NTNU – Trondheim
Norwegian University of
Science and Technology

Statistical Methods for Multiple Testing in Genome-Wide Association Studies

Kari Krizak Halle

Master of Science in Physics and Mathematics
Submission date: June 2012
Supervisor: Mette Langaas, MATH

Norwegian University of Science and Technology
Department of Mathematical Sciences

Sammendrag

I genetiske assosiasjonsstudier ønsker man å studere mulige sammenhenger mellom genetiske markører og sykdom. For hver genetiske markør utføres en hypotesetest. Siden antallet genetiske markører er stort (i størrelsesorden hundretusener) snakker vi her om fagfeltet multippel testing. En populær strategi i multippel testing er å estimere et effektivt antall tester og deretter bruke metoder basert på uavhengige tester for å kontrollere den totale type I feilen. Fokuset i denne masteroppgaven har vært å studere ulike metoder for å estimere effektivt antall uavhengige tester. Metodene har blitt anvendt på et stort datasett fra TOP studien ved Universitetet i Oslo og Oslo Universitetssykehus der man har studert sykdommene schizofreni og bipolar lidelse. Korrelasjon mellom de genetiske markørene er sentral i de ulike metodene, og i denne masteroppgaven har vi studert metoder basert på enten haplotype eller genotype korrelasjon mellom markørene.

Abstract

In Genome-Wide Association Studies (GWAS) the aim is to look for association between genetic markers and phenotype (disease). For each genetic marker we perform an hypothesis test. Since the number of markers is high (in the order of hundred thousands), we use multiple hypothesis tests. One popular strategy in multippel testing is to estimate an effective number of independent tests, and then use methods based on independent tests to control the total type I error. The focus of this thesis has been to study different methods for estimating the effective number of independent tests. The methods are applied to a large data set on bipolar disorder and schizophrenia in Norwegian individuals from the TOP study at the University of Oslo and Oslo University Hospital (OUS). A key feature of these methods is the correlation between the genetic markers. The methods considered in this thesis are based on either haplotype or genotype correlation and one focus of this thesis has been to study the difference between haplotype and genotype correlation.

Contents

Preface

This thesis completes my Master in Industrial Mathematics at the Norwegian University of Science and Technology. In this thesis, statistical models for GWAS data are presented and applied to a real data set, and I would like to tank Professor Dr. med Ole A. Andreassen at the TOP study for making the TOP8 data set available to me.

I would also like to thank my cousin Thomas Løften and my classmate Christian Page for proofreading this thesis.

This work has been really exciting and a great learning experience for me and I wish to continue working within biostatistics. I would like to thank my supervisor, Associate Professor Mette Langaas at Department of Mathematical Sciences at NTNU, for inspiration, motivation and excellent guidance.

Trondheim, June 7, 2012

Kari Krizak Halle

Chapter 1

Introduction

A Genome-Wide Association Study (GWAS) is used to identify genetic variations that may have influence on health and disease. A GWAS include scanning the complete set of DNA of many people with the goal to find genetic variations that are associated with a particular disease (National Human Genome Research Institute 2011). Association between genetic variants and disease can be assessed using hypothesis testing. The data analyzed are often available as genotype data, but commonly used tests are based on haplotype data.

In a GWAS, many hypotheses need to be evaluated, and therefore the generalization of the theory for single hypothesis testing to multiple hypotheses testing is of importance. Resampling procedures are considered as the gold standard in multiple testing problems within this field. One approximation to resampling procedures is the use of the Šidák method for independent tests and define an estimate of the effective number of independent tests. Several researchers have worked on this problem. In this thesis, different methods for estimating the effective number of independent tests will be considered and tested on a data set from the TOP study (TOP 2012*c*). To compare with the gold standard, we will also use the minP resampling procedure to control the familywise error rate, FWER.

In Chapter 2 we present some background in biology. The data analyzed in this thesis is presented in Chapter 3. Since data are available as genotype data and many commonly used methods are based on haplotype data we have in Chapter 4 and 5 compared haplotype and genotype correlation for both a theoretical grid and a real data set. Chapter 6 and 7 will focus on hypothesis testing in general and on the theory for multiple testing. Methods for estimating the effective number of independent tests are presented in Chapter 8, and applied to the TOP data in Chapter 9. In Chapter 10, applications for the whole genome are discussed. Finally, the thesis ends with discussion and conclusion in Chapter 11.

Chapter 2

Background in biology

2.1 DNA

A DNA molecule is built up of two intertwined chains which form a double helical structure. The chains consists of nucleotides, which contains a phosphate group, a deoxyribose sugar molecule and one of four nitrogenous bases. The four possible nitrogenous bases in a DNA molecule are adenine, thymine, cytosine and guanine, and they are usually named only with a capital letter, A, T, C or G (Griffiths, Gelbart, Lewontin & Miller 2002, p. 4). The two nucleotide chains that form the helix structure are held together by weak bonds between one base from each chain. The two bases connected by weak bonds forms different base pairs. There are only two different base pairs in the DNA molecule, A-T and C-G, because between these bases there are only two possibilities for weak bonds to occur.

A genome is the total amount of DNA in an organism, built up of long DNA molecules. Human cells contains in total 46 chromosomes, which form 23 pairs of chromosomes, and each chromosome carry a different set of genes. A gene is a region of the chromosomal DNA that is involved in the production of proteins, and a gene contains information for one protein. A protein is built up of a chain of amino acids, which is called a polypeptide (Griffiths et al. 2002, p. 5). Any gene may exist in different form in different individuals.

2.2 SNP - Single nucleotide polymorphisms

A single nucleotide polymorphism (SNP) is a variation in a DNA sequence (Human Genome Project Information 2011). This variation occurs when a single nucleotide in the DNA sequence is changed. For example when the base adenine is altered with the base thymine. A SNP will change a subsequence of the DNA, for example if the DNA sequence AAGGCTAA is changed to ATGGCTAA, i.e. the second base adenine is altered with thymine, we see that we have a SNP in this DNA sequence. The entire human genome consists of about $3 \cdot 10^9$ bases, and the SNPs occurs often, about every 100 to 300 bases along the entire genome. A variation in a DNA sequence, when a single base is altered must occur in at least 1% of the population to be considered as a SNP. In about 2/3 of the SNPs the two bases that are altered are cytosine and guanine. Some of the SNPs may have influence on the risk for a person to develop a particular disease. Some of the SNPs occur in non-coding regions of the genome, which means a region of the genome that does not code for production of proteins. The SNPs that occur in the coding regions of the genome may influence genes that are involved in production of proteins, and may then have some influence in the risk for getting different diseases.

2.3 Biological definitions

Allele One particular gene may exist in different forms in different individuals. Alleles are different forms of the same gene that can exist at a particular locus (Griffiths et al. 2002, p. 654).

Gamete A gamete is a reproductive cell with haploid chromosome number (Thompson & Thompson 1980, p. 353). This means that they consist of only one copy of each chromosome.

Genotype A genotype is an unordered set of alleles present at one locus (Thompson & Thompson 1980, p. 353). For a locus with alleles A and a, the possible genotypes are AA, Aa and aa.

Haplotype A haplotype is an ordered set of alleles from closely linked loci. The alleles in a haplotype are usually inherited together (Thompson & Thompson 1980, p. 353). For a person having alleles A at one locus and b at an neighboring locus, the haplotype is denoted Ab.

Hardy-Weinberg equilibrium When the frequency distribution of the genotypes AA, Aa and aa is stable at $p^2, 2pq$ and q^2, the locus is in Hardy-Weinberg equilibrium. (Griffiths et al. 2002, p. 564).

Locus A locus is the position of a gene on a chromosome (Thompson & Thompson 1980, p. 157).

Minor allele frequency (MAF) Minor allele frequency (MAF) is the frequency of the rarer allele (Ziegler & König 2010, p. 98).

Phenotype Phenotypes are groups that are used for characterization of organisms by physiology (Griffiths et al. 2002, p. 7). Examples of phenotypes are “blue eyes” and “blood type B”.

2.4 Hardy-Weinberg disequilibrium

When allele and genotype frequencies are estimated it is of interest to look for non-random association between the the two alleles at a given locus. For a locus X with alleles A and a, we define an indicator variable as (Weir 2008)

$$X_i = \begin{cases} 1 & \text{if allele is A} \\ 0 & \text{if allele is a} \end{cases}, \quad i = 1, 2$$

where the subscript $i = 1, 2$ denote the first and second gamete at the locus, respectively.

Using this indicator variable, we define the following probability

$$P(X_i = 1) = p_A, \quad i = 1, 2,$$

and we have the expected values

$$\begin{aligned} E(X_i) &= p_A, \quad i = 1, 2 \\ E(X_1 X_2) &= P_{AA}. \end{aligned} \tag{2.1}$$

The variance of the random variable X_i is then

$$\text{Var}(X_i) = p_A(1 - p_A). \tag{2.2}$$

From Equation (2.1) and (2.2) we get

$$\begin{aligned} \text{Cov}(X_1, X_2) &= P_{AA} - p_A^2 \\ \text{Corr}(X_1, X_2) &= \frac{P_{AA} - p_A^2}{p_A(1 - p_A)}. \end{aligned}$$

The correlation $\text{Corr}(X_1, X_2)$ is referred to as the within-population inbreeding coefficient f_A (Weir 2008)

$$\text{Corr}(X_1, X_2) = \frac{P_{AA} - p_A^2}{p_A(1 - p_A)} = f_A. \tag{2.3}$$

Rewriting Equation (2.3) we observe that the genotype frequencies in the general case then can be parameterized as

$$\begin{aligned} P_{AA} &= p_A^2 + f_A p_A p_a \\ P_{Aa} &= 2p_A p_a(1 - f_A) \\ P_{aa} &= p_a^2 + f_A p_A p_a. \end{aligned} \tag{2.4}$$

The upper and lower bound for the inbreeding coefficient, f_A, are found from Equation (2.4) as

$$\begin{aligned} P_{AA} &= p_A^2 + f_A p_A p_a > 0 \\ P_{Aa} &= 2p_A p_a(1 - f_A) > 0 \\ P_{aa} &= p_a^2 + f_A p_A p_a > 0. \end{aligned}$$

Rewriting these inequalities gives the bounds for f_A as (Weir 2008)

$$-\min(p_A/p_a, p_a/p_A) \leq f_A \leq 1.$$

When we assume random mating in a very large population, genotype frequencies are the products of allele frequencies. Let the two alleles at one locus be A, a, the expected genotype frequencies under random mating is then given by

$$\begin{aligned} P_{AA} &= p_A^2 \\ P_{Aa} &= 2p_A p_a \\ P_{aa} &= p_a^2. \end{aligned}$$

Hardy-Weinberg disequilibrium describes departures from these frequencies, and can be described using a disequilibrium coefficient denoted D_A, given by Weir (2008)

$$D_A = f_A p_A(1 - p_A).$$

Equation (2.4) can then be rewritten as

$$\begin{aligned} P_{AA} &= p_A^2 + D_A \\ P_{Aa} &= 2p_A p_a - 2D_A \\ P_{aa} &= p_a^2 + D_A. \end{aligned}$$

The upper and lower bound for the Hardy-Weinberg disequilibrium coefficient D_A is found from the following inequalities

$$\begin{aligned} P_{AA} &= p_A^2 + D_A \geq 0 \\ P_{Aa} &= 2p_A p_a - 2D_A \geq 0 \\ P_{aa} &= p_a^2 + D_A \geq 0. \end{aligned}$$

These inequalities gives

$$\begin{aligned} p_A^2 + D_A &\geq 0 \\ D_A &\geq -p_A^2 \\ p_a^2 + D_A &\geq 0 \\ D_A &\geq -p_a^2, \end{aligned}$$

and then

$$D_A \geq -\min\{-p_A^2, p_a^2\}.$$

The upper bound for D_A are found from the inequality

$$2p_Ap_a - 2D_A \geq 0,$$

which gives the inequality

$$D_A \leq p_Ap_a.$$

The upper and lower bounds for the Hardy-Weinberg disequilibrium coefficient, D_A, can be summarized as

$$\max\{-p_A^2, -p_a^2\} \leq D_A \leq p_Ap_a.$$

We see that the disequilibrium coefficient D_A depends on the allele frequencies and the maximal range for D_A is $[-0.25, 0.25]$ since the maximum product of allele frequencies is obtained when $p_A = p_a = 0.5$.

Chapter 3

The TOP study

3.1 TOP

The Thematic Organized Psychosis Research study (TOP study) was started at the University of Oslo (UIO) in 2003 (TOP 2012*c*), and Professor Dr.med Ole A. Andreassen is the head of the study (TOP 2012*b*). The goal of the TOP study is to obtain information about the causes for severe mental disorders with focus on schizophrenia and bipolar disorder. In 2012 the TOP project was appointed K.G. Jebsen Centre for Psychotic Research (TOP 2012*b*). The K.G. Jebsen Centre for Psychotic Research is a cooperation project between the University in Oslo (UiO), the University in Bergen (UiB) and Oslo University Hospital (OUS). The centre has different projects and partners both in Norway and abroad.

The TOP study started in 2003 including patients from the University Hospitals in Oslo in (TOP 2012*c*). Today the database also includes individuals from other parts of the country, about 1100 individuals with disease and around 500 healthy individuals in a control group (TOP 2012*a*). The information about the data in the TOP study are collected in different ways, in the clinic, neurophysological tests, MR and genetic analysis (TOP 2012*a*).

3.2 Schizophrenia and bipolar disorder

The lifetime risk of the severe mental disorder schizophrenia is nearly 1% (Athanasiu, Mattingsdal, Kähler, Brown, Gustadsson, Agartx & et. al 2010). Persons that are affected with schizophrenia may hear voices that other people do not hear and can believe that other persons are able to read their minds or to control their thoughts (National Institute of Mental Health 2012*b*).

Bipolar disorder (BD) is a severe mental illness (Djurovic, Gustafsson, Mattingsdal, Athanasiu, Bjella, Tesli & et. al 2010) with also is known as a manic-depressive illness (National Institute of Mental Health 2012*a*). At least half of the individuals affected with bipolar disorder develop the disease before age 25. The first symptoms of a bipolar disorder may be misunderstood as symptoms of many separate problems, not as a part of a larger problem or disorder (National Institute of Mental Health 2012*a*).

3.3 The TOP8 data

We have been given permisson by Professor Dr.med Ole A. Andreassen at the TOP study, to analyze the TOP8 data set. The TOP8 data set consists of all samples from the previous TOP studies, including studies in schizophrenia (Athanasiu et al. 2010) and bipolar disorder (Djurovic et al. 2010). The data set contains as shown in Table 3.1 data for a total number of 1551 individuals, 1124 individuals with disease and 417 individuals not affected by schizophrenia or bipolar disorder. Among the 1551 individuals the disease status is missing for ten of the individuals. Among the individuals in the study, there are 770 males and 780 females, and missing values for ten of the individuals.

The cases in the TOP study had to satisfy some predefined criteria (TOP 2012*a*). These criteria are

- Psychotic disorder
- Age 16-65
- The disease are not caused by organic disease or drugs
- The patient must be able to give informed consent
- The patient must be able to speak and understand a Scandinavian language

The sample analyzed in the TOP study was genotyped using Affymetrix Genome-Wide Human SNP Array 6.0 (Athanasiu et al. 2010). The preprocessing of the genotype data includes removing of individuals and SNPs with high percentage of missing genotype data. All SNPs with minor allele frequency below 1% was removed from the study. The SNPs in the study was also tested for Hardy-Weinberg disequilibrium removing all SNPs with p-value < 0.01.

The aim of the data analysis in this thesis is not to arrive at medical findings, but to use real data to compare different methods. This has proven especial impact

Table 3.1: The TOP8 study

Sex		Affected / Unaffected	
Male	770	Affected	1124
Female	780	Unaffected	417
Missing	1	Missing	10
Total	1551	Total	1551

for the result for comparing the LD and CLD correlation which is described in Chapter 4 and 5.

Chapter 4

Linkage disequilibrium and correlation

In this chapter, correlation between SNPs based both on haplotypes and genotypes will be presented. Haplotype correlation will be presented in Section 4.1 and genotypecorrelation will be presented in Section 4.2. Haplotype and genotype correlation will be compared theoretically in Section 4.3.

We consider two biallelic loci X and Y with alleles A, a and B, b respectively. We define two random variables, X_i and Y_i, describing the alleles at locus X and Y respectively, and let subscript $i = 1$ indicate the first gamete at the locus, and subscript $i = 2$ indicates the second gamete. The random variables are illustrated in Figure 4.1. We define the random variables as (Weir 2008)

$$X_i = \left\{ \begin{array}{ll} 1 & \text{if allele is A} \\ 0 & \text{if allele is a} \end{array} \right. , \quad i = 1, 2 \tag{4.1}$$

and

$$Y_i = \left\{ \begin{array}{ll} 1 & \text{if allele is B} \\ 0 & \text{if allele is b.} \end{array} \right. , \quad i = 1, 2. \tag{4.2}$$

For the two random variables in Equation (4.1) and (4.2) we define six probabilities that describes frequencies of alleles and combinations of alleles at different loci and on different gametes. These probabilities are defined as (Weir 2008)

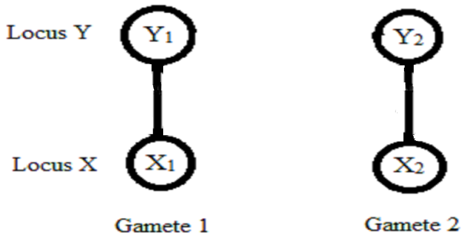


Figure 4.1: Two biallelic loci, X and Y

$$\begin{aligned}
P(X_i = 1) =& p_A, \quad i = 1, 2 \\
P(Y_i = 1) =& p_B, \quad i = 1, 2 \\
P(X_i = 1, Y_i = 1) =& P_{AB}, \quad i = 1, 2 \\
P(X_i = 1, Y_j = 1) =& P_{A/B}, \quad i, j = 1, 2, i \neq j \\
P(X_1 = 1, X_2 = 1) =& P_{AA} \\
P(Y_1 = 1, Y_2 = 1) =& P_{BB}.
\end{aligned} \tag{4.3}$$

Figure 4.2 illustrates the situation where we consider alleles and combinations of alleles at different loci and on different gametes. The Hardy-Weinberg disequilibrium coefficients are described in Section 2.4, and the linkage disequilibrium measures D_{AB} and $D_{A/B}$ will be defined later in Section 4.1 and 4.2.

4.1 Haplotype correlation and linkage disequilibrium

From the probabilities defined in Equation (4.3) we see that the expected values of the random variables defined in Equation (4.1) and (4.2) are (Weir 2008)

$$\begin{aligned}
E(X_i) =& p_A, \quad i = 1, 2, \quad \text{and} \\
E(Y_i) =& p_B, \quad i = 1, 2
\end{aligned}$$

where the subscript i indicates the first or second gamete for the individual.

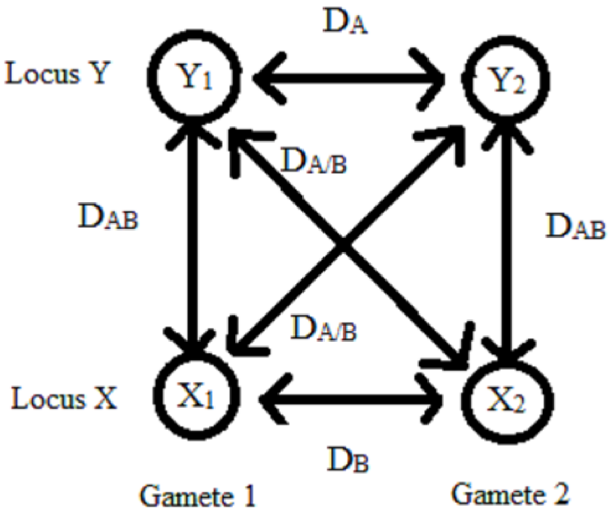


Figure 4.2: Figure illustrating Hardy-Weinberg disequilibrium, linkage disequilibrium and composite linkage disequilibrium between alleles at two loci X and Y with alleles A, a and B, b respectively. D_A and D_B are the Hardy-Weinberg disequilibrium coefficients for locus X and Y respectively. D_{AB} is the linkage disequilibrium measure between two alleles at different loci on the same gamete. $D_{A/B}$ describes linkage disequilibrium between two alleles that are both at different loci and on different gametes. The linkage disequilibrium measures D_{AB} and $D_{A/B}$ will be defined later in this chapter.

The variances of X_i and Y_i are

$$\begin{aligned} \text{Var}(X_i) =& p_A(1-p_A), \quad i=1,2 \quad \text{and} \\ \text{Var}(Y_i) =& p_B(1-p_B), \quad i=1,2 \end{aligned} \tag{4.4}$$

where the subscript i indicates the first or second gamete for the individual.

For the random variables X_i and Y_i we have the following expected values

$$\begin{aligned} E(X_i^2) &= 0^2 \cdot P(X_i=0) + 1^2 \cdot P(X_i=1) &&= p_A \quad , i=1,2 \\ E(Y_i^2) &= 0^2 \cdot P(Y_i=0) + 1^2 \cdot P(Y_i=1) &&= p_B \quad , i=1,2, \end{aligned}$$

and

$$E(X_iY_i) = 0 \cdot 0 \cdot P(X_i = 0 \cap Y_i = 0) + 0 \cdot 1 \cdot P(X_i = 0 \cap Y_i = 1) + \\ 1 \cdot 0 \cdot P(X_i = 1 \cap Y_i = 0) + 1 \cdot 1 \cdot P(X_i = 1 \cap Y_i = 1) = P_{AB}, i = 1, 2.$$

The covariance between the random variables X_i and Y_i is given by (Weir 2008)

$$\begin{aligned}\text{Cov}(X_i, Y_i) =& E(X_iY_i) - E(X_i)E(Y_i)\\ =& P_{AB} - p_Ap_B. \end{aligned} \tag{4.5}$$

From the variances in Equation (4.4) and the covariance calculated in Equation (4.5) the correlation between the random variables X_i and Y_i is given by

$$\text{Corr}(X_i, Y_i) = \frac{P_{AB} - p_Ap_B}{\sqrt{p_A(1-p_A)p_B(1-p_B)}}. \tag{4.6}$$

$\text{Corr}(X_i, Y_i)$ measures the correlation between the alleles at two different loci, X and Y, on the same gamete.

Linkage disequilibrium

For two loci, X and Y, with alleles A, a and B, b respectively, the four possible combinations of alleles are AB, ab, Ab and aB with probabilities given by P_{AB}, P_{ab}, P_{Ab} and P_{aB} respectively. Linkage disequilibrium (LD) measures non random association between alleles. The linkage disequilibrium measure D was by Lewontin & Kojima (1960) defined as

$$D = P_{AB}P_{ab} - P_{Ab}P_{aB}.$$

The linkage disequilibrium measure D describes the difference between the observed haplotype frequency and the expected haplotype frequency under equilibrium, when the alleles A and B are inherited independently. D can also be written in terms of allelic and haplotypic frequencies as

$$D = P_{AB} - p_Ap_B, \tag{4.7}$$

where P_{AB} is the probability for haplotype AB. For two loci X and Y, two alleles A and B are in linkage equilibrium when $D = 0$. This means that the estimated haplotype frequency equals the expected haplotype frequency under the equilibrium condition when the alleles A and B are inherited independently. Two alleles are in LD when $D \neq 0$, which means that the estimated haplotype frequency differs from the expected haplotype frequency under equilibrium. Linkage disequilibrium is affected by the activity of recombination (Kulle, Frigessi, Edvardsen, Kristensen & Wojnowski 2008).

From Equation (4.5) and (4.7) we observe that the linkage disequilibrium measure D defined by Lewontin & Kojima (1960) represents the covariance between the random variables X_i and Y_i defined in Equation (4.1) and (4.2) since

$$D = P_{AB} - p_Ap_B = \text{Cov}(X_i, Y_i).$$

Haplotype frequencies in terms of D

Since linkage disequilibrium is defined as the difference between the observed haplotype frequency and expected haplotype frequency under equilibrium we can in general write

$$\begin{aligned} P_{AB} &= p_A p_B + D_{AB} \\ P_{Ab} &= p_A p_b + D_{Ab} \\ P_{aB} &= p_a p_B + D_{aB} \\ P_{ab} &= p_a p_b + D_{ab} \end{aligned} \tag{4.8}$$

where D_{xy} denotes the linkage disequilibrium between alleles x and y.

We know that

$$\begin{aligned} p_A + p_a &= 1 \text{ and} \\ p_B + p_b &= 1. \end{aligned} \tag{4.9}$$

Adding the equations for P_{AB} and P_{Ab} from Equation (4.8) and using the result in Equation (4.9) gives

$$\begin{aligned} P_{AB} + P_{Ab} &= p_A p_B + D_{AB} + p_A p_b + D_{Ab} \\ &= p_A(p_B + p_b) + D_{AB} + D_{Ab} \\ &= p_A + D_{AB} + D_{Ab} \\ &= p_A. \end{aligned} \tag{4.10}$$

From Equation (4.10) we see that

$$D_{Ab} = -D_{AB}.$$

Similarly, we get

$$D_{aB} = -D_{AB}.$$

Adding the equations for P_{AB} and P_{ab} give

$$\begin{aligned} P_{Ab} + P_{ab} &= p_A p_b + D_{Ab} + p_a p_b + D_{ab} \\ &= p_b(p_A + p_a) + D_{Ab} + D_{ab} \\ &= p_b + D_{Ab} + D_{ab} \\ &= p_b. \end{aligned} \tag{4.11}$$

From Equation (4.11) we get

$$D_{ab} = -D_{Ab} = D_{AB}.$$

From these equations we observe that

$$D_{AB} = D_{ab} = -D_{Ab} = -D_{aB} = D.$$

Then, Equation (4.8) can be rewritten as

$$\begin{aligned} P_{AB} &= p_A p_B + D \\ P_{Ab} &= p_A p_b - D \\ P_{aB} &= p_a p_B - D \\ P_{ab} &= p_a p_b + D. \end{aligned}$$

We will use $D_{AB} = D$ to denote the linkage disequilibrium between two alleles at different loci on the same gamete.

The LD measure D'

The LD measure D' (Lewontin 1964) is a normalized measure where the measure D is normalized by using the maximum possible deviation from equilibrium given the observed allele frequencies, denoted by D_{max} . The measure D' is given by

$$D' = \frac{|D|}{D_{\max}},$$

where $D_{\max}$ is given by

$$D_{\max} = \begin{cases} \min\{p_A p_b, p_a p_B\} & \text{for} \quad D > 0 \\ \min\{p_A p_B, p_a p_b\} & \text{for} \quad D < 0 \end{cases} \tag{4.12}$$

We then observe that

$$0 \leq D' \leq 1.$$

For positive linkage disequilibrium, $D > 0$ we have from Equation (4.7)

$$P_{AB} > p_A p_B.$$

This means that the observed haplotype frequency is greater than the expected haplotype frequency under independence (the equilibrium condition). Then, $D > 0$ indicates that the probability of haplotype AB is greater than the probability of the haplotype under the equilibrium condition.

For negative linkage disequilibrium, $D < 0$, Equation (4.7) give

$$P_{AB} < p_A p_B.$$

This means that the observed haplotype frequency is less than the expected haplotype frequency under independence (the equilibrium condition). $D < 0$ gives that the probability to inherit the haplotype AB is less than the probability for inheriting alleles A and B under the equilibrium condition.

Gametic correlation coefficient, ρ_{LD}

The LD measure D can also be scaled by the square root of the product of all allelic frequencies, which gives the gametic correlation coefficient as (Weir 1996, p. 137)

$$\rho_{LD} = \frac{D}{\sqrt{p_A p_b p_a p_B}}.$$

By using

$$p_a = 1 - p_A$$

and

$$p_b = 1 - p_B$$

we can rewrite ρ_{LD} as

$$\rho_{LD} = \frac{D}{\sqrt{p_A(1-p_A)p_B(1-p_B)}}. \tag{4.13}$$

We see that the range for ρ_{LD} is

$$-1 \leq \rho_{LD} \leq 1.$$

Comparison of the LD measures

All the LD measures described above, D, D' and ρ_{LD}, include the the difference between the observed haplotype frequency and the expected haplotype frequency under the equilibrium condition. We have seen that the LD measures have different ranges. The range of D depends on the observed allele frequencies, which is not a desirable property. The measures D' and ρ_{LD} can take values in the interval [-1,1]. We also observe that when we have a situation with rare alleles, but a small value of LD between them, we can get D' equal to one and a small value of ρ_{LD}. This shows that using the correlation coefficient ρ_{LD} may be a better choice for a situation with rare alleles because it is more easy to interpret.

4.2 Genotype correlation and composite linkage disequilibrium

A genotype is an unordered sequence of the two alleles present at both gametes at one locus. The random variables defined in Equation (4.1) and (4.2) describes alleles present at one locus and one gamete. We observe that the genotype at each locus can be described by a sum of the random variables representing each of the alleles present at one locus. At locus X we observe that the genotype may be given by the sum

$$X' = X_1 + X_2$$

where X_1 and X_2 are the random variables defined in Equation (4.1) representing the alleles present at the first and second gamete respectively. For locus Y, the genotype is defined by the new variable Y',

$$Y' = Y_1 + Y_2$$

where Y_1 and Y_2 are the random variables defined in Equation (4.2) representing the alleles present at the first and second gamete respectively.

These new variables will be given as

$$X' = \begin{cases} 0 & \text{if genotype is } aa \\ 1 & \text{if genotype is } Aa \\ 2 & \text{if genotype is } AA \end{cases} \tag{4.14}$$

and

$$Y' = \begin{cases} 0 & \text{if genotype is } bb \\ 1 & \text{if genotype is } Bb \\ 2 & \text{if genotype is } BB \end{cases} . \tag{4.15}$$

For the variables defined in Equation (4.14) and (4.15) we find the following probabilities

$$\begin{array}{llll} P(X'=2) & = P(X_1=1, X_2=1) & & = P_{AA} \\ P(X'=1) & = P(X_1=1, X_2=0) & = P(X_1=0, X_2=1) & = P_{Aa} \\ P(X'=0) & = P(X_1=0, X_2=0) & & = P_{aa} \\ P(Y'=2) & = P(Y_1=1, Y_2=1) & & = P_{BB} \\ P(Y'=1) & = P(Y_1=1, Y_2=0) & = P(Y_1=0, Y_2=1) & = P_{Bb} \\ P(Y'=0) & = P(Y_1=0, Y_2=0) & & = P_{bb} \end{array}$$

We observe that the expected values of the variables X' and Y' are

$$\begin{aligned} E(X') = E(X_1 + X_2) = E(X_1) + E(X_2) = 2p_A \\ E(Y') = E(Y_1 + Y_2) = E(Y_1) + E(Y_2) = 2p_B. \end{aligned} \tag{4.16}$$

The expected values can also be written as

$$E(X') \quad = 0 \cdot P(X'=0) + 1 \cdot P(X'=1) + 2 \cdot P(X'=2) \quad = 1 \cdot P_{Aa} + 2 \cdot P_{AA}$$

and

$$E(Y') \quad = 0 \cdot P(Y'=0) + 1 \cdot P(Y'=1) + 2 \cdot P(Y'=2) \quad = 1 \cdot P_{Bb} + 2 \cdot P_{BB}.$$

These equations gives

$$\begin{aligned} p_A &= P_{AA} + \frac{1}{2} P_{Aa} \quad \text{and} \\ p_B &= P_{BB} + \frac{1}{2} P_{Bb}. \end{aligned} \tag{4.17}$$

Using Equation (4.3) we also see that

$$\begin{aligned} E(X'Y') =& E[(X_1+X_2)(Y_1+Y_2)] \\ =& E[X_1Y_1 + X_1Y_2 + X_2Y_1 + X_2Y_2] \\ =& E[X_1Y_1] + E[X_1Y_2] + E[X_2Y_1] + E[X_2Y_2] \\ =& P_{AB} + P_{A/B} + P_{A/B} + P_{AB} \\ =& 2(P_{AB} + P_{A/B}). \end{aligned} \tag{4.18}$$

In the most general case we do not assume Hardy-Weinberg equilibrium, which means that the alleles present at different gametes at the same locus in general are not independent of each other. Using Equation (4.4) we see that the variance of X' and Y' is

$$\begin{aligned} \mathrm{Var}(X') =& \mathrm{Var}(X_1 + X_2) \\ =& \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + 2\mathrm{Cov}(X_1, X_2) \\ =& 2p_A(1-p_A) + 2(P_{AA} - p_A^2) \\ =& 2[p_A(1-p_A) + D_A], \end{aligned}$$

where D_A is the Hardy-Weinberg disequilibrium coefficient

$$D_A = P_{AA} - p_A^2$$

defined in Section 2.4.

Similar calculations show that the variance of Y' is given by

$$\mathrm{Var}(Y') = 2[p_B(1-p_B) + D_B]$$

where D_B is the Hardy-Weinberg disequilibrium coefficient

$$D_B = P_{BB} - p_B^2.$$

The covariance between X' and Y' is from Equation (4.16) and (4.18)

$$\begin{aligned}\mathrm{Cov}(X', Y') =& 2(P_{AB} + P_{A/B}) - 2p_A 2p_B \\ =& 2[P_{AB} + P_{A/B} - 2p_A p_B]\end{aligned} \tag{4.19}$$

and the correlation is then given by

$$\begin{aligned}\mathrm{Corr}(X', Y') =& \frac{\mathrm{Cov}(X', Y')}{\sqrt{\mathrm{Var}(X')}\sqrt{\mathrm{Var}(Y')}} \\ =& \frac{2[P_{AB} + P_{A/B} - 2p_A p_B]}{\sqrt{2[p_A(1-p_A) + D_A]}\sqrt{2[p_B(1-p_B) + D_B]}} \\ =& \frac{P_{AB} + P_{A/B} - 2p_A p_B}{\sqrt{[p_A(1-p_A) + D_A]}\sqrt{[p_B(1-p_B) + D_B]}}.\end{aligned} \tag{4.20}$$

We observe that $\mathrm{Corr}(X', Y')$ given in Equation (4.20) measures genotype correlation between two loci X and Y, where the genotypes are represented by the sums of random variables, X' and Y'.

Generalization of linkage disequilibrium to the two gamete case

Consider two biallelic loci, X and Y, with alleles A, a and B, b respectively as defined in Equation (4.1) and (4.2). We have in total ten possible haplotype combinations of the four alleles present at these two loci. These frequencies are given in Table 4.1.

We use the notation P_{xy}^{xy} to denote the haplotypes present at both gametes for loci X and Y, where the subscript indicates the haplotype present at one gamete and the superscript indicates the haplotype present at the other gamete. For example, P_{AB}^{Ab} indicates that the haplotype present at one gamete is AB and the haplotype present at the other gamete is Ab.

From the ten possible haplotype combinations described in Table 4.1, only nine of these probabilities can be directly observed from genotypic data. In general it is not possible to distinguish between the double heterozygotes (AB, ab) and (Ab, aB), we can only observe the total frequency for these two double heterozygotes as shown in Table 4.2.

The probabilities in Table 4.2 are related to the probabilities defined in Equation

Table 4.1: Possible haplotype frequencies for two biallelic loci.

Haplotype	Frequency
(AB, AB)	P_{AB}^{AB}
(AB, Ab)	P_{AB}^{Ab}
(Ab, Ab)	P_{Ab}^{Ab}
(AB, aB)	P_{aB}^{AB}
(Ab, aB)	P_{Ab}^{aB}
(AB, ab)	P_{ab}^{AB}
(Ab, ab)	P_{ab}^{Ab}
(aB, aB)	P_{aB}^{aB}
(aB, ab)	P_{ab}^{aB}
(ab, ab)	P_{ab}^{ab}

Table 4.2: Possible haplotype pairs for two biallelic loci.

		Locus Y		
		BB	Bb	bb
	AA	P_{AB}^{AB}	P_{Ab}^{AB}	P_{Ab}^{Ab}
Locus X	Aa	P_{aB}^{AB}	$P_{Ab}^{aB} + P_{ab}^{AB}$	P_{ab}^{Ab}
	aa	P_{aB}^{aB}	P_{ab}^{aB}	P_{ab}^{ab}

(4.3). From Table 4.2 we observe that we have the following relationships

$$\begin{aligned} P_{AA} &= P_{AB}^{AB} + P_{Ab}^{AB} + P_{Ab}^{Ab} \\ P_{Aa} &= P_{aB}^{AB} + P_{Ab}^{aB} + P_{ab}^{AB} + P_{ab}^{Ab} \\ P_{aa} &= P_{aB}^{aB} + P_{ab}^{aB} + P_{ab}^{ab} \end{aligned}$$

which gives as in Equation (4.17)

$$p_A = P_{AA} + \frac{1}{2} P_{Aa}.$$

Similarly, we get

$$\begin{aligned} P_{BB} &= P_{AB}^{AB} + P_{aB}^{AB} + P_{aB}^{aB} \\ P_{Bb} &= P_{Ab}^{AB} + P_{Ab}^{aB} + P_{ab}^{AB} + P_{ab}^{aB} \\ P_{bb} &= P_{Ab}^{Ab} + P_{ab}^{Ab} + P_{ab}^{ab} \end{aligned}$$

and as in Equation (4.17)

$$p_B = P_{Bb} + \frac{1}{2}P_{Bb}.$$

According to Weir (1996, p. 122) and the proof in Appendix B we have that

$$P_{AB} = P^{AB}_{AB} + \frac{1}{2}(P^{AB}_{Ab} + P^{AB}_{aB} + P^{AB}_{ab}).$$

The LD measure D as defined in Equation (4.7) describes the association between two alleles on two different loci on the same gamete. The generalization of the LD measure D to the case where we consider two loci on two gametes was described by Weir (1996, p. 125). Considering two loci and two gametes we also need to take into account possible disequilibrium between alleles that are both at different loci and on different gametes. This measure of linkage disequilibrium is denoted $D_{A/B}$ and is defined by Weir (1996, p. 122) by introducing the non-gametic frequency, denoted $P_{A/B}$. The non-gametic frequency, $P_{A/B}$ describes the frequency of alleles A and B at different loci and on different gametes.

The non-gametic frequency, $P_{A/B}$, is by Weir (1996, p. 122) given by

$$P_{A/B} = P^{AB}_{AB} + \frac{1}{2}(P^{AB}_{Ab} + P^{AB}_{aB} + P^{aB}_{Ab}),$$

which is proved in Appendix B.

We also observe that the sum of P_{AB} and $P_{A/B}$ can be written as

$$P_{AB} + P_{A/B} = 2P^{AB}_{AB} + P^{AB}_{Ab} + P^{AB}_{aB} + \frac{1}{2}(P^{AB}_{ab} + P^{aB}_{Ab}).$$

The digenic disequilibrium, $D_{A/B}$, is defined as (Weir 1996, p. 122)

$$D_{A/B} = P_{A/B} - p_A p_B. \tag{4.21}$$

We see that the digenic disequilibrium is related to the disequilibrium measure D_{AB} defined for the one gamete case. Both D_{AB} and $D_{A/B}$ measure the difference between the observed haplotype frequency and the expected frequency under the equilibrium condition.

Composite linkage disequilibrium correlation

Weir (1996, p. 126) defined a composite linkage disequilibrium measure, denoted Δ_{AB},

$$\begin{aligned}\Delta_{AB} =& P_{AB} + P_{A/B} - 2p_A p_B \\ =& D_{AB} + D_{A/B} \end{aligned} \tag{4.22}$$

where $D_{A/B}$ is the digenic linkage disequilibrium measure for non random association between alleles at different gametes and different loci as defined in Equation (4.21).

The composite linkage disequilibrium correlation was by Weir (1996, p. 137) defined as

$$\rho_{CLD} = \frac{\Delta_{AB}}{\sqrt{[p_A(1-p_A) + D_A]}\sqrt{[p_B(1-p_B) + D_B]}} \tag{4.23}$$

where D_A and D_B are the Hardy-Weinberg disequilibrium coefficients as defined in Section 2.4.

From Equation (4.19) and (4.22) we observe that

$$\begin{aligned}\text{Cov}(X', Y') &= 2(P_{AB} + P_{A/B} - 2p_A p_B) \\ &= 2\Delta_{AB}. \end{aligned} \tag{4.24}$$

Equation (4.24) shows that the composite linkage disequilibrium measure Δ_{AB} equals half the covariance between the genotypic variables X' and Y' defined in Equation (4.14) and (4.15).

From Equation (4.20) and (4.23) we observe the following relationship

$$\begin{aligned}\rho_{CLD} &= \text{Corr}(X_1 + X_2, Y_1 + Y_2) \\ &= \text{Corr}(X', Y'). \end{aligned} \tag{4.25}$$

This shows that the composite linkage disequilibrium correlation describes the genotype correlation for alleles at different loci and on different gametes.

The upper and lower bound for the linkage disequilibrium coefficient D is given in Equation (4.12) and for the composite linkage disequilibrium measure Δ_{AB} as described in Equation (4.22), Hamilton & Cole (2004) described the upper and lower bounds for the composite linkage disequilibrium measure Δ_{AB} as

$$\max(-2p_A p_B, -2p_a p_b) \leq \Delta_{AB} \leq \min(2p_A p_b, 2p_a p_B).$$

4.3 LD vs. CLD correlation

Numerical example

We will use the example from Weir (1996, p. 123) to illustrate the difference between LD and CLD correlation. We denote the frequency of genotypes AA, Aa and aa by P_A^A, P_A^a, P_a^a, respectively, and similarly for loci with alleles B and b.

Table 4.3: Genotypic frequencies for two alleles at each of two loci.

		Locus Y			
		BB	Bb	Bb	
Locus X	AA	$P_{AB}^{AB} = 0.20$	$P_{AB}^{Ab} = 0.18$	$P_{Ab}^{Ab} = 0.02$	$P_A^A = 0.40$
	Aa	$P_{aB}^{AB} = 0.26$	$P_{Ab}^{aB} = 0.04, P_{ab}^{AB} = 0.08$	$P_{ab}^{Ab} = 0.02$	$P_A^a = 0.40$
	aa	$P_{aB}^{aB} = 0.04$	$P_{ab}^{aB} = 0.10$	$P_{ab}^{ab} = 0.06$	$P_a^a = 0.20$
	Total	$P_B^B = 0.50$	$P_B^b = 0.40$	$P_b^b = 0.10$	1

Table 4.4: Haplotypic frequencies for two alleles at each of two loci.

	A/B	A/b	aB	ab	
AB	$P_{AB}^{AB} = 0.20$	$\frac{1}{2}P_{Ab}^{AB} = 0.09$	$\frac{1}{2}P_{aB}^{AB} = 0.13$	$\frac{1}{2}P_{ab}^{AB} = 0.04$	$P_{AB} = 0.46$
Ab	$\frac{1}{2}P_{AB}^{Ab} = 0.09$	$P_{Ab}^{Ab} = 0.02$	$\frac{1}{2}P_{aB}^{Ab} = 0.02$	$\frac{1}{2}P_{ab}^{Ab} = 0.01$	$P_{Ab} = 0.14$
a/B	$\frac{1}{2}P_{AB}^{aB} = 0.13$	$\frac{1}{2}P_{ab}^{AB} = 0.04$	$P_{aB}^{aB} = 0.04$	$\frac{1}{2}P_{ab}^{aB} = 0.05$	$P_{a/B} = 0.26$
a/b	$\frac{1}{2}P_{aB}^{Ab} = 0.02$	$\frac{1}{2}P_{Ab}^{ab} = 0.01$	$\frac{1}{2}P_{aB}^{ab} = 0.05$	$P_{ab}^{ab} = 0.06$	$P_{a/b} = 0.14$
	$P_{A/B} = 0.44$	$P_{A/b} = 0.16$	$P_{aB} = 0.24$	$P_{ab} = 0.16$	1

Table 4.3 and 4.4 shows an numerical example of genotypic frequencies for two alleles at each of two loci. The two tables shows the same example, but Table 4.4 is rewritten for use in estimation of the composite linkage disequilibrium correlation as defined in Equation (4.20). From these two tables, we can easily set up the estimates for LD and CLD correlation as defined in Equations (4.6) and (4.20). We get

$$\begin{aligned} P_{AB} &= P_{AB}^{AB} + \frac{1}{2}(P_{AB}^{Ab} + P_{aB}^{AB} + P_{ab}^{AB}) \\ &= 0.20 + \frac{1}{2}(0.18 + 0.26 + 0.08) \\ &= 0.46 \end{aligned}$$

and

$$\begin{aligned} P_{AB} + P_{A/B} &= 2P_{AB}^{AB} + P_{AB}^{Ab} + P_{aB}^{AB} + \frac{1}{2}(P_{Ab}^{aB} + P_{ab}^{AB}) \\ &= 2 \cdot 0.20 + 0.18 + 0.26 + \frac{1}{2}(0.08 + 0.04) \\ &= 0.9. \end{aligned}$$

The Hardy-Weinberg coefficients are calculated from Table 4.3 as

$$D_A = P_{AA} - p_A^2 = 0.40 - (0.60)^2 = 0.04$$

and

$$D_B = P_{BB} - p_B^2 = 0.50 - (0.70)^2 = 0.01.$$

We see that

$$\begin{aligned} \rho_{LD} &= \frac{P_{AB} - p_A p_B}{\sqrt{p_A p_B p_a p_b}} \\ &= \frac{0.46 - 0.60 \cdot 0.70}{\sqrt{0.60 \cdot 0.40 \cdot 0.70 \cdot 0.30}} \\ &= 0.1781742 \end{aligned}$$

and

$$\begin{aligned} \rho_{CLD} &= \frac{P_{AB} + P_{A/B} - 2p_A p_B}{\sqrt{(p_A p_a + D_A)(p_B p_b + D_B)}} \\ &= \frac{0.9 - 2 \cdot 0.60 \cdot 0.70}{\sqrt{(0.60 \cdot 0.40 + 0.04)(0.70 \cdot 0.30 + 0.01)}} \\ &= 0.2417469. \end{aligned}$$

From this example we observe that the CLD correlation, ρ_{CLD}, is more extreme than the LD correlation ρ_{LD}. We want to investigate if this is a general finding, and then use this to decide which measure of correlation, ρ_{LD} or ρ_{CLD}, we want to use in multiple testing correction problems.

MAF and LD correlation

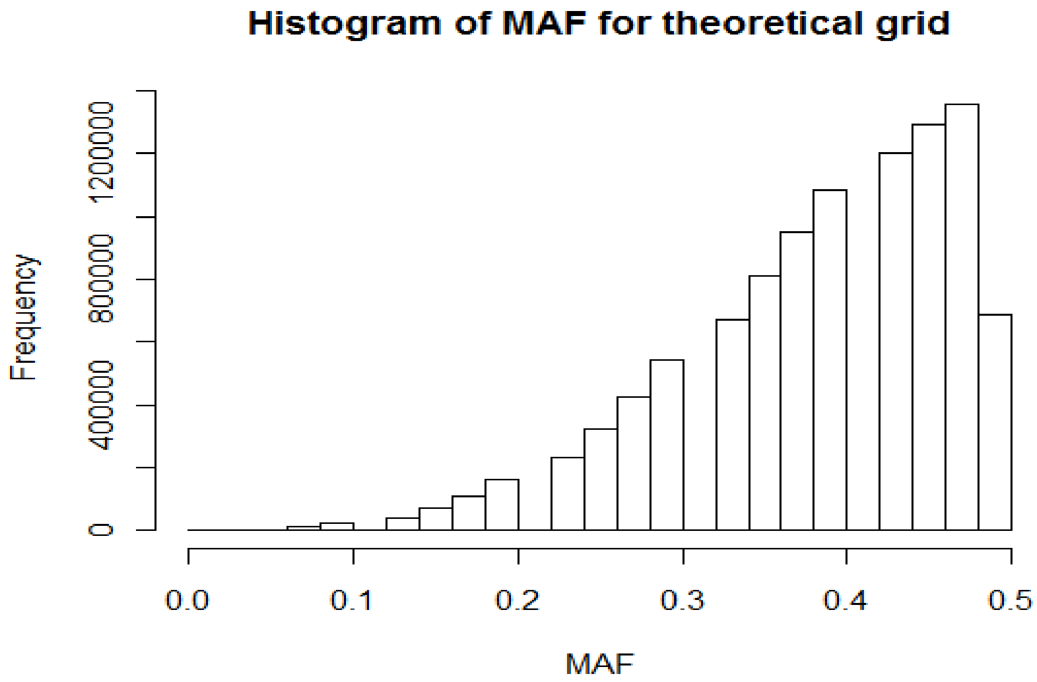


Figure 4.3: Histogram of MAF for the theoretical grid.

We do not know the real distribution of the minor allele frequencies (MAF) for SNPs in a general population, and therefore we implemented in R (R Development Core Team 2011) a theoretical grid which not is realistic because the grid includes all possible combinations of probabilities. The R code is shown in Appendix D. The distribution of the minor allele frequencies for the theoretical grid is shown in Figure 4.3, and in Chapter 5 we will see how the distribution of the minor allele frequencies will be for a real data set, chromosome 22 of the TOP8 data.

CLD vs. LD correlation in the general case

To study the CLD correlation compared to the LD correlation we used the theoretical grid for all combinations of probabilities to estimate the haplotype and genotype correlation, ρ_{LD} and ρ_{CLD}, as defined in Equation (4.13) and (4.23).

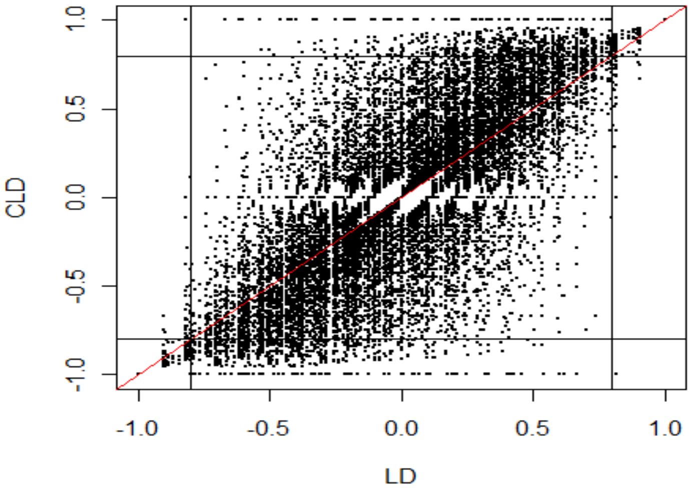


Figure 4.4: Plot of ρ_{CLD} vs ρ_{LD} using the theoretical grid. The horizontal and vertical lines are plotted at ρ_{LD} and ρ_{CLD} equal to 0.8 in absolute value.

From Figure 4.4 we see the CLD correlation as defined in Equation (4.23) plotted against the LD correlation as defined in Equation (4.13). The horizontal lines and vertical lines are all plotted at the correlation values equal to 0.8 in absolute value. From Figure 4.4 we clearly see that the CLD correlation is more extreme than the LD correlation, since we have more points where $|\rho_{CLD}| > 0.8$ than points where $|\rho_{LD}| > 0.8$. This plot indicates that the CLD correlation is more extreme than the LD correlation.

Table 4.5: Summary statistics for LD and CLD correlation

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
LD	-1.000	-0.2357	0.000	$-2.218 \cdot 10^{-05}$	-0.2357	1.000
CLD	-1.000	-0.3440	0.000	$4.814 \cdot 10^{-12}$	-0.3440	1.000

Table 4.5 shows the summary statistics for the LD and CLD correlation calculated based on the theoretical grid. The results for the first and third quantile in Table 4.5 shows indicates that the CLD correlation is more extreme than the LD correlation.

Table 4.6: Summary statistics for the absolute difference between LD and CLD correlation

LD,CLD	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
LD,CLD> 0.01	-0.9524	-0.2319	-0.0951	-0.1167	-0.0024	0.7071
CLD,LD< -0.01	-0.9524	-0.2319	-0.0951	-0.1167	-0.0024	0.7071

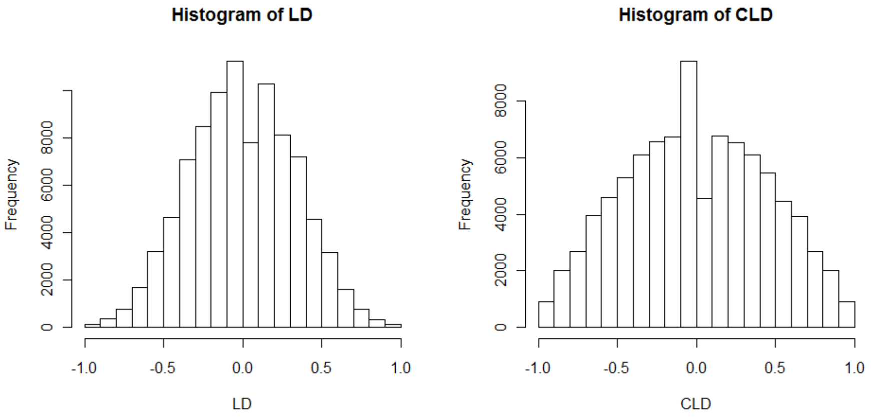
Table 4.7: LD vs. CLD correlation for the theoretical matrix

ρ_{CLD}, ρ_{LD}	$\rho_{CLD} - \rho_{LD}$	proportion
$\rho_{CLD}, \rho_{LD} > 0.01$	$\rho_{CLD} - \rho_{LD} > 0$	0.7554417
$\rho_{CLD}, \rho_{LD} > 0.01$	$\rho_{CLD} - \rho_{LD} < 0$	0.2414901
$\rho_{CLD}, \rho_{LD} < -0.01$	$\rho_{CLD} - \rho_{LD} > 0$	0.7554417
$\rho_{CLD}, \rho_{LD} < -0.01$	$\rho_{CLD} - \rho_{LD} < 0$	0.2414901

For the case where both correlations are positive and greater than 0.01, the results of Table 4.6 shows that the mean difference is equal to -0.1167, and for the case when both correlation are negative with value less than 0.01 we see that the mean difference is also equal to -0.1167. For both cases we also observe that the CLD correlation is greater than the LD correlation in approximately 75% of the cases, as shown in Table 4.7.

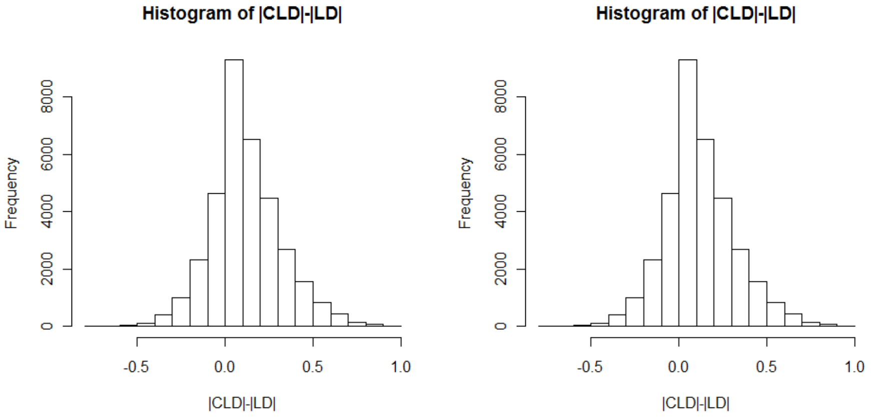
From Figure 4.5a and 4.5b we observe that for our theoretical grid the CLD correlation ρ_{CLD} tend to be more extreme than the LD correlation because the histograms shows more extreme values for the CLD correlation in Figure 4.5b than for the LD correlation in Figure 4.5a.

From Figure 4.6a and 4.6b we see histogram of the difference between the absolute values of the CLD and the LD correlation. Figure 4.6a shows histogram for the difference when both the CLD and LD correlation are positive and in Figure 4.6b we see the histogram for the difference when both the CLD and LD correlation are negative. From both these figures, we observe that the CLD correlation tend to be more extreme than the LD correlation.



(a) (b)

Figure 4.5: (a) Histogram of LD correlation for the theoretical grid. (b) Histogram of CLD correlation for the theoretical grid. From these figures we see that for the theoretical grid, we have more extreme values of CLD correlation compared to the results using LD correlation, which indicates that the CLD correlation is more extreme than the LD correlation.



(a) (b)

Figure 4.6: (a) Histogram of difference $|\rho_{CLD}| - |\rho_{LD}|$ when both correlations are positive. (b) Histogram of difference $|\rho_{CLD}| - |\rho_{LD}|$ when both correlations are negative. From these figures we observe that the CLD correlation is more extreme than the LD correlation.

LD vs. CLD

We have seen that for the theoretical grid, the CLD correlation is more extreme than the LD correlation in approximately 75% of the cases. For comparison we also investigated the difference between the squared correlation measures for LD and CLD correlation, ρ_{LD}^2 and ρ_{CLD}^2, respectively and we observed similar relationship, ρ_{CLD}^2 was more extreme than ρ_{LD}^2 in approximately 75% of the cases. From Figure 4.3 we see the distribution of the minor allele frequencies for the theoretical grid. We do not know how the distribution of the minor allele frequencies for a general population, and in Chapter 5 we will see the distribution of the minor allele frequencies for chromosome 22 of the TOP data. In Section 5.5 we will look at how the LD and CLD correlations can be estimated based on the observed data from the TOP8 data set, to see which estimated values of ρ_{CLD} and ρ_{LD} we will have for a real data set, instead of using a theoretical grid with all possible combinations.

Chapter 5

Estimation

In previous chapter we have showed that SNP dependence can be assessed by calculating the linear correlation between SNPs, based on either haplotypes or genotypes. In this chapter, methods for estimating LD and CLD correlation based on observed genotype data will be presented. We will also present some methods for estimating haplotypes and haplotype blocks.

According to Weir (1996, p. 137), linkage disequilibrium correlation (haplotype correlation) is given by

$$\rho_{LD} = \frac{D}{\sqrt{p_A(1-p_A)p_B(1-p_B)}}$$

as described in Section 4.1.

The composite linkage disequilibrium correlation (genotype correlation) is defined as (Weir 1996, p. 137)

$$\rho_{CLD} = \frac{\Delta_{AB}}{\sqrt{(p_A(1-p_A)+D_A)(p_B(1-p_B)+D_B)}}$$

which is described in Section 4.2. D_A and D_B are the Hardy-Weinberg disequilibrium coefficients defined in Section 2.4.

5.1 Estimating LD correlation from genotype data

Consider two loci X and Y with alleles A, a and B, b respectively. The possible pairwise haplotype combinations of these alleles are AB, aB, Ab and ab.

From Equation (4.7) the linkage disequilibrium measure D is estimated by

$$\begin{aligned}\hat{D} &= \hat{P}_{AB}\hat{P}_{ab} - \hat{P}_{Ab}\hat{P}_{aB}\\ &= \hat{P}_{AB} - \hat{p}_A\hat{p}_B.\end{aligned}$$

The LD correlation as described in Section 4.1 is then estimated by

$$\hat{\rho}_{LD} = \frac{\hat{P}_{AB} - \hat{p}_A\hat{p}_B}{\sqrt{\hat{p}_A\hat{p}_a\hat{p}_B\hat{p}_b}}. \tag{5.1}$$

For two biallelic loci there are in total nine possible observable genotypes. We will denote the observed genotype counts, $n_1, ..., n_9$, as shown in Table 5.1. Here n be the number of individuals in the study.

Table 5.1: Table for observed genotype counts

		Locus Y			
		BB	Bb	bb	Total
	AA	n_1	n_2	n_3	n_{AA}
Locus X	Aa	n_4	n_5	n_6	n_{Aa}
	aa	n_7	n_8	n_9	n_{aa}
	Total	n_{BB}	n_{Bb}	n_{bb}	n

The corresponding genotype frequencies are given by

$$p_i = \frac{n_i}{n}, i = 1, ..., 9. \tag{5.2}$$

Assuming HWE as described in Section 2.4, the probabilities $p_1, ..., p_9$ defined in Equation (5.2) can be written as shown in Table 5.2.

From Equation (5.1) we observe that the only unknown parameter is the haplotype frequency P_{AB}, which can be estimated using maximum likelihood estimation.

In general, the likelihood function for a 9-nomial distribution is given by

$$\frac{n!}{n_1! \cdots n_9!}\prod_{i=1}^{9} p_i^{n_i}. \tag{5.3}$$

Table 5.2: Genotypic frequencies for two loci assuming HWE.

$p_1 =$	P_{AB}^2
$p_2 =$	$P_{AB}P_{Ab}$
$p_3 =$	P_{Ab}^2
$p_4 =$	$P_{AB}P_{aB}$
$p_5 =$	$P_{AB}P_{ab} + P_{Ab}P_{aB}$
$p_6 =$	$P_{Ab}P_{ab}$
$p_7 =$	P_{aB}^2
$p_8 =$	$P_{aB}P_{ab}$
$p_9 =$	P_{ab}^2

From Table 5.2 and Equation (5.3) we see that the likelihood function for our parameters $\theta = (P_{AB}, P_{Ab}, P_{aB}, P_{ab})$ given the observed data can be written as

$$\begin{aligned} L(\theta|n_1, ..., n_9) \propto\, & P_{AB}^{2n_1}(P_{AB}P_{Ab})^{n_2}P_{Ab}^{n_3}(P_{AB}P_{aB})^{n_4}\cdot \\ & (P_{AB}P_{ab} + P_{Ab}P_{aB})^{n_5}(P_{Ab}P_{ab})^{n_6}P_{aB}^{2n_7}(P_{aB}P_{ab})^{n_8}P_{ab}^{2n_9}. \end{aligned}$$

The log-likelihood function can then be written as (Foulkes 2009, p. 68)

$$\begin{aligned} \text{log}L(\theta|n_1, ..., n_9) \propto\, & (2n_1 + n_2 + n_4)\text{log}P_{AB} + (2n_3 + n_2 + n_6)\text{log}P_{Ab} \\ & + (2n_7 + n_4 + n_8)\text{log}P_{aB} \\ & + (2n_9 + n_8 + n_6)\text{log}P_{ab} + n_5\text{log}(P_{AB}P_{ab} + P_{Ab}P_{aB}) \end{aligned}$$

From Table 5.3 we have the following relationships

$$\begin{aligned} P_{Ab} &= p_A - P_{AB} \\ P_{aB} &= p_B - P_{AB} \\ P_{ab} &= 1 - P_{AB} - P_{Ab} - P_{aB}. \end{aligned}$$

Then, we can write the log-likelihood function as

$$\begin{aligned} \text{log}L(P_{AB}|n_1, ..., n_9) \propto\, & (2n_1 + n_2 + n_4)\text{log}P_{AB} + (2n_3 + n_2 + n_6)\text{log}(p_A - P_{AB}) \\ & + (2n_7 + n_4 + n_8)\text{log}(p_B - P_{AB}) \\ & + (2n_9 + n_8 + n_6)\text{log}(P_{AB} - P_{Ab} - P_{aB}) \\ & + n_5\text{log}(P_{AB}(P_{AB} - P_{Ab} - P_{aB}) + (p_A - P_{AB})(p_B - P_{AB})), \end{aligned} \tag{5.4}$$

where we observe that the only unknown parameter is the haplotype frequency P_{AB}. We use maximum likelihood estimation with this log-likelihood function to estimate P_{AB}, and then we get the estimate of ρ_{LD},

$$\begin{aligned}\hat{\rho}_{LD} &= \frac{\hat{D}}{\sqrt{\hat{p}_A(1-\hat{p}_A)\hat{p}_B(1-\hat{p}_B)}}\\ &= \frac{\hat{P}_{AB}-\hat{p}_A\hat{p}_B}{\sqrt{\hat{p}_A(1-\hat{p}_A)\hat{p}_B(1-\hat{p}_B)}}.\end{aligned}$$

The Pearson correlation coefficient

Pearson's correlation coefficient is denoted by r. The correlation coefficient for two random variables X and Y is in general given by

$$\begin{aligned}\rho_{X,Y} =&\mathrm{Corr}(X,Y)\\ =&\frac{\mathrm{Cov}(X,Y)}{\sigma_X\sigma_Y}\\ =&\frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X\sigma_Y},\end{aligned}$$

where μ_X and μ_Y are the expected values of X and Y respectively and σ_X and σ_Y are the standard deviations of X and Y respectively.

Pearson's product moment sample correlation, r

$$\begin{aligned}r =&\frac{s_{XY}}{s_Xs_Y}\\ =&\frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}\end{aligned}$$

where $\bar{X},\bar{Y}$ and s_X^2, s_Y^2 are the sample mean and variance of the observed variables X and Y, respectively. Pearson's correlation coefficient takes values in $[-1,1]$.

The estimated haplotype frequencies for the SNP data, defined in Equation (4.1) and (4.2), can be represented as shown in Table 5.3.

Table 5.3: Table for estimated haplotype frequencies

		Locus Y		
		B	b	Total
	A	$\hat{P}_{AB}$	$\hat{P}_{Ab}$	$\hat{p}_A$
Locus X	a	$\hat{P}_{aB}$	$\hat{P}_{ab}$	$\hat{p}_a$
	Total	$\hat{p}_B$	$\hat{p}_b$	1

The data analyzed for each SNP are binary data and can in general be summarized as shown in Table 5.4 where n is the number of individuals in the study.

Table 5.4: Binary data

		Locus Y		
		1	0	Total
	1	a	b	$a+b$
Locus X	0	c	d	$c+d$
	Total	$a+c$	$b+d$	$2n$

The Φ-coefficient is the Pearson correlation coefficient for binary data and from Table 5.4 we can estimate the Φ-coefficient as

$$\Phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}.$$

Following the notation in Table 5.3, the Φ-coefficient is given by

$$\Phi = \frac{\hat{P}_{AB}\hat{P}_{ab} - \hat{P}_{Ab}\hat{P}_{aB}}{\sqrt{\hat{p}_A\hat{p}_a\hat{p}_B\hat{p}_b}}. \tag{5.5}$$

From Equation (5.5) we observe that the pairwise haplotype correlation as defined in Equation (4.6) is the Pearson correlation coefficient for binary data.

When allele counts are directly observed as given in Table 5.1, the haplotype phase is often unknown. Then, the haplotype frequency P_{AB} cannot be estimated directly as a proportion of AB haplotypes among all haplotypes in the sample, and therefore we used maximum likelihood estimation to estimate $\hat{P}_{AB}$. If the haplotype phase is not ambiguous, the Pearson correlation coefficient can be used to estimate $\hat{\rho}_{LD}$. For the data analyzed in this thesis, the haplotype phase are ambiguous, and if we want to use the Pearson correlation coefficient to estimate $\hat{\rho}_{LD}$, we need to use the EM algorithm or other strategies and impute values for unobserved data. The introduced uncertainty in the haplotype estimation can be taken into account as in Kulle et al. (2008).

5.2 Estimating CLD correlation from genotype data

According to Weir (1996, p. 122) and the proof in Appendix B, the gametic disequilibrium can be estimated directly from the observed genotypic frequencies,

$$\hat{P}_{AB} = \hat{P}^{AB}_{AB} + \frac{1}{2}\left(\hat{P}^{AB}_{Ab} + \hat{P}^{AB}_{aB} + \hat{P}^{AB}_{ab}\right). \tag{5.6}$$

The nongametic frequency, $\hat{P}_{A/B}$, is according to Weir (1996, p. 122) and the proof in Appendix B estimated by

$$\hat{P}_{A/B} = \hat{P}^{AB}_{AB} + \frac{1}{2}\left(\hat{P}^{AB}_{Ab} + \hat{P}^{AB}_{aB} + \hat{P}^{aB}_{Ab}\right). \tag{5.7}$$

The sum of Equation (5.6) and (5.7) gives

$$\hat{P}_{AB} + \hat{P}_{A/B} = 2\hat{P}^{AB}_{AB} + \hat{P}^{AB}_{Ab} + \hat{P}^{AB}_{aB} + \frac{1}{2}\left(\hat{P}^{AB}_{ab} + \hat{P}^{Ab}_{aB}\right)$$

which can be estimated directly from the observed data as shown in Table 5.1 and 5.2.

From the numerical coding of the random variables defined in Equation (4.14) and (4.15), we observe that the CLD correlation as defined in Equation (4.23) is the Pearson correlation coefficient with the numerical coding $0, 1, 2$. The numerical coding represents the wild type allele homozygote, heterozygote and variant type allele homozygote, respectively.

The estimated covariance between observed pairs of two random variables $(X_i, Y_i), i = 1, ..., n$, where n is the number of observations, representing genotypes at different locus defined as in Equation (4.14) and (4.15) is given by

$$\widehat{\text{Cov}}(\mathbf{X}, \mathbf{Y}) = \frac{1}{n}\sum X_i Y_i - \frac{1}{n^2}\sum X_i \sum Y_i \tag{5.8}$$

where

$$\mathbf{X} = (X_1, ..., X_n)$$

and

$$\mathbf{Y} = (Y_1, ..., Y_n).$$

Following the notation introduced in Table 5.1 we observe that Equation (5.8) gives

$$\widehat{\mathrm{Cov}}(\mathbf{X}, \mathbf{Y}) = \frac{1}{n}\left(2n_4 + n_5 + 4n_1 + 2n_2 - \frac{1}{n^2}(n_{Aa} + 2n_{AA})(n_{Bb} + 2n_{BB})\right).$$

From Equation (4.24) we have observed that

$$\widehat{\mathrm{Cov}}(\mathbf{X}, \mathbf{Y}) = 2\Delta_{AB}.$$

The allele frequencies, p_A and p_B are estimated by

$$\begin{aligned}\hat{p}_A =& \frac{(2n_{AA} + n_{Aa})}{2n} \\ \hat{p}_B =& \frac{(2n_{BB} + n_{Bb})}{2n}.\end{aligned}$$

The empirical variance of $\mathbf{X}$ is then given by

$$\begin{aligned}\widehat{\mathrm{Var}}(\mathbf{X}) =& \frac{1}{n}\sum X_i^2 - \left(\frac{\sum X_i}{n}\right)^2 \\ =& \frac{1}{n}\left(n_{Aa} + 4n_{AA}\right) - \left(\frac{n_{Aa} + 2n_{AA}}{n}\right)^2 \\ =& \frac{n_{Aa} + 2n_{AA}}{n} + \frac{2n_{AA}}{n} - \left(\frac{n_{Aa} + 2n_{AA}}{n}\right)^2 \\ =& 2\hat{p}_A + 2\hat{p}_{AA} - 4\hat{p}_{A^2} \\ =& 2[\hat{p}_A(1 - \hat{p}_A)] + \hat{P}_{AA} - \hat{p}_{A^2} \\ =& 2[\hat{p}_A(1 - \hat{p}_A) + \hat{D}_A].\end{aligned}$$

Similarly, we get the empirical variance of $\mathbf{Y}$,

$$\widehat{\mathrm{Var}}(\mathbf{Y}) = 2[\hat{p}_B(1 - \hat{p}_B) + \hat{D}_B].$$

The estimate of the composite linkage disequilibrium correlation is then

$$\hat{\rho}_{CLD} = \frac{\hat{\Delta}_{AB}}{\sqrt{(\hat{p}_A(1 - \hat{p}_A) + \hat{D}_A)(\hat{p}_B(1 - \hat{p}_B) + \hat{D}_B)}}.$$

5.3 Estimating LD and CLD correlation using R

We have implemented in R a function for estimating both the LD and CLD correlation (Appendix D), ρ_{LD} and ρ_{CLD}. As described by Gao, Starmer & Martin (2008), the CLD correlation is estimated using numerical coding $0, 1, 2$ of the three possible genotypes, aa, aA and AA, respectively. Since we do not know which of the alleles at each SNP that are considered as the high risk allele, our numerical coding of the data is based on the observed frequency of the different genotypes in the data, where we assume the less common allele to be the high risk allele. The homozygote with the assumed high risk allele, AA is coded as 2. The heterozygote genotypes are coded as 1, and the most common genotype are coded as 0. For some of the SNPs we observed only two different genotypes, and then the most common genotype was coded as 0, and the least common genotype was coded as 1. For chromosome 22 in the TOP8 data, we observed one SNP where the observed frequency for two of the genotypes was equal. The numerical coding, $0, 1, 2$ was for this SNP chosen in alphabetical order.

We compared our function for estimating the LD and CLD correlation (Appendix D) to the LD function from the genetics package (Warnes, with contributions from Gregor Gorjanc, Leisch, & Man. 2011) in R. We have observed that our function for estimating the LD correlation gives a small difference in the results compared to the function from the genetics package. Looking at the description of the LD function in the genetics package (Warnes et al. 2011), we see that this function includes all information about the alleles in each SNP, not considering for which pair of the SNPs we have pairwise complete observations. Our implemented function as described in Appendix D, takes into account if the observations for the SNPs for all persons in the study are pairwise complete. Using this procedure, we will loose some information about the allele frequencies at each SNP, but since we want to compare the ρ_{LD} and ρ_{CLD} correlation and the ρ_{CLD} correlation is based on pairwise complete observations, we decide to also use pairwise complete observations in the estimation of the LD correlation, $\hat{\rho}_{LD}$, between the SNPs.

5.4 LD vs. CLD correlation on a real data set

In Section 4.3, we investigated the difference between the LD and CLD correlation based on a theoretical grid of all possible combinations of the haplotype frequencies $p_1, ..., p_9$ as described in Table 5.2. We then estimated the CLD and LD correlations for the SNPs on chromosome 22 in the TOP8 data, and the summary statistics for the two measures of correlations are given by Table 5.5.

Table 5.5: Summary statistics for LD and CLD correlation for the TOP8 data.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
LD	-0.9961000	-0.0180600	-0.0001673	0.0007110	0.0179200	0.9998000
CLD	-0.9962000	-0.0180300	-0.0001713	0.0006997	0.0178500	1.0000000

We compared the LD and CLD correlations for the two cases when both are positive and both are negative, to see if one of the correlation measures seems to be more extreme than the other and compare to the theoretical results as described in Table 4.5, 4.6 and 4.7. We compared the CLD correlation and the LD correlation and we observe from the results in Table 5.6 that the LD correlation is more extreme than the CLD correlation in approximately 52% of the cases.

Table 5.6: LD vs CLD correlation for the TOP8 data.

ρ_{CLD}, ρ_{LD}	$\rho_{CLD} - \rho_{LD}$	proportion
$\rho_{CLD}, \rho_{LD} > 0.01$	$\rho_{CLD} - \rho_{LD} > 0$	0.480584
$\rho_{CLD}, \rho_{LD} > 0.01$	$\rho_{CLD} - \rho_{LD} < 0$	0.519416
$\rho_{CLD}, \rho_{LD} < -0.01$	$\rho_{CLD} - \rho_{LD} > 0$	0.515839
$\rho_{CLD}, \rho_{LD} < -0.01$	$\rho_{CLD} - \rho_{LD} < 0$	0.484161

One great advantage with the use of the CLD correlation instead of the LD correlation is that estimation of the CLD correlation is less computationally intensive. The CLD correlation is estimated directly from the observed genotypes, and estimating LD correlation we need to estimate haplotype frequencies as discussed in Section 5.1. In the TOP8 data, chromosome 22 contained 8928 SNPs for 1551 individuals. We estimated the CLD correlation matrix for this chromosome using the `cor` function in R (R Development Core Team 2011), and the computation time was approximately 20 minutes [1]. The LD correlation matrix for the same chromosome was estimated using maximum likelihood estimation as described in Appendix D and the computation time was approximately 4 days.

[1] 4 CPU cors, 1.8 GHz Intel i7

MAF and LD correlation

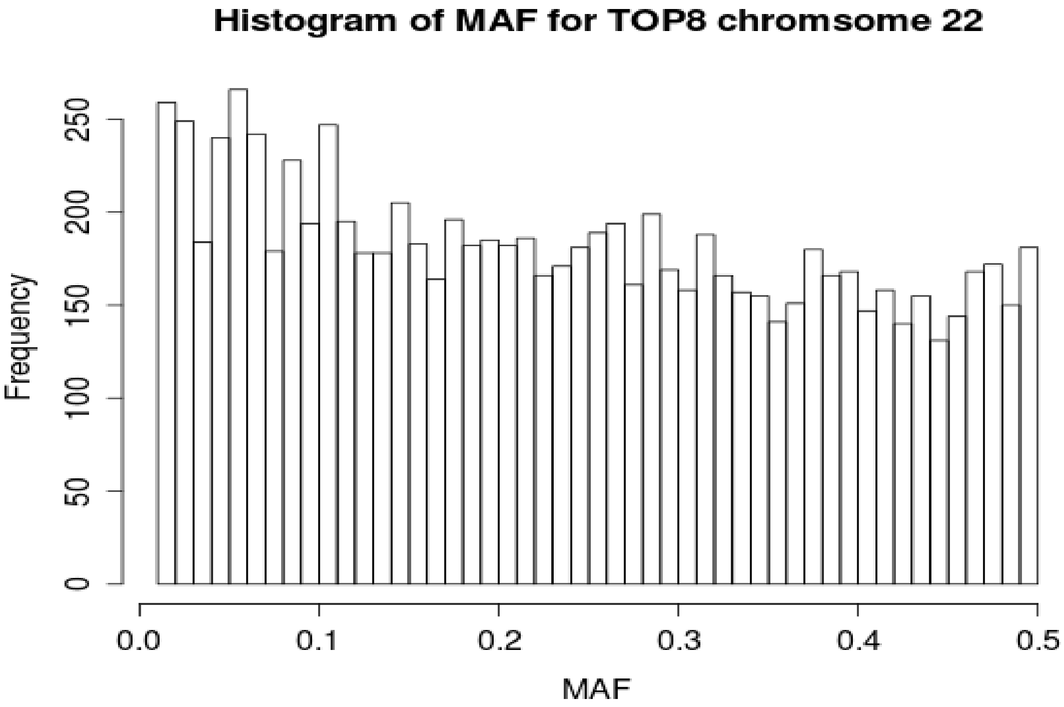


Figure 5.1: Histogram of MAF for chromosome 22 of the TOP8 data.

From Figure 5.1 we see the distribution of the minor allele frequencies for chromosome 22 of the TOP8 data.

From Figure 5.2 we observe that the maximal LD correlation between SNPs depends on the MAF for the different SNPs, and from Table 5.7 we observe that the smallest maximal value of the LD correlation is obtained when one SNP has MAF 0.01 and the other SNP has MAF 0.5, the maximal LD correlation is then equal to 0.10050. From Table 5.7 we also observe that the maximal LD correlation between two SNPs is obtained when the MAF for both SNPs are equal, the maximal LD correlation is then equal to 1.

From Figure 5.3 we observe that the maximal LD correlation between different pairs of SNPs in the TOP8 data are strongly dependent on the MAF for the different SNPs.

Table 5.7: MAF and maximal LD correlation

MAF	0.01	0.05	0.1	0.2	0.3	0.4	0.5
0.01	1.00000	0.43809	0.30151	0.20100	0.15352	0.12309	0.10050
0.05	0.43809	1.00000	0.68825	0.45883	0.35044	0.28098	0.22942
0.1	0.30151	0.68825	1.00000	0.66667	0.50918	0.40825	0.33333
0.2	0.20101	0.45883	0.66667	1.00000	0.76376	0.61237	0.50000
0.3	0.15352	0.35044	0.50918	0.76376	1.00000	0.80178	0.65465
0.4	0.12309	0.28098	0.40825	0.61237	0.80178	1.00000	0.81650
0.5	0.10050	0.22942	0.33333	0.50000	0.65465	0.81650	1.00000

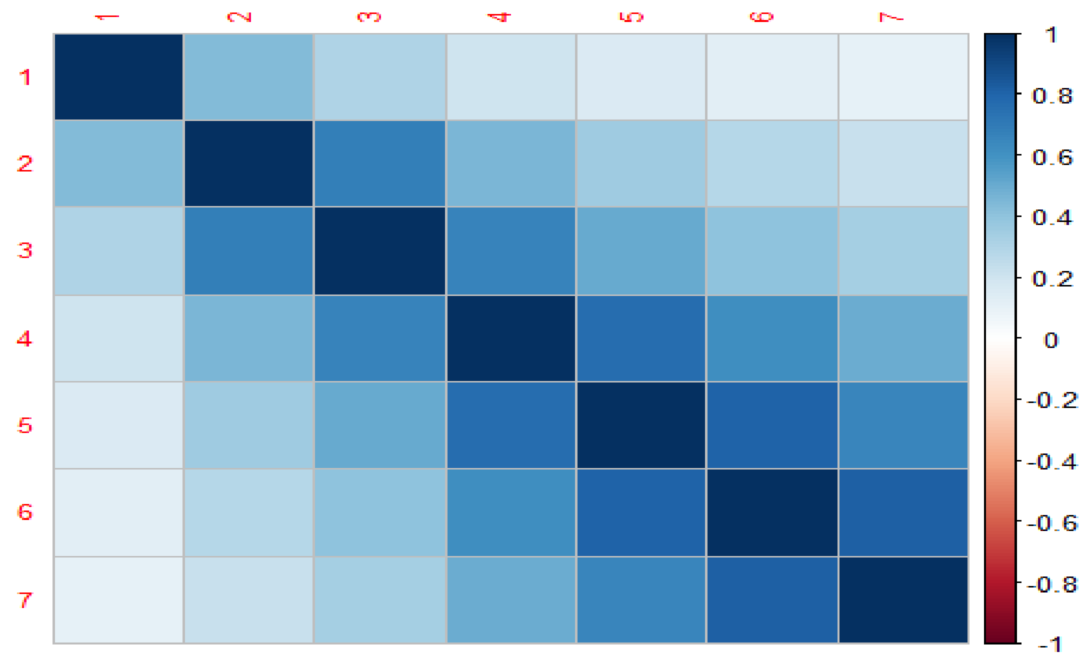


Figure 5.2: Plot of maximal LD correlation for a grid of MAF between 0.01 and 0.5. The MAF in the interval 0.01-0.5 are indexed by 1-7, respectively. We observe that the maximal LD correlation between SNPs depends on the MAF for the different SNPs.

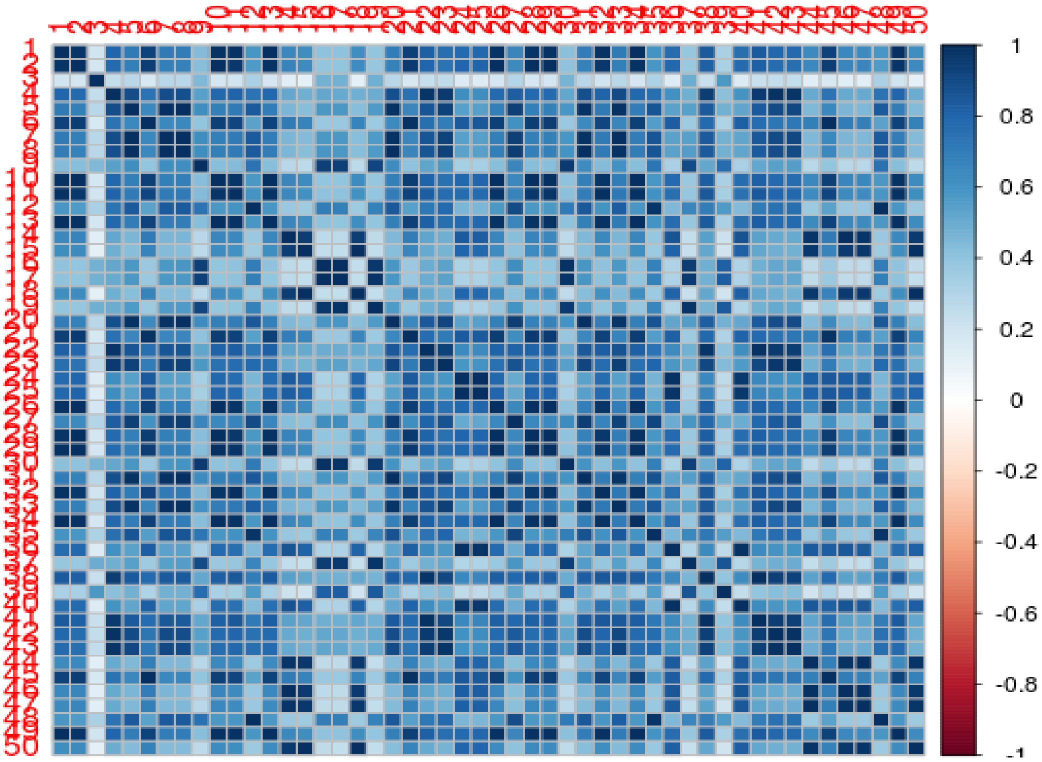


Figure 5.3: Correlation plot for the 50 first SNPs on chromosome 22 in the TOP8 data using maximal value of LD correlation between the pairs of SNPs. From this plot we observe that the maximal LD correlation between two SNPs is strongly dependent on the MAF for each SNP.

5.5 Estimating haplotypes

There exists several software programs for estimating haplotypes along the genome, for example the programs HAPLOVIEW (Barret, Fry, Maller & Daly 2005) and PHASE (Stephens & Scheet 2005, Stephens, Smith & Donnelly 2001). The default algorithm for estimating the haplotypes in HAPLOVIEW is the algorithm described by Gabriel (2002).

The method by Gabriel (2002) for estimating haplotypes is based on confidence intervals for the scaled linkage disequilibrium measure D', described in Section 4.1. The history of recombination between a pair of SNPs can be estimated using the scaled LD measure, D'. When we have a sample with rare alleles or we have only a small number of samples, it is known (Gabriel 2002) that the values of the LD measure D' will fluctuate upward. Therefore, the method by Gabriel (2002) is based on confidence intervals for D' rather than points estimates.

The method by Gabriel (2002) classifies the pair of SNPs into three categories, "strong LD", "historical evidence of recombination" and "others". Pairs of SNPs are classified as "strong LD" when the one sided upper bound for the confidence interval for D' is > 0.98 and the lower confidence bound is > 0.7. If the upper confidence bound for D' is < 0.9, the pairs of SNPs are classified as "strong evidence for historical recombination".

The method of Gabriel (2002) define a haplotype block as a region where only a small proportion (5%) of the SNP pairs show "strong evidence of historical recombination", and a haplotype block is found by counting the number of SNP pairs over a region which show "strong evidence of historical recombination".

Chapter 6

Hypothesis testing

In this chapter, we will present some background for single hypothesis testing. Test for HWE will be presented in Section 6.3, and a test for association between genotype and phenotype will be presented in Section 6.4.

Definition. *The two complementary hypotheses in a hypothesis testing problem are the null hypothesis, H_0 and the alternative hypothesis H_1 (Casella & Berger 2002, p. 373).*

Let θ denote a population parameter, then a general hypothesis testing problem can be written as

$$H_0 : \theta \in \Theta_0 \quad \text{and} \quad H_1 : \theta \in \Theta_0^C \tag{6.1}$$

where Θ_0 are a subset of the parameter space and Θ_0^C are the complement of Θ_0.

While testing the hypothesis in Equation (6.1), two types of errors are possible, type I error and type II error. Type I error describes the probability of rejecting H_0 while H_0 is true. Type II error describes the probability of accepting H_0 while the alternative hypotehsis H_1 is true.

Definition. *Type I error is defined as the probability of erroneously rejecting a true null hypothesis,*

$$P(\textit{type I error}) = \alpha.$$

Definition. *Type II error is defined as the probability of not rejecting H_0 when H_0 is false,*

$$P(\textit{type II error}) = \beta.$$

Type I errors is named false positives, type II errors are named false negatives. The two types of errors for a single hypothesis test can be summarized in Table 6.1 (Casella & Berger 2002).

Table 6.1: Single hypothesis testing set-up

	Not reject H_0	Reject H_0
H_0 true	Correct	Type I error
H_0 false	Type II error	Correct

6.1 P-values

Let $\mathbf{X} = (X_1, ..., X_n)$ be independent and identical distributed variables.

Definition. *A p-value $p(\mathbf{X})$ is a test statistic satisfying $0 \leq p(x) \leq 1$ for every sample point x. Small values give evidence that H_1 is true. A p-value is valid if, for every $\theta \in \Theta_0$ and every $0 \leq \alpha \leq 1$,*

$$P_\theta(p(\mathbf{X}) \leq \alpha) \leq \alpha$$

where α is the significance level (Casella & Berger 2002, p. 397).

The p-value gives information about the probability of observing what we have observed or more extreme given that the null-hypothesis H_0 is true. When the p-value is small, we reject the null hypothesis. If the p-value is less than the significance level, then we reject H_0, i.e the probability of rejecting the null hypothesis is less than or equal to the given significance level.

If $P_\theta(p(\mathbf{X}) \leq \alpha) = \alpha$, the p-value is called an exact p-value. The probability distribution of the p-value is then the uniform distribution.

6.2 Power of a test

The power of a single hypothesis test is the probability of rejecting the null hypothesis given that it is false. Let $\mathbf{X} = (X_1, ..., X_n)$ be independent and identically distributed variables. The power function of a hypothesis test with rejection region R is defined by (Casella & Berger 2002, p. 383)

$$\beta(\theta) = P_\theta(\mathbf{X} \in R).$$

6.3 Test for Hardy-Weinberg equilibrium

When assuming Hardy-Weinberg equilibrium (HWE) in a case-control study, HWE should only be assumed for the control group, not for the cases in the study. Assumption of HWE among the cases may lead to erroneous conclusions about association between genotype and phenotype, and because of this, association between genotype and phenotype can be seen as non-random mating.

Goodness-of-fit test for HWE

Consider a biallelic locus, X, with alleles A and a. The frequencies for the genotypes at the locus can be summarized as shown in Table 6.2.

Table 6.2: Genotype frequencies for locus X

AA	Aa	aa	Total
P_{AA}	P_{Aa}	P_{aa}	1

The allele frequencies, p_A and p_a, are given by

$$p_A = P_{AA} + \frac{1}{2} P_{Aa}$$

and

$$p_a = P_{aa} + \frac{1}{2} P_{Aa}.$$

The sum of the allele frequencies is

$$p_A + p_a = 1.$$

We use the χ^2 test for deviation to test for HWE. Assume data for locus X for n individuals. The observed and expected counts for locus X are summarized in Table 6.3.

The χ^2 test statistic for deviation is given by

$$\chi^2 = \frac{(n_{AA} - np_A^2)^2}{np_A^2} + \frac{(n_{Aa} - 2np_A(1-p_A))^2}{2np_A(1-p_A)} + \frac{(n_{aa} - n(1-p_A)^2)^2}{n(1-p_A)^2}.$$

This test statistic is χ^2 distributed with one degree of freedom, and is used to test the null hypothesis of Hardy-Weinberg equilibrium.

Table 6.3: Test for HWE

Genotype	AA	Aa	aa	Total
Observed	n_{AA}	n_{Aa}	n_{aa}	n
Expected	np_A^2	$2np_A(1-p_A)$	$n(1-p_A)^2$	n

6.4 Test for association between genotype and phenotype

The presentation in this section is based on Langaas & Bakke (2012). When considering biallelic markers, there are several possible genetic models for association between genotype and phenotype to consider (Ziegler & König 2010, p. 30). The three most popular different genetic models are the recessive, additive and dominant models, and are based on the number of the high risk allele at the loci. For biallelic markers, we index the three genotypes as aa, aA and AA where A is assumed to be the high risk allele. We use the numerical coding $0, 1, 2$ for the genotypes aa, aA and AA respectively.

Table 6.4: SNP data

	0	1	2	Total
Case	x_0	x_1	x_2	n_1
Control	y_0	y_1	y_2	n_2
Total	m_0	m_1	m_2	n

For a given biallelic SNP in the study we can set up a 2×3 contingency table as shown in Table 6.4. The total number of cases is given by n_1 and the total number of controls is n_2. The number of individuals with genotype $i, i = 0, 1, 2$ is given by $m_i, i = 0, 1, 2$ respectively. The total number of individuals in the study is

$$n = n_1 + n_2 = m_0 + m_1 + m_2.$$

Each individual $i, i = 1, ..., n$ is denoted by $z = (x_0, ..., y_2)$.

Genetic models

We denote the prevalence, $P(\text{case})$, of the disease by

$$\text{prevalence} = P(\text{case}) = k.$$

This can not be observed in a case-control study.

We define $f_i, i = 0, 1, 2$ to be the penetrance, i.e. the probability that the individual belongs to the case group given that the individual has genotype i,

$$f_i = P(\text{case}|\text{genotype } i). \tag{6.2}$$

When testing for association between genotype and phenotype, the null hypothesis of no association between genotype and phenotype can be expressed in terms of the penetrances, $f_i, i = 0, 1, 2$ as defined in Equation (6.2),

$$H_0 : f_0 = f_1 = f_2. \tag{6.3}$$

The null hypothesis in Equation (6.3) can also be given in terms of the conditional probabilities of having genotype i given the disease status of the individual. We define p_i as the conditional probability for an individual having genotype i given that the individual belongs to the case group, and q_i as the conditional probability of having genotype i given that the individual belongs to the control group,

$$p_i = P(\text{genotype } i|\text{case}),$$

and

$$q_i = P(\text{genotype } i|\text{control}).$$

Using Bayes' rule, $P(A|B) = P(B|A)P(A)/P(B)$, we have

$$\begin{aligned} f_i &= P(\text{case}|\text{genotype } i) \\ &= \frac{P(\text{genotype } i|\text{case})P(\text{case})}{P(\text{genotype } i)} \\ &= \frac{kp_i}{g_i}, \end{aligned}$$

and

$$\begin{aligned} 1 - f_i &= P(\text{control}|\text{genotype } i) \\ &= \frac{P(\text{genotype } i|\text{control})P(\text{control})}{P(\text{genotype } i)} \\ &= \frac{(1-k)q_i}{g_i}. \end{aligned}$$

The equations for f_i and $1 - f_i$ can be rewritten as

$$p_i = \frac{f_i g_i}{k} \text{ and } q_i = \frac{(1-f_i)g_i}{(1-k)}.$$

When the null hypothesis defined in Equation (6.3) is true, all p_i/q_i will be equal, and since both probabilities p_i and q_i sum to one, we have $p_i = q_i, i = 0, 1, 2$. The equivalent form of the null hypothesis in Equation (6.3) is given by

$$H_0 : p_0 = q_0, \quad p_1 = q_1, \quad p_2 = q_2.$$

When testing for association between genotype and phenotype, the alternative hypothesis will be different for the recessive, additive and dominant models. In the recessive model, we assume that two copies of the high risk allele at a locus is necessary for having the disease. The alternative hypothesis for the recessive model can then be given by

$$H_1 : f_0 = f_1 < f_2, \text{or } H_1 : p_0/q_0 = p_1/q_1 < p_2/q_2.$$

In the additive model, we assume that genotype aA gives an increased risk of the disease compared to the risk when having genotype aa, but a smaller risk than when having genotype AA. The alternative hypothesis for the additive model is then given by

$$H_1 : f_0 < f_1 < f_2, \text{ or } H_1 : p_0/q_0 < p_1/q_1 < p_2/q_2.$$

In the dominant model, we assume that individuals having one or two copies of the high risk allele will be affected by the disease. The alternative hypothesis for the dominant model is given by

$$H_1 : f_0 < f_1 = f_2, \text{ or } H_1 : p_0/q_0 < p_1/q_1 = p_2/q_2.$$

The genetic models are illustrated in Figure 6.1, where the y-axis represents the probability of disease.

The Cochran-Armitage test for trend

The Cochran-Armitage test for trend (CATT$_s$) is often used to test for association between genotype and phenotype. The CATT$_s$ statistic can, following the notation by Langaas & Bakke (2012), be written as

$$\text{CATT}_s = \frac{\sum_{i=0}^{2} s_i(n_2 x_i - n_1 y_i)}{\sqrt{n_1 n_2 (\sum_{i=0}^{2} s_i^2 m_i - \frac{1}{n}(\sum_{i=0}^{2} s_i m_i)^2}},$$

where s_0, s_1, s_2 are scores describing the genetic model. The absolute value of the CATT$_s$ statistic is invariant under linear transformation of the scores, and the

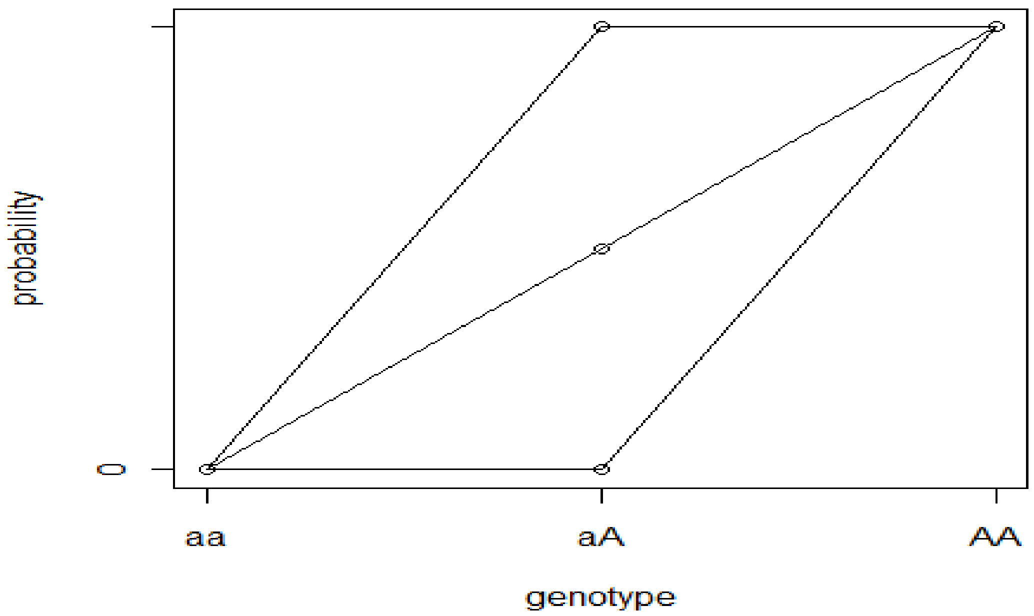


Figure 6.1: Figure illustrating the three different genetic models, recessive, additive and dominant model.

scores for the CATT$_s$ test statistic are $(s_0, s_1, s_2) = (0, s, 1)$. The index s denotes the chosen genetic model, and the recessive, additive and dominant model are denoted by $s = 0, 1/2, 1$ respectively. This test statistic asymptotically has the standard normal distribution under H_0, and the squared statistic then asymptotically has the chi-square distribution with one degree of freedom.

The CATT$_s$ statistic can also be expressed in terms of the Pearson correlation coefficient. Let r be the Pearson correlation coefficient between the score vector and the disease status vector and n be the number of individuals in the study. Then, the CATT$_s$ statistic is given by

$$\mathrm{CATT}_s = \sqrt{n}r.$$

MAX3 test

The MAX3 statistic is the maximum of the CATT_s statistics for the recessive, additive and dominant model, and is given by

$$\max(\text{CATT}_0, \text{CATT}_{1/2}, \text{CATT}_1).$$

For the data analyzed in this thesis, we do not know which of the alleles at each locus that is assumed to be the high risk allele. When the potential high risk allele is unknown, we can use

$$\text{MAX3} = \max(|\text{CATT}_0|, |\text{CATT}_{1/2}|, |\text{CATT}_1|). \tag{6.4}$$

which will cover all possible combinations of genetic models when the high risk allele is unknown. The p-value for the MAX3 statistic defined in Equation (6.4) is given by

$$P(\text{MAX3} < t) = P(|\text{CATT}_0| < t, |\text{CATT}_{1/2}| < t, |\text{CATT}_1| < t).$$

Conditioning on sufficient statistic

When analyzing a GWAS data set, the row sums will in general be different for the different SNPs because there are a different number of missing data for each SNP. The column sums $M = (m_0, m_1, m_2)$ will be different for most of the SNPs. Conditioning on the column sums, $M = (m_0, m_1, m_2)$, we get a trivariate hypergeometric probability

$$P(Z = z|M = (m_0, m_1, m_2)) = \frac{\binom{m_0}{x_0}\binom{m_1}{x_1}\binom{m_2}{x_2}}{\binom{n}{n_1}},$$

where n_1 is the number of cases and n is the total number of individuals in the study.

Under the null hypothesis of no association between genotype and phenotype, the column margins $M = (m_0, m_1, m_2)$ are sufficient statistics for the genotype frequencies (g_0, g_1, g_2). Conditioning on the sufficient statistics, the column margins, gives the conditional p-value (Langaas & Bakke 2012), denoted the C p-value

$$\begin{aligned} p(z_{\text{obs}}) &= P(T(Z) \geq T(z_{\text{obs}})|M = (m_0, m_1, m_2)) \\ &= \sum_{T(z) \geq T(z_{\text{obs}})} P(Z = z|M = (m_0, m_1, m_2)), \end{aligned} \tag{6.5}$$

where the sum $T(Z) \geq T(z_{\text{obs}})$ is over all possible tables with column margin $M = (m_0, m_1, m_2)$ where the MAX3 test observator $T(Z)$ is larger than or equal to the observable MAX3 test observator $T(z_{\text{obs}})$.

In a general situation we consider $r \times c$ contingency tables of nonnegative integers where r and c are positive integers. Following the notation introduced in Table 6.4 the number of tables having row sums $(n_1, ..., n_r)$ are (Bakke & Langaas 2012)

$$\prod_{i=1}^{r} \binom{n_i + c - 1}{c - 1}.$$

In this thesis, we consider a case-control study where the data for each SNP can be represented by a 2×3 contingency table as shown in Table 6.4. The number of 2×3 tables with given row sums (n_1, n_2) and column sums (m_0, m_1, m_2) are given by (Bakke & Langaas 2012)

$$\binom{n_1+2}{2} - \binom{n_1 - m_0 + 1}{2} - \binom{n_1 - m_1 + 1}{2} - \binom{n_1 - m_2 + 1}{2} + \binom{n_1 - m_0 - m_1}{2} + \binom{n_1 - m_0 - m_2}{2} + \binom{n_1 - m_1 - m_2}{2}$$

The maximal number of tables with given row sums as a function of the column sums, for $2 \times c$ tables when n is the lesser of the two row sums, for 2×2 tables is $n + 1$ (Bakke & Langaas 2012). For 2×3 tables, the maximum number of tables is

$$\binom{n+2}{2} - 3\binom{n - m + 1}{2} + r \max(n - m, 0),$$

where m and r are unique integers such that $n = 3m + r$ and $0 \leq r < 3$. When conditioning on the column margins, $M = (m_0, m_1, m_2)$ we observe that the possible number of contingency tables is reduced. The calculation of the C p-value in Equation (6.5) includes then a sum over a small number of tables compared to the maximal possible number of contingency tables, and then the estimation of the C p-value is less computational intensive.

For the TOP data analyzed in this thesis we have in the order of $n_1 = 1100$ cases and $n_2 = 400$ controls, which gives the maximum number of conditional tables equal to 80601. The maximum number of tables without conditioning on the row sums are $4.9 \cdot 10^{10}$, which shows that conditioning on the column margins gives substantial decrease in the computational complexity of the problem.

Chapter 7

Multiple testing

Analyzing experimental data often involve many simultaneous hypothesis tests. For each null hypothesis, an individual test is performed, and the significance level is usually set to $\alpha = 0.05$. This means that the probability of making a type I error is at most 5%. For the multiple testing problem, where in total m hypotheses are tested, the total type I error rate could be larger than 5% when we do not adjust for multiple testing. The goal in multiple testing problems is to control the total type I error rate at a given significance level. In a multiple testing problem, we have m hypotheses, $H_{0i}, i = 1, ..., m$ to be evaluated simultaneously. The multiple testing problem for a total number of m hypotheses can be summarized in Table 7.1 (Benjamini & Hochberg 1995).

Table 7.1: Multiple testing set-up

	Not reject H_0	Reject H_0	All
H_0 true	U	V	m_0
H_0 false	T	S	$m - m_0$
Total	$m - R$	R	m

In Table 7.1, V represents the number of type I errors, the number of erroneously rejected null hypotheses and T represents the number of type II errors, the number of hypotheses that are not rejected when H_0 is false. The total number of hypotheses in the multiple testing problem are denoted by m, and m_0 represents the number of true null hypotheses. The two only known variables in Table 7.1 are m and R.

7.1 Type I error rates

The definition of the familywise error rate (FWER), is given by the number of type I errors among all hypotheses. Assume V is the number of type I errors. Then, (Ge, Dudoit & Speed 2003, p. 7)

$$\text{FWER} = P(V > 0).$$

Strong control means control of the type I multiple error rate under any combination of true and false hypotheses (Ge et al. 2003, p. 8). Weak control means control of the type I multiple error rate under the complete null hypothesis. If the type I multiple error rate can be controlled under any combination of true and false hypotheses, it follows that the type I multiple error rate also can be controlled under the complete null hypothesis. This means that strong control implies weak control. The complete null hypothesis is the hypothesis that assume all m null hypotheses are true. The complete null hypothesis is denoted by H_0^C.

Some researchers distinguish between two different types of FWER, named FWEC and FWEP (Westfall & Young 1993, p. 9). A familywise error rate gives the probability of rejecting one or more true null hypotheses. FWEC is the familywise error rate calculated under the complete null hypothesis, i.e. when all subhypotheses H_{0i} are assumed to be true. The FWEC is given by

$$\text{FWEC} = P(\text{Reject at least one } H_{0i} | \text{all } H_{0i} \text{ are true}).$$

The FWEP is the familywise error rate calculated under the partial null hypothesis, i.e. assuming that only a subset of the null hypotheses are true. Then,

$$\text{FWEP} = P(\text{Reject at least one } H_{0i}, i = j_1, ..., j_t | H_{0j_1}, ..., H_{0j_t} \text{ are true}).$$

From this expression we see that the FWEP depends upon which subset of null hypotheses that are true.

We also have similar expressions for the type I error rate that can be used, i.e. the per comparison error rate

$$\text{PCER} = E(V)/m$$

where m is the total number of tests.

The false discovery rate, FDR, is defined as

$$\text{FDR} = E\left(\frac{V}{R} \cdot I(R > 0)\right).$$

Here we must use the indicator function since we need to consider the special case where R, the total number of rejections can be zero.

FWER vs FDR

Comparing FWER and FDR we see that the FWER and FDR are equivalent if all the null hypotheses are true (Benjamini & Hochberg 1995). If only a subset of the hypotheses are true the FDR is less than the FWER. The differences between the FWER and FDR becomes larger when the number of non-true null hypotheses increase (Benjamini & Hochberg 1995), which also will give increase in power.

Comparing the definitions of FWER and FDR, we see that one difference is that FWER focuses on probabilities while FDR focuses on expectations. Researchers that use the FWER criterion to control the type I error for the multiple testing problem are interested in the probability of erroneously reporting any result as statistically significant. For researchers using the FDR criterion, the most important is the proportion of false positives among all rejected hypotheses.

The Bonferroni procedure is the most known method for control of the FWER, while the Benjamini Hochberg step-up procedure is the most popular procedure for control of FDR. When analyzing SNP data, FWER is considered as the gold standard, but for gene expression data, the FDR is considered as the gold standard.

7.2 Distribution for the smallest p-value

Consider testing one hypothesis H_0 vs. H_1 using significance level $\alpha = 0.05$. The probability of declaring one test significant at 5% level is exact 0.05 when the p-values are exact. Assume that we divide the interval between from 0 to 1 into equal-sized intervals of length 0.05. Then we will have 20 intervals between 0 and 1. The probability for one p-value to fall in one of these intervals is then exact 0.05, and the probability of declaring one hypothesis test as significant at 5% level is exact 0.05. In multiple testing we assume that m tests are performed. The probability of declaring at least one of the m tests significant at level 0.05 is, assuming independent tests and under the complete null hypothesis,

$$\begin{aligned} P\left(\min_i p_i \leq 0.05\right) &= 1 - P\left(\min_i p_i > 0.05\right) \\ &= 1 - P(\text{all } p_i > 0.05) \\ &= 1 - (1 - 0.05)^m. \end{aligned}$$

In general, this can be written as

$$F(\alpha) = 1 - (1 - \alpha)^m$$

and then,

$$f(\alpha) = F'(\alpha) = m(1 - \alpha)^{m-1}$$

which has the form of a beta distribution $f(\alpha) \sim \text{Beta}(1, m)$ (Westfall & Young 1993, p. 8).

7.3 Adjusted p-values

In multiple testing, we often use adjusted p-values instead of the raw p-values. Raw p-values, p_i, are the lowest nominal level to reject H_0. The adjusted p-value $\tilde{p_i}$ is the nominal level of the simultaneous test procedure at which H_{0i} is just rejected, given the values of all test statistics involved. For any multiple testing procedure which controls FWER or FDR, the adjusted p-values can be defined as (Westfall & Young 1993, p. 11)

$$\tilde{p_i} = \inf\{\alpha \in [0,1] | H_{0i} \text{ is rejected at nominal level FWER/FDR } = \alpha\}.$$

When a multiple number of tests is performed and all hypotheses with adjusted p-value below α are rejected, FWER/FDR will control the type I error at level α.

For a multiple testing problem considering m hypotheses simultaneously we will correct for multiple testing using a individual significance level, α_p, for each individual test. We use a multiple testing correction method to find the significance level, α_p, for each individual test. Using α_p for the individual tests, the total type I error rate will be controlled at level α.

7.4 Single-step procedures for control of FWER

There exists three different groups of commonly used multiple testing procedures (Ge et al. 2003, p. 12), single-step, step-down and step-up procedures. In single-step procedures each individual hypothesis is evaluated using rejection regions that are independent of the result of the other hypotheses. In step-up and step-down methods, the rejection region depends on the result of the other hypotheses. In step-down procedures, the test statistics are ordered based on the most significant test statistic, and step-up procedures starts with the least significant test statistic. Step-down and step-up procedures are less conservative procedures than the single-step procedures for control of FWER. For a single-step method, only one p-value cutoff is used, while for step-down and step-up methods different cutoffs are used based on the rank of the p-value in question. We will only consider single-step procedures in this thesis.

The Bonferroni method

The Bonferroni method is the simplest single-step method multiple testing procedure for FWER control (Westfall & Young 1993, p. 44). The Bonferroni method gives strong control of the FWER at significance level α. In multiple testing of m tests, the Bonferroni method rejects the null hypothesis H_{0i} when the p-value p_i is less than $\alpha_p = \alpha/m$. The Bonferroni single-step adjusted p-value is given by

$$\tilde{p_i} = \min(mp_i, 1).$$

The Bonferroni single-step method (Westfall & Young 1993, p. 44) is given by

$$\begin{aligned} P(\text{Reject at least one } H_i|H_0^C) &= P\left(\min_{1\leq i\leq m} p_i \leq \alpha/m|H_0^C\right) \\ &\leq \sum_{i=1}^{m} P(p_i \leq \alpha/m|H_0^C) \end{aligned}$$

where the inequality above is named Bonferroni's inequality.

Bonferroni and strong control

To show that Bonferroni's method controls the FWER we use Booles's inequality. For m events $A_i, i = 1, ..., m$ Boole's inequality can be written as

$$P(\cup_{i=1}^{m} A_i) \leq \sum_{i=1}^{m} A_i.$$

We let A_i denote the event

$$\tilde{p_i} \leq \alpha.$$

Using the Bonferroni adjusted p-value we see that (Ge et al. 2003, p. 12)

$$\text{FWER} = P(V > 0) \leq P[\cup_{i=1}^{m_0}\{\tilde{p_i} \leq \alpha\}] \leq \sum_{i=1}^{m_0} P(\tilde{p_i} \leq \alpha) \leq \sum_{i=1}^{m_0} \alpha/m = m_0\alpha/m \leq \alpha$$

given that the number of true null hypotheses is equal to m_0.

The Šidák method

The Šidák method is derived by assuming that all the individual tests are independent. Assume that the total significance level is α. The Šidák single-step adjusted p-value is given by

$$\tilde{p}_i = 1 - (1-\alpha)^m,$$

and the significance level for the individual tests is given by

$$\alpha_p = 1 - (1-\alpha)^{1/m}$$

where m is the total number of tests (Westfall & Young 1993, p. 44).

Šidák's method and strong control

The Šidák method provides strong control of the FWER.

$$\begin{aligned} P(V=0) &= P(\cap_{i=1}^{m_0}\{\tilde{p}_i \geq \alpha\}) \\ &= \prod_{i=1}^{m_0} P(\tilde{p}_i \geq \alpha) \\ &= \prod_{i=1}^{m_0} P(p_i \geq 1-(1-\alpha)^{1/m}) \\ &= \{(1-\alpha)^{1/m}\}^{m_0}. \end{aligned}$$

The FWER is in Section 7.1 defined as

$$\text{FWER} = P(V>0) = 1 - P(V=0),$$

and this gives

$$\begin{aligned} \text{FWER} &= P(V>0) \\ &= 1 - P(V=0) \\ &= 1 - (1-\alpha)^{m_0/m} \\ &\leq \alpha. \end{aligned}$$

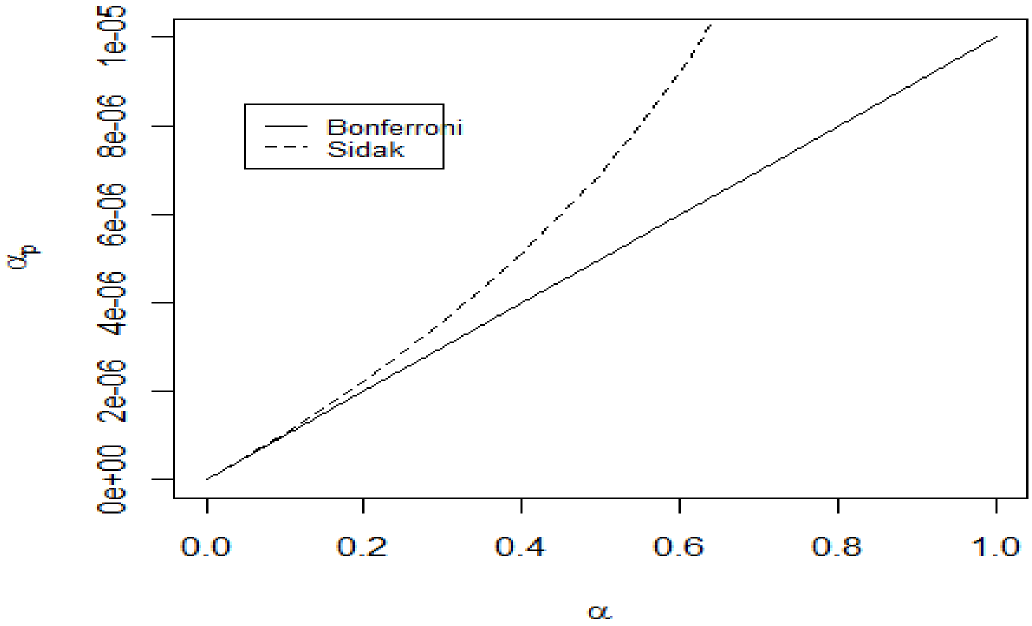


Figure 7.1: Plot of Bonferroni and Šidák correction for different values of α. FWER control at level α and α_p is the individual significance level.

Bonferroni vs. Šidák

Figure 7.1 shows the Bonferroni correction and Šidák correction for $m = 10000$ tests, and for different p-values in the interval [0,1]. This plot shows that the two procedures for multiple testing corrections are approximately equal for small values of α. Bonferroni and Šidák correction gives approximately equal results up to about $\alpha = 0.1$. The plot of the Bonferroni correction shows a straight line as expected, and the plot for the Šidák correction is as we would expect nonlinear.

minP single-step procedure

The minP adjusted p-values are by Westfall & Young (1993, p. 46) defined as

$$\tilde{p}_i = P\left(\min_{1\leq l\leq m} P_l \leq p_i | H_0^C\right)$$

where H_0^C denotes the complete null hypothesis as defined in Section 7.1, and P_l

is the random variable for the raw p-value of the lth hypothesis.

Resampling procedures based on minP adjusted p-values will provide weak control of the FWER under all conditions. We have

$$\begin{aligned}\text{FWER} &= P(V > 0)\\ &= 1 - P(V = 0)\\ &= 1 - P(\text{all } p_i > \alpha_p)\\ &= 1 - P(\min p_i > \alpha_p)\\ &= P(\min p \leq \alpha_p)\\ &\leq \alpha.\end{aligned}$$

The minP single-step procedure is a less conservative procedure than the Bonferroni procedure and if the data are independent less conservative than the Šidák procedure. An alternative to the minP procedure is the maxT single-step procedure. The maxT adjusted p-values, $\tilde{p}_i$, are based on the tests statistics, $T_i, i = 1, ..., m$, and is by Westfall & Young (1993, p. 50) defined as

$$\tilde{p}_i = P\left(\max_{1\leq j\leq k} |T_j| \geq |t_i| | H_0^C\right).$$

In this thesis, we do not have complete observations of our data. We have different number of missing data for each SNP and the test statistics, $T_i, i = 1, ..., m$, will then not be identically distributed. Therefore we do not consider maxT p-values, but use the more computationally intensive minP procedure.

minP and resampling

Resampling procedures are considered the gold standard in multiple testing problems. For case-control data as described in Table 6.4, we permuted the disease status vector, and for each permutation of the data, all m p-values was calculated and the minimum p-value was recorded. We repeated B times, to get B $\min P_j, j = 1, ..., B$ values. As estimate for α_p we use the $\alpha \cdot B$ order statistic in the minP distribution. The resampling algorithm was implemented using the C p-value as described in Section 6.4. We used $B = 100000$ resampled data sets in the resampling procedure.

Chapter 8

Multiple correction methods based on an effective number of independent tests, M_{eff}

When performing hypothesis testing with a large number of SNPs, the correlation structure among the SNPs in the data set needs to be considered. We have in Section 7.4 presented different methods to control the FWER when multiple hypotheses are considered. The Šidák method assumes independent tests, while the Bonferroni method allows for any correlation structure between dependent tests. The effective number of independent tests are denoted as M_{eff}. Methods based on M_{eff} use Šidák correction where the number of tests, m, are replaced with M_{eff}. Bonferroni and Šidák method keep the total error rate at a nominal level, α, by adjusting the error rates for each test at level α_p. We have seen in Section 7.4 that the formula for the individual significance level, α_p, for the Bonferroni method is

$$\alpha_p = \alpha/m$$

and using Šidák method α_p is given by

$$\alpha_p = 1 - (1-\alpha)^{1/m}$$

where m is the number of tests and α is the FWER significance level chosen.

Different methods to estimate the effective number of independent tests, M_{eff}, have been studied by researchers, for example Cheverud (2001), Nyholt (2004), Gao et al. (2008) and Moskvina & Schmidt (2008). The methods described by Nyholt (2004), Gao et al. (2008) and Moskvina & Schmidt (2008) for a single chromosome will be presented in the next sections. Application of the methods to genome-wide estimates will be discussed in Chapter 10.

8.1 The Cheverud-Nyholt method

The method of Nyholt (2004) is based on the method of Cheverud (2001). The idea of the method of Cheverud (2001) was to use the variance of the eigenvalues of the correlation matrix to construct M_{eff}. The method of Cheverud (2001) and Nyholt (2004) is based on spectral decomposition (SpD) of the pairwise correlation matrix between the SNPs. The difference between the methods of Cheverud (2001) and Nyholt (2004) is that the correlation matrix in the method of Cheverud (2001) is based on genotype data and the correlation matrix in the method of Nyholt (2004) is based on haplotype data as described in Section 4.1, thus information on phenotypes is not needed. Nyholt (2004) improved the method of Cheverud (2001) by removing all SNPs in perfect LD except one before estimating M_{eff} (Nyholt 2005).

Cheverud (2001) estimated the effective number of independent tests by

$$M_{\text{eff}} = m\left(1-(m-1)\frac{\text{Var}(\lambda)}{m^2}\right), \tag{8.1}$$

where $\text{Var}(\lambda_{\text{obs}})$ is the variance of the eigenvalues of the correlation matrix based on genotypic data. In the method of Nyholt (2004), the estimate of the effective number of tests in Equation (8.1) was rewritten as

$$M_{\text{eff}} = 1+(m-1)\left(1-\frac{\text{Var}(\lambda)}{m}\right) \tag{8.2}$$

where $\text{Var}(\lambda)$ is the variance of the eigenvalues of the correlation matrix based on pairwise linkage disequilibrium between the SNPs. The significance level for the individual tests, α_p, are found by using the Šidák method which gives

$$\alpha_p = 1-(1-\alpha)^{1/M_{\text{eff}}}$$

where M_{eff} is the effective number of tests.

When the SNPs are independent the correlation between SNPs are zero. Then all the eigenvalues $\lambda_1, \ldots.., \lambda_m$ are equal to 1. The variance of the eigenvalues is zero ($\text{Var}(\lambda) = 0$), so by using Equation (8.2), $M_{\text{eff}} = m$. The other special case is when there are perfect correlation between the SNPs. For this case, the first eigenvalue of the correlation matrix is equal to m, and the others are equal to zero. This means that $\text{Var}(\lambda) = m$ and the effective number of independent tests from Equation (8.2) is $M_{\text{eff}} = 1$.

The formula for the effective number of tests in Equation (8.2) is based on an interpolation of the two extreme cases (Salyakina, Seaman, Browning, Dudbridge

& Muller-Myhsok 2005), when the correlation between the SNPs are zero and when we have perfect correlation between the SNPs.

The general formula for linear interpolation between two points in the (x, y) plane is given by

$$y = y_a + (a - x_a)\frac{(y_b - y_a)}{(x_b - x_a)}.$$

In the method of Nyholt (2004), let x describe the variance of the eigenvalues, and let y describe the effective number of independent tests. The interpolation between the two points in the (x, y) plane $(0, m)$ (zero correlation) and $(m, 1)$ (perfect correlation) can be used to define M_{eff} as follows

$$\begin{aligned} M_{\text{eff}} &= m + \text{Var}(\lambda)\frac{(1-m)}{m} \\ &= 1 + (m-1) + \text{Var}(\lambda)\frac{(1-m)}{m} \\ &= 1 + (m-1)\left(1 - \frac{\text{Var}(\lambda)}{m}\right), \end{aligned}$$

which is the known form of the M_{eff} used in the method of Nyholt (2004).

8.2 Moskvina's alternative formulation for Nyholt's method

Moskvina & Schmidt (2008) gave an alternative formulation of Nyholt's method. This alternative method shows that the numerical expense and uncertainty with calculating eigenvalues of a large matrix can be avoided by direct calculation from the correlation matrix. Let $\mathbf{C} = (r_{jk}), j, k = 1, ..., m$ be the correlation matrix with eigenvalues $\lambda_1, ..., \lambda_m$. The average of these eigenvalues is

$$\frac{1}{m}\sum_{j=1}^{m}\lambda_j = \frac{1}{m}\text{trace}(\mathbf{C}) = \frac{1}{m}\sum_{j=1}^{m} r_{jj} = 1.$$

This gives the estimated variance as

$$\widehat{\text{Var}}(\lambda) = \frac{1}{m-1}\sum_{j=1}^{m}(\lambda_j - 1)^2 = \frac{1}{m-1}\left(\sum_{j=1}^{m}\lambda_j^2 - m\right). \tag{8.3}$$

Inserting the estimated variance from Equation (8.3) into the method of Nyholt (2004) as given in Equation (8.2) gives

$$M_{\text{eff}} = m + 1 - \frac{1}{m}\sum_{j=1}^{m}\lambda_j^2.$$

The eigenvalues of the correlation matrix, $\mathbf{C}$, are denoted $\lambda_j, j = 1, ..., m$ and the eigenvalues of the squared matrix, $\mathbf{C}^2$, are then $\lambda_j^2, j = 1, ..., m$. This gives

$$\sum_{j=1}^{m}\lambda_j^2 = \text{trace}(\mathbf{C}^2) = \sum_{j=1}^{m}\sum_{k=1}^{m} r_{jk}^2.$$

This shows that M_{eff} can be computed directly from the correlation coefficients by

$$M_{\text{eff}} = 1 + \frac{1}{m}\sum_{j=1}^{m}\sum_{k=1}^{m}(1 - r_{jk}^2).$$

8.3 The method of Gao et al. (2008)

The method of Gao et al. (2008), the simple$\mathcal{M}$ method, use the composite linkage disequilibrium (CLD) correlation to calculate the pairwise correlation matrix. The CLD correlation is described in Section 4.2. The method use the eigenvalues of the correlation matrix to estimate the effective number of tests, M_{eff}, and then the Bonferroni correction with M_{eff} to estimate α_p. Later Šidák correction was used in the method of Gao et al. (2008) instead of Bonferroni correction (Gao, Becker, Becker, Starmer & Province 2010). Since Šidák's method is based on independent tests this seems more appropriate than using the Bonferroni method, although for small α the two methods give approximately the same results as shown in Section 7.4.

The eigenvalues from the CLD correlation matrix, $\lambda_1, ..., \lambda_m$, are sorted in decreasing order,

$$\lambda_1 \geq \lambda_2 \geq ..... \geq \lambda_m.$$

The sum of the diagonal elements of $\mathbf{C}$ (called the total variance) is given by

$$\text{trace}(\mathbf{C}) = \sum_i \lambda_i$$

where $\mathbf{C}$ is the correlation matrix, and λ_i are the eigenvalues of the correlation matrix.

Suppose we have a $n \times m$ matrix with the numerical coding $0, 1, 2$ for each SNP. The standardized matrix where all columns has mean zero and standard deviation equal to one is denoted by $\mathbf{Z}$. The idea in principal components analysis (PCA) is to find a number of $q < m$ linear combinations which best represents the original data (Ripley 1996, p. 289). The principal components are found by taking the singular value decomposition $\mathbf{Z} = \mathbf{UDV}^T$, as described in Appendix C, where $\mathbf{D}$ is the diagonal matrix of the eigenvalues of $\mathbf{Z}$, $\mathbf{D} = \text{diag}(\lambda_1, ..., \lambda_m)$. The principal components are then the columns of the $\mathbf{ZV}$ matrix (Ripley 1996, p. 289).

The proportion of the total variance accounted for by each of the principal components equals

$$\delta_i = \frac{\lambda_i}{\sum_{j=1}^{m} \lambda_j}.$$

Because $\sum_{j=1}^{m} \lambda_j = m$, we see that

$$\delta_i = \frac{\lambda_i}{m}.$$

This is the ratio of the eigenvalue to the sum of all eigenvalues in the matrix, i.e. the ratio of the eigenvalue to the trace of the diagonal matrix of eigenvalues.

In principal components analysis, the first principal component explain most of the variation in the data. The second principal component is normal to the first principal component and will explain most of the remaining variation in the data after the first principal component is found. This means that to explain a given percent of the total variation in the data, only the first x eigenvalues are needed given a predetermined cutoff. This gives

$$\frac{\sum_{i=1}^{x} \lambda_i}{m} > c$$

where c is the predetermined cutoff.

In general we want to find the number of eigenvalues, x, such that we are able to explain a given percent of the variation for the data. The effective number of independent tests from Gao et al. (2008) is given by

$$M_{\text{eff}} = x.$$

Gao et al. (2010) used Šidák correction to calculate the significance level for each individual test,

$$\alpha_p = 1 - (1 - \alpha)^{(1/M_{\text{eff}})}.$$

The number of nonzero eigenvalues

The method of Gao et al. (2008) was originally defined as calculating the eigenvalues from the matrix of pairwise CLD correlations between the SNPs. Gao et al. (2008) analyzed a relatively small data set with 1723 SNPs and 500 persons. The method was originally described without dividing the correlation matrix into blocks, but analyzing the data set in Gao et al. (2008) the data were divided into smaller blocks, all of size ~ 133 SNPs. It was not written explicitly by Gao et al. (2008) why this block size was used. Alternative methods for choosing block sizes was also discussed by Gao et al. (2008), and using a software program called HAPLOVIEW (Barret et al. 2005), the data set was divided into blocks of size about $100 - 140$ SNPs.

The data analyzed in Gao et al. (2008) are divided into blocks with the justification that there is a problem with calculating eigenvalues efficiently when the number of SNPs is large. We observe from the theory in Appendix C that for an $n \times p$ matrix, where $n \leq p$, the maximal number of nonzero eigenvalues is equal to $n - 1$. Without using blocks we see that for the data set with 1723 SNPs and 500 persons analyzed in Gao et al. (2008), the maximal number of nonzero eigenvalues will be equal to $500 - 1 = 499$. The results of Gao et al. (2008) using blocks shows that the effective number of independent tests using ~ 133 SNPs in each block is $M_{\text{eff}} = 1132$, which seems to have been found from the sum of the estimates for each block.

From Gao et al. (2010) we see that in the Illumnia 1M data there are $n = 656$ individuals. From command line 81 in the R code of Gao X. (2012) we observe that the method used block size equal to 133, but in the article it is not explicitly given which block size that is used. As explained above the maximal number of nonzero eigenvalues in this correlation matrix is $n-1 = 655$, so because the chosen block size equal to 133 is less than $n - 1 = 655$, the problem with the number of nonzero eigenvalues is avoided in Gao et al. (2010).

We observe that the problem with the number of nonzero eigenvalues in the case when $n \leq p$ can be avoided by dividing the correlation matrix into blocks of size at most equal to n. For each block of size n the maximal number of nonzero eigenvalues then is equal to $n - 1$, and assuming independence between blocks, the sum of the M_{eff}'s for each block will be an estimate for the total number of independent tests, M_{eff}.

8.4 The method of Moskvina and Schmidt (2008)

The method by Moskvina & Schmidt (2008) use the following estimate of the overall type I error probability, FWER,

$$\alpha \leq 1 - (1 - \alpha_p)^{M_{\text{eff}}},$$

where α_p is the individual significance level and M_{eff} is the estimate of the effective number of independent tests. The estimate M_{eff} is given by

$$M_{\text{eff}} = 1 + \sum_{j=2}^{m} \kappa_j,$$

where

$$\kappa_j = \frac{1}{\log(1-\alpha_p)} \log \left(1 - \frac{1}{(1-\alpha_p)} \sqrt{\frac{2}{\pi}} \int_{-\sigma}^{\sigma} e^{-x^2/2} \Phi \left(\frac{r_j x - \sigma}{\sqrt{1 - r_j^2}} \right) dx \right). \quad (8.4)$$

In Equation (8.4), $r_j = \max_{1 \leq k \leq j-1} |r_{kj}|$, where r_{kj} is the pairwise haplotypic Pearson's correlation coefficient between SNP at locus k and SNP at locus j as described in Section 4.1. $\Phi(x)$ is the cumulative distribution function of the standard normal distribution and σ is the $(1 - \alpha_p/2)$ quantile.

For $\alpha_p \leq 0.01$, Moskvina & Schmidt (2008) gave the approximation

$$\kappa_j \approx \sqrt{1 - r_j^{-1.31 \cdot \log_{10} \alpha_p}}. \quad (8.5)$$

We now go through the presentation of the method of Moskvina & Schmidt (2008) considering a case-control study where we for each SNP can set up a table as shown in Table 8.1.

Table 8.1: Data for SNP at locus X

	0	1	Total
Case	x_0	x_1	n_1^*
Control	y_0	y_1	n_2^*
Total	m_0	m_1	n^*

In Table 8.1 $n_1^* = 2n_1, n_2^*$ and $n^* = 2n$, where n is the total number of individuals and n_1 and n_2 is the total number of cases and controls, respectively. We consider a biallelic SNP with alleles A and a, which in Table 8.1 is denoted by 1 and 0 respectively, assuming that A is the high risk allele. The total number of allele a and allele A among the $2n$ gametes are m_0 and m_1, respectively.

We define the probabilities

$$p_1 = P(\text{allele A}|\text{case})$$

and

$$p_2 = P(\text{allele a}|\text{control}).$$

For each SNP, the information about each individual can be expressed by a vector

$$Z = (z_i) = (1, 0, 0, 1, 1, ...),$$

of length n^*. Each individual is represented with two elements representing each gamete, indicating whether the high risk allele is present or not present.

We have the following estimators

$$\hat{p}_1 = \frac{x_1}{n_1^*}, \hat{p}_2 = \frac{y_1}{n_2^*},$$

for cases and controls, respectively.

We will test for difference in the frequency of the high risk allele between the case and control group using an allelic test. The null hypothesis for the allelic test is that there is no difference in frequency between the two groups,

$$H_0 : p_1 = p_2.$$

For the high risk allele, A, we have the following estimator for the total population

$$\hat{p} = \frac{x_1 + y_1}{n^*} \approx p \text{ under } H_0.$$

We use the test-statistic

$$T = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p(1-p)\left(\frac{1}{n_1^*} + \frac{1}{n_2^*}\right)}} \approx N(0, 1) \text{ under } H_0.$$

Rewriting the estimators gives

$$\hat{p}_1 - \hat{p}_2 = \hat{p}_1 - \frac{y_1}{n_2^*} = \hat{p}_1 - \frac{m_1 - x_1}{n_2^*} = \hat{p}_1 - \frac{n^* p - n_1^* \hat{p}_1}{n_2^*} = \frac{n^*(\hat{p}_1 - p)}{n_2^*}.$$

The test-statistic T can then be rewritten as

$$T = \frac{n^*(\hat{p}_1 - p)}{n_2^* \sqrt{p(1-p)\left(\frac{1}{n_1^*} + \frac{1}{n_2^*}\right)}}.$$

Table 8.2: Data for SNP at locus Y

	0	1	Total
Case	$x_0^{'}$	$x_1^{'}$	n_1^*
Control	$y_0^{'}$	$y_1^{'}$	n_2^*
Total	$m_0^{'}$	$m_1^{'}$	n^*

We now turn to another SNP, at locus Y, and we use similar notation as shown in Table 8.2. $n_1^* = 2n_1, n_2^*$ and $n^* = 2n$, where n is the total number of individuals and n_1 and n_2 is the total number of cases and controls, respectively. The estimators are

$$\hat{p}_1^{'} = \frac{x_1^{'}}{n_1^*}, \hat{p}_2^{'} = \frac{y_1^{'}}{n_2^*},$$

and

$$\hat{p}^{'} = \frac{x_1^{'} + y_1^{'}}{n^*} \approx p^{'} \text{ under } H_0.$$

The test observator is as for locus X

$$\begin{aligned} T' &= \frac{\hat{p}_1^{'} - \hat{p}_2^{'}}{\sqrt{p(1-p)\left(\frac{1}{n_1^*} + \frac{1}{n_2^*}\right)}} \\ &= \frac{n^*(\hat{p}_2 - p)}{n_1^{'} \sqrt{p(1-p)\left(\frac{1}{n_1^*} + \frac{1}{n_2^*}\right)}} \approx N(0,1) \text{ under } H_0. \end{aligned}$$

Since the test-statistics T and T' are linear combinations of $\hat{p}_1$ and $\hat{p}_2$, respectively, we have

$$\text{Corr}(T, T') = \text{Corr}(\hat{p}_1, \hat{p}_1^{'}). \tag{8.6}$$

Under H_0, we have

$$\text{Corr}(\hat{p}_1, \hat{p}_1^{'}) = \text{Corr}(\hat{p}, \hat{p}^{'}). \tag{8.7}$$

From Equation (8.4), we have $\hat{p} = \frac{\sum z_i}{n^*}$, which gives

$$\operatorname{Corr}(\hat{p}, \hat{p}^{'}) = \operatorname{Corr}\left(\sum z_i, \sum z_i^{'}\right). \tag{8.8}$$

We also have

$$\operatorname{Cov}\left(\sum_{i=1}^{n^*} z_i, \sum_{i=1}^{n^*} z_i^{'}\right) = n^* \operatorname{Cov}(z_i, z_i^{'})$$

and the variance

$$\operatorname{Var}\left(\sum_{i=1}^{n^*} z_i\right) = \sum_{i=1}^{n^*} \operatorname{Var}(z_i) = n^* \operatorname{Var}(z_i)$$

and

$$\operatorname{Var}\left(\sum_{i=1}^{n^*} z_i^{\prime}\right) = \sum_{i=1}^{n^*} \operatorname{Var}(z_i^{\prime}) = n^* \operatorname{Var}(z_i^{'})$$

Then,

$$\begin{aligned}
\operatorname{Corr}\left(\sum_{i=1}^{n^*} z_i, \sum_{i=1}^{n^*} z_i^{'}\right) &= \frac{n^* \operatorname{Cov}(z_i, z_i^{'})}{\sqrt{n^* n^* \operatorname{Var}(z_i) \operatorname{Var}(z_i^{'})}} \\
&= \frac{\operatorname{Cov}(z_i, z_i^{'})}{\sqrt{\operatorname{Var}(z_i) \operatorname{Var}(z_i^{'})}} \\
&= \operatorname{Corr}(z_i, z_i^{'}). \qquad (8.9)
\end{aligned}$$

We are interested in the correlation between the two test statistics at two loci, X and Y. From Equation (8.6), (8.7), (8.8) and (8.9), we observe that

$$\operatorname{Corr}(T, T') = \operatorname{Corr}(\hat{p}_1, \hat{p}_1^{'}) = \operatorname{Corr}(\hat{p}, \hat{p}^{'}) = \operatorname{Corr}\left(\sum_{i=1}^{n^*} z_i, \sum_{i=1}^{n^*} z_i^{'}\right) = \operatorname{Corr}(z_i, z_i^{'}) = \rho.$$

Under H_0 both T and T' approximately follows a $N(0,1)$ distribution, i.e. the expected values and variances are given by

$$\begin{aligned}
E(T) &= 0 \\
E(T') &= 0 \\
\operatorname{Var}(T) &= 1 \\
\operatorname{Var}(T') &= 1.
\end{aligned}$$

This gives

$$(T, T') \sim \text{binormal}(0, 1, 0, 1, \rho).$$

We accept H_0 for loci X when $T \in [-\sigma, \sigma], \sigma = z_{\alpha/2}$ and H_0 for loci Y when $T' \in [-\sigma, \sigma], \sigma = z_{\alpha/2}$. H_0 for both loci X and Y are accepted when

$$(T, T') \in [-\sigma, \sigma] \times [-\sigma, \sigma], \text{ where } \sigma = z_{\alpha/2}.$$

We assume that each of the individual tests has significance level α_p. We assume the null hypothesis, H_0, that no marker is associated with the disease status (Moskvina & Schmidt 2008). We let $O_j, j = 1, ..., m$ denote the event that the allelic test for the jth marker does not give a significant result at level α_p. The probability of event O_j is $P(O_j) = 1 - \alpha_p$.

The total type I error probability when testing m hypotheses simultaneously is given by

$$\begin{aligned}
\alpha &= 1 - P(O_1 \cap \cdots \cap O_m) \\
&= 1 - P(O_1)P(O_2|O_1)P(O_3|O_1 \cap O_2) \cdots P(O_m|O_1 \cap \cdots O_{m-1}) \\
&\leq 1 - P(O_1)P(O_2|O_1)P(O_3|O_2) \cdots P(O_m|O_{m-1}) \\
&= 1 - \frac{P(O_1 \cap O_2)P(O_2 \cap O_3) \cdots P(O_{m_1} \cap O_m)}{P(O_2) \cdots P(O_{m-1})}. \qquad (8.10)
\end{aligned}$$

From Moskvina & Schmidt (2008) we have

$$P(O_j|O_1 \cap ... \cap O_{j-1}) \geq P(O_j|O_k)$$

for any $k < j$, which explains the inequality in Equation (8.10).

Moskvina & Schmidt (2008) use the maximal correlation between a SNP and the previous markers, $r_j = \max_{1 \leq k \leq j-1} |r_{kj}|$. From the inequality in Equation (8.10) we observe that using maximal correlation, we choose the maximal $P(O_j|O_k), k = 1, ..., j-1$ which is closest to the value of $P(O_j|O_1 \cap \ldots \cap O_{j-1})$.

We will now work further with $P(O_X \cap O_Y)$ to be inserted into the numerator of Equation (8.10). The acceptance probability for locus X and locus Y is given by

$$P(O_X \cap O_Y) = P(\text{accept } H_0 \text{ for locus X} \cap \text{ accept } H_0 \text{ for locus Y}).$$

This means that both T and T' need to be inside the acceptance region $[-\sigma, \sigma]$. This results in a bivariate two dimensional integral for both T and T' for the SNPs. We insert the estimate of ρ_j with the absolute value $|r_j|$.

$$P(O_X \cap O_Y) = \frac{1}{2\pi\sqrt{1-r_j^2}} \int_{-\sigma}^{\sigma}\int_{-\sigma}^{\sigma} \exp\left(-\frac{1}{2(1-r_j^2)}(x^2 - 2r_jxy + y^2)\right) dydx$$

It can be shown that this integral equals

$$= \frac{1}{2\pi\sqrt{1-r_j^2}} \int_{-\sigma}^{\sigma} e^{-x^2/2} \left(\Phi\left(\frac{r_jx+\sigma}{\sqrt{1-r_j^2}}\right) - \Phi\left(\frac{r_jx-\sigma}{\sqrt{1-r_j^2}}\right)\right) dx.$$

Further, it can be shown that

$$\begin{aligned} \frac{1}{2\pi\sqrt{1-r_j^2}} \int_{-\sigma}^{\sigma} e^{-x^2/2} \left(\Phi\left(\frac{r_jx+\sigma}{\sqrt{1-r_j^2}}\right) - \Phi\left(\frac{r_jx-\sigma}{\sqrt{1-r_j^2}}\right)\right) dx \\ = 1 - \alpha_p - \sqrt{\frac{2}{\pi}} \int_{-\sigma}^{\sigma} e^{-x^2/2} \left(\Phi\left(\frac{r_jx-\sigma}{\sqrt{1-r_j^2}}\right)\right) dx, \end{aligned}$$

which gives

$$P(O_X \cap O_Y) = 1 - \alpha_p - \sqrt{\frac{2}{\pi}} \int_{-\sigma}^{\sigma} e^{-x^2/2} \left(\Phi\left(\frac{r_jx-\sigma}{\sqrt{1-r_j^2}}\right)\right) dx.$$

From this, we have from Equation (8.10),

$$\begin{aligned} \alpha &\le 1 - \frac{P(O_1 \cap O_2)P(O_2 \cap O_3)\cdots P(O_{m_1} \cap O_m)}{P(O_2)\cdots P(O_{m-1})} \\ &= 1 - \frac{\prod_{j=2}^{m}\left(1 - \alpha_p - \sqrt{\frac{2}{\pi}}\int_{-\sigma}^{\sigma} e^{-x^2/2}\Phi\left(\frac{r_jx-\sigma}{\sqrt{1-r_j^2}}\right)dx\right)}{(1-\alpha_p)^{m-2}} \\ &= 1 - (1-\alpha_p)\prod_{j=2}^{m}\left(1 - \alpha_p - \sqrt{\frac{2}{\pi}}\int_{-\sigma}^{\sigma} e^{-x^2/2}\Phi\left(\frac{r_jx-\sigma}{\sqrt{1-r_j^2}}\right)dx\right). \end{aligned}$$

This shows that the method by Moskvina & Schmidt (2008) controls the FWER,

$$\alpha \le 1 - (1-\alpha)\prod_{j=2}^{m}\left(1 - \alpha - \sqrt{\frac{2}{\pi}}\int_{-\sigma}^{\sigma} e^{-x^2/2}\Phi\left(\frac{r_jx-\sigma}{\sqrt{1-r_j^2}}\right)dx\right).$$

8.5 The Beta-distribution method

Dudbridge & Gusnanto (2008) described a method for estimating the effective number of tests based on the beta distribution. If it exists an effective number of tests, M_{eff}, the minimum p-value should follow a beta distribution

$$\text{Beta}(1, M_{\text{eff}})$$

as described in Section 7.2. The general Beta(a, b) distribution is given by (Casella & Berger 2002, p. 623)

$$f(x|a,b) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}, 0 \leq x \leq 1, a > 0, b > 0,$$

where the constant $B(a, b)$ is defined in terms of gamma functions and is given by

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

For the general Beta(a, b) distribution, the expected value and variance are given by (Casella & Berger 2002, p. 623)

$$E(X) = \frac{a}{a+b} \text{ and } \text{Var}(X) = \frac{ab}{(a+b)^2(a+b+1)}.$$

Assume independent data $x_i, i = 1..., B$, of B observations of the minimum p-values, and let $\bar{x}$ and s^2 denote the sample mean and variance, respectively.

Moment estimators

The moment estimators for the Beta(a, b) distribution are found by solving the equations

$$\bar{x} = \frac{a}{a+b} \text{ and } s^2 = \frac{ab}{(a+b)^2(a+b+1)}$$

where $\bar{x}$ and s^2 are the sample mean and variance of the observations, respectively. The moment estimators are then given by

$$\hat{a} = \bar{x}\left(\frac{\bar{x}(1-\bar{x})}{s^2} - 1\right)$$

and

$$\hat{b} = (1-\bar{x})\left(\frac{\bar{x}(1-\bar{x})}{s^2} - 1\right),$$

where $\bar{x}$ and s^2 are the sample mean and variance of the observations, respectively. Dudbridge & Koeleman (2004) tested the null-hypothesis whether $a = 1$, and under this null hypothesis, the method of moments estimate of b is

$$\hat{b} = \frac{1-\bar{x}}{\bar{x}}.$$

Maximum likelihood estimator

The maximum likelihood estimator is asymptotically efficient (Casella & Berger 2002, p. 472), and therefore we want to use the maximum likelihood estimator to estimate the parameter b of the Beta$(1, b)$ distribution. The likelihood-function for the Beta$(1, b)$ distribution is given by

$$L(b|x) = \prod_{i=1}^{B} b(1-x_i)^{b-1}.$$

This gives the maximum likelihood estimator for b

$$\hat{b} = -\frac{B}{\sum_{i=1}^{B} \ln(1-x_i)},$$

where B is the number of observations in the data set used.

This method can only be used if a set of B minimum p-values are available. The role of the method is a qualitatively assessment of the distribution of the minimum p-values.

8.6 Other methods for estimating M_{eff}

The method of Chen and Liu (2011)

The method of Chen & Liu (2011) consists of three steps

1. For each SNP $i, i = 1, ..., m$, we estimate the absolute CLD coefficient between this SNP and any of the other SNPs $|r_{ij}|, j \neq i$.

2. Calculate $R_i = \sum_{j=1}^{m} |r_{ij}|^k, i = 1, 2, ..., m$, where the positive constant k is a statistical test-dependent parameter.

3. Estimate the effective number of independent tests by

$$M_{\text{eff}} = \sum_{i=1}^{m} \frac{1}{R_i}.$$

As described by Chen & Liu (2011), the statistical test-dependent parameter is equal to $k = 7$ when the statistical test is the Cochran-Armitage test for trend as described in Section 6.1. When the statistical test used is Pearson's χ^2-test with 2 degrees of freedom, then the parameter is equal to $k = 3$. No explanation of how k is found is given by Chen & Liu (2011).

The method of Li and Ji (2005)

The method of Li & Ji (2005) is as the method of Cheverud (2001) and Nyholt (2004) based on the eigenvalues of the correlation matrix. Li & Ji (2005) considered a total number of m tests, where the m tests contains $c, 1 \leq c \leq m$ copies of m/c independent tests. The eigenvalues of the correlation matrix are

$$\begin{aligned} \lambda_i &= c, \quad i = 1, ..., m/c \\ \lambda_i &= 0, \quad i = (m/c + 1, ..., m). \end{aligned}$$

From the method of Nyholt (2004) we then get

$$M_{\text{eff}} = 1 + (m-1)\left(1 - \frac{\text{Var}(\lambda_{\text{obs}})}{m}\right),$$

which gives

$$M_{\text{eff}} = m + 1 - c.$$

As described above we have a total of m/c independent tests, and then we observe that

$$r = \frac{m+1-c}{m/c} = \frac{c(m+1-c)}{m} \geq 1, \quad 1 \leq c \leq m. \tag{8.11}$$

From Equation (8.11) we observe that for $1 \leq c \leq m$, the method of Cheverud (2001) and Nyholt (2004) will overestimate the effective number of independent tests and give conservative results.

The method of Li & Ji (2005) is based on decomposition of the eigenvalues into an integral part and an nonintegral part. The integral part of the eigenvalue represents identical tests, and the nonintegral part represents partially correlated tests. Li & Ji (2005) described an estimate for the effective number of independent tests, M_{eff}, as

$$M_{\text{eff}} = \sum_{i=1}^{m} f(|\lambda_i|),$$

where

$$f(x) = I(x \geq 1) + (x - \lfloor x \rfloor), x \geq 0. \tag{8.12}$$

In Equation (8.12), $I(x \geq 1)$ is the indicator function and $\lfloor x \rfloor$ is the floor function. In the method by Li & Ji (2005) perfectly correlated tests will be counted as $I(x \geq 1)$, and partially correlated tests will be counted as $(x - \lfloor x \rfloor)$.

The method of Galwey (2009)

The method of Galwey (2009) is an improvement of the method of Li & Ji (2005). Compared to the method of Li & Ji (2005), the method of Galwey (2009) will give more weight to the fractional part of the eigenvalues, than to the integer part. We let $\lambda_i, i = 1, ..., m$ denote the eigenvalues of the correlation matrix. The method of Galwey (2009) for estimating the effective number of independent tests can be set up as

$$M_{\text{eff}} = \frac{\left(\sum_{i=1}^{m} \sqrt{\lambda_i}\right)^2}{\sum_{i=1}^{m} \lambda_i}.$$

As for the method of Gao et al. (2008) described in Section 8.3, we use $\sum_{i=1}^{m} \lambda_i = m$, and rewrite the method of Galwey (2009) as

$$M_{\text{eff}} = \frac{\left(\sum_{i=1}^{m} \sqrt{\lambda_i}\right)^2}{m}.$$

In general, when we have complete observations of our data, the correlation matrix will be positive semidefinite, and hence all eigenvalues will be positive. For the data analyzed in this thesis, we do not have complete observations, we have different number of missing data for different SNPs. Therefore, the correlation matrix is calculated based on pairwise complete observations, which means that the matrix will not be positive semidefinite, and some of the eigenvalues may then be zero. The problem with negative eigenvalues is avoided in the method of Galwey (2009) by assuming that all negative eigenvalues are small in absolute value, and therefore are set equal to zero.

8.7 Comparing the different methods

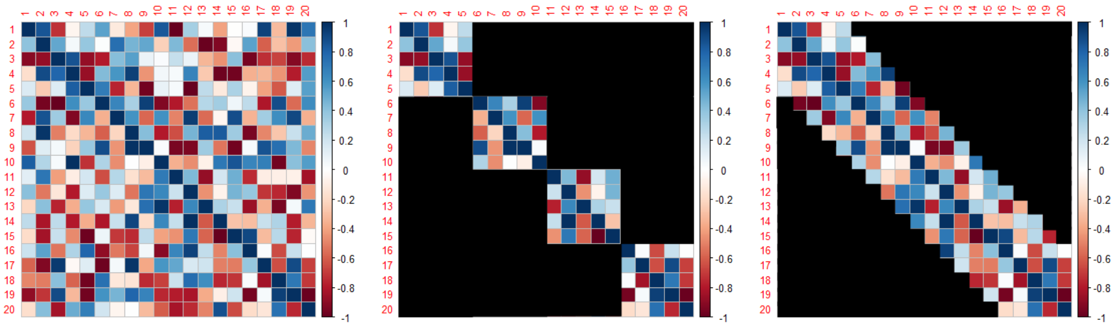
The methods of Nyholt (2004), Gao et al. (2008) and Moskvina & Schmidt (2008) are described in Section 8.1, 8.3 and 8.4, respectively. The method of Nyholt (2004) use the whole correlation matrix, the method of Gao et al. (2008) use blocks of predetermined size and the method of Moskvina & Schmidt (2008) use a sliding window around each SNP as illustrated in Figure 8.1a, 8.1b and 8.1c, respectively. Chen & Liu (2011) gave three desired properties for a method to calculate M_{eff}. These properties are

1. When all tests are completely independent, then all the m eigenvalues are equal to 1, and the variance of the eigenvalues is then $\text{Var}(\lambda) = 0$. This gives $M_{\text{eff}} = m$
2. When all tests are completely correlated, one eigenvalue is equal to m, the other are equal to 0. In this situation, $\text{Var}(\lambda) = m$. The effective number of independent tests is then $M_{\text{eff}} = 1$.
3. When the m tests is composed of $c, 1 \leq c \leq m$ copies of m/c independent tests, the effective number of independent tests is m/c.

We have observed that the methods of Nyholt (2004) and Moskvina & Schmidt (2008) are described for the two different cases when the SNPs are perfectly correlated and when the SNPs are completely independent, which means that these two methods satisfies the first and second property described above. It has been shown by Salyakina et al. (2005) and Li & Ji (2005) that the method by Nyholt (2004) gives conservative results for the effective number of independent tests when the SNPs are partially correlated, which means that the methods by Cheverud (2001) and Nyholt (2004) do not satisfy the third property.

Chen & Liu (2011) observed that the method by Gao et al. (2008) does not satisfy the second property, they observed that when all tests are completely independent, the method by Gao et al. (2008) will always underestimate the effective number of independent tests for all predetermined cutoff's $c < 1$. In Section 8.3 we have observed that the method of Gao et al. (2008) depends on the block size used and will give conservative results for partially correlated SNPs, which means that the method of Gao et al. (2008) also does not satisfy the third property as described above.

The method of Chen & Liu (2011) and Li & Ji (2005) satisfies all the three properties described above.



(a)

(b)

(c)

Figure 8.1: (a) The method of Nyholt (2004). (b) The method of Gao et al. (2008). (c) The method of Moskvina and Schmidt (2008). Figure illustrating the difference between the methods of Nyholt (2004), Gao et al. (2008) and Moskvina and Schmidt (2008). The method of Nyholt use the eigenvalues of the whole correlation matrix, the method of Gao uses blocks of fixed size, and the method of Moskvina uses a window around each SNP marker.

We observed that the methods of Nyholt (2004) and Gao et al. (2008) estimates the effective number of independent tests, M_{eff}, and then uses the method of Šidák to find the individual significance level. The method of Moskvina & Schmidt (2008) first estimates the individual significance level, α_p, and then uses the method of Šidák to find the estimate the effective number of independent tests.

An important contribution of the method of Moskvina & Schmidt (2008) is that the correlation between the test-observators are equal to the estimated Pearson correlation between the SNPs. Han, Kang & Eskin (2009) showed that this relationship between the test-observators and the correlation between SNPs in general not will be as for the method of Moskvina & Schmidt (2008). We have observed that the method of Chen & Liu (2011) as the method of Moskvina & Schmidt (2008) depends on the statistical test used, but Chen & Liu (2011) does not describe how the statistical test-dependent parameter is found. Based on the observations described in this chapter, we have observed that the M_{eff} estimate using the method of Moskvina & Schmidt (2008) and Chen & Liu (2011) depends on the statistical test used, and the M_{eff} estimate using the other methods does not depend on the statistical test used. One interesting question is whether the M_{eff} estimate should be dependent or independent of the statistical test used.

Chapter 9

TOP8 - Data analysis

In this chapter the multiple testing correction methods presented in Chapter 8 will applied to chromosome 22 of the TOP data. The different methods are implemented using estimates of either LD or CLD correlation,

$$\hat{\rho}_{LD} = \frac{\hat{P}_{AB} - \hat{p}_A\hat{p}_B}{\sqrt{\hat{p}_A\hat{p}_a\hat{p}_B\hat{p}_b}},$$

and

$$\hat{\rho}_{CLD} = \frac{\hat{\Delta}_{AB}}{\sqrt{(\hat{p}_A(1-\hat{p}_A)+\hat{D}_A)(\hat{p}_B(1-\hat{p}_B)+\hat{D}_B)}},$$

as described in Chapter 5.

9.1 TOP8 - chromosome 22

The data from the TOP study for chromosome 22 contains information about 1551 individuals and 8928 SNPs. Chromosome 22 was the smallest chromosome of all 22 chromosomes in the TOP data. We have different numbers of missing data for each SNP and therefore we used pairwise complete observations to calculate the CLD correlation matrix. When the correlation matrix is calculated based on only pairwise complete observations, the correlation matrix may not be positive semidefinite, which may give negative eigenvalues.

According to the observations in Section 8.3, the maximum number of nonzero eigenvalues of the correlation matrix are equal to $1551 - 1 = 1550$. From Figure 9.1 we see that the first 1550 eigenvalues of the CLD correlation matrix are positive and decreasing. The rest of the in total 8928 eigenvalues of the CLD correlation

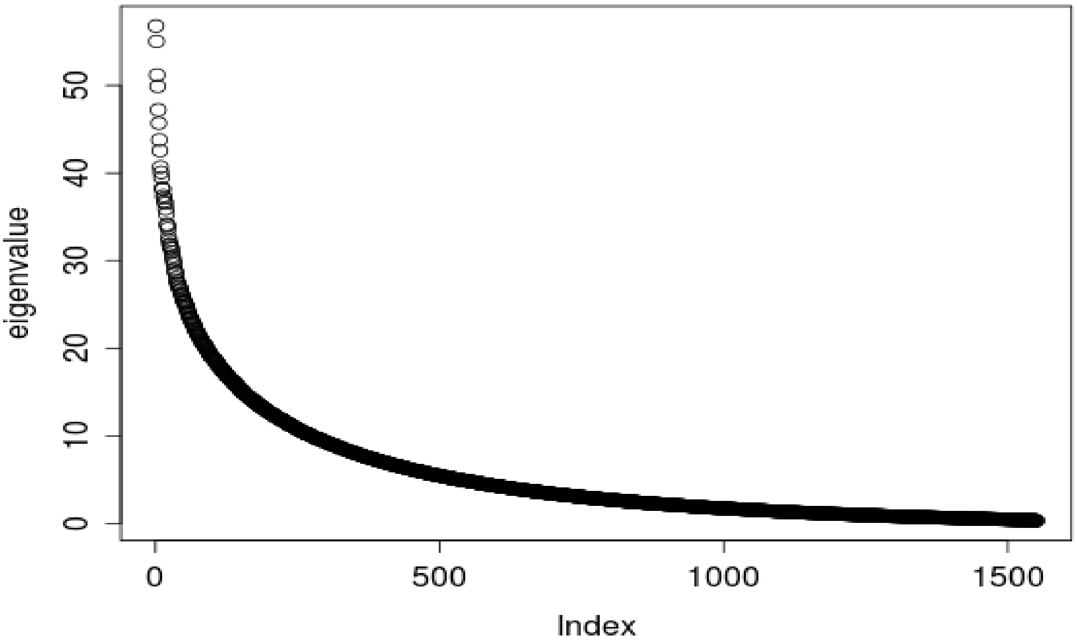


Figure 9.1: Plot of the first 1550 eigenvalues of the ρ_{CLD} correlation matrix. We see that all the 1550 first eigenvalues are positive and decreasing.

matrix are small with both positive and negative signs as expected.

In Galwey (2009) the problem with negative eigenvalues was avoided by assuming the negative eigenvalues to be small in absolute value, and therefore set equal to zero. The methods described in Chapter 8 are in our analysis implemented by setting the negative eigenvalues equal to zero.

9.2 Estimates of the effective number of independent tests, M_{eff}

We used the methods by Nyholt (2004), Gao et al. (2008) and Moskvina & Schmidt (2008) to compare the estimate of the effective number of independent tests. The method of Nyholt (2004) was impemented as originally described, and according to the observations in Section 8.3 we implemented the method of Gao et al. (2008) using blocks of fixed size. The method of Moskvina & Schmidt (2008) was implemented both exact and by the approximation formula for individual significance level less than or equal to $\alpha_p = 0.01$. R-code for the different methods are given in Appendix D.

The Cheverud-Nyholt method

The method by Nyholt (2004) was as shown in Section 8.1 implemented using the variance of the eigenvalues of the pairwise LD correlation matrix.

As described in Section 8.3, the maximal number of nonzero eigenvalues for a $n \times p$, $n < p$ matrix is equal to $n - 1$. For the data analyzed in this thesis, $n = 1551$. Using only the first $n - 1 = 1550$ eigenvalues, the effective number of tests is

$$M_{\text{eff}} = 8872.378.$$

The significance level threshold for the individual tests are found using the method of Šidák as described in Section 7.4.

The estimate $M_{\text{eff}} = 8872.378$ gives the individual significance level

$$\alpha_p = 1 - (1 - 0.05)^{(1/8872.378)} = 5.78 \cdot 10^{-6}.$$

Using the CLD matrix and the first $n - 1 = 1550$ eigenvalues we get

$$M_{\text{eff}} = 8872.384,$$

and the individual significance level

$$\alpha_p = 1 - (1 - 0.05)^{(1/8872.384)} = 5.78 \cdot 10^{-6},$$

which we observe is the same result as obtained when using the LD correlation matrix.

Moskvina's alternative formulation of Nyholt's method

Moskvina & Schmidt (2008) showed that the estimate of effective number of independent tests defined by Nyholt (2004) can be calculated directly from the correlation coefficients of the pairwise correlation matrix.

We implemented Moskvina's alternative formulation of Nyholt's method in R, and for chromosome 22 in the TOP8 data, the estimate of the effective number of independent tests using the LD correlation matrix was

$$M_{\text{eff}} = 8913.441,$$

with the corresponding individual significance level, $\alpha_p = 1 - (1 - 0.05)^{1/8913.441} = 5.75 \cdot 10^{-6}$.

Using the CLD correlation matrix, we get the estimate

$$M_{\text{eff}} = 8913.508,$$

with the corresponding individual significance level, $\alpha_p = 1 - (1 - 0.05)^{1/8913.441} = 5.75 \cdot 10^{-6}$.

We observe that there is a difference in the results using the method of Nyholt (2004) with the 1550 first eigenvalues and Moskvina's alternative formulation of Nyholt's method. Because we have different proportion of missing data for each SNP, the correlation matrix based on pairwise complete observations may not be positive semidefinite, giving some negative eigenvalues. The method of Nyholt (2004) use the eigenvalues of the correlation matrix while Moskvina's alternative formulation of the method uses the whole correlation matrix, which may explain the difference in the results between the method of Nyholt (2004) and Moskvina's alternative formulation of Nyholt's method.

Moskvina's alternative formulation is a less computationally intensive method than the original method by Nyholt (2004), but this formulation requires the whole correlation matrix and may hence not be a preferable method for estimating M_{eff}. The method of Nyholt (2004) use only the eigenvalues of the correlation matrix.

The method of Gao

The method of Gao et al. (2008) was implemented using $c = 99.5\%$ as the cutoff value, indicating that the result should explain 99.5% of the variation in the data.

Implementing the method by Gao et al. (2008) without using blocks gives

$$M_{\text{eff}} = 1351$$

as the estimate for the effective number of independent tests.

Since the method of Gao et al. (2008) is highly dependent on the block size, as discussed in Section 8.3, we implemented the method by Gao et al. (2008) for different block sizes up to block size equal to the number of individuals minus one, $n - 1 = 1550$.

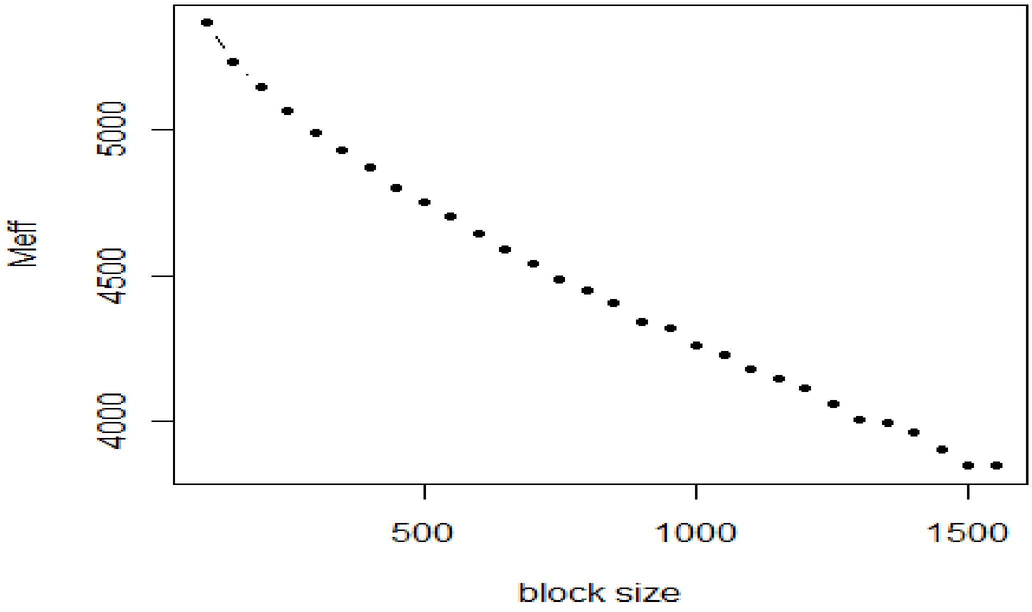


Figure 9.2: Plot of the estimate M_{eff} for the method of Gao (2008) for different block sizes, b up to $b = 1550$. We observe that the effective number of tests decreases when the block size are increasing and that the method of Gao (2008) gives estimates of M_{eff} in a relatively large interval.

From Figure 9.2 we see the estimate of the effective number of independent tests using the method described by Gao et al. (2008) using blocks of fixed size. We observe that using different block sizes in the method by Gao et al. (2008) will give results in a relatively large interval for the estimate of M_{eff}.

Table 9.1: The Gao estimate for M_{eff} for different block sizes

block size	M_{eff}, CLD
100	5361
150	5230
200	5145
250	5063
300	4987
350	4926
400	4871
450	4796
500	4749
550	4700
600	4641
650	4590
700	4540
750	4487
800	4447
850	4406
900	4339
950	4321
1000	4259
1050	4231
1100	4178
1150	4150
1200	4118
1250	4059
1300	4008
1350	3997
1400	3963
1450	3903
1500	3850
1550	3854

Using the Šidák method as described in Section 7.4 we find the significance level α_p for the individual tests for different block sizes. From Table 9.1 the estimated effective number of tests for block size $b = 100$ is $M_{\text{eff}} = 100$, and using $\alpha = 0.05$ the individual significance level threshold is

$$\begin{aligned}\alpha_p &= 1 - (1-\alpha)^{1/M_{\text{eff}}}\\ &= 1 - (1-0.05)^{1/5361}\\ &= 9.57 \cdot 10^{-6}.\end{aligned}$$

For block size equal to $b = n - 1 = 1550$ the estimated effective number of tests using the method of Gao et al. (2008) is

$$\begin{aligned}\alpha_p &= 1 - (1-\alpha)^{1/M_{\text{eff}}}\\ &= 1 - (1-0.05)^{1/3854}\\ &= 1.33 \cdot 10^{-5}.\end{aligned}$$

We observe that the individual significance level, α_p, varies from $9.57 \cdot 10^{-6}$ to $1.33 \cdot 10^{-5}$ using different block sizes in the method of Gao et al. (2008).

The method of Moskvina and Schmidt

From Section 8.4 we have observed that the method of Moskvina & Schmidt (2008) gives the same results as the methods of Nyholt (2004) and Gao et al. (2008) for the extreme cases when we have complete correlation or complete independence. When all markers are completely independent, all $\kappa_j = 1$ and $M_{\text{eff}} = m$ and when all markers are perfectly correlated, then all $\kappa_j = 0$ and $M_{\text{eff}} = 1$.

We implemented the method by Moskvina & Schmidt (2008) as shown in Appendix D, both exact and using window size w. The significance level α_p for each of the individual tests was determined by estimating M_{eff} for different values of α_p, where the individual significance level, α_p, is reduced from the starting value until the estimate of the overall type I error probability, $\alpha \leq 1 - (1-\alpha_p)^{M_{\text{eff}}}$, passes the predetermined level, here $\alpha = 0.05$.

The results in Table 9.2 using the method by Moskvina & Schmidt (2008) with the approximation formula and the LD correlation matrix gives estimates for the effective number of independent tests in the interval $6249.893 - 6269.99$, and the individual significance level in the interval $8.18 - 8.20 \cdot 10^{-6}$. Using the CLD matrix we get the estimate of the effective number of independent tests in the

Table 9.2: Results for the method of Moskvina using the LD and CLD correlation matrix

w	M_{eff}, LD	α_p, LD	M_{eff}, CLD	α_p, CLD
20	6269.99	$8.18 \cdot 10^{-6}$	6252.445	$8.20 \cdot 10^{-6}$
25	6259.273	$8.19 \cdot 10^{-6}$	6241.887	$8.21 \cdot 10^{-6}$
30	6255.688	$8.19 \cdot 10^{-6}$	6238.231	$8.22 \cdot 10^{-6}$
35	6254.354	$8.20 \cdot 10^{-6}$	6237.008	$8.22 \cdot 10^{-6}$
40	6251.881	$8.20 \cdot 10^{-6}$	6234.495	$8.22 \cdot 10^{-6}$
45	6251.438	$8.20 \cdot 10^{-6}$	6234.058	$8.22 \cdot 10^{-6}$
50	6250.96	$8.20 \cdot 10^{-6}$	6233.581	$8.22 \cdot 10^{-6}$
55	6250.279	$8.20 \cdot 10^{-6}$	6232.900	$8.22 \cdot 10^{-6}$
60	6249.94	$8.20 \cdot 10^{-6}$	6232.571	$8.22 \cdot 10^{-6}$
65	6249.905	$8.20 \cdot 10^{-6}$	6232.427	$8.23 \cdot 10^{-6}$
70	6249.895	$8.20 \cdot 10^{-6}$	6232.419	$8.23 \cdot 10^{-6}$
75	6249.893	$8.20 \cdot 10^{-6}$	6232.415	$8.23 \cdot 10^{-6}$

interval $6232.415 - 6252.445$ and the individual significance level in the interval $8.20 - 8.23 \cdot 10^{-6}$.

From Table 9.2 we observe that the method by Moskvina & Schmidt (2008) will give approximately equal results for the effective number of tests when the window size is larger than $w = 35$, both using the LD and the CLD matrix. We also observe that the individual significance level, α_p, is approximately equal for the different window sizes and both LD and CLD correlation matrix.

Using the LD correlation matrix and the exact formula we observe that the FWER is controlled at level $\alpha \leq 0.05$ when the individual significance level is equal to

$$\alpha_p = 8.19 \cdot 10^{-6},$$

and the estimate of the effective number of tests is given by

$$\begin{aligned} M_{\text{eff}} &= \frac{\log(1-\alpha)}{\log(1-\alpha_p)} \\ &= \frac{\log(1-0.05)}{\log(1-8.19 \cdot 10^{-6})} \\ &= 6262.892. \end{aligned}$$

Using the CLD correlation matrix and the exact formula we observe that the FWER is controlled at level $\alpha \leq 0.05$ when the individual significance level is equal to

$$\alpha_p = 8.21 \cdot 10^{-6},$$

and the estimate of the effective number of tests is given by

$$\begin{aligned} M_{\text{eff}} &= \frac{\log(1-\alpha)}{\log(1-\alpha_p)} \\ &= \frac{\log(1-0.05)}{\log(1-8.21 \cdot 10^{-6})} \\ &= 6247.635. \end{aligned}$$

Comparing the results from the method by Moskvina & Schmidt (2008) using both the exact method and the approximation formula shows that when using window size in the interval $w = 20 - 25$, the approximation gives result for M_{eff} close to the result for the exact formula, both using the LD and the CLD correlation.

The method of Chen and Liu

The method of Chen & Liu (2011) is implemented in R as shown in Appendix D. Since we do not know which parameter k to use with the C p-value, we used different values of the parameter k. With $k = 3$ and using the CLD correlation matrix we get the estimate

$$M_{\text{eff}} = 4161.581.$$

The individual significance level are found using the method of Sidak as described in Section 7.4 and total type I error rate, $\alpha = 0.05$,

$$\begin{aligned} \alpha_p &= 1 - (1-0.05)^{(1/4161.581)} \\ &= 1.232536 \cdot 10^{-5}. \end{aligned}$$

The method of Li and Ji

The method by Li & Ji (2005) is based on the eigenvalues of the correlation matrix between the SNPs, and decomposes the eigenvalues into an integral part and an nonintegral part as shown in Section 8.6. The negative eigenvalues of the correlation matrix are set to zero, since the negative eigenvalues are assumed to be small in absolute value.

Implementing the method of Li & Ji (2005) as described in Section 8.6 using the LD correlation matrix gives the M_{eff} estimate

$$M_{\text{eff}} = 2653.512,$$

with the corresponding individual significance level

$$\alpha_p = 1 - (1 - 0.05)^{(1/2653.512)} = 1.93 \cdot 10^{-5}.$$

The method of Galwey

We implemented the method of Galwey (2009) as described in Section 8.6. Using the LD correlation matrix, the estimate of the effective number of tests is

$$M_{\text{eff}} = 1712.295,$$

with the corresponding individual significance level

$$\alpha_p = 1 - (1 - 0.05)^{(1/1712.295)} = 3.0 \cdot 10^{-5}.$$

9.3 Resampling

We implemented a resampling procedure based on the C p-value as described in Section 6.4 for estimating the significance level threshold, α_p for the individual tests. The R code is shown in Appendix D. For each resampled data set we permuted the disease status vector, by randomly drawing, without replacement, n_1 cases and n_2 controls from a total of n individuals, and the minimum C p-value was recorded. The significance level threshold α_p for the individual tests are found by determining the 5% quantile of the empirical distribution of the minimum C p-values.

From the results shown in Figure 9.4 we get the significance level threshold, α_p, for the individual tests from the $0.05 \cdot 100000$ order statistic equal to

$$\alpha_p = 9.71 \cdot 10^{-6}.$$

Using the Šidák method, the corresponding estimate of the effective number of independent tests, M_{eff}, is

$$\begin{aligned} M_{\text{eff}} &= \frac{\log(1-\alpha)}{\log(1-\alpha_p)} \\ &= \frac{\log(1-0.05)}{\log(1-9.71 \cdot 10^{-6})} \\ &= 5284.572. \end{aligned} \tag{9.1}$$

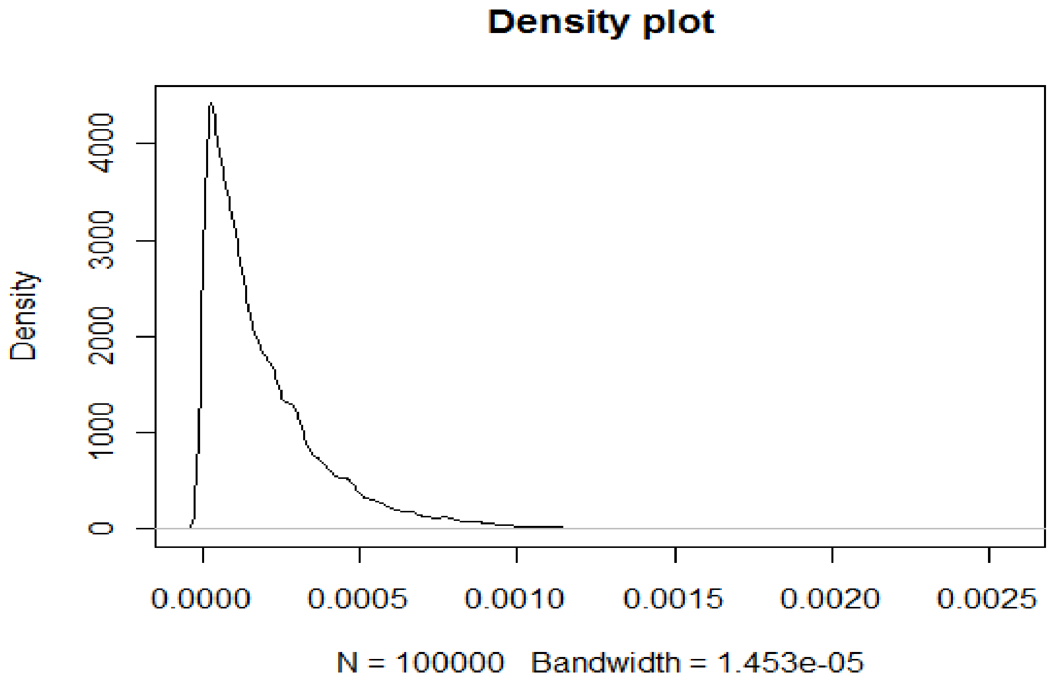


Figure 9.3: Density plot of the minP C p-values for 100000 permutations of the case-control data in the TOP study

Number of permutations in the resampling procedure

To investigate the number of permutations needed in a resampling procedure for obtaining suitable results, we used the empirical distribution of the 100000 minP C p-values obtained from our resampling procedure as described in Section 7.4. We sampled subsets of size 1000 and 10000 of the minP C p-values and investigated the 5% quantile of these subsets compared to the results of the other methods.

The observed resampled distribution shown in Figure 9.3 is a skewed distribution with high density for the smallest values, and therefore, using only 1000 permutations, we may get results from the resampling procedure that are more conservative than the Bonferroni procedure. This shows that more than 1000 permutations of the data may be necessary to obtain suitable results for α_p.

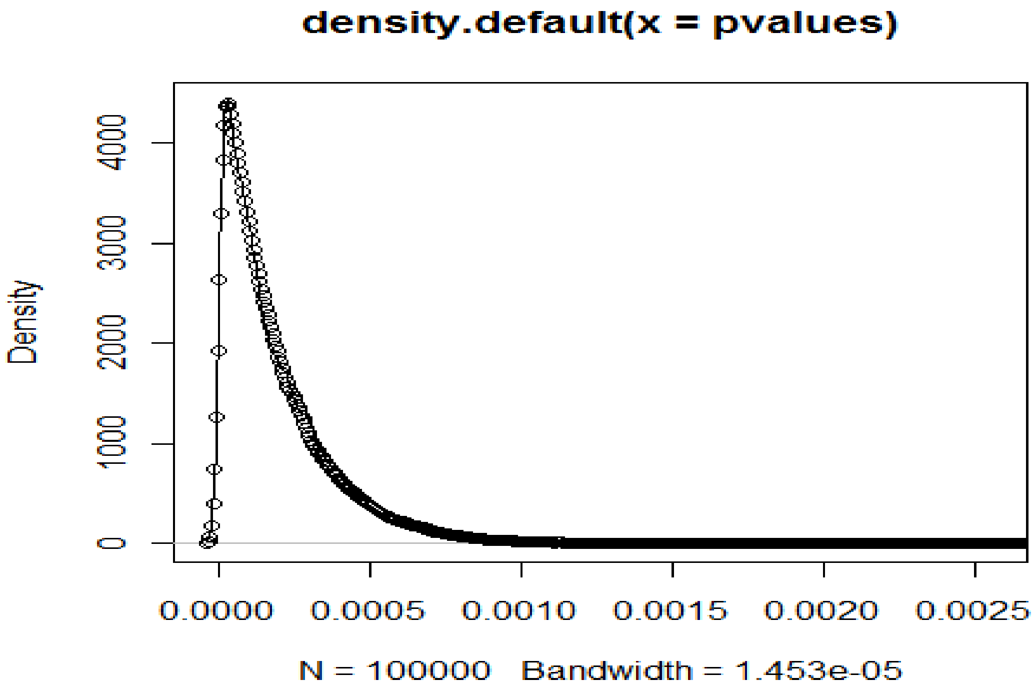


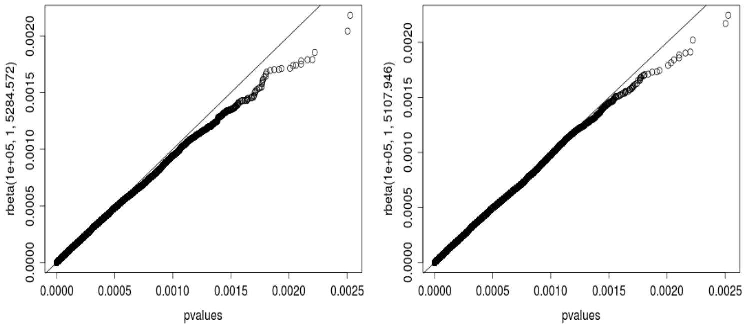
Figure 9.4: Beta(1,5284.572) distribution fitted to the minP values from the resampling. The line represents the theoretical Beta(1,5284.572) distribution and the open points represents the minP p-values from the resampling as described above.

Beta distribution and min p-values

We fitted a Beta$(1, M_{\text{eff}})$ distribution to the resampled minP p-values. From Equation (9.1) we get that the estimate of the effective number of independent tests using the resampling procedure equal to $M_{\text{eff}} = 5284.572$. Figure 9.4 shows density plot for our observed resampled distribution and the open points represents a theoretical Beta$(1, 5284.572)$ distribution.

The resampling procedure was implemented as shown in Appendix D. The C p-value was calculated using the MaXact package in R (Tian & Xu 2009). The sample mean, $\bar{x}$, of the minP p-values is $\bar{x} = 0.0001957351$, and from Section 8.5, we get the method of moments estimate of b when $a = 1$,

$$\hat{b} = \frac{1 - 0.0001957351}{0.0001957351} = 5107.946.$$



(a) (b)

Figure 9.5: (a) QQplot for the minP p-values from the resampling procedure, plotted against a Beta$(1, 5284.572)$ distribution based on 100000 observations. (b) QQplot for the results of the beta method, plotted against a Beta$(1, 5107.946)$ distribution distribution based on 100000 observations.

The individual significance level is found using the method of Šidák as described in Section 7.4,

$$\begin{aligned}\alpha_p &= 1 - (1-\alpha)^{(1/M_{\text{eff}})}\\ &= 1 - (1-0.05)^{(1/5107.946)}\\ &= 1.00 \cdot 10^{-5}.\end{aligned}$$

Maximum-likelihood estimation of Beta$(1, b)$ distribution gives

$$\hat{b} = -\frac{B}{\sum_{i=1}^{B} \ln(1-x_i)} = 5107.916,$$

where B is the number of permutations of the data.

From the QQ plots in Figure 9.5a and 9.5b we observe neither the method of moments estimator or the maximum likelihood estimator seems to give a reasonable fit to a beta distribution.

9.4 Summary of the results

In this chapter, the multiple correction methods described in Chapter 8 have been applied to real data from the TOP study and estimated results for the different

Table 9.3: Summary of the results

Method	M_{eff}	α_p
Bonferroni	8928	$5.60 \cdot 10^{-6}$
Nyholt	8913.441	$5.75 \cdot 10^{-6}$
Moksvina, $w = 25$	6259.273	$8.19 \cdot 10^{-6}$
Moskvina, $w = 45$	6251.438	$8.20 \cdot 10^{-6}$
Gao, $b = 100$	5361	$9.57 \cdot 10^{-6}$
Resampling	5284.572	$9.71 \cdot 10^{-6}$
Beta method	5107.946	$1.00 \cdot 10^{-5}$
Gao, $b = 500$	4749	$1.08 \cdot 10^{-5}$
Chen and Liu	4161.581	$1.23 \cdot 10^{-5}$
Gao, $b = 1550$	3854	$1.33 \cdot 10^{-5}$
Li and Ji	2653.512	$1.93 \cdot 10^{-5}$
Galwey	1712.295	$3.00 \cdot 10^{-5}$

methods are shown in Table 9.3. The most conservative multiple testing procedure is the Bonferroni method, described in Section 7.4.

The method of Nyholt (2004) and Moskvina & Schmidt (2008) are based on LD correlation between the SNPs, and the method og Gao et al. (2008) and Chen & Liu (2011) are based on CLD correlation between the SNPs. From Section 5.4 we have observed that for chromosome 22 the CLD correlation is more extreme than the LD correlation in approximately 52% of the cases, and the average difference $|CLD| - |LD|$ is equal to 0.1059.

Resampling based methods are considered as the gold standard in multiple testing problems and from Table 9.3 we observe that the resampling-based method as described in Section 7.4 using the C p-value gives $\alpha_p = 9.71 \cdot 10^{-6}$ as the individual significance threshold. According to the observations in Section 8.3 we have observed that the results of the method of Gao et al. (2008) is strongly dependent on the block size used. From Table 9.3 we observe that using different block sizes the method of Gao et al. (2008) gives estimated results that are both approximately equal to the results of the resampling based method and results that are anti-conservative compared to the resampling based method. These results indicates that the method of Gao et al. (2008) is not a preferable method for estimating M_{eff}.

The method of Moskvina & Schmidt (2008) use a sliding window around each SNP marker, and we expect this procedure to take more of the LD structure between the SNPs into account than using either the full correlation matrix or smaller blocks of fixed size as the method of Nyholt (2004) and Gao et al. (2008). For the TOP8 data, the total number of elements in the correlation matrix is

$$N = \frac{8928 \cdot 8927}{2} = 39850128.$$

Using the method of Moskvina & Schmidt (2008) with window size $w = 20$, we observe that the number of correlations needed are

$$\begin{aligned} N &= 0 + 1 + 2 + ... + 19 + 20 \cdot (8928 - 20) \\ &\frac{19 \cdot 20}{2} + 20 \cdot (8928 - 20) \\ &= 178350. \end{aligned}$$

These results shows that using the method of Moskvina & Schmidt (2008) with a small window size gives a less compuationally intensive problem than using a method which requires the whole correlation matrix. We have also shown that the method of Moskvina & Schmidt (2008) using a fixed window size gives results close to the result of the resampling procedure.

Comparing the theory and results for the different methods for estimating M_{eff}, we will prefer to use the method of Moskvina & Schmidt (2008) to estimate the effective number of independent tests because the method seems to be robust with respect to the use of window size w, and the results are close to the results using the resampling-based minP procedure.

The TOP8 data

Calculating the C p-value for each SNP on chromosome 22 analyzed in the TOP8 data we found the smallest p-value equal to $5.55 \cdot 10^{-5}$. Comparing this result to the results in Table 9.3 we observe that the smallest p-value are greater than the individual significance level threshold for all the multiple correction methods considered in Chapter 8, and hence no significant result for chromosome 22 of the TOP8 data are discovered in this thesis.

Chapter 10

Application to GWAS

In Chapter 8, we have presented different methods for estimating the effective number of independent tests for a single chromosome. Applications of the methods of Nyholt (2004), Gao et al. (2008) and Moskvina & Schmidt (2008) from one single chromosome to the whole genome will be presented in Section 10.1, and the method of Dudbridge & Gusnanto (2008) for estimating a genome-wide significance threshold based on a permutation test will be presented in Section 10.2. An alternative method for estimating the genome-wide significance level based on the effective ratio of the M_{eff} estimate against the total number of SNPs on each chromosome will be presented in Section 10.3. In the following sections, we will denote the effective number of independent tests for the whole genome and individual chromosomes by $M_{\text{eff,g}}$ and M_{eff}, respectively.

10.1 From M_{eff} per chromosome to $M_{\text{eff,g}}$ for the whole genome

Cheverud (2001) described two different alternatives for estimating the genome-wide significance threshold. When all chromosomes are in linkage equilibrium, the genome-wide effective number of independent tests can be found by summing the different M_{eff} estimates for the 22 chromosomes,

$$M_{\text{eff,g}} = \sum_{i=1}^{22} \{M_{\text{eff}}\}_{\text{chromosome i}}.$$

and then use $M_{\text{eff,g}}$ in the method of Šidák as described in Section 7.4 to find the genome-wide individual significance threshold, α_p.

The other alternative described by Cheverud (2001) for estimating the effective number of tests is to construct one correlation matrix for the whole genome, and use the variance of the eigenvalues in the method of Cheverud (2001) and Nyholt (2004) as described in Section 8.1, but this will be a very computational intensive problem.

Since SNPs on different chromosomes are expected to be in linkage disequilibrium in general populations (Gao et al. 2008), the method of Gao et al. (2008) and Li & Ji (2005) estimates the genome-wide effective number of independent tests by the sum of the different M_{eff} estimates for all 22 chromosomes.

From the results of Moskvina & Schmidt (2008) it is not written explicitly how the genome-wide individual significance level, α_p, is found. For the data analyzed by Moskvina & Schmidt (2008), the sum of the chromosome-specific M_{eff} estimates using window-size equal to $w = 20$ is $M_{\text{eff,g}} = 298518.6$. For this window-size, Moskvina & Schmidt (2008) estimated the genome-wide significance level $\alpha_p = 1.68 \cdot 10^{-7}$ and the genome-wide effective number of tests $M_{\text{eff,g}} = 306981$. We observe that the $M_{\text{eff,g}}$ estimate using the method of Moskvina & Schmidt (2008) is not equal to the sum of the chromosome-specific M_{eff} estimates. The result may indicate that the genome-wide significance level for the method of Moskvina & Schmidt (2008) is estimated based on one single correlation matrix for the genome and using a small window-size for this correlation matrix. These results shows that the method of Moskvina & Schmidt (2008) does not seem to be additive as for other methods where the estimate of $M_{\text{eff,g}}$ is found by summing the chromosome-specific M_{eff} estimates. Comparing with the other methods, the resampling procedure is also not a additive method for estimating $M_{\text{eff,g}}$. The method of Moskvina & Schmidt (2008) and resampling procedures are based on estimating the individual significance level, α_p, and then use the method of Šidák to find the estimate of the effective number of indepenent tests.

10.2 Genome-wide significance level

Resampling procedures are considered as the gold standard in multiple testing correction. Dudbridge & Gusnanto (2008) described a permutation method for estimating one single genome-wide significance threshold using a resampling procedure. For each resampled data set in the method of Dudbridge & Gusnanto (2008), half the individuals in the data set was classified as cases and the other half as controls. For each permutation the 1000 smallest p-values was recorded.

The genome-wide significance level threshold, α_p, was by Dudbridge & Gusnanto

(2008) set equal to the 5% quantile point of the distribution of the observed minimum p-values. In the method of Dudbridge & Gusnanto (2008), α_p represents the genome-wide individual significance level given a marker density. For a region with low SNP densities, Dudbridge & Gusnanto (2008) expected the SNPs to be independent, and the 5% significance level was by Dudbridge & Gusnanto (2008) found using the method of Bonferroni as described in Section 7.4. In a region with high SNP density, Dudbridge & Gusnanto (2008) expected the 5% significance level to converge to an asymptote, giving the individual significance level for the whole genome.

10.3 Effective ratio

We define the ratio between the effective number of independent tests for a chromosome, M_{eff}, and the total number of SNPs, m, on the chromosome as

$$\text{effective ratio} = \frac{M_{\text{eff}}}{m}.$$

Similarly, we have the effective ratio for the whole genome

$$\text{effective ratio} = \frac{M_{\text{eff,g}}}{m}$$

where m is the total number of SNPs on the genome.

We denote the effective ratio for the whole genome by M_g, and the effective ratio for the individual chromosomes by M_c.

Table 10.1: Effective ratio for the Illumnia 1M data analyzed by Gao et. al. (2010)

Method	M_g	M_c	mean(M_c)
10000 permutations	0.49	$0.49 - 0.56$	0.53
simple$\mathcal{M}$	0.53	$0.50 - 0.60$	0.54
$K_{\text{eff}}, w = 20$	0.68	$0.65 - 0.72$	0.68

Table 10.1, 10.2 and 10.3 shows the effective ratio calculated from the results of Gao et al. (2008), Gao et al. (2010) and Moskvina & Schmidt (2008). We observe that the mean effective ratio for the 22 chromosomes are approximately equal to the effective ratio for the whole genome. The results of the method of Gao et al. (2008) gives the genome-wide effective ratio as 0.53 and 0.60 for the Illumnia 1M

Table 10.2: Effective ratio for the Affymetrix 500K data analyzed by Gao et. al. (2010)

Method	M_g	M_c	mean(M_c)
10000 permutations	0.54	$0.50 - 0.62$	0.57
simple$\mathcal{M}$	0.60	$0.57 - 0.68$	0.61
$K_{\text{eff}}, w = 20$	0.67	$0.66 - 0.74$	0.70

Table 10.3: Effective ratio for the data analyzed by Moskvina and Schmidt (2008)

Method	M_g	M_c	mean(M_c)
1000 permutations	0.60	$0.46 - 0.70$	0.60
$K_{\text{eff}}, w = 10$	0.66	$0.64 - 0.71$	0.67
$K_{\text{eff}}, w = 20$	0.65	$0.63 - 0.70$	0.66

and Affymetrix 500K data analyzed by Gao et al. (2010), respectively. The results using the method of Moskvina & Schmidt (2008) gives the genome-wide effective ratio as $0.68, 0.67$ and 0.65 for window size equal to $w = 20$. For window size $w = 10$, the estimated genome-wide effective ratio was equal to 0.65.

The TOP8 data

From Section 8.3 and 9.2 we observed that the estimate of M_{eff} using the method of Gao et al. (2008) is highly dependent on the block size used. Therefore, the effective ratio for the method of Gao et al. (2008) is not a preferable method for estimating M_{eff}.

For the method of Moskvina & Schmidt (2008) using window size $w = 25$ the estimated effective number of independent tests for chromosome 22 of the TOP8 data was $M_{\text{eff}} = 6259.273$, which gives the effective ratio

$$\text{effective ratio} = \frac{M_{\text{eff}}}{m} = \frac{6259.273}{8928} = 0.70108.$$

We observe that this estimated effective ratio is approximately equal to the effective ratio for the method of Moskvina & Schmidt (2008) given in Table 10.1, 10.2 and 10.3.

We implemented the minP single-step procedure based on the conditional p-value as defined in Section 6.4 to control the FWER. In Section 9.3, we estimated the individual significance level and the corresponding effective number of tests as 5284.572. This gives the effective ratio

$$\begin{aligned}\text{effective ratio} &= \frac{M_{\text{eff}}}{m}\\ &= \frac{5284.572}{8928}\\ &= 0.59191.\end{aligned}$$

The results from Gao et al. (2010) and Moskvina & Schmidt (2008) showed that the effective ratio for the permutation procedure was 0.49, 0.54 and 0.60, respectively. Moskvina & Schmidt (2008) used only 1000 permutations of the data, Gao et al. (2010) used 10000 permutations. The number of resampled data sets may influence the precision in the result, but this topic was not considered in this thesis.

Effective ratio as a method for estimating M_{eff} and $M_{\text{eff,g}}$

In this section, we have presented the effective ratio for individual chromosomes and for the whole genome, and we have observed from the data analyzed by Gao et al. (2010) and Moskvina & Schmidt (2008) that the effective ratio seems to be stable for the different chromosomes. Based on these observations, two interesting questions occurs, does it exists an effective ratio for individual chromosomes and the whole genome, and what is the value of the effective ratio. If it exists such an effective ratio, we can determine the effective number of tests for a GWAS or a single chromosome by multiplying the total number of SNPs by the effective ratio, which gives a result that is not dependent of the statistical test used. Considering the effective ratio as a method for estimating M_{eff} and $M_{\text{eff,g}}$, it is reasonable to think that the effective ratio will depend on the density of SNPs along the chromosomes.

Chapter 11

Discussion and Conclusion

In this thesis, different methods for estimating the effective number of independent tests have been presented and some possible applications of the methods in a GWAS have been discussed. In a multiple testing problem, the correlation between pairs of SNPs is of importance, and in this thesis we have considered correlation between SNPs based both on haplotypes and genotypes.

LD vs. CLD correlation

We observed that using the theoretical grid as described in Section 4.3 the CLD correlation was more extreme than the LD correlation in approximately 75% of the cases, and based on the TOP8 data, the LD correlation was more extreme than the CLD correlation in approximately 52% of the cases (see Section 5.4). In Chapter 5 we observed that the maximal LD correlation between SNPs depends on the minor allele frequencies for the different SNPs. We investigated the distribution for the minor allele frequencies, both based on a theoretical grid and for chromosome 22 of the TOP8 data. We observed that the distribution for the minor allele frequencies is different for the theoretical grid and for the TOP8 data, which may explain our opposite results when comparing LD and CLD correlation. More work on this topic is needed.

When analyzing GWAS data, the haplotype phase is in general unknown. Using LD correlation we need to estimate the haplotype frequency P_{AB} as defined in Equation (4.3). This can be done by using maximum likelihood estimation, but for all individuals Hardy-Weinberg equilibrium must be assumed. The CLD correlation does not assume HWE and can be estimated directly from the observed genotype data. As shown in Section 5.5 estimating the CLD correlation matrix is a less computationally intensive problem than estimating the LD correlation matrix.

Our observations in Chapter 4 and 5 give no reasons for choosing LD correlation in favor of CLD correlation when analyzing GWAS data.

Methods for estimating M_{eff}

The gold standard for multiple testing problem within this field is resampling-based methods, e.g. minP procedure, but these methods are time consuming and computationally intensive. Cheverud (2001) was the first to propose a method for estimating the effective number of independent tests, M_{eff}, and then applied the method of Šidák to find the individual significance level. Different methods for estimating M_{eff} for individual chromosomes have been presented in Chapter 8 and applied to the TOP8 data in Chapter 9.

As described in Section 8.3 we have observed that the result of the method of Gao et al. (2008) is highly dependent on the block size used and therefore we will not recommend this method. Based on our observations comparing the different methods, the method of Moskvina & Schmidt (2008) is a more preferable method for estimating M_{eff}. The method of Moskvina & Schmidt (2008) is robust with respect to window size w and the results are close to the result of the resampling procedure, minP.

GWAS

When considering GWAS, we have two possible views for the individual significance level for each test when considering the whole genome. One view is to consider the same significance level for all chromosomes, the other view is to consider chromosome-specific individual significance levels. Dudbridge & Gusnanto (2008) described a method for estimating one genome-wide individual significance level, while Cheverud (2001) suggested that each individual chromosome should be tested at individual chromosome-specific thresholds since the densities of markers will be different for different chromosomes. Discussion of these two alternatives can be a topic for future work. Is it more sensible to estimate one significance level for the whole genome or should we estimate chromosome specific individual significance levels?

Different views of multiple testing correction

A different view of multiple testing have been introduced by Dudbridge & Gusnanto (2008). This new approach suggests that in a multiple testing problem, we should correct not only for the collected markers, but also for the uncollected markers. Dudbridge & Gusnanto (2008) estimated one single individual significance threshold by subsampling SNPs at different SNP densities. The current view as presented in this thesis is that the effective number of independent tests depends on the number of SNPs genotyped, and eventually on a effective ratio. Because of the development of the genotyping technology, we expect the number of uncollected markers to decrease. As pointed out by Han et al. (2009), the different views of multiple testing correction will converge as the number of uncollected markers goes to zero.

Estimating M_{eff} and $M_{\text{eff,g}}$ or use effective ratio?

In Chapter 10 an alternative method for estimating the genome-wide effective number of independent tests based on an effective ratio was presented. If there exists such an effective ratio, the genome-wide effective number of independent tests can be found by multiplying the total number of SNPs with the effective ratio as described in Section 10.3. Using the different methods for estimating the effective number of independent tests presented in Chapter 8, we get an estimate of M_{eff} for each chromosome, which are used to find the estimate of the effective number of independent tests for the whole genome.

Based on our observations, two interesting questions is if there exists such an effective ratio and what is the value of the effective ratio. Another interesting question for future work is whether chromosome-specific individual significance levels can be estimated using the effective ratio or if we should estimate the effective number of tests based on the methods presented in Chapter 8.

Conclusion

In this thesis, different methods for estimating the effective number of independent tests have been studied and tested on a large data set on schizophrenia and bipolar disorder from the TOP study. The different methods were compared both theoretically and when applied to the TOP8 data. The different methods and the results were also compared to the resampling-based minP procedure. The methods were tested using either haplotype or genotype correlation. According to the literature, resampling-based procedures are considered as the gold standard for multiple testing problems within this field. Due to computational complexity we would like to do a less time consuming method by using a method for estimating the effective number of independent tests. But, as shown in this thesis, estimating the full LD correlation matrix will be very time consuming. Based on our observations in this thesis, we will recommend to use the method of Moskvina & Schmidt (2008) since this method is robust with respect to window size and the result is close to the result of the minP procedure.

Bibliography

Athanasiu, L., Mattingsdal, M., Kähler, A. K., Brown, A., Gustadsson, O., Agartx, I. & et. al, I. G. (2010), 'Gene variants associated with schizophrenia in a norwegian genome-wide study are replicated in a large european cohort', *Journal of Psychiatric Research* **44**, 748–753.

Bakke, Ø. & Langaas, M. (2012), 'The number of $2 \times c$ tables with given margins'. submitted.

Barret, J., Fry, B., Maller, J. & Daly, M. (2005), 'Haploview: analysis and visualization of ld and haplotype maps', *Bioinformatics* **21**, 263–265.

Benjamini, Y. & Hochberg, Y. (1995), 'Controlling the false discovery rate: A practical and powerful approach to multiple testing', *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289–300.

Casella, G. & Berger, R. L. (2002), *Statistical inference*, 2. edn, Duxbury Thomson Learning.

Chen, Z. & Liu, Q. (2011), 'A new approach to account for the correlations among single nucleotide polymorphisms in genome-wide association studies', *Human Heredity* **72**, 1–9.

Cheverud, J. M. (2001), 'A simple correction for multiple comparisons in interval mapping genome scans', *Heredity* **87**, 52–58.

Djurovic, S., Gustafsson, O., Mattingsdal, M., Athanasiu, L., Bjella, T., Tesli, M. & et. al, I. A. (2010), 'A genome-wide association study of bipolar disorder in norwegian individuals, followed by replication in icelandic sample', *Journal of Affective Disorders* **126**, 312–316.

Dudbridge, F. & Gusnanto, A. (2008), 'Estimation of significance thresholds for genomewide association scans', *Genetic Epidemiology* **32**, 227–234.

Dudbridge, F. & Koeleman, B. P. C. (2004), 'Efficient computation of significance levels for multiple associations in large studies of correlated data, including genomewide association studies', *Am. J. Hum. Genet.* **75**, 424–435.

Foulkes, A. S. (2009), *Applied Statistical Genetics with R For Population Based Studies*, Springer.

Gabriel (2002), 'The structure of haplotype blocks in the human genome', *Science* **296**, 2225–2229.

Galwey, N. W. (2009), 'A new measure of the effective number of tests, a practical tool for comparing families of non-independent significance tests', *Genetic Epidemiology* **33**, 559–568.

Gao X. (2012), 'simpleM', `http://simplem.sourceforge.net/`.

Gao, X., Becker, L. C., Becker, D. M., Starmer, J. D. & Province, M. A. (2010), 'Avoiding the high bonferroni penalty in genome-wide association studies', *Genetic Epidemiology* **34**, 100–105.

Gao, X., Starmer, J. & Martin, E. R. (2008), 'A multiple testing correction method for genetic association studies using correlated single nucleotide polympor-phisms', *Genetic Epidemiology* **32**, 361–369.

Ge, Y., Dudoit, S. & Speed, T. P. (2003), 'Resampling-based multiple testing for microarray data analysis', *Sociedad de Estadistica e Investigacion Operativa Test* **12**, 1–77.

Griffiths, A. J. F., Gelbart, W. M., Lewontin, R. C. & Miller, J. H. (2002), *Modern genetic analysis - Integrating genes and genomes*, 2. edn, W. H. Freeman and Company.

Hamilton, D. & Cole, D. (2004), 'Standardizing a composite measure of linkage disequilibrium', *Annals of Human Genetics* **68**, 234–239.

Han, B., Kang, H. M. & Eskin, E. (2009), 'Rapid and accurate multiple testing correction and power estimation for millions of correlated markers', *PLoS Genetics* **0**, 1–13.

Human Genome Project Information (2011), 'SNP Fact Sheet', `http://www.ornl.gov/sci/techresources/Human_Genome/faq/snps.shtml`.

Kulle, B., Frigessi, A., Edvardsen, H., Kristensen, V. & Wojnowski, L. (2008), 'Accounting for haplotype phase uncertainty in linkage disequilibrium estimation', *Genetic Epidemiplogy* **32**, 168–178.

Langaas, M. & Bakke, Ø. (2012), 'Methods for calculating statistical significance for discrete genotype-phenotype case-control data'. in preparation.

Lewontin, R. C. (1964), 'The interaction of selection and linkage. i. general considerations; heterotic models', *Genetics* **49**, 49–67.

Lewontin, R. C. & Kojima, K. (1960), 'The evolutionary dynamics of complex polymorphisms', *Evolution* **14**, 458–472.

Li, J. & Ji, L. (2005), 'Adjusting multiple testing in multilocus analyses using the eigenvalues of the correlation matrix', *Heredity* **95**, 221–227.

Moskvina, V. & Schmidt, K. M. (2008), 'On multiple-testing correction in genome-wide association studies', *Genetic Epidemiology* **32**, 567–573.

National Human Genome Research Institute (2011), 'Genome-Wide Association Studies', `http://www.genome.gov/20019523`.

National Institute of Mental Health (2012*a*), 'What is bipolar disorder?', `http://www.nimh.nih.gov/health/publications/bipolar-disorder/what-is-bipolar-disorder.shtml`.

National Institute of Mental Health (2012*b*), 'What is schizophrenia?', `http://www.nimh.nih.gov/health/publications/schizophrenia/what-is-schizophrenia.shtml`.

Nyholt, D. R. (2004), 'A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other', *Am. J. Hum. Genet.* **74**, 765–769.

Nyholt, D. R. (2005), 'Evaluation of nyholt's procedure for multiple testing correction - authors reply', *Human Heredity* **60**, 61–62.

R Development Core Team (2011), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Ripley, B. D. (1996), *Pattern Recognition and Neural Networks*, Cambridge University Press.

Salyakina, D., Seaman, S. R., Browning, B. L., Dudbridge, F. & Muller-Myhsok, B. (2005), 'Evaluation of nyholt's procedure for multiple testing correction', *Human Heredity* **60**, 19–25.

Stephens, M. & Scheet, P. (2005), 'Accounting for decay of linkage disequilibrium in haplotype inference and missing data imputation', *American Journal of Human Genetics* **76**, 449–462.

Stephens, M., Smith, N. J. & Donnelly, P. (2001), 'A new statistical method for haplotype reconstruction from population data', *American Journal of Human Genetics* **68**, 978–989.

Thompson, J. S. & Thompson, M. W. (1980), *Genetics in Medicine*, W.B. Saunders Company.

Tian, J. & Xu, C. (2009), *MaXact: Exact max-type Cochran-Armitage trend test(CATT).* R package version 0.1.
URL: *http://CRAN.R-project.org/package=MaXact*

TOP (2012*a*), 'Deltakerinformasjon', `http://www.med.uio.no/klinmed/forskning/grupper/top/Deltakere/`.

TOP (2012*b*), 'K. G. Jebsen senter for psykoseforskning', `http://www.med.uio.no/klinmed/forskning/grupper/top/Jebsen/om-senteret.html`.

TOP (2012*c*), 'TOP', `http://www.med.uio.no/klinmed/forskning/grupper/top/Mer%20om%20TOP/`.

U.S. Department of Health and Human Services (2011), 'Genome-Wide Association Studies', `http://gwas.nih.gov/`.

Warnes, G., with contributions from Gregor Gorjanc, Leisch, F., & Man., M. (2011), *genetics: Population Genetics.* R package version 1.3.6.
URL: *http://CRAN.R-project.org/package=genetics*

Weir, B. S. (1996), *Genetic Data Analysis II: Methods for discrete population genetic data*, Sinauer Associates, Inc.

Weir, B. S. (2008), 'Linkage disequilibrium and association mapping', *Annual Review of Genomics and Human Genetics* **9**, 129–142.

Westfall, P. H. & Young, S. S. (1993), *Resampling-Based Multiple Testing: Examples and Methods for p-value Adjustment*, Wiley.

Ziegler, A. & König, I. R. (2010), *A Statistical Approach to Genetic Epidemiology*, 2. edn, Wiley-Blackwell.

Appendix A

Notation

m	Total number of tests
n	Total number of individuals
α	Experimentwide significance level, FWER
α_p	Significance level for individual tests
M_{eff}	Effective number of independent tests
LD,ρ_{LD}	Linkage disequilibrium correlation
CLD,ρ_{CLD}	Composite linkage disequilibrium correlation
r	Pearson correlation coefficient
D_A	Hardy-Weinberg disequilibrium
Δ_{AB}	Composite linkage disequilibrium
p_A	Frequency of allele A
P_{AB}	Frequency of haplotype AB
$f(x\|\theta)$	Probability density function
$L(\theta\|\mathbf{x})$	Likelihood function
FWER	Family-wise error rate
FDR	False discovery rate

Appendix B

Proof of P_{AB} and $P_{A/B}$ from Weir (1996)

Proof. From Section 4.1 we have

$$P_{AB} = E(X_1Y_1) = E(X_2Y_2).$$

We define

$$P(h_1, h_2) = P(\text{haplotype } h_1 \text{ on gamete 1 and haplotype } h_2 \text{ on gamete 2})$$

Then, we have

$$E(X_1Y_1) = \sum P(AB, xy), xy = \{AB, Ab, aB, ab\}$$

and

$$E(X_2Y_2) = \sum P(xy, AB), xy = \{AB, Ab, aB, ab\}.$$

We get

$$\begin{aligned}2P_{AB} &= \sum P(AB, xy) + \sum P(xy, AB) \\ &= P(AB, AB) + P(AB, Ab) + P(AB, aB) + P(AB, ab) + \\ &P(AB, AB) + P(Ab, AB) + P(aB, AB) + P(ab, AB)\end{aligned} \tag{B.1}$$

According to Weir (1996) we have

$$\begin{aligned}
P_{AB}^{AB} &= P(AB, AB)\\
P_{Ab}^{AB} &= P(AB, Ab) + P(Ab, AB)\\
P_{aB}^{AB} &= P(AB, aB) + P(aB, AB)\\
P_{ab}^{AB} &= P(AB, ab) + P(ab, AB)\\
P_{Ab}^{aB} &= P(Ab, aB) + P(aB, Ab).
\end{aligned}$$

Equation (B.1) can then be rewritten as

$$2P_{AB} = P_{AB}^{AB} + P_{Ab}^{AB} + P_{aB}^{AB} + P_{ab}^{AB},$$

and then

$$P_{AB} = P_{AB}^{AB} + \frac{1}{2}(P_{Ab}^{AB} + P_{aB}^{AB} + P_{ab}^{AB}),$$

which we observe is the formula for P_{AB} according to Weir (1996, p. 122).

Similarly, for $P_{A/B}$ as defined in Section (ref) we have

$$P_{A/B} = E(X_1Y_2) = E(X_2Y_1),$$

where

$$E(X_1Y_2) = \sum P(Ay, xB), xy = \{AB, Ab, aB, ab\}$$

and

$$E(X_2Y_1) = \sum P(xB, Ay), xy = \{AB, Ab, aB, ab\}.$$

This gives

$$\begin{aligned}
2P_{A/B} &= \sum P(Ay, xB) + \sum P(xB, Ay)\\
&= P(AB, AB) + P(AB, aB) + P(Ab, aB) + P(Ab, AB)\\
&P(AB, AB) + P(AB, Ab) + P(aB, AB) + P(aB, Ab)\\
&= 2P_{AB}^{AB} + P_{Ab}^{AB} + P_{aB}^{AB} + P_{Ab}^{aB},
\end{aligned}$$

and then, according to Weir (1996, p. 122) we have

$$P_{A/B} = P_{AB}^{AB} + \frac{1}{2}(P_{Ab}^{AB} + P_{aB}^{Ab} + P_{Ab}^{aB}).$$

□

Appendix C

Singular value decomposition

Singular value decomposition (SVD) is defined as

$$Z = UDV^T$$

where Z is a $n \times p$ matrix, U is a $n \times p$ matrix, D is a $p \times p$ matrix and V is a $p \times p$ matrix (Ripley 1996, p. 289). The elements of the matrix U is the eigenvectors of ZZ^T, the elements of V are the eigenvectors of Z^TZ and the matrix D is a diagonal matrix of the singular values of Z, which is equal to

$$\sqrt{\text{diag}(\text{eigenvalues}(ZZ^T))}$$

The matrices ZZ^T and Z^TZ have the same eigenvalues but different eigenvectors. This can be seen by

$$\begin{aligned} Z^TZe &= \lambda e \\ ZZ^TZe &= Z\lambda e \\ ZZ^Te^* &= \lambda e^* \end{aligned}$$

where $e^* = Ze$.

We will use this result were the matrix Z is a centered and scaled version of the genotype matrix where n equals the number of persons and p equals the number of SNP's. Since the data are centered the maximal number of nonzero eigenvalues are equal to $n - 1$.

We must assume that the matrix Z of the genotypes are centered and scaled. The estimate for correlation is given by

$$\hat{\rho} = \frac{1}{n-1}(Z^TZ)$$

Appendix D

R-code

Estimating LD and CLD correlation

```
correlation <- function(numgeno1,numgeno2)
{
  nvec <- rep(0,9)
  nNA <- 0
  nnotNA <- 0

  for(i in 1:1551)  #0=AA;1=Aa,2=aa
  {
    if((is.na(numgeno1[i])==TRUE)||(is.na(numgeno2[i])==TRUE))
    {
      nNA <- nNA +1
      i <- i+1
    }
    else{
      nnotNA <- nnotNA +1
      if(numgeno1[i]==0) #AA
      {
        if(numgeno2[i]==0)
        {
          nvec[1] <- nvec[1]+1
        }
        if(numgeno2[i]==1)
        {
          nvec[2] <- nvec[2]+1
        }
        if(numgeno2[i]==2)
        {
          nvec[3] <- nvec[3]+1
        }
      }
      if(numgeno1[i]==1) #Aa
      {
        if(numgeno2[i]==0)
        {
          nvec[4] <- nvec[4]+1
        }
        if(numgeno2[i]==1)
        {
          nvec[5] <- nvec[5]+1
        }
        if(numgeno2[i]==2)
        {
          nvec[6] <- nvec[6]+1
        }
      }
      if(numgeno1[i]==2) #aa
      {
        if(numgeno2[i]==0)
        {
          nvec[7] <- nvec[7]+1
        }
        if(numgeno2[i]==1)
        {
          nvec[8] <- nvec[8]+1
        }
        if(numgeno2[i]==2)
        {
          nvec[9] <- nvec[9]+1
        }
      }
    }
  }
  #   BB       Bb       bb
  #AA nvec[1]  nvec[2]  nvec[3]
  #Aa nvec[4]  nvec[5]  nvec[6]
```

```
#aa nvec[7]  nvec[8]  nvec[9]

p <- rep(NA,9)
n <- nnotNA
for(i in 1:9)
{
  p[i] <- nvec[i]/n
}

PAA <- (p[1]+p[2]+p[3])
PAa <- (p[4]+p[5]+p[6])
Paa <- (p[7]+p[8]+p[9])
PBB <- (p[1]+p[4]+p[7])
PBb <- (p[2]+p[5]+p[8])
Pbb <- (p[3]+p[6]+p[9])

pA <- PAA + (1/2)*PAa
pa <- 1-pA
pB <- PBB + (1/2)*PBb
pb <- 1-pB

Dmin <- max(-pA * pB, -pa * pb)
pmin <- pA * pB + Dmin
Dmax <- min(pA * pb, pB * pa)
pmax <- pA * pB + Dmax

loglik <- function(pAB,...){(2 * nvec[1]+nvec[2]+nvec[4])*log(pAB)+
 (2*nvec[3]+nvec[2]+nvec[6])*log(pA - pAB) +
 (2*nvec[7]+nvec[4]+nvec[8])*log(pB - pAB)+(2*nvec[9]+nvec[8]+
   nvec[6])*log(1 - pA - pB + pAB)+nvec[5] *
    log(pAB*(1 - pA - pB + pAB)+(pA - pAB)*(pB - pAB))
}
solution <- optimize(loglik, lower = pmin + .Machine$double.eps,
      upper = pmax - .Machine$double.eps, maximum = TRUE)

pAB <- solution$maximum

 #P_{AB} + P_{A/B}:
 PABAdB <- 2*p[1]+p[2]+p[4]+(1/2)*p[5]   #p5 - (AB,ab) and (Ab,aB)
```

```
   D <- pAB - (pA*pB)
   DA <- PAA - (pA^2)
   DB <- PBB - (pB^2)

    rhoLD <- D/sqrt(pA*pa*pB*pb)
    rhoCLD <- (PABAdB - (2*pA*pB))/(sqrt((pA*pa+DA)*(pB*pb+DB)))

  return(list("rhoLD"=rhoLD,"rhoCLD"=rhoCLD))
}
```

The method of Nyholt

```
nyholt <- function(matr)
{
Mvec <- NA
Nyholt <- NA
    Mvec <- dim(matr)[2]
    thislambda <- eigen(matr)$values

    Nyholt<- 1+(Mvec-1)*(1- var(thislambda)/Mvec)
    return(Nyholt)
}
```

The method of Gao

```
gao <- function(mat,M,r)
{
  blokk <- seq(r,M,r)
  Mvec <- M
  k <- length(blokk)

  Mgaovec <- rep(NA,k)
  blokkmat1 <- mat[1:blokk[1],1:blokk[1]]
  e <- eigen(blokkmat1)
  lambda <- e$values
  percexplny <- cumsum(lambda)/sum(lambda)
  Mgaovec[1] <- min((1:length(lambda))[percexplny >=varexpl])

  for(j in 2:k)
  {
    a <- blokk[j-1]+1
    b <- blokk[j]
    blokkmat <- mat[a:b,a:b]
    e <- eigen(blokkmat)
    lambda <- e$values
    percexplny <- cumsum(lambda)/sum(lambda)
    Mgaovec[j] <- min((1:length(lambda))[percexplny >=varexpl])
  }
  a <- blokk[j]+1
  b <- 8928
  blokkmat <- mat[a:b,a:b]
  e <- eigen(blokkmat)
  lambda <- e$values
  percexplny <- cumsum(lambda)/sum(lambda)
  Mgaovec[j+1] <- min((1:length(lambda))[percexplny >=varexpl])

  Meff = sum(Mgaovec)

  return(Meff)
}
```

The method of Moskvina

```
moskvina <- function(mat,M,w,alpha)
{
  r <- 0
  sumKappa <- 0
  Keff <- 0
  kappavec <- rep(NA,M-1)
  corr <- rep(NA,M-1)

  for(i in 1:M)
  {
     if(i<=w)
     {
       r <- max(abs(mat[i,1:(i-1)]))
     }
     if(i>w)
     {
      r <- max(abs(mat[i,(i-w):(i-1)]))
     }
    kappavec[i] <- sqrt(1-r^(-1.31*log10(alpha)))
    sumKappa <- sumKappa+ kappavec[i]
    corr[i] <- max(abs(mat[i,1:(i-1)]))
  }
  Keff <- 1 + sumKappa
  Pn <- 1 - (1-alpha)^(Keff)
  return(list("Keff"=Keff,,"fwer"=Pn))
}
```

The method of Moskvina without approximation

```
s<-numgenods
c<-corrmat
m<-8928
r<-rep(0,length(s[1,]))

for(i in 2:m)
{
  r[i]<-max(abs(corrmat[i,1:(i-1)]))
}
const<-sqrt(2/pi)

Pi<-function(alfa){
  b<-1-alfa
  sigma<-qnorm(1-alfa/2)
  prod<-1

  for(i in 2:m){
    const2<-1/sqrt(1-r[i]^2)
    prod<-prod*(1-const/b*integrate(function(x)exp(-x^2/2)*
      pnorm((r[i]*x-sigma)*const2),-sigma,sigma)$value)
  }
  1-b*prod
}
```

Resampling procedure to control FWER

```
library(MaXact)

B <- 100000
ninner <- 100
nouter <- B/ninner

geno <- dget(paste(datadir,"numgenods22.dd",sep=""))
disease <- read.table(paste(datadir,"TOP8chr22Disease.txt",sep=""))
geno <- geno[disease!=-9,]
dis <- disease[disease!=-9]
nsnp <- dim(geno)[2]

mmat <- matrix(0,ncol=3,nrow=nsnp)
mmat[,1] <- apply(geno==0,2,sum,na.rm=TRUE)
mmat[,2] <- apply(geno==1,2,sum,na.rm=TRUE)
mmat[,3] <- apply(geno==2,2,sum,na.rm=TRUE)

bigminP <- NULL
set.seed(123)
for (i in 1:nouter)
  {
    minP <- rep(NA,ninner)
    for (j in 1:ninner)
      {
        newstatus <- sample(dis,replace=F)
        minP[j] <- min(calcCpvalfrommat3(newstatus,geno,nsnp,mmat))
      }
    cat(minP,file="minP.res","\n",append=TRUE)

    bigminP <- c(bigminP,minP)
  }

write.table(bigminP,"bigminP.res")
```