genealogical links from genotypes. *Human Genetics,* 117**,** 188-199.

STAPLES, J., QIAO, D. D., CHO, M. H., et al. 2014. PRIMUS: Rapid Reconstruction of Pedigrees from Genome-wide Estimates of Identity by Descent. *American Journal of Human Genetics,* 95**,** 553-564.

STRACHAN, T. R., A 2011. *Human Molecular Genetics,* Abingdon, Garland Sciencem,Taylor & Francis Group, LLC.

TALIUN, D., GAMPER, J. & PATTARO, C. 2014. Efficient haplotype block recognition of very long and dense genetic sequences. *Bmc Bioinformatics,* 15.

THOMPSON, E. A. 1976. Inference of Genealogical Structure. *Social Science Information Sur Les Sciences Sociales,* 15**,** 477-526.

THOMPSON, E. A. 1986. *Pedigree Analysis in Human Genetics,* Baltimore, The Johns Hopkins University Press.

THOMPSON, E. A. 2013. Identity by Descent: Variation in Meiosis, Across Genomes, and in Populations. *Genetics,* 194**,** 301-326.

THORNTON, T. & MCPEEK, M. S. 2010. ROADTRIPS: Case-Control Association Testing with Partially or Completely Unknown Population and Pedigree Structure. *American Journal of Human Genetics,* 86**,** 172-184.

THORNTON, T., TANG, H., HOFFMANN, T. J., et al. 2012. Estimating Kinship in Admixed Populations. *American Journal of Human Genetics,* 91**,** 122-138.

WANG, J. L. 2002. An estimator for pairwise relatedness using molecular markers. *Genetics,* 160**,** 1203-1215.

WEIR, B. S., ANDERSON, A. D. & HEPLER, A. B. 2006. Genetic relatedness analysis: modern data and new challenges. *Nature Reviews Genetics,* 7**,** 771-780.

WIJSMAN, E. M. 2012. The role of large pedigrees in an era of high-throughput sequencing. *Human genetics,* 131**,** 1555-1563.

WRIGHT, S. 1922. Coefficients of inbreeding and relationship. *The American Naturalist* 56**,** 330-338.

WTCCC2. 2008. Available: http://www.wtccc.org.uk/ccc2/.

XUE, Y., WANG, Q., LONG, Q., et al. 2009. Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. *Curr Biol,* 19**,** 1453-7.