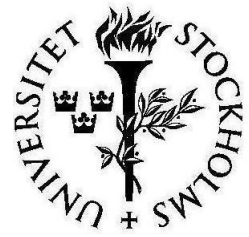

Mathematical Statistics
Stockholm University

Statistical Methods for Assessing Genetic Association in the Presence of Linkage

Gudrun Jonasdottir

Research Report 2005:6
Licentiate thesis

ISSN 1650-0377

Postal address:

Mathematical Statistics
Dept. of Mathematics
Stockholm University
SE-106 91 Stockholm
Sweden

Internet:

http://www.math.su.se/matstat

Mathematical Statistics
Stockholm University
Research Report **2005:6**,
http://www.math.su.se/matstat

Statistical Methods for Assessing Genetic Association in the Presence of Linkage

Gudrun Jonasdottir*

May 2005

Abstract

This thesis is concerned with the analysis of association between genetic markers and disease. We consider a scenario where it is known that a genetic region of interest has a tendency to be transmitted intact from parent to offspring. The region is said to be linked. The hope is that a mutation involved in the causal pathway of the disease is contained in the linked region, and that we can pinpoint its exact location through association analysis.

We describe and assess existing methodologies, parametric and non-parametric, for the testing and estimation of association in the presence of linkage. Many genetic association studies have complex ascertainment schemes. We develop a novel score test of association in the presence of linkage for binary traits that takes ascertainment, as well as population stratification, into account.

KEY WORDS: Association, Linkage, Variance Components Model, Family-Based Association Test, Random Effects, Score Test, Retrospective Likelihood.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: gudrunj@math.su.se.

Acknowledgements

First of all, I would like to thank my main supervisor, Juni Palmgren, for the opportunities she has given me and especially for letting me spend a whole semester at Harvard School of Public Health. I would also like to thank my co-supervisor, Keith Humphreys, who in his stressful days, managed to find time to help me get back on track.
I would like to thank all my colleagues at the Department of Medical Epidemiology and Biostatistics (MEB) at Karolinska Institutet and at the Department of Mathematics, divistion of Mathematical Statistics (Matstat), at Stockholm University. Especially the Biostat group at MEB; Rino, Paul, Annica, Arief, Mark, Anna J, Åsa, Monica, Yudi, Alexander, Marie, Sven, Chuen Seng, Anna T, Ben, and John; and the computer wiz'es at MEB, Olle and Denny, and Patricia, Elinore, Ola Hössjer, Hedvig, Andreas and Örjan at Matstat. I would also like to thank everyone at the Department of Biostatistics at Harvard School of Public Health for making me feel welcome during my stay there. Maria Grünewald deserves a special "thank you" as she is both a friend and a colleague: Thanks for all the good advice, talks and chocolate.
My collaborators at Neurotec in Huddinge, led by Jan Hillert, have taught me almost everything I know about genetics. We have exchanged knowledge and collaborated on papers, and it has been a lot of fun. Special thanks to Kristina, Frida, Boel, Eva, Iza, Helena, Thomas, Kerstin and Jan.
I would like to thank all my friends, new and old ones, for forcing me to practice my social skills. Special thanks to my good friends Catherine, Ida, Johanna and Mary. A very special "thank you" to my girlfriend Sophia for letting me be a part of (and making me take part of) her life.

Without the love and endless support from my father, Jonas, and my mother, Hrafnhildur, I would not have made it this far. They made it happen through the genetical predisposition to be stubborn and a study-positive environment. That is why this thesis is dedicated to them, my beloved father and mother.

Financial support from SSF, STINT and VR is gratefully acknowledged.

Contents

Notation

$\boldsymbol{H_0}$ Null hypotheses

$\boldsymbol{H_1}$ Alternative hypotheses

$\boldsymbol{Y}$ Trait

$\boldsymbol{M}$ Offspring marker data

$\boldsymbol{g}$ Parental marker data

$\boldsymbol{v}$ Inheritance vector

P Probability

L Likelihood

I Fisher Information matrix

S Score function

θ Recombination fraction

r^2 Correlation coefficient

A/a Marker alleles (biallelic)

D/d Disease susceptibility alleles (biallelic)

n Number of families

Abbreviations

DNA Deoxyribonucleic acid

SNP Single Nucleotide Polymorphism

LD Linkage Disequilibrium

QTL Quantitative Trait Locus

VCM Variance Components Model

FBAT Family-Based Association Test

RL Rabinowitz Laird

GRE Gamma Random-Effects

HWE Hardy-Weinberg Equilibrium

IBD Identity-By-Descent

IBS Identity-By-State

DS Disease Susceptibility

MGRR Matched Genotype-based Relative Risk

GHRR Genotype-based Haplotype Relative Risk

HHRR Haplotype-based Haplotype Relative Risk

TDT Transmission Disequilibrium Test

LRT Likelihood Ratio Test

ML Maximum Likelihood

GLM Generalized Linear Model

GLMM Generalized Linear Mixed Model

IWLS Iterative re-Weighted Least Squares

1 Introduction

This thesis is concerned with statistical methodology for finding genes that are associated with a trait, either binary or continuous. Attention is restricted to a specific scenario where it is known that the genetic region of interest has a tendency to be transmitted intact from parent to offspring. We say that there is *linkage* in that region and that we analyse *association in the presence of linkage*. The hope is that a mutation involved in the causal pathway of the disease is contained in the linked region, and that we can pinpoint its exact location. We may see this as a conditional analysis which requires specific statistical methodology. The null and alternative hypotheses can be written as

$$\begin{aligned} H_0 &: \text{ Linkage, but no association.} \\ H_1 &: \text{ Linkage and association.} \end{aligned} \tag{1.1}$$

We discuss two existing methodologies; one non-parametric, and one parametric, to test and to model association in the present of linkage. The parametric model is a normal mixed effects model for testing association in the presence of linkage for continuous traits, the Variance Components Model (VCM) of Fulker, Cherny, Sham & Hewitt (1999). It is designed to take into account that the study population may be stratified into subgroups with different genetic backgrounds (*population stratification*), but it does not take the ascertainment scheme into account. The non-parametric test statistic, the *Family-Based Association Test* (FBAT), takes the ascertainment scheme into account, as well as the population stratification. The FBAT has the additional advantage that it tests for association in the presence of linkage for both continuous and binary traits.

In this thesis, we develop a new score test for association in the presence of linkage for binary traits, which takes ascertainment and population stratification into account.

In Section 2 we present material which serves as a background to the topic of this thesis. Section 2.1 defines some fundamental genetic concepts. In Section 2.2 we define Linkage Disequilibrium and Linkage, and discuss their

relatedness. In Section 2.3, as means of a historical background, we describe tests of hypotheses closely related, but not identical, to the null and alternative hypotheses in (1.1). The aims and the structure of the main part of the thesis are presented in Section 3. Section 4 describes two existing methods for testing association in the presence of linkage in family-based studies: the VCM in Section 4.1 and the FBAT in Section 4.2. In Section 5 we give a short introduction to the Generalized Linear Mixed Models and in Section 6 the novel score test from a gamma random effects model is developed for testing association in the presence of linkage for binary traits. Section 7 is a discussion. Two papers have been extracted from the main text and are attached in Appendix C.

2 Background

2.1 Genetic Primer

Humans carry genetic information in double helix strings of nucleotides, called DNA (deoxyribonucleic acid). These strings of DNA are called *chromosomes*. There are 46 human chromosomes, forming 22 pairs and two sex chromosomes. If a certain location (*locus*, pl. *loci*) on a chromosome carries information on a specific trait (for example eye colour) then the complimentary location on the other chromosome in the pair also carries information about the same trait (eye colour).

In the simplest case when two individuals mate, one randomly selected chromosome gets transmitted from each parent to the offspring. In the more complex case, the chromosomes in a pair recombine at (supposedly) random locations, and form new chromosomes, that then are transmitted from the parent to the offspring. The probability of a recombination occurring between two loci depends on chromosome length, chromosome type and sex. The *recombination fraction* is defined, such that a recombination fraction of 1/2 means that the probability of recombination is 1, whilst a recombination fraction of 0 means that the probability of recombination is zero.

Most human DNA is identical for the whole population, but at some loci different variants exist. Variants at a locus are called *alleles* and if there are only two alleles in the population it is said that the locus is *biallelic*. Each human carry two alleles, one at each chromosome. The unordered combination of alleles is called the *genotype*, whilst the ordered combination is called the *haplotype*. By ordered we mean that the alleles at different loci can be differentiated with respect to the chromosome on which they are carried. Sometimes the alleles are referred to as the maternal and paternal allele, implying knowledge of the transmission of alleles from the mother and father of the offspring. If the variant is defined on a single nucleotide the variant is called a *Single Nucleotide Polymorphism* (SNP). We often make the simplifying assumption that the genotypic frequencies are the product of allele frequencies. When that assumption is met, we say that the locus is in *Hardy-Weinberg Equilibrium* (HWE). It can be shown, theoretically, that in a closed population a locus reach HWE after a single round of random

mating, assuming no stratification or admixture.

In genetic association studies interest lies in finding a locus involved in a disease or trait. Such loci are called *Disease Susceptibility* (DS) loci. In order to pinpoint the location of the DS locus we find the genotypes of a set of variant loci. This procedure is called *genotyping* and the loci we genotype are called *markers*. Assume for simplicity that the DS locus is biallelic and denote the alleles D and d, d being the disease susceptibility allele. We denote the (unknown) frequency of the disease susceptibility allele by p_d ($p_D = 1 - p_d$). If it is also assumed that the marker locus is biallelic and denote the alleles by A and a. We denote the frequency of a in the population by p_a ($p_A = 1 - p_a$). Our hope is that at least one of the markers will be in such close vicinity to the DS, or that one of the markers is the DS, locus that no or little recombination occurs between the loci, and that we therefore see a co-transmission of the two. The desired endpoint, from a biological point of view, is to find the exact location and to investigate the biological consequence of the disease causing variant.

There are many different approaches for the testing and estimation of genetic association. Some base the analysis on studies of unrelated individuals, for example *case-control studies*. Others base the analysis on a cluster (or clusters) of related individuals, so called *family-based studies*. Both areas have seen an explosive methodological advancement. This thesis focus on studies of nuclear families, but most methods discussed are easily adapted to more complex family structure (or *pedigrees*).

We will often use graphs to represent family structure of observed data. Figure (2.1) summarises the form of the graphs we have used.

When studying co-transmission of alleles in families, it is common to refer to allele similarities between siblings in terms of *Identity-By-State* (IBS) and *Identical-By-Descent* (IBD). If two sibs share an allele IBS this means that they both have the same allele type, but if they share it IBD then it means that they share the same allele from the same parental chromosome. Siblings can share, either 0,1 or 2 alleles IBD. When the mode of inheritance is known we enumerate the parental alleles, so that the paternal alleles are enumerated as *1* and *2*, and the maternal alleles are enumerated as *3* and *4*. We can also keep track of co-transmission by defining an inheritance vector where each

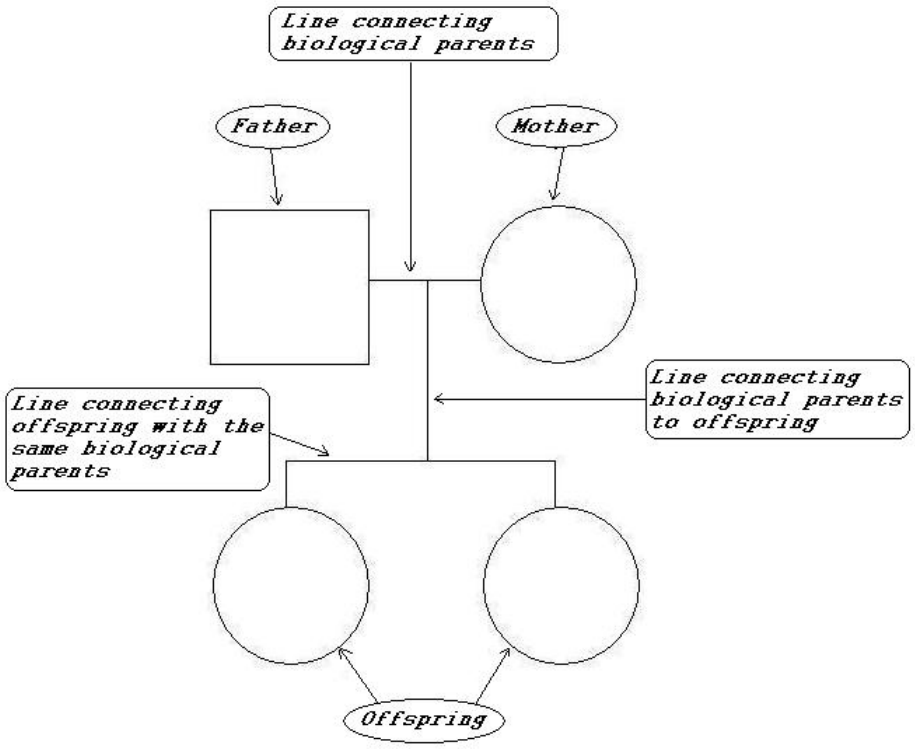

Figure 2.1: Form of graphs used, in this thesis, to represent family (or pedigree). Any additional information about genotype and trait, will be given under the corresponding circle or box. Note that boxes and circles are used to denote male and female sex, respectively. Since sex is not a factor in any of the analysis made in this thesis, we arbitrarily denote the offspring as females.

sib contributes to two cells in the inheritance vector. The first cell indicates which paternal (*1* or *2*) allele was transmitted and the second cell indicates which maternal (*3* or *4*) allele was transmitted. So an offspring contributes the vector;

$$\begin{cases} (1,1) & \text{if alleles } \textit{1} \text{ and } \textit{3} \text{ were transmitted,} \\ (0,1) & \text{if alleles } \textit{2} \text{ and } \textit{3} \text{ were transmitted,} \\ (1,0) & \text{if alleles } \textit{1} \text{ and } \textit{4} \text{ were transmitted,} \\ (0,0) & \text{if alleles } \textit{2} \text{ and } \textit{4} \text{ were transmitted.} \end{cases}$$

See figure (2.2) for an example with two sib pairs.

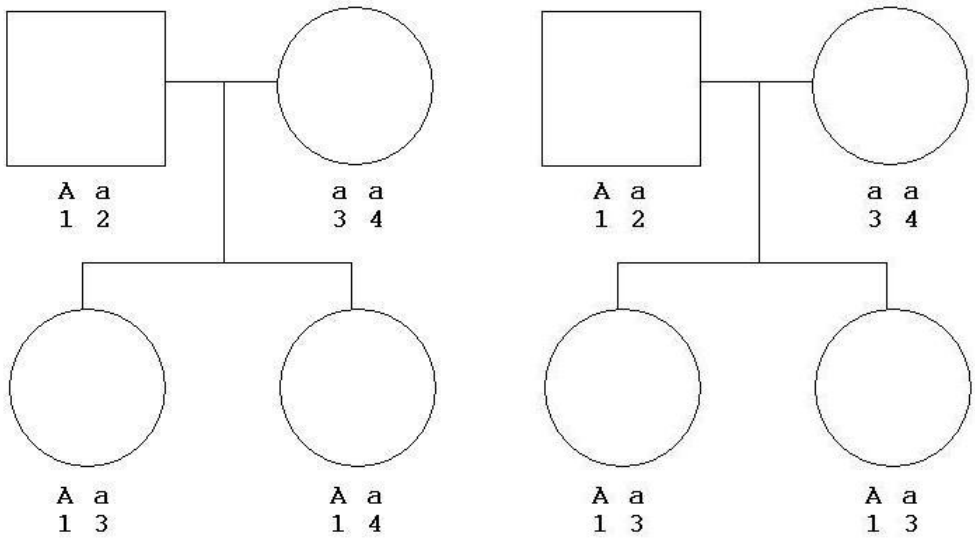

Figure 2.2: Example to illustrate the concepts of Identical-By-State (IBS), Identical-By-Descent (IBD) and inheritance vector. The sib pair to the left share allele a IBS, whilst the sib pair to the right share allele a IBD. Both sib pairs share allele A IBD. The inheritance vectors are $(1,1,1,0)$ and $(1,1,1,1)$ for the left and the right sib pair, respectively

2.2 Definitions and the Relatedness of LD and Linkage

The concepts of *Linkage* Disequilibrium (LD) and *Linkage*, which are key to the subject of this work, are closely related. We will, in this section, define and discuss the relatedness between the two concepts.

Linkage is defined as a concept of biology, more specifically a concept related to the transmission of genes from parents to offspring. It is defined to be the non-random co-inheritance of alleles at two loci. In terms of recombination, linkage between loci means that the recombination fraction θ is less than 0.5. For example, assume that the genotypes of parent 1, parent 2 and the offspring are AD/ad, ad/ad and aD/ad respectively (figure 2.3). In this example, we know that the offspring has to inherit D from parent 1 and from that we can deduct that there has occurred a recombination in parent 1, which means that the two loci in this example are not linked.

The purpose of linkage analysis is to use information about transmission and disease status to identify loci, either in themselves functional, or linked to functional loci.

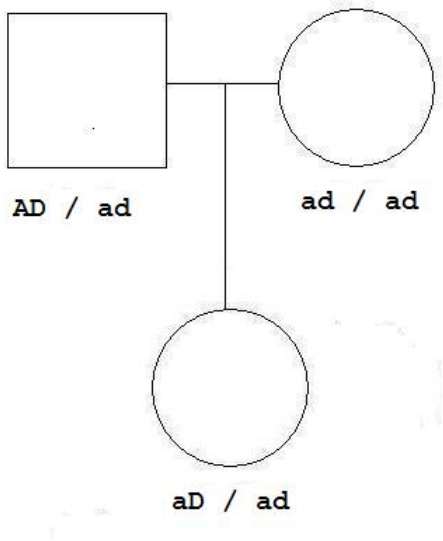

Figure 2.3: Example where two loci are unlinked

In contrast to linkage, LD is a population level concept. Two loci are said to be in LD if their alleles are statistically dependent. Let p_{AD}, p_A ($p_a = 1 - p_A$) and p_D ($p_d = 1 - p_D$) be population frequencies of haplotype AD, and alleles A and D, respectively. In mathematical terms, the marker locus and disease locus are said to be in LD if, $p_{AD} \neq p_A \cdot p_D$. A measure of LD is therefore

	A	a	Total
D	0.5	0	0.5
d	0	0.5	0.5
Total	0.5	0.5	1

Table 2.1: Example of perfect LD

usable for a test of independence. One measure of LD is the correlation coefficient,

$$r^2 = \frac{(p_{AD} - p_A p_D)^2}{\sqrt{p_A p_a p_D p_d}} \ . \tag{2.1}$$

For example, assume that the frequency of haplotypes AD, Ad, aD and ad are 0.5, 0, 0 and 0.5 respectively (see Table (2.1). The allele frequencies in this example are 0.5 for all allele. The LD measure r^2 will then be 1, which indicates perfect LD.

In Population-Based association studies, we test for LD between a set of marker loci and a putative disease locus by comparing the distribution of alleles across trait values. In Family-Based studies of association, our aim is to estimate and test LD, but not by comparing allele frequencies. There can be a number of reasons why an association between marker and trait is found:

1. The marker locus may be linked to a disease locus.

2. The marker locus may be a DS locus.

3. There may be population stratification or population admixture in the study population.

In association studies (1) or (2) are what we hope to find, (3) is *spurious association*.

2.3 A Review of Statistical Methods for Testing Association and Linkage in Family-Based Studies

The topic of the present thesis is testing and estimation of association in the presence of linkage. Many of the Family-Based tests developed today are, however, either for analysing association or linkage separately, or association and linkage jointly. Although these latter tests are not the topic of this thesis, we provide a short review of these tests with the purpose of providing a general background of statistical methods for analysing family-based association and linkage studies.

We focus predominantly on the scenario where two parents and an affected offspring have been successfully genotyped at one biallelic marker. We assume that the trait of interest is binary and that all offspring in the sample are affected. We assume that there exists an underlying disease locus with alleles D and d and that only offspring with genotype dd are affected, i.e. $\mathrm{P}(Y = 1|dd)$=1 and $\mathrm{P}(Y = 1|Dd)$=$\mathrm{P}(Y = 1|DD)$=0. Let n be the number of families which are in the study.

Now, we will suppose that parent 1 carries alleles *1* and *2* and transmits allele *1* to it's affected offspring, and that parent 2 carries alleles *3* and *4* and transmits *3* to it's affected offspring. The alleles of the affected offspring are *1* and *3*. Rubinstein, Walker, Carpenter, Carrier, Krassner, Falk & Ginsberg (1981) and Falk & Rubinstein (1987) propose using the two non-transmitted alleles, *2* and *4*, to define a pseudo control individual (Figure 2.4). We use T_{ij} to denote the number of pairs of cases and pseudo-controls which carry specific genotypes where: index i denotes the allele carried by the case, and index j denotes the allele carried by the pseudo-control. That is, i and j take values 1 for allele A (i.e. genotypes AA and Aa) or 2 for allele a (i.e. genotype aa), and $\sum_{i,j} T_{ij} = n$.

Rubinstein et al. (1981) suggest treating the two genotypes, of the affected offspring and the pseudo-control offspring, as being dependent (matched).

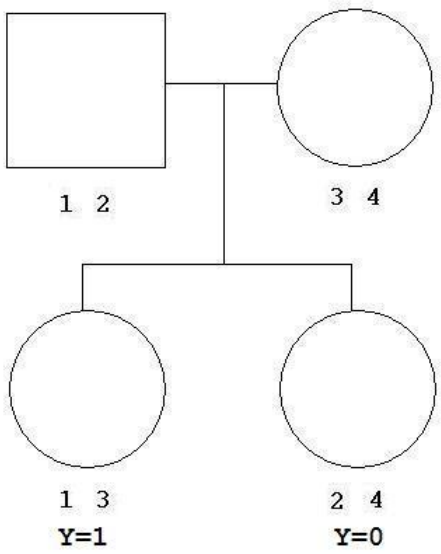

Figure 2.4: Family-trio with one affected child and a pseudo-control

	Control		
Case	A present	A absent	Total
A present	T_{11}	T_{12}	$T_{1\cdot}$
A absent	T_{21}	T_{22}	$T_{2\cdot}$
Total	$T_{\cdot 1}$	$T_{\cdot 2}$	n

Table 2.2: Matched cases (transmitted) and controls (non-transmitted) on a genotype level

The data can then be summarized as in table 2.2. The data can be analyzed using Mc Nemar's test (McNemar 1947). In this setting Rubinstein et al. (1981) renamed the test as the *Matched Genotype-based Relative Risk*

(MGRR) test. The test statistic is written as

$$MGRR = \frac{(T_{12} - T_{21})^2}{T_{12} + T_{21}} \tag{2.2}$$

Falk & Rubinstein (1987) suggest breaking down the matching in Table 2.2. Instead of looking at the pairs of genotypes, they separate the genotypes carried by cases and pseudo-controls. The un-matched data can be summarized as in Table 2.3.

	A present	A absent	Total
Case	$T_{1\cdot}$	$T_{2\cdot}$	n
Control	$T_{\cdot 1}$	$T_{\cdot 2}$	n
Total	$T_{1\cdot} + T_{\cdot 1}$	$T_{2\cdot} + T_{\cdot 2}$	2n

Table 2.3: Non-matched cases (transmitted) and controls (non-transmitted) on a genotype level

Falk & Rubinstein (1987) propose analyzing this data using the *Genotype-based Haplotype Relative Risk* (GHRR) test statistic,

$$GHRR = \frac{(T_{12} - T_{21})^2}{(2T_{11} + T_{12} + T_{21})(T_{12} + T_{21} + 2T_{22})/2n} \ . \tag{2.3}$$

The difference between the MGRR and the GHRR tests lies in the variance estimator of $(T_{12} - T_{21})^2$, i.e. in the denominator of (2.2) and (2.3). Both test statistics are tests of linkage and association, jointly.

An alternative approach, to the MGRR and GHRR tests, which analyse transmission of genotypes, is one which considers transmitted alleles individually, as haplotypes. Ott (1989) and Terwilliger (1992) propose a matched analysis, parallel to the MGRR, for transmitted (*1* or *3*) and non-transmitted (*2* or *4*) alleles (Figure 2.4). Instead of considering pairs of transmitted genotypes, it is possible to consider pairs of case (transmitted) and pseudo-control (non-transmitted) alleles. Let t_{ij} denote the number of pairs of alleles defined

by: index i denotes the allele transmitted to the case, and index j denotes the allele transmitted to the pseudo-control. That is, i and j takes possible values 1 for allele A and 2 for allele a, and $\sum_{ij} t_{ij} = 2n$. The data can be summarized as in Table 2.4.

	A	a	Total
Transmitted allele	$t_{1\cdot}$	$t_{2\cdot}$	n
Non-transmitted allele	$t_{\cdot 1}$	$t_{\cdot 2}$	n
Total	$t_{1\cdot} + t_{\cdot 1}$	$t_{2\cdot} + t_{\cdot 2}$	4n

Table 2.4: Non-matched cases (transmitted) and controls (non-transmitted) at the allele level

To analyse the data in table 2.4, Ott (1989) and Terwilliger (1992) proposed using the *Haplotype-based Haplotype Relative Risk* (HHRR) test statistic, a test similar in spirit to the test statistic (2.2),

$$HHRR = \frac{(t_{12} - t_{21})^2}{(2t_{11} + t_{21} + t_{12})(t_{12} + t_{21} + 2t_{22})/4n} \ . \tag{2.4}$$

The HHRR test was the precursor of the *Transmission Disequilibrium Test* (TDT). Instead of considering pairs of alleles carried by cases and pseudo-controls (HHRR), the data can be categorized into alleles transmitted among cases and pseudo-controls separately (Table 2.5).

	Non-transmitted alleles		
Transmitted allele	A	a	Total
A present	t_{11}	t_{12}	$t_{1\cdot}$
A absent	t_{21}	t_{22}	$t_{2\cdot}$
Total	$t_{\cdot 1}$	$t_{\cdot 2}$	2n

Table 2.5: Matched cases (transmitted) and controls (non-transmitted) on an allele level

The data in Table 2.5 can be analysed using the TDT test statistic

$$TDT = \frac{(t_{12} - t_{21})^2}{t_{12} + t_{21}} \ . \tag{2.5}$$

The TDT was first introduced as a test of linkage only (as were the tests MGRR, GHRR and HHRR). The test has, however, been shown to be a test for linkage and association simultaneously (see Sham (1998) and on page 20 herein). If all individuals come from a single pedigree with a common ancestor, then the TDT will be a test of linkage only and as generations go by the amount of LD will increase (Sham 1998). As with the MGRR and the GHRR test statistic, the difference between the MGRR and the GHRR test statistic lies in the different variances of the matched and unmatched test statistics ((2.4) and (2.5)] respectively).

Ott (1989) derived the transmission probabilities in Table 2.5 for the specific scenario we described in the first paragraph of Section 2.3. For simplicity of notation, we let $q = p_A$ and $p = p_d$. We let θ denote the recombination fraction and δ be a measure of LD, equal to $p_{Ad}p_{aD} - p_{ad}p_{AD}$, were p_{Ad}, p_{aD}, p_{ad} and p_{AD} are frequencies of haplotypes Ad, aD, ad and AD, respectively. We do not need to make any assumptions about distribution of the trait, conditional on genotype, since they cancel out in the derivations of the transmission probabilities (Ott 1989). Assume that we are interested in deriving the transmission probabilities for Parent 1 (Figure 2.4). The transmission probabilities are defined as the probability of the parent transmitting allele *1*, and not allele *2*, conditional on the offspring being affected. Since the affected child has to have disease genotype dd, the Parent 1 has to transmit a haplotype consisting of alleles *1* and d, and Parent 2 also has to transmit the d allele. If Parent 1 is heterozygote at both loci, then the transmission probabilities will include the recombination fraction, since the transmitted haplotype is not necessarily equal to the haplotype carried by the parent. The transmission probabilities as derived by Ott (1989) are presented in Table 2.6. The expected squared difference between t_{12} and t_{21} can be expressed in terms of the difference between the expected squared difference between

the proportions in the off-diagonals in table 2.6 (equation (2.6)),

$$\left\{\left(q+\frac{\delta}{p}\right)(1-q)-\frac{\theta\delta}{p}-\left(1-q-\frac{\delta}{p}\right)-\frac{\theta\delta}{p}\right\}^2=\left\{\frac{\delta}{p}(1-2\theta)\right\}^2 \quad . \qquad (2.6)$$

Equation (2.6) equals zero if and only if θ equals 0.5 or δ equals zero, i.e. only if there is no linkage or if there is no association. Alternatively, equation (2.6) differs from zero only if there is linkage and association. Therefore, it is clear that the TDT and HHRR test are tests of linkage and association jointly.

Transmitted	Non-transmitted A	Non-transmitted a	Total
A	$(q+\frac{\delta}{p})q$	$(q+\frac{\delta}{p})(1-q)-\theta\frac{\delta}{p}$	$q+(1-\theta)\frac{\delta}{p}$
a	$(1-q-\frac{\delta}{p})q+\theta\frac{\delta}{p}$	$(1-q-\frac{\delta}{p})(1-q)$	$1-q-(1-\theta)\frac{\delta}{p}$
Total	$q+\theta\frac{\delta}{p}$	$1-q-\theta\frac{\delta}{p}$	1

Table 2.6: Transmission probabilities by Ott

All of the tests which have been presented up until now, are based on test statistics which, under a null hypotheses of linkage but no association, are asymptotically distributed as a χ^2 variable with one degree of freedom.

The TDT has been used extensively in family based studies of association and linkage. Many extensions of the original TDT have been proposed. For example, extensions to allow for multiple markers, multiple alleles, multiple sibs, quantitative traits etc, have been described in the statistical literature in the last one and a half decade. We mention some of these below, with references.

- **Allowing for multiple sibs:** the Sib-TDT, or S-TDT is a test of association in the presence of linkage that allows for multiple sibs, but at least one has to be affected (Spielman, McGinnis & Ewens 1993).

- **Multiple alleles and multiple sibs with no missing parental genotypes:** TDT-SE (Spielman & Ewens 1996). As pointed out by Sham (1997), Schaid (1996), Lazzeroni & Lange (1998), the TDT-SE does not account for covariances and tends to be anti-conservative.

- **Multiple markers:** ETDT (Extended-TDT) Sham & Curtis (1995); Self, Longton, Kopecky & Liang (1991); Schaid (1996)

- **Missing parental genotypes:** Sham & Curtis (1995)

- **Unresolved haplotype phase and missing parental genotypes:** Clayton (1999)

- **Method based on haplotypes - similarity measures:** Clayton & Jones (1999)

- **Allowing for genotyping errors:** Chen & Deng (2001); Gordon, Heath, Liu & Ott (2001)

- **Continuous traits:** Allison (1997), Rabinowitz (1997), Abecasis, Cardon & Cookson (2000a) and Abecasis, Cookson & Cardon (2000b).

- **Gene-environment interaction:** Cordell, Barratt & Clayton (2004)

- **General pedigrees:** Martin, Monks, Warren & Kaplan (2000)

Other approaches

Thomson (1995) proposes comparing frequencies of transmitted alleles to affected offspring, to frequencies of non-transmitted alleles. The test is called the *Affected Family Based Controls* (AFBAC) (Thomson 1995).

Bickeboller & Clerget-Darpoux (1995) propose using a test statistic based on the marginal differences $d_i = t_{i+} - t_{+i}$, where $t_{i+} = \sum_{j=1}^{k} t_{ij}$ and $t_{i+} = \sum_{j=1}^{k} t_{ji}$.

Sham & Curtis (1995) propose a logistic regression model of the probability that a particular marker allele is transmitted by a heterozygous parent. Curtis (1997) looks at siblings as controls and developed a test based on conditional logistic regression model. These methods extend easily to include gene-environment interaction.

Boehnke & Langefeld (1998) propose non-parametric tests for discordant sib pairs, but their tests are not for more than two sibs. Horvath & Laird (1998) look at any number of sibs, using a different non-parametric test than Boehnke & Langefeld (1998).

Horvath & Laird (1998) describe a sign-test for discordant sib ships.

Clayton (1999), and Whittemore & Tu (2000) developed a likelihood-based theory for testing association.

Continuous traits

In previous sections, we have described methods for testing linkage and association to binary traits. The TDT was first developed for binary traits, but Allison (1997) and Rabinowitz (1997) have generalised the TDT for continuous traits.

3 Aims and Structure of Thesis

The aim of this thesis is to study statistical methods for analysing association in the presence of linkage in family-based studies. Note the difference between the null hypotheses of no association and **no linkage** in Section 2.3 and the null hypotheses of no association **in the presence of linkage** in the sequel of this work. Analysis of binary as well as continuous traits are relevant and needed. Most methods, however, have been developed for continuous traits, with the exception of Family-Based Association Test (FBAT) (Horvath & Laird (1998), Lake & Laird (2004)), which can handle both binary and continuous traits.

We develop a novel parametric model for binary traits, based on a *Gamma-Random-Effects* (GRE) model. The GRE model is similar in spirit to the VCM and to an recently proposed method for time to event data (Zhong & Li 2004). We develop a likelihood-based Score test for testing association in the presence of linkage which deals with all of the common sources of nuisance; linkage, population stratification and ascertainment. In Paper I an early version of the GRE is presented, where ascertainment is not dealt with and where a simplifying assumption is made in the likelihood calculations.

We also discuss two existing methods for analysing association in the presence of linkage for continuous traits: the *Variance Components Model* (VCM) (Fulker et al. 1999) and the Family-Based Association Test (FBAT) (Horvath & Laird (1998), Lake & Laird (2004)). The VCM uses a parametric framework for estimating and testing the degree of association, while simultaneously modelling linkage in the covariance structure. The FBAT is a non-parametric test statistic for association that allows for linkage by using an empirical variance estimate. In Paper II, we study the VCM properties under different types of mis-specification. We also, suggest simulations that could be carried out to further investigate interesting features of the VCM.

4 Tests of Association in the Presence of Linkage

It is reasonably common for researchers to use a two-stage strategy for genetic association mapping. In the first stage (i) large regions of the genome are sequenced and linkage analysis is used to identify regions of potential linkage. In the second stage (ii) marker density is increased in the identified regions and methods to analyse association are applied.

There are a number of advantages of the approach:

1. Linkage analysis is typically performed on a less dense map, thereby potentially decreasing the number of not associated loci investigated.

2. Testing association in the presence of linkage may increase the power for finding association (in comparison to "association only" analysis).

3. Search for association is possible without prior biological hypotheses.

When using this strategy multiple testing needs to be handled appropriately. Tests may not be independent, which makes correction of p-values complex.

One of the advantages with a family design is that it is straightforward to construct valid tests for association that are robust against population stratification.

There have been two lines of methodological development. One is the Variance Components Model (VCM) (Fulker et al. 1999) and the other is the Family Based Association Test (FBAT) (Rabinowitz 1997). In section 4.1 we present the parametric Variance Components Model (VCM), which has been described for multivariate normal traits. The VCM allows for a fixed effect of genotypes on the trait (association) and co-variability within families, which is a function of IBD-sharing (linkage). In section 4.2 we present the second line of methodological development, the non-parametric Family-Based

Association Test (FBAT) for association, which treats linkage as nuisance through a robust empirical within family covariance matrix.

In describing the VCM and the FBAT we assume n independent nuclear families, consisting of parents and their offspring. Both methods, however, extend to arbitrary pedigrees.

We let i denote family ($i = 1, 2, ..., n$) and j offspring within a family i ($j = 1, 2, ..., J_i$). We have marker genotype data on parents (denoted $\boldsymbol{g}_i$) and offspring (denoted $\boldsymbol{M}_i$) and trait information on offspring (denoted $\boldsymbol{Y}_i$). The trait, $\boldsymbol{Y}_i$, is either a vector of binary random variables (e.g. disease status yes/no) or continuous random variables (e.g. BMI, insulin level etc). Unless otherwise indicated, we use $\boldsymbol{\mu}_i$ to denote the expected value of $\boldsymbol{Y}_i$.

4.1 Variance Components Models (VCM)

The original form of the Variance Components Model (VCM) is the well known ANOVA (Fisher 1925) model. The VCM is aimed at data collected from studies of many small families and is therefore most appropriate if the underlying genetic effect is polygenic or oligogenic. We also assume that the probability of ascertainment does not depend on the trait values. Such studies are not appropriate for identifying low penetrant genes. The VCM has a long history of quantifying the importance of the genetic components for quantitative traits. They were used prior to the availability of high throughput marker genotype data to asses expected genotype similarity between related individuals, for example in twin-studies (Neal & Cardon 1992).

Almasy & Blangero (1998) developed Variance Components methodology to asses linkage between a marker and a quantitative trait. The name 'QTL analysis' (QTL = *Quantitative Trait Loci*), is used as a common name for all analysis on quantitative traits, including Variance Components analysis.

Fulker et al. (1999) were the first to propose the use of the VCM to analyse association and linkage jointly. The VCM that Fulker et al. (1999) propose, is in effect a Generalized Linear Mixed Model (GLMM) with an identity link function and normally distributed random effects. We write the model for the trait, $\boldsymbol{Y}_i$, as linear in terms of the fixed effect, which is a function

of the offspring genotypes, and the random effects. The random effects are partitioned into parts accounting for individual specific and family specific effects, as well as an effect accounting for the marker locus. The random effects are assumed to be normally distributed. As well as being able to test for association in the presence of linkage, it is possible to test whether the QTL is functional or merely in Linkage Disequilibrium (LD) with a trait locus, and whether there is population stratification (Fulker et al. 1999). Sham, Cherny, Purcell & Hewitt (2000) and Sham, Cherny & Abecasis (2002) developed this Variance Components model further to include dominance effects in the mean. These authors have also described a simple method to calculate approximative power of tests based on the model they described (Sham et al. 2000). We proceed by describing the random effects, as described in (Fulker et al. 1999), and the fixed effect in turn.

4.1.1 The Random Effects

Consider the case where three random effects affect the quantitative trait vector in a sibship:

- A non-shared random effect, $e_{ij} \sim N(0, \sigma_N^2)$. A random effect, unique for each sib j in a family i that describes the environmental component for that sib.

- A shared random effect, $s_{ij} \sim N(0, \sigma_S^2)$, equal for all sibs $j = 1, 2, ..., J_i$ in family i.

- A random effect for the additive QTL effect $a_{ij} \sim N(0, \sigma_A^2)$.

The co-variability between two sibs, j and j', in family i depend on genetic similarities of the QTL in terms of the expected proportion of alleles shared Identical-By-Descent (IBD), $\hat{\pi}_{jj'}$. The variance of the QTL random effect, σ_A^2, captures the proportion of the total variance attributable to the QTL. Following Sham et al. (2000, page 1617), we partition the polygenic effect into the shared and environmental random effects. We are not interested in

quantifying the polygenic random effect per se, and it will therefore make no difference if we assume it to be partitioned into σ_N^2 and σ_S^2.

The random effects are assumed to act independently on the offspring trait. The model can be written in terms of the three random effects, e_{ij}, s_{ij} and a_{ij}, and the expected value as a fixed effect, μ_{ij},

$$Y_{ij} = \mu_{ij} + a_{ij} + s_{ij} + e_{ij} \ . \tag{4.1}$$

Let $\mathbf{1}(j = j')$ be a variable indicating wether $j = j'$ or not. The covariance between Y_{ij} and $Y_{ij'}$ can be written in terms of the covariances between the random effects,

$$\begin{cases} \text{cov}(a_{ij}, a_{ij'}) = \hat{\pi}_{jj'}\sigma_A^2 \\ \\ \text{cov}(e_{ij}, e_{ij'}) = \mathbf{1}(j = j')\sigma_N^2 \\ \\ \text{cov}(s_{ij}, s_{ij'}) = \sigma_S^2 \end{cases} . \tag{4.2}$$

The covariance between two sibs j and j' in family i is thus

$$\begin{cases} \text{cov}(Y_{ij}, Y_{ij'}) = \hat{\pi}_{jj'}\sigma_A^2 + \sigma_S^2 \ , \text{ if } j \neq j' \\ \\ \text{cov}(Y_{ij}, Y_{ij'}) = \sigma_A^2 + \sigma_N^2 + \sigma_S^2 \ , \text{ if } j = j' \end{cases} \tag{4.3}$$

Linkage is accounted for by letting the sib trait correlation depend on IBD allele sharing. Let $\mathbf{\Sigma}_i$ denote the covariance matrix for family i. For a sib pair i, the covariance matrix is written

$$\mathbf{\Sigma}_i = \begin{pmatrix} \sigma_N^2 + \sigma_S^2 + \sigma_A^2 & \sigma_S^2 + \hat{\pi}\sigma_A^2 \\ \sigma_S^2 + \hat{\pi}\sigma_A^2 & \sigma_N^2 + \sigma_S^2 + \sigma_A^2 \end{pmatrix} .$$

Note that, in this model for the covariance, there is an implicit assumption that the marker is in full linkage with the true DS locus, i.e. no recombination

has occurred between the two loci. Thus σ_A^2 is the variance component accounting for the variance at the QTL, but in constructing the covariance matrix, we estimate $\hat{\pi}_{jj'}$ at the marker. There are approaches that avoid this assumption, by conditioning on the alleles at the DS locus and modelling the biological parameters directly (Hössjer 2005).

4.1.2 The Fixed Effects

Consider a biallelic QTL with alleles A and a. We assume a co-dominant model, implying that the mean effects of genotype AA, Aa and aa are $-a$, 0 and a, respectively. We write the mean of Y_{ij}, μ_{ij}, in terms of an overall mean μ and the allele effect a,

$$\mu_{ij} = \mu + aX_{ij} \ . \tag{4.4}$$

with X_{ij} 1,0 or -1 for genotypes AA, Aa and aa, respectively. Model (4.4) can include other covariates. We reparameterise the mean by splitting X_{ij} into two components, $\boldsymbol{X}_{bi}$ and $\boldsymbol{X}_{wi}$:

$$\begin{cases} \boldsymbol{X}_i = [X_{ij}]_{j=1,2,\ldots,J_i} \\ \overline{X}_i = \sum_{j=1}^{J_i} X_{ij} \\ \boldsymbol{X}_{bi} = [\overline{X}_i]_{j=1,2,\ldots,J_i} \\ \boldsymbol{X}_{wi} = \boldsymbol{X}_i - \boldsymbol{X}_{bi} \end{cases}$$

Note that $\boldsymbol{X}_{bi} + \boldsymbol{X}_{wi} = \boldsymbol{X}_i$. With this notation, $a \cdot \boldsymbol{X}_{bi}$ is a vector of the mean allele effect for family i, and $a \cdot \boldsymbol{X}_{wi}$ is a vector capturing the difference from $a \cdot \boldsymbol{X}_{bi}$ for each offspring j. Since population stratification only affect the mean allele effect in family i, Fulker et al. (1999) propose a mean model where $a\boldsymbol{X}_{ij}$ is split into $a_b X_{bi}$ and $a_w X_{wi}$, where the parameter a_w is protected against population stratification.

To illustrate, the elements of $\boldsymbol{X}_i$, $\boldsymbol{X}_{bi}$ and $\boldsymbol{X}_{wi}$ for a sib pair are given in

Table 4.1. As an example, consider a sib pair with genotypes AA and Aa, for sib 1 and 2 respectively,

$$\begin{aligned} \mu_{i1} &= \mu + \tfrac{1}{2}a_b + \tfrac{1}{2}a_w \\ \mu_{i2} &= \mu + \tfrac{1}{2}a_b - \tfrac{1}{2}a_w \end{aligned} \quad .$$

If there is no population stratification, then $a_b = a_w = a$ and the mean genotype effects is a for sib 1 and 0 for sib 2, consistent with our additive allele mean model (4.4).

For a general family i, we write,

$$\boldsymbol{\mu}_i = \mu + a_b \boldsymbol{X}_{bi} + a_w \boldsymbol{X}_{wi} = \begin{pmatrix} \mathbf{1} & \boldsymbol{X}_{bi} & \boldsymbol{X}_{wi} \end{pmatrix} \begin{pmatrix} \mu \\ a_b \\ a_w \end{pmatrix} = \mathbf{X}_i' \beta \ . \qquad (4.5)$$

Genotype				
Sib 1	Sib 2	$\boldsymbol{X}^T$	$\boldsymbol{X}_b^T$	$\boldsymbol{X}_w^T$
AA	AA	(1,1)	(1,1)	(0,0)
AA	Aa	(1,0)	($\frac{1}{2}$,$\frac{1}{2}$)	($\frac{1}{2}$,-$\frac{1}{2}$)
AA	aa	(1,-1)	(0,0)	(1,-1)
Aa	AA	(0,1)	($\frac{1}{2}$,$\frac{1}{2}$)	(-$\frac{1}{2}$,-$\frac{1}{2}$)
Aa	Aa	(0,0)	(0,0)	(0,0)
Aa	aa	(0,-1)	(-$\frac{1}{2}$,-$\frac{1}{2}$)	($\frac{1}{2}$,-$\frac{1}{2}$)
aa	AA	(-1,1)	(0,0)	(-1,1)
aa	Aa	(-1,0)	(-$\frac{1}{2}$,-$\frac{1}{2}$)	(-$\frac{1}{2}$,$\frac{1}{2}$)
aa	aa	(-1,-1)	(-1,-1)	(0,0)

Table 4.1: The elements of $\boldsymbol{X}$, $\boldsymbol{X}_b$ and $\boldsymbol{X}_w$ for all possible sib pairs. Here $\boldsymbol{X}$ is the vector of X_j's for the two sibs. The elements of $\boldsymbol{X}_b$ are the mean of the elements in $\boldsymbol{X}$ and $\boldsymbol{X}_w = \boldsymbol{X} - \boldsymbol{X}_b$. Note that all indexes i have been omitted and that T denotes transposition of the vectors.

4.1.3 Likelihood Inference

We assume that the probability of ascertainment does not depend on the trait values and use a prospective likelihood. The trait vector of the offspring in family i, $\boldsymbol{Y}_i$, is multivariate normal with mean $\boldsymbol{\mu}_i$ (4.5), variance $\boldsymbol{\Sigma}_i$ (4.3) and likelihood

$$\prod_{i=1}^{n}(2\pi)^{-J_i/2}|\boldsymbol{\Sigma}_i|^{-1/2}\exp\left(-\frac{1}{2}(\boldsymbol{Y}_i-\boldsymbol{\mu}_i)'|\Sigma_i|^{-1}(\boldsymbol{Y}_i-\boldsymbol{\mu}_i)\right)\ . \tag{4.6}$$

The null and alternative hypotheses in (1.1) for testing association in the presence of linkage are formulated as

$$\begin{aligned} H_0:&\ \text{Linkage and no association, } \sigma_A^2>0 \text{ and } a_w=0\ .\\ H_1:&\ \text{Linkage and association, } \sigma_A^2>0 \text{ and } a_w\neq 0\ . \end{aligned} \tag{4.7}$$

Let ψ denote the vector of parameters, $(\mu, a_b, a_w, \sigma_N^2, \sigma_S^2, \sigma_A^2)$. Based on the likelihood (4.6), the Score function, $S(\psi)$, and the Fisher Information matrix, $I(\psi)$, we can construct a Score test or a Likelihood Ratio test. Both estimation and testing are then possible.

We use the Newton-Raphson algorithm, where updated parameters estimates at iteration $i+1$ ($\hat{\psi}^{i+1}$) are given by,

$$\hat{\psi}^{i+1}=\hat{\psi}^{i}+\left(I(\psi^i)\right)^{-1}S(\psi^i)\ .$$

The iteration stops when $|\hat{\psi}^{i+1}-\hat{\psi}^{i+1}|$ is smaller than some predefined small δ. A Likelihood Ratio test of the null hypotheses is given by,

$$\frac{L(\hat{\mu},\hat{a}_b,a_w=0,\hat{\sigma}_A^2,\hat{\sigma}_N^2,\hat{\sigma}_S^2)}{L(\hat{\psi})}\ ,$$

where $\hat{\mu}$ $\hat{a}_b$ $\hat{\sigma}_A^2$, $\hat{\sigma}_N^2$, $\hat{\sigma}_S^2$, and $\hat{\psi}$ are maximum likelihood estimates of μ a_b

σ_A^2, σ_N^2, σ_S^2, and $\hat{\psi}$, respectively. The LRT is χ^2 distributed with 1 degree of freedom.

4.1.4 Simulation Study of the Variance Components Model

We study the behavior of the estimates of a_b and a_w (Section 4.1) under different scenarios. Continuous traits in nuclear families with two sibs were simulated under different values of recombination fractions and LD-structure; see Paper I in Appendix C.

4.2 Family-Based Association Tests (FBAT)

Lake, Blacker & Laird (2000) introduced a statistic for testing association in the presence of linkage. This represents an extension of the FBAT statistic (Rabinowitz & Laird 1999). The FBAT is a non-parametric test statistic that uses the statistical concept of sufficiency to deal with missing parental genotypes.

We first describe the FBAT statistic (Rabinowitz & Laird 1999). Let $T(Y_{ij})$ be a function of the trait, Y_{ij}. We let $X(M_{ij})$ denote some score (possibly a vector) of the offspring genotype, M_{ij}. For example, $X(M_{ij})$ may be equal the number of A alleles in genotype M_{ij}. For simplicity of notation, we will write T_{ij} and X_{ij} in place of $T(Y_{ij})$ and $X(M_{ij})$, respectively. Rabinowitz & Laird (1999) propose the following score statistic for testing association,

$$S = \sum_i S_i = \sum_{i=1}^{n} \sum_{j=1}^{J_i} T_{ij} X_{ij} \ . \tag{4.8}$$

The product in the sum, $T_{ij}X_{ij}$, can be viewed as an interaction term between offspring trait and offspring genotype, and the score is a summation of these terms for all individuals in the n families.

Rabinowitz & Laird (1999) propose calculating the expected value of the family score, S_i, by conditioning on the sufficient statistic of the parental

genotypes, $\boldsymbol{g}_i$, and trait ($\boldsymbol{Y}_i$). By conditioning on trait and the sufficient statistic for the parental genotype, they design a valid test for association, regardless of genetic model and population admixture or stratification. Rabinowitz & Laird (1999) present an algorithm for finding the sufficient statistic of the parental genotypes, and calculating the conditional probabilities of the possible sibship genotype vector, given the sufficient statistic for parental genotype. The algorithm can be divided into five steps

- Step 1: Find all phased mating types, compatible with the observed marker data: $\boldsymbol{g}_1, ..., \boldsymbol{g}_k$.

- Step 2a: Find the set of offspring genotypes consistent with phased mating type $\boldsymbol{g}_l$ ($l = 1, ..., k$): $\gamma_1, ..., \gamma_k$. Let γ be the intersection $\gamma_1 \cap ... \cap \gamma_k$, i.e. the minimal set of offspring genotypes consistent with all mating types.

- Step 2b: From the genotypes in γ, construct all possible sets of offspring genotypes (of the same size as the observed sibship). Choose those that give the exact same set of phased mating types as the observed sibship genotypes (as derived in Step 1): $m_1, ..., m_h$.

- Step 3: Compute the probability of offspring genotype m_f ($f = 1, ..., h$), conditional on parental mating type. This will give a $h \times k$ matrix.

- Step 4: Consider only offspring genotypes where $P(m_f|g)$ ($f = 1, .., h$) is proportional to $P(m_1|g)$ (where m_1 is the observed vector of offspring genotype), for all mating types g: $m_1^*, ..., m_{h'}^*$ ($\subset m_1, ..., m_h$).

- Step 5: Compute the conditional probabilities for each vector $m_1^*, ..., m_{h'}^*$, given g: $P_{\text{cond}}(m_r^*)$ ($r = 1, ..., h'$).

We let $\phi = (\xi(\boldsymbol{g}_i), \boldsymbol{Y}_i)$, where $\xi(\boldsymbol{g}_i)$ is the sufficient statistic for the parental genotypes $\boldsymbol{g}_i$. From the RL-algorithm, we can calculate

$$E(X_{ij}|\phi) = \sum_{r=1}^{h'} X(m_{rj}) P_{\text{cond}}(m^*_{rj})$$

The expected value under the null hypotheses of S_i follows straightforwardly, $E(S_i|\phi) = \sum_{j=1}^{J_i} T_{ij} E(X_{ij}|\phi)$.

Lake et al. (2000) show that the FBAT statistic can be used for testing for association in the presence of linkage. The expected values under the null hypotheses of the test statistics are the same, and

$$S_L = \sum_{i=1}^{n} (S_i - E(S_i|\phi)) \tag{4.9}$$

is a valid test statistic for testing association in the presence of linkage (Lake et al, 2000). However, the covariance of the statistic will not be the same, so instead Lake et al. (2000) propose using a robust covariance estimator (White (1980), Liang & Zeger (1986)),

$$\Sigma_L = \sum_{i=1}^{n} (S_i - E(S_i|\phi))(S_i - E(S_i|\phi))' \ . \tag{4.10}$$

The robust variance estimator accounts for the co-variability among siblings, thereby adjusting for linkage. To test for association in the presence of linkage, we use the expected value S_L and the covariance Σ_L to construct a Z statistic or a χ^2 statistic, both assuming approximate normality. Since the expected value of S_L is zero, the Z statistic takes the form

$$Z_L = \Sigma_L^{-1} Z_L \tag{4.11}$$

The Lake extension of FBAT is valid under any genetic model and population stratification / admixture (Rabinowitz & Laird 1999). It also deals with missing marker data, through the conditioning on $\xi(\boldsymbol{g}_i)$.

5 Generalized Linear Mixed Models (GLMM)

Let $\boldsymbol{Y}_i$ $(i = 1...n)$ be a vector of random variables taking observed value $\boldsymbol{y}_i$ $(i = 1...n)$. Let $\boldsymbol{X}_i$ be a $n \times k$ matrix of predictors. The Generalized Linear Mixed Model (GLMM) ((Liang & Zeger 1986)) can be seen as an extension of the Generalized Linear Model ((McCullagh & Nelder 1989)) in that it allows for clusters of dependencies among the $\boldsymbol{Y}_i$. The GLMM can be defined, similarly to the GLM, in steps:

1. Let $\boldsymbol{\mu}_i$ $(i = 1, ..., n)$ be the conditional mean of the response $\boldsymbol{Y}_i$ (for individual i), $E(\boldsymbol{Y}_i|\boldsymbol{X}_i, \boldsymbol{\beta}, \mathbf{b})$, and let $h(\cdot)$ be a twice differentiable, continuous function. The conditional mean can then be expressed as

$$\boldsymbol{h}(\boldsymbol{\mu}_i) = \boldsymbol{\gamma}_i = \mathbf{X_i}\boldsymbol{\beta} + \boldsymbol{Z}_i\mathbf{b} \ , \tag{5.1}$$

   where $\boldsymbol{\beta}$ are fixed effects, common for all individuals $i = 1, ..., n$, and $\mathbf{b}$ are random effects. $\boldsymbol{X}$ and $\boldsymbol{Z}$ are design matrices for the fixed effects and random effects respectively.

2. Assume that conditional on the random effects $\mathbf{b}$, $\boldsymbol{Y}_i$ is a response following some random distribution $P(\boldsymbol{Y}_i = \boldsymbol{y}_i|\boldsymbol{b})$ with mean $\boldsymbol{\mu}_i$ and variance $\sigma^2\mathbf{I}$.

3. Assume that the random effects $\mathbf{b}$ follow some random distribution $P(\mathbf{b} = b)$ with mean zero and variance D, and with parameter $\boldsymbol{\theta}$.

The likelihood can be formulated, using Bayes formula in terms of the conditional distribution of $\boldsymbol{Y}_i$ and the distribution of the random parameter $\boldsymbol{b}$

$$p(\boldsymbol{Y}_i|\boldsymbol{\beta}, \boldsymbol{\theta}) = \int p(\boldsymbol{Y}, \boldsymbol{b}|\boldsymbol{\beta}, \boldsymbol{\theta})d\boldsymbol{b} = \int p(\boldsymbol{Y}_i|\boldsymbol{\beta}, \boldsymbol{b})p(\boldsymbol{b}|\boldsymbol{\theta})d\boldsymbol{b} \ . \tag{5.2}$$

A problem with [5.2] is that, although conceptually easy, it will typically not take a closed form when integrating over the random effect. Some special

cases exist for some specific combination of distributions of $\boldsymbol{Y}_i|\boldsymbol{b}$ and $\boldsymbol{b}$. We maximize 5.2 with respect to $\boldsymbol{\beta}$. The value of $\boldsymbol{\beta}$ which maximizes 5.2 is called the Maximum Likelihood (ML) estimate of $\boldsymbol{\beta}$, and is denoted $\hat{\boldsymbol{\beta}}$.

Assume that the distribution of $\boldsymbol{Y}_i$, conditional on the random effect, comes from the exponential family. The exponential family is typically expressed as

$$p(\boldsymbol{Y}_i|\boldsymbol{\gamma}_i, \boldsymbol{b}) = c(\boldsymbol{Y}_i, \psi) \exp\left(\frac{S(\boldsymbol{Y}_i)\boldsymbol{\gamma}_i - a(\boldsymbol{\gamma}_i)}{\psi}\right) , \tag{5.3}$$

where $c(\cdot)$ and $a(\cdot)$ are some functions, $S(\boldsymbol{Y}_i)$ is the sufficient statistic for $\boldsymbol{Y}_i$ and ψ is a dispersion parameter. The the conditional mean can be expressed in terms of the derivative of the canonical term, $a(\gamma_i)$,

$$E(\boldsymbol{Y}_i|\boldsymbol{\gamma}_i, \mathbf{b}) = \boldsymbol{\mu}_i = a'(\boldsymbol{\gamma}_i) \tag{5.4}$$

and the variance can be expressed in terms of the second derivative of a

$$Var(\boldsymbol{Y}_i|\boldsymbol{\gamma}_i, \mathbf{b}) = v(\boldsymbol{\mu}_i) = a"(\boldsymbol{\gamma}_i) \tag{5.5}$$

The link function h is the inverse function of a' [5.4].

5.1 Maximum-Likelihood Estimation

The likelihood in Equation [5.2] is generally difficult to optimize, the main difficulty being the integration over $\mathbf{b}$. Closed form solutions for the conditional likelihood may exist. Except in some special cases, maximum-likelihood estimates of the fixed parameters are therefore often most easily found using an iterative algorithm. Two of the most commonly used algorithms are the Expectation-Maximization (EM) algorithm and the Iterative re-Weighted Least Squares (IWLS) algorithm. In Section 6 we describe a special case where a closed form solution for the conditional likelihood can be found for

a binary trait. Another example of the GLMM, for continuous traits, is the Variance Components Model described in Section 4.1.

5.2 A Variance Components Model as a GLMM

Zhong & Li (2004) propose a variance components model for survival data, based on log-gamma random effects for modelling association in the presence of linkage. They formulate IBD sharing in terms of random inheritance vectors on the linear predictor and they propose a GLMM where the log(-log) of the probability of trait, given the marker data and the inheritance vector, is linearly dependent on log-gamma random effects and a fixed genotype effect. On the basis of this model they derive the corresponding joint survival function of age of onset for the sibs within a sibship. In Section 6 we develop a related approach, also based on inheritance vectors, but for binary traits.

6 Gamma Random Effects Model for Binary Traits

6.1 Notation

Assume the following notation:

- Y_{ij} = phenotype of offspring j in family i.
- X_{ij} = genotype score for offspring j in family i.
- M_{ij} = marker genotypes for offspring j in family i.
- $\boldsymbol{g}_i$ = marker genotypes for parents in family i.
- v_{ij} = inheritance vector for offspring j in family i.
- β = association parameter.

When denoting a vector for a family, the index j will be omitted. For example, the vector of Y_{ij}'s for family i will be denoted $\boldsymbol{Y}_i$.

6.2 The Model

Our aim is to evaluate the probability of trait, conditional on the observed marker data on parents and offspring. However, to capture both association and linkage, we need to write our model in terms of the offspring marker data and the inheritance vector. We can then write

$$P(\boldsymbol{Y}_i = \boldsymbol{y}_i|\boldsymbol{M}_i, \boldsymbol{g}_i) = \sum_v P(\boldsymbol{Y}_i = \boldsymbol{y}_i|\boldsymbol{M}_i, v)P(v|\boldsymbol{M}_i, \boldsymbol{g}_i) \ , \qquad (6.1)$$

were the summation is over all possible inheritance vectors, v, given the observed marker data. We continue by describing a model for $P(\boldsymbol{Y}_i = \boldsymbol{y}_i|\boldsymbol{M}_i, \boldsymbol{v}_i)$, that is for the probability of an offspring's trait, conditional on the alleles which it carries and from whom the alleles were transmitted. We use $\epsilon_{p_{ij}}$ and $\epsilon_{m_{ij}}$ to denote the effect of the alleles transmitted to offspring j in family i, where p_{ij} denotes the paternal allele and m_{ij} denotes the maternal allele. The paternal alleles, p_{ij}, can take possible values *1* and *2* and the maternal alleles, m_{ij}, can take possible values *3* and *4*. Thus, offspring can carry paternal and maternal alleles *1* and *3*, *1* and *4*, *2* and *3* or *2* and *4*. We let a_{jk} be an indicator variable representing whether offspring j has inherited allele k or not. Note that both the transmitted alleles, p_{ij} and m_{ij}, and the indicator variables a_{jk} are functions of the inheritance vector v_{ij}

To illustrate the use of the notation we have introduced, we consider an example, based on a family with two offspring, depicted in Figure 6.1. At a marker of interest, the father carries alleles 1 = A and 2 = a, and the mother carries alleles 3 = a and 4 = a. Assume that the mode of transmission is known, and that alleles 1 and 3 were transmitted to offspring $j = 1$*, whilst alleles 1 and 4 were transmitted to offspring* $j = 2$*. Thus, the inheritance vector* $\boldsymbol{v}_i$ *takes value* $(1, 1, 1, 0)$*, whilst the vectors of indicator variables* $a_1 = (a_{11}, a_{12}, a_{13}, a_{14})$ *and* $a_2 = (a_{21}, a_{22}, a_{23}, a_{24})$ *take values* $(1, 0, 1, 0)$ *and* $(1, 0, 0, 1)$*, respectively. The transmission effects are* ϵ_1 *and* ϵ_3 *for offspring* 1*, and* ϵ_1 *and* ϵ_4 *for offspring* 2*. The marker data is* $\boldsymbol{g}_i = (Aa, aa)$ *for the parents and* $\boldsymbol{M}_i = (Aa, Aa)$ *for the offspring. We further assume that offspring* 1 *is affected,* $Y_1 = 1$*, and that offspring* 2 *is unaffected,* $Y_2 = 0$*.* ■

We let $X(M_{ij})$ denote the score of genotype M_{ij}. For example, $X(M_{ij})$ may be defined as the number of A alleles in genotype M_{ij}. To simplify notation, we will write X_{ij} in place of $X(M_{ij})$. Let $\boldsymbol{\beta}$ denote the association parameter vector, describing the effect of the genotype score X_{ij}. We recall that $\epsilon_{p_{ij}}$ and $\epsilon_{m_{ij}}$ are functions of the inheritance vector and write

$$p_{ij} = P_\beta(Y_{ij} = 1|\epsilon_{m_{ij}}, \epsilon_{p_{ij}}, M_{ij}) =$$

$$= \exp\left(-\epsilon_{m_{ij}} \exp\left(X_{ij}\beta\right)\right) \exp\left(-\epsilon_{p_{ij}} \exp\left(X_{ij}\beta\right)\right) \ , \qquad (6.2)$$

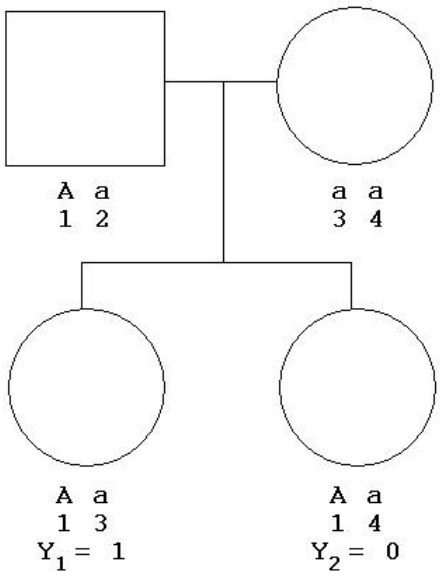

Figure 6.1: Example of a nuclear pedigree, with two sibs (of which one is affected). The mode of inheritance is assumed known.

or equivalently,

$$\log(-\log p_{ij}) = \log(\epsilon_{m_{ij}} + \epsilon_{p_{ij}}) + X_{ij}\beta \ . \tag{6.3}$$

The transmission effects $\epsilon_{m_{ij}}$ $\epsilon_{p_{ij}}$ act multiplicatively on the offspring trait probability, and the effect of each transmitted allele is multiplied by a term involving the association parameter vector $\boldsymbol{\beta}$. Following Zhong & Li (2004), we assume that the transmission effects $\epsilon_{m_{ij}}$ and $\epsilon_{p_{ij}}$ are independent and gamma distributed with scale $\alpha/2$ and shape λ. The density function of $\epsilon_{m_{ij}}$ and $\epsilon_{p_{ij}}$ is denoted $f(\epsilon)$. The association parameter $\boldsymbol{\beta}$ is assumed to be fixed for all individuals in the population. Let ψ denote the vector of parameters (β, α, λ). Thus, the model in Equation (6.3) is a Generalized Linear Mixed Model, with a log(-log) link, and a linear predictor composed of a log-gamma ($= \log(\epsilon_{m_{ij}} + \epsilon_{p_{ij}})$) distributed random effect and fixed effects

$X_{ij}\beta$. The model (6.3) can easily incorporate and additional random effect for shared familial effects (which will also account for contributions from unlinked genetic loci) as is done by Sham et al. (2002).

If we assume that the offspring in family i are independent, conditional on the random effects and $\boldsymbol{M}_i$, then,

$$P_\beta(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, ..., Y_{iJ_i} = y_{iJ_i} | \boldsymbol{M}_i, \boldsymbol{\epsilon}_{m_i}, \boldsymbol{\epsilon}_{p_i}) =$$

$$= \prod_{j=1}^{J_i} P(Y_{ij} = y_{ij} | M_{ij}, \epsilon_{m_{ij}}, \epsilon_{p_{ij}}) \ . \tag{6.4}$$

For the example depicted in Figure 6.1, if we assume that X_{ij} is the count of the number of A alleles, then $X_{i1} = X_{i2} = 1$. In this case,

$$P(Y_{i1} = 1, Y_{i2} = 0 | \epsilon_1, \epsilon_3, \epsilon_4, \boldsymbol{M}_i) =$$

$$= \exp\left(-(\epsilon_1 + \epsilon_3)\exp\left(\beta\right)\right) - \exp\left(-(2\epsilon_1 + \epsilon_3 + \epsilon_4)\exp\left(\beta\right)\right) \quad \blacksquare$$

Let $\pi_{\boldsymbol{y}_i}$ denote the probability $P_\psi(\boldsymbol{Y}_i = \boldsymbol{y}_i | \boldsymbol{M}_i, \boldsymbol{v}_i)$, which appears on the right hand side of (6.1). To obtain $\pi_{\boldsymbol{y}_i}$ we integrate (6.4) over the random effects $\epsilon_1, \epsilon_2, \epsilon_3$ and ϵ_4,

$$\pi_{\boldsymbol{y}_i} = \int_\epsilon \prod_{j=1}^{J_i} P(Y_{ij} = y_{ij} | \epsilon_{m_{ij}}, \epsilon_{p_{ij}}) f(\epsilon) d\epsilon \ . \tag{6.5}$$

In general, evaluation of (6.5) is cumbersome. Conaway (1990) shows that under particular assumptions for $P(\boldsymbol{Y}_i = \boldsymbol{y}_i | \boldsymbol{M}_i, \epsilon_{p_{ij}}, \epsilon_{m_{ij}})$ and $f(\epsilon)$ (6.5) is tractable. With our choice of model for $P(\boldsymbol{Y}_i = \boldsymbol{y}_i | \boldsymbol{M}_i, \epsilon_{p_{ij}}, \epsilon_{m_{ij}})$ and $f(\epsilon)$ $\pi_{\boldsymbol{y}_i}$ may be evaluated in terms of marginal probabilities and the joint probability $P(Y_{i1} = 1, ..., Y_{iJ_i} = 1 | \boldsymbol{M}_i, \boldsymbol{v}_i)$. Following Conaway (1990), we let T be a subset of the offspring $(1, 2, ..., J_i)$ in family i. We use π^*_T to denote the

probability of $\{Y_{ij} = 1\}_{j \in T}$. To obtain π_T^* we integrate $P_\beta(Y_{ij} = 1, \forall j \in T | \boldsymbol{M}_i, \boldsymbol{\epsilon}_{m_i}, \boldsymbol{\epsilon}_{p_i})$ over ϵ_1, ϵ_2, ϵ_3 and ϵ_4,

$$= \int_{\epsilon_4} \int_{\epsilon_3} \int_{\epsilon_2} \int_{\epsilon_1} \prod_{j \in T} P(Y_{ij} = 1 | M_{ij}, \epsilon_{m_{ij}}, \epsilon_{p_{ij}}) \cdot$$

$$\cdot f(\epsilon_1) f(\epsilon_2) f(\epsilon_3) f(\epsilon_4) d\epsilon_1 d\epsilon_2 d\epsilon_3 d\epsilon_4 =$$

$$= E\left(\exp\left\{ - \sum_{j \in T} \exp(X_j \beta) \cdot (\epsilon_{m_j} + \epsilon_{p_j}) \right\} \right) . \tag{6.6}$$

For simplicity of exposition, let $-\exp(X_j\beta) = c_j$. We can then write (6.6) as,

$$E\left(\exp\left\{ \sum_{k=1}^{4} \epsilon_k \sum_{j \in T} c_j \cdot a_{jk} \right\} \right) = \prod_{k=1}^{4} E\left(\exp\left\{ \epsilon_k \sum_{j \in T} c_j \cdot a_{jk} \right\} \right) . \tag{6.7}$$

Equation (6.7) is the product of four gamma distributed mgf's. Hence,

$$\pi_T^* = \prod_{k=1}^{4} \left(\frac{\lambda}{\lambda + h_{kT}(\beta)} \right)^{\alpha/2} , \tag{6.8}$$

where $h_{kT}(\beta) = \sum_{j \in T} \exp(X_{ij}\beta) a_{jk}$. Let Ψ denote all possible offspring subsets T, including the empty set $\{\emptyset\}$, ordered as $\{\{\emptyset\}, \{1\}, \{2\}, \{1,2\}, \{3\}, ...,, \{1, 2, ..., J_i\}\}$. Let $\boldsymbol{\pi}^*$ denote the vector of values $\{\pi_T^*\}_{T \in \Psi}$. The vector $\boldsymbol{\pi}^*$ contains the marginal probabilities and the probability of all offspring being affected. Let $\boldsymbol{\pi}$ denote a vector containing all possible outcomes of $\pi_{(y_1 y_2 ... y_{J_i})}$, ordered so that $\pi_{(11...1)}$ comes first and the consecutive items are such that the left subscript changes fastest.

We return to the example depicted in Figure 6.1. In this case,

$$\boldsymbol{\pi} = \begin{pmatrix} P\Psi(Y_1 = 1, Y_2 = 1|\boldsymbol{M}_i, \boldsymbol{v}_i) \\ P\Psi(Y_1 = 0, Y_2 = 1|\boldsymbol{M}_i, \boldsymbol{v}_i) \\ P\Psi(Y_1 = 1, Y_2 = 0|\boldsymbol{M}_i, \boldsymbol{v}_i) \\ P\Psi(Y_1 = 0, Y_2 = 0|\boldsymbol{M}_i, \boldsymbol{v}_i) \end{pmatrix} ,$$

and

$$\boldsymbol{\pi}^* = \begin{pmatrix} 1 \\ P_\psi(Y_1 = 1|\boldsymbol{M}_i, \boldsymbol{v}_i) \\ P_\psi(Y_2 = 1|\boldsymbol{M}_i, \boldsymbol{v}_i) \\ P_\psi(Y_1 = 1, Y_2 = 1|\boldsymbol{M}_i, \boldsymbol{v}_i) \end{pmatrix} = \begin{pmatrix} 1 \\ \left(\frac{\lambda}{\lambda+\beta}\right)^\alpha \\ \left(\frac{\lambda}{\lambda+\beta}\right)^\alpha \\ \left(\frac{\lambda}{\lambda+\beta}\right)^\alpha \left(\frac{\lambda}{\lambda+2\beta}\right)^{\alpha/2} \end{pmatrix} .$$

It is easily shown that all joint probabilities, $\boldsymbol{\pi}$, may be written in terms of the elements in $\boldsymbol{\pi}^*$. For example, the probability of the observed outcome $(Y_1 = 1, Y_2 = 0)$ *can be obtained as* $P_\psi(Y_1 = 1|\boldsymbol{M}_i, \boldsymbol{v}_i) - P_\psi(Y_1 = 1, Y_2 = 1|\boldsymbol{M}_i, \boldsymbol{v}_i)$, which equals

$$\left(\frac{\lambda}{\lambda + \exp(\beta)}\right)^\alpha \left(1 - \left(\frac{\lambda}{\lambda + \exp(2\beta)}\right)^{\alpha/2}\right) \quad \blacksquare$$

Conaway (1990) notes that,

$$\boldsymbol{\pi}^* = \boldsymbol{A} \cdot \boldsymbol{\pi} \Leftrightarrow \pi = \boldsymbol{A}^{-1} \cdot \boldsymbol{\pi}^* \ . \tag{6.9}$$

The matrix $\boldsymbol{A}$ is the J_i-factorial design matrix. See Appendix B for a description of the general k-factorial design matrix. Any $\boldsymbol{A}$ may therefore be derived from knowing only the size of the sibship.

For our example, with two sibs,

$$\boldsymbol{A} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} , \tag{6.10}$$

is the design matrix in a 2-factorial trial. ■

Now, if we let $\{c_T^Y\}_{T\in\Psi}$ be the row in $\boldsymbol{A}^{-1}$ corresponding to the observed outcome of $\boldsymbol{Y}_i$, we can write,

$$\pi_{\boldsymbol{y}_i} = P(\boldsymbol{Y}_i = \boldsymbol{y}_i | \boldsymbol{M}_i, \boldsymbol{v}_i) = \sum_{T\in\Psi} c_T^Y \pi_T^* \ . \tag{6.11}$$

Note that, stating that the probability of the trait depends on offspring marker data and the inheritance vector implies that (if the inheritance vector is unknown) the probability depends on the parental genotypes. Also note that the GRE, as described above, models the probability in (6.2) one marker at the time (single-point).

6.3 Testing Association in the Presence of Linkage

Our aim is to test association in the presence of linkage. Therefore we assume that there is linkage both in the null and the alternative hypotheses and association in the alternative but not the null. The null and alternative hypotheses in (1.1) can be formulated in the following way,

$$\begin{aligned} H_0 &: \text{ Linkage but no association, } \theta \neq 1/2 \text{ and } \beta = 0 \\ H_1 &: \text{ Linkage and association, } \theta \neq 1/2 \text{ and } \beta \neq 0. \end{aligned} \tag{6.12}$$

In Family-based studies, families are typically ascertained if at least one sib in the family is affected. Since families are ascertained on trait, likelihood

analysis based on (6.1) will not be valid. To deal with this problem, we continue by working with a conditional retrospective likelihood (Section 6.3.1), based on the conditional retrospective probability of $(\boldsymbol{M}_i|\boldsymbol{Y}_i, \boldsymbol{g}_i)$. Likelihood analysis based on the conditional retrospective likelihood deals with both the ascertainment on trait (by conditioning on trait vector; *retrospective*) and population stratification (by conditioning on parental genotypes; "*conditional*"). We derive a score test for testing our null hypotheses of association in the presence of linkage (Sections 6.3.2-6.3.4), based on the conditional retrospective likelihood.

6.3.1 The Conditional Retrospective Likelihood

The conditional retrospective likelihood is written in terms of the probability of $(\boldsymbol{M}_i|\boldsymbol{Y}_i, \boldsymbol{g}_i)$. The likelihood contribution for family i is written as

$$L_i = L_i(\beta, \alpha, \lambda) = P(\boldsymbol{M}_i|\boldsymbol{Y}_i, \boldsymbol{g}_i) \ , \tag{6.13}$$

and the likelihood for the data set $i = 1, 2, ..., n$ is simply the product of all such family contributions, $L = \prod_i L_i$. We can write the likelihood contribution from family i in terms of $P(\boldsymbol{Y}_i|\boldsymbol{M}_i, \boldsymbol{g}_i)$ by applying Bayes rule on $P(\boldsymbol{M}_i|\boldsymbol{Y}_i, \boldsymbol{g}_i)$,

$$L_i = \frac{P(\boldsymbol{Y}_i|\boldsymbol{M}_i, \boldsymbol{g}_i)P(\boldsymbol{M}_i|\boldsymbol{g}_i)}{\sum_M P(\boldsymbol{Y}_i|M, \boldsymbol{g}_i)P(\boldsymbol{M}_i|\boldsymbol{g}_i)} \ . \tag{6.14}$$

Since our interest lies in the vector of parameters $\varphi = (\beta, \alpha, \lambda)$, and especially in β, we can write the log of likelihood (6.14) as proportional to

$$log(P(\boldsymbol{Y}_i|\boldsymbol{M}_i, \boldsymbol{g}_i)) - log\left(\sum_M P(\boldsymbol{Y}_i|M, \boldsymbol{g}_i)P(M|\boldsymbol{g}_i)\right) \ . \tag{6.15}$$

We can continue in two ways:

1. Let $P(\boldsymbol{Y}_i|\boldsymbol{g}_i, \boldsymbol{M}_i) = \sum_v P(\boldsymbol{v}_i|\boldsymbol{M}_i, \boldsymbol{g}_i)P(\boldsymbol{Y}_i|\boldsymbol{M}_i, \boldsymbol{g}_i, \boldsymbol{v}_i) =$
$= \sum_v P(\boldsymbol{v}_i|\boldsymbol{M}_i, \boldsymbol{g}_i)P(\boldsymbol{Y}_i|\boldsymbol{M}_i, \boldsymbol{v}_i)$.

2. Assume that the inheritance vector is known and substitute $P(\boldsymbol{Y}_i|\boldsymbol{M}_i, \boldsymbol{v}_i)$ for $P(\boldsymbol{Y}_i|\boldsymbol{M}_i, \boldsymbol{g}_i)$.

Zhong & Li (2004) propose the second option in a slightly different setting (survival outcomes), assuming known inheritance vectors to derive a score for the retrospective likelihood. To deal with the fact that inheritance vectors are in practice generally unknown they propose averaging a score over the distribution of the inheritance vectors. The first option is correct from a probabilistic perspective, whilst the second option obtains an ad-hoc, approximate, score.

6.3.2 General Form for the Score Test

Let $\nu = (\lambda, \beta)$. A natural efficient score for the ith family is defined as

$$S_i(\nu) = \frac{\partial l_i(0,\nu)}{\partial \beta} - I_{\beta\nu}(0,\nu) I_{\nu\nu}^{-1} \frac{\partial l_i(0,\nu)}{\partial \nu} \ , \tag{6.16}$$

where $\frac{\partial l_i(0,\nu)}{\partial \beta} = \frac{\partial l_i}{\partial \beta}|_{\beta=0}$, $\frac{\partial l_i(0,\nu)}{\partial \nu} = \frac{\partial l_i}{\partial \nu}|_{\beta=0}$, $I_{\beta\nu}(0,\nu) = \frac{\partial^2 l_i}{\partial \varphi}|_{\beta=0}$ and $I_{\nu\nu}(0,\nu) = \frac{\partial^2 l_i}{\partial \nu}|_{\beta=0}$.

It is easily shown that $\frac{\partial l_i(0,\nu)}{\partial \nu} = 0$ by noting that $\frac{\partial}{\partial \nu} P(\boldsymbol{Y}_i|\boldsymbol{M}_i, \boldsymbol{g}_i)$ is independent of $\boldsymbol{M}_i$ when $\beta = 0$ (see Appendix A). Hence, we can write the score as,

$$S_i(\nu) = \frac{\partial l_i(0,\nu)}{\partial \beta} =$$

$$= \frac{\frac{\partial}{\partial \beta} P(\boldsymbol{Y}_i|\boldsymbol{M}_i, \boldsymbol{g}_i)}{P(\boldsymbol{Y}_i|\boldsymbol{M}_i, \boldsymbol{g}_i)} - \frac{\frac{\partial}{\partial \beta} \sum_M P(\boldsymbol{Y}_i|M, \boldsymbol{g}_i)P(M|\boldsymbol{g}_i)}{\sum_M P(\boldsymbol{Y}_i|M, \boldsymbol{g}_i)P(M|\boldsymbol{g}_i)} \ .$$

$P(\boldsymbol{Y}_i|\boldsymbol{M},\boldsymbol{g}_i)$ does not depend on $\boldsymbol{M}$ when $\beta = 0$ and $\sum_M P(M|\boldsymbol{g}_i) = 1$, which leads to

$$S_i(\nu) = \frac{\frac{\partial}{\partial\beta}P(\boldsymbol{Y}_i|\boldsymbol{M}_i,\boldsymbol{g}_i)}{P(\boldsymbol{Y}_i|\boldsymbol{M}_i,\boldsymbol{g}_i)}|_{\beta=0} - \sum_M P(M|\boldsymbol{g}_i)\frac{\frac{\partial}{\partial\beta}P(\boldsymbol{Y}_i|\boldsymbol{M}_i,\boldsymbol{g}_i)}{P(\boldsymbol{Y}_i|\boldsymbol{M}_i,\boldsymbol{g}_i)}|_{\beta=0} \ . \qquad (6.17)$$

6.3.3 An Approximate Score

Following Zhong & Li (2004) we substitute $P(\boldsymbol{Y}_i|\boldsymbol{M}_i,\boldsymbol{v}_i)$ for $P(\boldsymbol{Y}_i|\boldsymbol{M}_i,\boldsymbol{g}_i)$ in the score (6.17).

$$\frac{\frac{\partial}{\partial\beta}P(\boldsymbol{Y}_i|\boldsymbol{M}_i,\boldsymbol{v}_i)}{P(\boldsymbol{Y}_i|\boldsymbol{M}_i,\boldsymbol{v}_i)}|_{\beta=0} - \sum_M P(M|\boldsymbol{g}_i)\frac{\frac{\partial}{\partial\beta}P(\boldsymbol{Y}_i|\boldsymbol{M}_i,\boldsymbol{v}_i)}{P(\boldsymbol{Y}_i|\boldsymbol{M}_i,\boldsymbol{v}_i)}|_{\beta=0} \ .$$

We denote this score as $S_i(\nu|\boldsymbol{v}_i)$ to emphasize that it depends on knowing the inheritance vector $\boldsymbol{v}_i$. Inserting the derivatives of $P(\boldsymbol{Y}_i|\boldsymbol{M}_i,\boldsymbol{v}_i)$, in Equation (6.17) obtains,

$$S_i(\nu|\boldsymbol{v}_i) = \left(\sum_{T\in\Psi} c_T^*\pi_T^*\right)^{-1}\left\{\Upsilon(\alpha,\lambda|\boldsymbol{M}_i) - \sum_M [P(M|\boldsymbol{g}_i)\Upsilon(\alpha,\lambda|M)]\right\} \ ,$$

where

$$\Upsilon(\alpha,\lambda|\boldsymbol{M}_i) = \frac{\alpha}{2}\sum_{T\in\psi} c_T^Y \sum_{k=1}^{4}\left[\left(\frac{\lambda}{\lambda + h_{kT}(0)}\right)^{-1} h'_{kT}(0)\pi_T^*\right] \ . \qquad (6.18)$$

$h'_{kT}(0)$ denotes the first derivative of $h_{kT}(\beta)$ with respect to β, inserting $\beta = 0$. Note that $h'_{kT}(0)$ depends on offspring marker data, whilst $h_{kT}(0)$ does not. Thus, $\Upsilon(\alpha,\lambda|\boldsymbol{M}_i)$ depends on the offspring marker data. To deal

with the reality of unknown inheritance vectors Zhong & Li (2004) suggests summing the score for family i, $S_i(\nu)$, over the inheritance vectors,

$$S_i(\nu) = \sum_v S_i(\nu|v)P(v|\boldsymbol{M}_i, \boldsymbol{g}_i) \ , \tag{6.19}$$

where $P(v|\boldsymbol{g}_i, \boldsymbol{M}_i)$ can be calculated by enumerating all possible inheritance vectors, given marker data on parents and offspring. The probability $P(v|\boldsymbol{g}_i, \boldsymbol{M}_i)$ is simply the reciprocal of the number of possible inheritance vectors.

6.3.4 The Correct Score

We derive the correct score by noting that the probability $P(\boldsymbol{Y}_i|\boldsymbol{M}_i, \boldsymbol{g}_i)$ can be written as $\sum_v P(\boldsymbol{Y}_i|\boldsymbol{M}_i, v)P(v|\boldsymbol{M}_i, \boldsymbol{g}_i)$. The score (6.17) can then be rewritten as,

$$S_i(\nu) = \frac{\frac{\partial}{\partial\varphi}\sum_v P(\boldsymbol{Y}_i|\boldsymbol{M}_i, v)P(v|\boldsymbol{M}_i, \boldsymbol{g}_i)}{\sum_v P(\boldsymbol{Y}_i|\boldsymbol{M}_i, v)P(v|\boldsymbol{M}_i, \boldsymbol{g}_i)}|_{\beta=0} -$$

$$\sum_M P(M|\boldsymbol{g}_i)\frac{\frac{\partial}{\partial\varphi}\sum_v P(\boldsymbol{Y}_i|M, v)P(v|M, \boldsymbol{g}_i)}{\sum_v P(\boldsymbol{Y}_i|M, v)P(v|M, \boldsymbol{g}_i)}|_{\beta=0} \ . \tag{6.20}$$

We find that after taking the derivative of $P(\boldsymbol{Y}_i|\boldsymbol{M}_i, \boldsymbol{v}_i)$ (derived in Appendix A1),

$$S_i(\nu) = \frac{\sum_v P(v|\boldsymbol{g}_i, \boldsymbol{M}_i)\Upsilon(\alpha, \lambda|\boldsymbol{M}_i)}{\sum_v P(v|\boldsymbol{g}_i, \boldsymbol{M}_i)\sum_{T\in\Psi} c_T^Y \pi_T^*(0)} -$$

$$-\sum_M P(M|\boldsymbol{g}_i)\frac{\sum_v P(v|\boldsymbol{g}_i, M)\Upsilon(\alpha, \lambda|\boldsymbol{M}_i)}{\sum_v P(v|\boldsymbol{g}_i, M)\sum_{T\in\Psi} c_T^Y \pi_T^*(0)} \ , \tag{6.21}$$

where $\Upsilon(\alpha, \lambda|\boldsymbol{M}_i)$ is described in (6.18).

6.3.5 Similarities and Differences

The difference between the scores seems to be in the denominator of the score. In the score by (Zhong & Li 2004) there is no summation over the inheritance vectors in the denominator. The weights of the score are therefore different. In the score we propose, the denominator is a weighted sum over all possible inheritance vectors, given parental and offspring genotypes. It should be possible to quantify the difference between the two scores ((6.19) and (6.21)).

Both scores (6.19) and (6.21) can be written as a sum of differences between observed and expected quantities,

$$\Omega_i(\nu|\boldsymbol{M}_i, \boldsymbol{g}_i) - E_v(\Omega_i(\nu|\boldsymbol{M}_i, \boldsymbol{g}_i)) \ .$$

6.4 Application to the GAW14 Simulated Data

We have applied the GRE to the GAW14 simulated data (Greenberg 2004). However, we did not use the retrospective likelihood. We calculated the mean of the likelihood, over all possible patterns of inheritance vectors. We compared this to two existing methods, the FBAT (first using -o, and then on the converted trait data, using -e), and to a GEE (using an exchangeable covariance structure).

7 Discussion

In this thesis we have presented statistical methods for testing and estimating association in the presence of linkage. We have described and evaluated methods for continuous traits and developed a novel approach for binary traits. Our approach is based on a Generalized Linear Mixed Model, which assumes log-gamma distributed random effects in a linear predictor. As noted, there is a lack of methods for dealing with binary traits. We have developed a score test, based on the retrospective likelihood, for testing association in the presence of linkage. The method controls for population stratification by conditioning on parental genotypes and, since it is based on the retrospective likelihood, it is valid under non random ascertainment.

One disadvantage of the GRE is that it is parametric, and its validity depends on how well the observed data follow the probability function in (6.11). An advantage with the underlying random effects distribution is that it allows for many distributional forms. We have not proposed how to extend our GRE method to deal with missing data, however, the standard *Expectation-Maximization* (EM) algorithm should be straightforward to apply.

In the immediate future work will be aimed at investigating the properties of the score described in Section 6.3.4. We are currently in the process of implementing this. The score described in Section 6 is based upon single marker data. A next logical step is to extend the GRE model approach beyond single marker data. We note that the score test described in Section 6 deals with population stratification by conditioning on parental genotypes. This is a somewhat different approach, to account for population stratification, than the mean model approach proposed by Fulker et al. (1999). Future work will be put into comparing different approaches to incorporate population stratification. Another direction for future research is to develop methodology for other types of outcomes, such as repeated binary events.

References

Abecasis, G., Cardon, L. & Cookson, W. (2000a), 'A general test of association for quantitative traits in nuclear families', *American Journal of Human Genetics* **66**, 279–292.

Abecasis, G., Cookson, W. & Cardon, L. (2000b), 'Pedigree tests of transmission disequilibrium', *European Journal of Human Genetics* **8**, 545–551.

Allison, D. (1997), 'Transmission-disequilibrium tests for quantitative traits', *American Journal of Human Genetics* **60**, 676–690.

Almasy, L. & Blangero, J. (1998), 'Multipoint quantitative-trait linkage analysis in general pedigrees', *American Journal of Human Genetics* **62**, 1198–1211.

Bickeboller, H. & Clerget-Darpoux, F. (1995), 'Statistical properties of the allelic and genotypic transmission/disequilibrium test for multiallelic markers', *Genetic Epidemiology* **12**, 865–870.

Boehnke, M. & Langefeld, C. (1998), 'Genetic association mapping based on discordant sib pairs: the discordant sib pair test', *American Journal of Human Genetics* **62**, 950–961.

Chen & Deng (2001), 'A general and accurate approach for computing the statistical power of the tdt for complex disease genes', *Genetic Epidemiology* **21**, 53–67.

Clayton, D. (1999), 'A generalization of the transmission/disequilibrium test for uncertain haplotype transmission', *American Journal of Human Genetics* **65**, 1170–1177.

Clayton, D. & Jones, H. (1999), 'Transmission/disequilibrium tests for extended marker haplotypes', *American Journal of Human Genetics* **65**, 1161–1169.

Conaway, M. (1990), 'A random effects model for binary data', *Biometrics* **46**, 317–328.

Cordell, H., Barratt, B. & Clayton, D. (2004), 'Case/pseudocontrol analysis in genetic association studies: a unified framework for detection of genotype and haplotype associations, gene-gene and gene-environment interactions, and parent-of-origin effects', *Genetic Epidemiology* **26**, 167–185.

Curtis, D. (1997), 'Use of siblings as controls in case-control association studies', *Annals of Human Genetics* **61**, 319–333.

Falk, C. & Rubinstein, P. (1987), 'Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculation', *Annals of Human Genetics* **51**, 227–233.

Fisher, R. (1925), *Statistical methods for research workers*, London: Oliver and Boyd.

Fulker, D., Cherny, S., Sham, P. & Hewitt, J. (1999), 'Combined linkage and association sib-pairs analysis for quantitative traits', *American Journal of Human Genetics* **64**, 259–267.

Gordon, Heath, Liu & Ott (2001), 'A transmission/disequilibrium test that allows for genotyping errors in the analysis of single-nucleotide polymorphism data', *American Journal of Human Genetics* **69**, 371–380.

Greenberg, D. (2004), 'Description of the gaw14 simulated data: http://www.gaworkshop.org/data.htm'.

Horvath, S. & Laird, N. (1998), 'A discordant-sibship test for disequilibrium and linkage: no need for parental data', *American Journal of Human Genetics* **63**, 1886–1897.

Hössjer, O. (2005), 'Conditional likelihood score functions for mixed models in linkage analysis', *Biostatistics* **6**(2), 313–332.

Lake, S., Blacker, D. & Laird, N. (2000), 'Family-based tests of association in the presence of linkage', *American Journal of Human Genetics* **67**, 1515–1525.

Lake, S. & Laird, N. (2004), 'Tests of gene-environment interaction for case-parent triads with general environment exposures', *Annals of Human Genetics* **65**, 55–64.

Lazzeroni, L. & Lange, K. (1998), 'A conditional inference framework for extending the transmission/disequilibrium test', *Human Heredity* **48**, 67–81.

Liang, K. & Zeger, S. (1986), 'Longitudinal data analysis using generalized estimating equations', *Biometrika* **73**, 13–22.

Martin, E., Monks, S., Warren, L. & Kaplan, N. (2000), 'A test for linkage and association in general pedigrees: the pedigree disequiilibrium test', *American Journal in Human Genetics* **67**, 146–154.

McCullagh, P. & Nelder, J. (1989), *Generalized Linear Models*, Vol. 37 of *Monographs on Statistics and Applied Probability*, second edn, CRC Press.

McNemar, Q. (1947), 'Note on the sampling error of the difference between correlated proportions of percentages', *Psychometrika* **12**, 153–157.

Neal, M. & Cardon, L. (1992), *Methodology for genetic studies of twins and families*, Kluwer Academic Publishers: Dordrecht.

Ott, J. (1989), 'Statistical properties of the haplotype relative risk', *Genetic Epidemiology* **6**(1), 127–130.

Rabinowitz, D. (1997), 'A transmission disequilibrium test for quantitative trait loci', *Human Heredity* **47**, 342–350.

Rabinowitz, D. & Laird, N. (1999), 'A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information', *Human Heredity* **50**, 211–223.

Rubinstein, P., Walker, M., Carpenter, C., Carrier, C., Krassner, J., Falk, C. & Ginsberg, F. (1981), 'Genetics of hla disease associations: the use of the haplotype relative risk (hrr) and the 'haplo-delta' (dh) estimates in juvenile diabetes from three racial groups', *Human Immunology* **3**, 384.

Schaid, D. (1996), 'General score tests for associations of genetic markers with disease using cases and their parents', *Genetic Epidemiology* **13**, 423–449.

Self, S., Longton, G., Kopecky, K. & Liang, K.-Y. (1991), 'On estimating hla/disease association with application to a study of aplastic anemia', *Biometrics* **47**, 53–61.

Sham, P. (1997), 'Transmission/disequilibrium tests for multiallelic loci (letter to the editor)', *American Journal of Human Genetics* **61**, 774–778.

Sham, P. (1998), *Statistics in Human Genetics*, John Wiley & Sons Inc., New York.

Sham, P., Cherny, S., Purcell, S. & Hewitt, J. (2000), 'Power of linkage versus association analysis of quantitative traits, by use of variance-components models, for sibship data', *American Journal of Human Genetics* **66**, 1616–1630.

Sham, P. & Curtis, D. (1995), 'An extended transmission/equilibrium test (tdt) for multi-allele marker loci', *Annals of Human Genetics* **59**, 323–336.

Sham, PC Purcell, S., Cherny, S. & Abecasis, G. (2002), 'Powerful regression-based quantitative-trait linkage analysis of general pedigrees', *American Journal of Human Genetics* **71**, 238–253.

Spielman, R. & Ewens, W. (1996), 'The tdt and other family-based tests for linkage disequilibrium', *American Journal of Human Genetics* **59**, 983–989.

Spielman, R., McGinnis, R. & Ewens, W. (1993), 'Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (iddm)', *American Journal of Human Genetics* **52**, 506–516.

Terwilliger, O. (1992), 'A haplotype-based 'haplotype relative risk' approach to detecting allelic associations', *Human Heredity* **42**, 337–346.

Thomson, G. (1995), 'Mapping disease genes: family based association studies', *American Journal of Human Genetics* **57**, 487–498.

White, H. (1980), 'A heteroscedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity', *Econometrica* **48**, 817–838.

Whittemore, A. & Tu, I.-P. (2000), 'Detection of disease genes by use of family data. i. likelihood-based theory', *American Journal of Human Genetics* **66**, 1328–1340.

Zhong, X. & Li, H. (2004), 'Score tests of genetic association in the presence of linkage based on the additive genetic gamma frailty model', *Biostatistics* **5**(2), 307–327.

A The First Derivatives of $P(Y_i|M_i, g_i)$

The probability $P(\boldsymbol{Y}_i|\boldsymbol{M}_i, \boldsymbol{g}_i)$ comprise of a linear summation over π^*_T. We therefore write the derivative of $P(\boldsymbol{Y}_i|\boldsymbol{M}_i, \boldsymbol{g}_i)$ in terms of the derivatives of π^*_T. Note that the first derivative of $h_{kT}(\beta)$ with respect to β is,

$$h'_{kT}(\beta) = \sum_{j \in T} X_j \exp(X_j \beta) a_{jk}$$

First derivative of π^*_T, with respect to β:

$$\frac{\partial \pi^*_T}{\partial \beta} = \frac{\partial}{\partial \beta}\left(\prod_{k=1}^{4}\left(\frac{\lambda}{\lambda + hk, T(\beta)}\right)^{\frac{\alpha}{2}}\right) =$$

$$= \sum_{k=1}^{4}\left[\frac{\alpha}{2}\left(\frac{\lambda}{\lambda + h_{kT}(\beta)}\right)^{\frac{\alpha}{2}-1} h_x(\beta) \cdot \prod_{k^* \neq k}\left(\frac{\lambda}{\lambda + h_{kT}(\beta)}\right)^{\frac{\alpha}{2}}\right] =$$

$$= \sum_{k=1}^{4}\left[\frac{\alpha}{2}\left(\frac{\lambda}{\lambda + h_{kT}(\beta)}\right)^{-1} h_x(\beta)\right] \cdot \prod_{k=1}^{4}\left(\frac{\lambda}{\lambda + h_{kT}(\beta)}\right)^{\frac{\alpha}{2}} =$$

$$= \sum_{k=1}^{4}\left[\frac{\alpha}{2}\left(\frac{\lambda}{\lambda + h_{kT}(\beta)}\right)^{-1} h_x(\beta)\right] \cdot \pi^*_T \tag{A.1}$$

First derivative of π^*_T, with respect to α:

$$\frac{\partial \pi^*_T}{\partial \alpha} = \frac{\partial}{\partial \alpha}\left(\prod_{k=1}^{4}\left(\frac{\lambda}{\lambda + h_{kT}(\beta)}\right)^{\frac{\alpha}{2}}\right) =$$

$$= \sum_{k=1}^{4}\left[\left(\frac{\lambda}{\lambda + h_{kT}(\beta)}\right)^{\frac{\alpha}{2}} + \frac{1}{2}\log\left(\frac{\lambda}{\lambda + h_{kT}(\beta)}\right) \prod_{k^* \neq k}\left(\frac{\lambda}{\lambda + h_{kT}(\beta)}\right)^{\frac{\alpha}{2}}\right]$$

$$= \sum_{k=1}^{4} \left[\frac{1}{2} \log \left(\frac{\lambda}{\lambda + h_{kT}(\beta)} \right) \right] \prod_{k=1}^{4} \left(\frac{\lambda}{\lambda + h_{kT}(\beta)} \right)^{\frac{\alpha}{2}}$$

$$= \sum_{k=1}^{4} \left[\frac{1}{2} \log \left(\frac{\lambda}{\lambda + h_{kT}(\beta)} \right) \right] \pi_T^* \tag{A.2}$$

First derivative of π_T^*, with respect to λ:

$$\frac{\partial \pi_T^*}{\partial \lambda} = \frac{\partial}{\partial \lambda} \left(\prod_{k=1}^{4} \left(\frac{\lambda}{\lambda + h_{kT}(\beta)} \right)^{\frac{\alpha}{2}} \right) =$$

$$= \sum_{k=1}^{4} \left[\frac{\frac{\alpha}{2} \lambda^{\frac{\alpha}{2}-1} (\lambda + h_{kT}(\beta))^{\frac{\alpha}{2}} - \lambda^{\frac{\alpha}{2}} \frac{\alpha}{2} (\lambda + h_{kT}(\beta))^{\frac{\alpha}{2}-1}}{(\lambda + h_{kT}(\beta))^{\alpha}} \cdot \prod_{k^* \neq k} \left(\frac{\lambda}{\lambda + h_{kT}(\beta)} \right) \right] =$$

$$= \sum_{k=1}^{4} \left[\frac{\frac{\alpha}{2} \lambda^{\frac{\alpha}{2}-1} h_{kT}(\beta)}{(\lambda + h_{kT}(\beta))^{\frac{\alpha}{2}+1}} \cdot \prod_{k^* \neq k} \left(\frac{\lambda}{\lambda + h_{kT}(\beta)} \right) \right] =$$

$$= \frac{\alpha}{2} \sum_{k=1}^{4} \left[\frac{h_{kT}(\beta)}{\lambda (\lambda + h_{kT}(\beta))^{\frac{\alpha}{2}-1}} \right] \cdot \prod_{k=1}^{4} \left(\frac{\lambda}{\lambda + h_{kT}(\beta)} \right)$$

$$= \frac{\alpha}{2} \sum_{k=1}^{4} \left[\frac{h_{kT}(\beta)}{\lambda (\lambda + h_{kT}(\beta))^{\frac{\alpha}{2}-1})} \right] \cdot \pi_T^* \tag{A.3}$$

Note that only the derivative with respect to β depends on $\boldsymbol{M}_i$ (through $h'(\beta) = \sum_{j \in T} X_j \exp X_{j\beta}$), when $\beta = 0$.

B The Design Matrix of a k-factorial Trial

$k = 2$

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \tag{B.1}$$

The columns in the design matrix (B.1) corresponds to intercept, Y_1, Y_2 and $Y_1 \cdot Y_2$, respectively.

$k = 3$

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{B.2}$$

The columns in the design matrix (B.2) corresponds to intercept, Y_1, Y_2, $Y_1 \cdot Y_2$, Y_3, $Y_1 \cdot Y_3$, $Y_2 \cdot Y_3$ and $Y_1 \cdot Y_2 \cdot Y_3$ respectively.

For a general k

The columns in the design matrix $\boldsymbol{A}$ corresponds to intercept, $Y_1, Y_2, Y_1 \cdot Y_2, ..., Y_1 \cdot Y_2 \cdot ... \cdot Y_k$, where the indexes are ordered as in Ψ.

C Papers

Paper I

Analysis of binary traits. Testing association in the presence of linkage.

G. Jonasdottir, J. Palmgren, K. Humphreys

Accepted for publication in BMC Genetics
(http://www.biomedcentral.com/bmcgenet/)

Abstract

Most methods for testing association in the presence of linkage, using family-based studies, have been developed for continuous traits. FBAT (Family Based Association Tests) is one of few methods appropriate for discrete outcomes. In this article we describe a new test of association in the presence of linkage for binary traits. We use a gamma random effects model where association and linkage are modelled as fixed effects and random effects, respectively. We have compared the gamma random effects model to an FBAT and a GEE-based alternative, using two regions in the GAW14 Simulated Data. One of these regions contained haplotypes associated with disease, and the other did not.

1 Background

Testing association in a region with confirmed linkage may increase the rate of false positives in family-based studies. In a linked region one expects similarity between related individuals. If unaccounted for, this similarity may be mistaken for association. Different remedies have been suggested, everything from using a robust variance estimator [1] for the general test statistic FBAT (Family Based Association Tests) [2] to a model-based approach where the linkage is modelled in the covariance structure [3] (VCM - Variance Components Model). The VCM has been developed for continuous traits, whilst FBAT can be used with both binary and continuous traits. In this article we concentrate on methods for testing association in the presence of linkage, using binary traits. We compare the program FBAT for binary traits to both the method described in section 2.1 and also a GEE (Generalised Estimating Equation) [4] approach. For the purpose of our comparisons we have used the simulated GAW14 data (Section 2.3). We have compared the three methods ability to pick up a signal in a region with association, as well as their ability to avoid signalling in a region with no association.

2 Methods

We consider a random effects model for binary events which is similar in spirit to the multivariate survival model in [5], which models association and linkage as fixed effects and random effects respectively. We use a result for random effects models for binary outcomes which has been described in [6]. It is shown that for gamma distributed random effects, the unconditional distribution of the outcome using a log-log link can be written as a sum of easily calculated terms. Analytical tractability is only achievable for a few other combinations of random effects distributions and link functions, such as the beta distribution with a log(-log) link [6]. The random effects model in [5] assigns one random effect for each of the two alleles of the father and one random effect for each of the two alleles of the mother. The notion of inheritance vector is used to describe the alleles for all family members jointly. The method presented here works for all sizes of sibships, and may also be easily adapted to extended pedigrees.

2.1 A Gamma Random Effects (GRE) model

Let $(Y_{i1}, Y_{i2}, ..., Y_{iJ_i})$ be the binary trait vector for family i and let j denote offspring ($j = 1, 2, ..., J_i$). We allow for different family sizes J_i. We use θ_{mj} and θ_{pj} to denote the effect of the transmitted alleles to offspring j, with $m_j = 1, 2$ the

maternal alleles and $p_j = 3, 4$ the paternal alleles, respectively. Conditional on the transmitted alleles, we write the probability of the trait for offspring j in family i as $P(Y_{ij} = 1|\theta_{m_j}, \theta_{p_j})$. We consider a model with a log(-log) link of the form

$$\log(-\log(P(Y_{ij} = 1|\theta_{m_j}, \theta_{p_j}))) = \log(\theta_{m_j} + \theta_{p_j}) + \boldsymbol{X}_j\boldsymbol{\beta} \ , \tag{1}$$

or equivalently

$$P(Y_{ij} = 1|\theta_{m_j}, \theta_{p_j}) = e^{-(\theta_{mj}e^{\boldsymbol{X}_j\boldsymbol{\beta}})}e^{-(\theta_{pj}e^{\boldsymbol{X}_j\boldsymbol{\beta}})}. \tag{2}$$

The effects θ of the transmitted alleles act multiplicatively on the offspring trait probability, and the effect of each transmitted allele is multiplied by a term involving the parameter vector $\boldsymbol{\beta}$ describing the fixed genetic effects. Following [7] and [8] we assume that the maternal and paternal alleles are independent and that each allele contributes an effect to the trait which is random and follows a Gamma distribution with scale $\alpha/2$ and shape λ. The model has a tractable closed form for the joint unconditional trait probabilities for the offsprings in a sibship. Let Ψ denote all ordered subsets of $1, 2, ..., J_i$, $\Psi = \{\{\emptyset\}, \{1\}, \{2\}, \{1, 2\}, \{3\}, ..., \{1, 2, ..., J_i\}\}$. Let π^*_T denote the joint unconditional probability of $Y_{ij} = 1$ for all $j \in T$, where $T \in \Psi$. Calculating the probability π^*_T requires integrating over θ_1, θ_2, θ_3, and θ_4. There is a tractable solution [6]. It turns out that

$$\pi^*_T = \prod_{k=1}^{4} \left(\frac{\lambda}{\lambda + \sum_{j\in T} \boldsymbol{X}'_j\boldsymbol{\beta}\boldsymbol{a}_k} \right)^{\alpha/2} \ . \tag{3}$$

The elements of vector $\boldsymbol{a}_k$, a_{jk}, indicate whether allele k has been transmitted to offspring j, $j = 1, 2, ..., J_i$. The probabilities for all $T \in \Psi$ can be placed in a vector $\boldsymbol{\pi}^*$. It has been shown [6] that the unconditional probability for all possible outcomes of $\boldsymbol{Y}$ can be written as,

$$\boldsymbol{\pi} = \boldsymbol{Z}^{-1} \ \boldsymbol{\pi}^* \ . \tag{4}$$

The matrix $\boldsymbol{Z}$ indicates all subsets of T. In order to get the probability of the observed Y_{ij} one needs only to pick the corresponding row in $\boldsymbol{\pi}$. In Table 1 an example of T, matrix $\boldsymbol{Z}$ and vector $\boldsymbol{\pi}$ for three sibs is given. The likelihood for the observed data, for families i ($i = 1, 2, ..., n$), is

$$\log L(\beta, \alpha, \lambda) = \sum_{i=1}^{n} \pi_i \ . \tag{5}$$

We used the statistical software **R** (version 1.9.1) [9] to implement the likelihood and maximize it with respect to the association parameter $\boldsymbol{\beta}$.

We have so far not described how to deal with incompletely observed inheritance vectors. In the context of testing association in the presence of linkage, Zhong et al [5] suggest using GENEHUNTER to obtain the distribution fo r inheritance vectors at any arbitrary point along the chromosome. In our single point analysis we treat all inheritance vectors compatible with the data as equally likely and construct a weighted mean of π_i. We return to the choice of weights in the discussion.

2.2 FBAT and GEE

We compare the GRE with FBAT (version 1.5.1) [2] and a GEE-based alternative [4]. For FBAT we assume a linear allele-dose model, and for the GEE-based alternative we assume a linear allele-dose on the logit scale and an exchangeable covariance structure.

We used FBAT option *-o* to find the optimal weight. We then applied the optimal weight to the phenotype score and used FBAT option *-e* to test our data. The function *gee* (in package *gee*) in **R** (version 1.9.1) was used for the GEE analysis. The *gee* package can be found at the **R** web page [9].

2.3 The GAW14 simulated data

For details concerning how the simulation was performed see [10].

All analyzes were performed with knowledge of the data simulation process. We chose to analyze the data with respect to trait A. Trait A is known to be associated with haplotypes in the Region **D3**, while markers in the D2 region are known to not be associated with trait A. For the purpose of our comparison we therefore chose to "purchase" markers in the D3 region (**B05T4135-B05T4142**) as well as markers from the D2 region (**B03T3048-B03T3067**). Our aim was to use regions D2 and D3 to gain some insight into the performance of the different methods. More specifically, we were not expecting a signal in Region D2, but were hoping for one in region D3.

The Aipotu population (one of four simulated populations) only consists of nuclear families, although these are of different sizes. For simplicity, we chose to concentrate on the Aipotu population and to only include families of maximum size six (*i.e.* two parents and at most four offspring).

We merged 10 (out of 100) replicates, in order to get a sample with reasonable power. This provided us with a total of 481 independent nuclear families. There was no missing data and we did not simulate any.

We selected the markers described above and analyzed each marker separately in a set of single-point analyzes. The method we have described can, however, be extended to multiple markers and a multi-point analysis.

3 Results

We analyzed the ten merged replicates in regions D2 and D3 and we were able to identify interesting markers in both regions. In region **D2**, all three methods (FBAT, GEE and GRE) indicated marker **B03T3056** as borderline significant with a p-value of around 0.01 (Figure 1). The peak was slightly less using FBAT. In Region **D3**, which harbored a haplotype based association in the simulated data, we were able to detect association with marker **B05T4136**. The detected association had a slightly smaller p-value when GEE and GRE (p-value ≈ 0.0001) were used, compared with the FBAT procedure (Figure 2).

4 Conclusions

In the simulated data Region **D2** harbored no locus associated with trait A. All three methods (FBAT, GEE and GEE) gave a signal for association with marker **B03T3056** with a p-value around 0.01. However, taking the multiple testing into account this p-value does not reach statistical significance. The results from all markers in the region are showed in Figure 1. Across the markers, no one method produced consistently higher/lower p-values than any other method.

In Region **D3**, association with trait A was simulated at the haplotype level. We still chose to perform single-point analyzes with each marker in turn. The GEE and the GRE turn out to be slightly better in detecting significant markers than FBAT.

The Gamma Random Effects model presented here seems to work well, compared to both GEE and FBAT. It would be useful to perform simulation studies to assess

validity and power of the three procedures under different genetic models. The GRE model is heavy on computational time, stemming from the fact that in spite of the closed form in (3) it is time consuming to evaluate and to maximize the likelihood.

A problem with the GRE model is how to handle the missing information on transmission. In our single-point algorithm we propose using a weighted sum (with equal weights) over all compatible inheritance vectors, given parental and offspring genotypes. Following Zhong et al [5] we compute the distribution over inheritance vectors without attention to phenotype. However, given that linkage is assumed, the probabilities of transmission are not invariant to offspring phenotypes. It would be useful to investigate the impact of using our suboptimal weights on the GAW data, and more generally in comparing the validity and power of the different approaches using simulations under different genetic models.

References

[1] Lake, SL, Blacker, D, Laird, NM: **Family-Based Tests of Association in the Presence of Linkage.** *Am J Hum Genet* 2000, **67**: 1515-1525.

[2] Rabinowitz, D, Laird, N: **A unified Approach to Adjusting Association Tests for Population Admixture with Arbitrary Pedigree Structure and Arbitrary Missing Marker Information.** *Hum Hered* 2000, **50**: 211-223.

[3] Fulker, DW, Cherny, SS, Sham, PC, Hewitt, JK: **Combined Linkage and Association Sib-Pair Analysis for Quantitative Traits.** *Am J Hum Genet* 1999, **64**: 259-267.

[4] Liang, KY, Zeger, SL: **Longitudinal data analysis using generalized estimating equations.** *Biometrika* 1986, **73**: 13-22.

[5] Zhong, X, Li, H: **Score tests of genetic association in the presence of linkage based on the additive genetic gamma frailty model.** *Biostatistics* 2004, **5(2)**: 307-327.

[6] Conaway, MR: **A random effects model for binary trait.** *Biometrics* 1990, **46(2)**: 317-328.

[7] Li, H: **The additive genetic gamma frailty model for linkage analysis.** *Annals of Human Genetics* 1999, **63**:455468.

[8] Li, H, Zhong, X: **Multivariate survival models induced by genetic frailties, with application to linkage analysis.** *Biostatistics* 2002, **3**:5775.

[9] **The R Project for Statistical Computing.** [http://www.r-project.org/]

[10] **GAW14 Data Description.** [http://www.gaworkshop.org/data.htm]

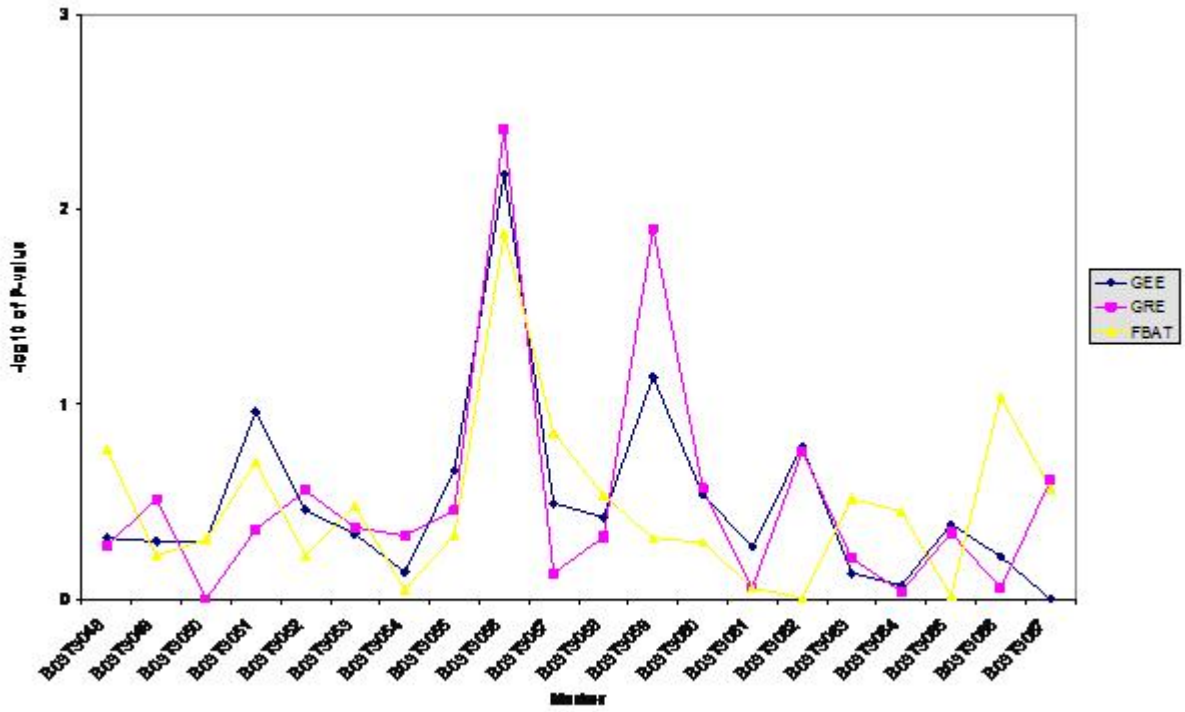

Fig. 1: Trait A region D2, -log10 of the p-values.

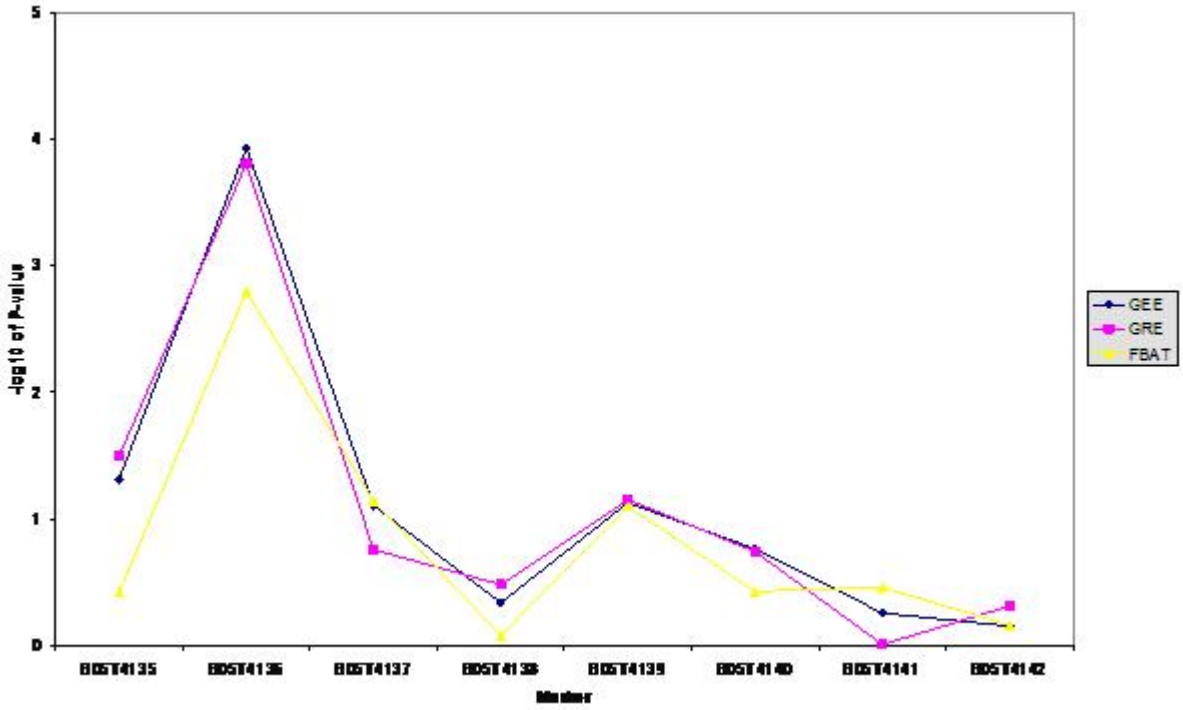

Fig. 2: Trait A region D3, -log10 of the p-values.

Table 1: **Matrices for J_i=3 offspring.**

T	$\boldsymbol{Z}$ matrix								$\boldsymbol{\pi}$ subscr		
$\emptyset$	1	1	1	1	1	1	1	1	1	1	1
1	1	0	1	0	1	0	1	0	0	1	1
2	1	1	0	0	1	1	0	0	1	0	1
1, 2	1	0	0	0	1	0	0	0	0	0	1
3	1	1	1	1	0	0	0	0	1	1	0
1, 3	1	0	1	0	0	0	0	0	0	1	0
2, 3	1	1	0	0	0	0	0	0	1	0	0
1, 2, 3	1	0	0	0	0	0	0	0	0	0	0

The column T denotes all ordered combinations of $1, 2, 3$. Each row in T corresponds in choosing a set off offspring ($j = 1, 2, 3$) and setting their trait Y_{ij} to one. π^*_T is the probability of offspring set T being affected.

The column $\boldsymbol{Z}$ denotes the matrix for which $\boldsymbol{\pi}^* = \boldsymbol{Z}\boldsymbol{\pi}$. The inverse of $\boldsymbol{Z}$ is used to calculate $\boldsymbol{\pi}$, given $\boldsymbol{\pi}^*$.

The column $\boldsymbol{\pi}$ denotes the subscripts for matrix $\boldsymbol{\pi}$. For example, row two corresponds to the outcome $Y_{i1} = 0, Y_{i1} = 1$ and $Y_{i3} = 1$.

Paper II

A Simulation Study of the Parameter Estimates in the Variance Components Model

G. Jonasdottir, J. Palmgren, K. Humphreys

Abstract

The variance components models described by Fulker et al (1999) constitute a flexible class of parametric models for testing and estimation of association and linkage between a quantitative trait and a marker locus. The trait vector for a family is modelled using a linear predictor with a fixed effect and three random effects, accounting for individual specific and family specific effects, as well as an effect accounting for the marker locus. The fixed effect is a function of the offspring genotypes and the random effects are assumed to be normally distributed. Examples of applications include tests for association while controlling for linkage, but it is also possible to test for population stratification and to assess if a marker is functionally related to the trait or merely in linkage disequilibrium with the trait locus. The main advantage with the Fulker et al (1999) approach is the flexibility by which it handles a wide variety of model specifications. This comes at the price of being dependent on the parameters in the fixed effects and on the parameters in the random effects distribution.

We study the properties of the VCM in a series of simulations. In these simulations we vary the degree of linkage disequilibrium between a marker locus and a Quantitative Trait Locus (QTL) in the founder generation, and we subsequently vary the recombination fraction between the marker locus and the QTL in the transmission of marker-QTL haplotypes from parents to offspring. We also simulate scenarios with varying degree of population stratification. We use these simulations to assess how well the estimates of the VCM perform under different types of scenarios.

1 Introduction

This article is concerned with finding association between a locus and a quantitative trait. A locus associated with a quantitative trait is called a Quantitative Trait Locus (QTL). We focus on the scenario where linkage has been found in a region and the researchers wish to continue by fine mapping using association methods.

We present a Variance Components Model (VCM) proposed by Fulker et al (1999) that allows for testing association, while modelling the trait co-variability within a family which is due to linkage. The model is aimed at data collected from studies of many small families and is therefore most appropriate if the underlying genetic effect is polygenic or oligogenic. In fact, since we are studying a continuously distributed trait we are inheritably assuming that many environmental and/or genetic factors affect the trait. We also assume that the probability of ascertainment does not depend on trait value, as is the case in Twin-registry studies (Neal & Cardon 1992). Such studies are not appropriate for identifying low penetrant genes.

We focus on the VCM since it is highly flexible. As well as being able to test for association in the presence of linkage, it is possible to test whether the QTL is functional or merely in Linkage Disequilibrium (LD) with a trait locus, and whether there is population stratification. With this model we may also quantify the proportion of the variance that is due to the Quantitative Trait Locus (QTL) (Fulker et al, 1999).

Consider a particular marker. Let θ denote the recombination fraction between the marker locus and the underlying QTL, and let $\boldsymbol{\beta}$ be a measure of association between the trait and the marker locus. The null and alternative hypothesis can be formulated as

H_0: Linkage but no association, $\theta \neq 1/2$ and $\boldsymbol{\beta} = 0$.

H_1: Linkage and association, $\theta \neq 1/2$ and $\boldsymbol{\beta} \neq 0$.

We study the properties of the VCM in a series of simulations. Specifically, we study what happens with the estimation and testing of the association parameter when varying the degree of linkage and linkage disequilibrium between a marker and a QTL in the study population.

2 The Model

For simplicity of exposition, we restrict ourselves to families with parents and two offspring. However, the proposed methodology can be extended for any type of pedigree. We let i denote family ($i = 1, 2, ..., n$) and j offspring within a family i ($j = 1, 2, ..., J_i$). For all families in the study, both parents and offspring have known genotypes at one marker locus, denoted $\boldsymbol{M}_i$ and $\boldsymbol{g}_i$, respectively, and all offspring have measured trait values, $\boldsymbol{Y}_i$. The marker as well as the underlying trait locus are bi-allelic, with alleles A/a and D/d, respectively.

Fulker et al (1999) propose model for the trait, $\boldsymbol{Y}_i$, which is linear in terms of a fixed effect and random effects. The random effects are partitioned into parts accounting for individual specific and family specific effects, as well as an effect accounting for the QTL; all assumed to be normally distributed. The fixed effect is a function of the offspring genotypes. This model allows for tests of association in the presence of linkage.

2.1 Variance Components Model (VCM)

The random effects are partitioned into three components, assumed independent of each other:

- A non-shared random effect, $e_{ij} \sim N(0, \sigma_N^2)$, unique for all offspring j in family i.
- A random effect shared by all offspring in a family i, $s_{ij} \sim N(0, \sigma_S^2)$.
- A random effect for the QTL, $a_{ij} \sim N(0, \sigma_A^2)$. For offspring j and j' in family i, $\text{cov}(a_{ij}, a_{ij'}) = \pi_{jj'}\sigma_A^2$, where $\hat{\pi}_{jj'}$ is the estimated proportion of alleles shared IBD, given parental genotypes.

We follow Sham et al (2000) and omit the random polygenic effect. It is assumed to be partitioned into the shared, and environmental, random effects (see Sham et al (2000), page 1617, for an example with a sib pair). As long as we are not interested in quantifying the polygenic random effect per se, it makes no difference if it is assumed to be partitioned into σ_N^2 and σ_S^2. The VCM can be written in terms of a fixed effect, μ_{ij}, and the three random effects,

$$Y_{ij} = \mu_{ij} + a_{ij} + s_{ij} + e_{ij} \ . \tag{1}$$

We write the expected covariance between offspring j and j' in terms of $\hat{\pi}_{jj'}$ and the parameters σ^2_N, σ^2_S and σ^2_A,

$$\begin{cases} \text{cov}(Y_{ij}, Y_{ij'}) = \hat{\pi}_{jj'}\sigma^2_A + \sigma^2_S \\ \\ \text{var}(Y_{ij}, Y_{ij'}) = \hat{\pi}_{jj'}\sigma^2_A + \sigma^2_N + \sigma^2_S \end{cases}$$

Thus, linkage enters the model by letting the within family correlation depend on IBD allele sharing. For a family with two offspring the expected covariance matrix is written as,

$$\Sigma = \begin{pmatrix} \sigma^2_N + \sigma^2_S + \sigma^2_A & \sigma^2_S + \hat{\pi}_{jj'}\sigma^2_A \\ \sigma^2_S + \hat{\pi}_{jj'}\sigma^2_A & \sigma^2_N + \sigma^2_S + \sigma^2_A \end{pmatrix} \ ,$$

2.2 The Mean

Consider a biallelic QTL with alleles A and a. Assume a co-dominant allele effect, implying that the mean effect of genotype AA, Aa and aa are $-a$, 0 and a, respectively. We write the mean of Y_{ij}, μ_{ij}, as a sum of an overall mean μ and an allele effect,

$$\mu_{ij} = \mu + aX_{ij} \ , \tag{2}$$

where X_{ij} equal 1,0 or -1 for genotypes AA, Aa and aa, respectively.

$$\begin{cases} \boldsymbol{X}_i = [X_{ij}]_{j=1,2,\ldots,J_i} \\ \\ \overline{X}_i = \sum_{j=1}^{J_i} X_{ij} \\ \\ \boldsymbol{X}_{bi} = [\overline{X}_i]_{j=1,2,\ldots,J_i} \\ \\ \boldsymbol{X}_{wi} = \boldsymbol{X}_i - \boldsymbol{X}_{bi} \end{cases}$$

Note that $\boldsymbol{X}_{bi} + \boldsymbol{X}_{wi}$ equals $\boldsymbol{X}_i$. With this reparameterisation, $a \cdot \boldsymbol{X}_{bi}$ is a vector of the mean allele effect for the offspring in family i, and $a \cdot \boldsymbol{X}_{wi}$ is a vector with the

difference from $a \cdot \boldsymbol{X}_{bi}$ for each offspring j. Fulker et al (1999) note that population stratification will only affect the mean allele effect within a family and therefore propose a mean model where aX_i is split into $a_b X_{bi}$ and $a_w X_{wi}$. The parameter a_w is robust against population stratification.

To illustrate, the elements of $\boldsymbol{X}_i$, $\boldsymbol{X}_{bi}$ and $\boldsymbol{X}_{wi}$ for a sib pair are given in Table 1. As an example, consider a sib pair with genotypes AA and Aa, for sib 1 and 2 respectively,

$$\begin{aligned} \mu_{i1} &= \mu + \tfrac{1}{2}a_b + \tfrac{1}{2}a_w \\ \mu_{i2} &= \mu + \tfrac{1}{2}a_b - \tfrac{1}{2}a_w \end{aligned} \quad .$$

If there is no population stratification, then $a_b = a_w = a$ and we have that the mean allele effects is a for sib 1 and 0 for sib 2. This is consistent with our additive allele mean model (2).

For a general family i, we write,

$$\boldsymbol{\mu}_i = \mu + a_b \boldsymbol{X}_{bi} + a_w \boldsymbol{X}_{wi} = \begin{pmatrix} \mathbf{1} & \boldsymbol{X}_{bi} & \boldsymbol{X}_{wi} \end{pmatrix} \begin{pmatrix} \mu \\ a_b \\ a_w \end{pmatrix} = \mathbf{X}_i' \beta \ . \tag{3}$$

2.3 Likelihood Inference

We assume that there is no ascertainment on trait and use a prospective likelihood, which is based on the probability of trait, given the observed data. We write the likelihood of the model as,

$$L(\mu, \Sigma) = \prod_{i=1}^{n} (2\pi)^{-1} |\Sigma_i|^{-1/2} \exp\left(-\frac{1}{2} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i)' |\Sigma_i|^{-1} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) \right) \ ,$$

where $\boldsymbol{\mu}_i$ is the mean of $\boldsymbol{Y}_i$ and Σ_i is a covariances matrix for the offspring in family i. The null and alternative hypothesis may be formulated as,

H_0: linkage and no association, $\sigma_A^2 > 0$ and $a_w = 0$.

H_A: linkage and association, $\sigma_A^2 > 0$ and $a_w \neq 0$.

Genotype				
Sib 1	Sib 2	$\boldsymbol{X}^T$	$\boldsymbol{X}_b^T$	$\boldsymbol{X}_w^T$
AA	AA	(1,1)	(1,1)	(0,0)
AA	Aa	(1,0)	($\frac{1}{2}$,$\frac{1}{2}$)	($\frac{1}{2}$,-$\frac{1}{2}$)
AA	aa	(1,-1)	(0,0)	(1,-1)
Aa	AA	(0,1)	($\frac{1}{2}$,$\frac{1}{2}$)	(-$\frac{1}{2}$,-$\frac{1}{2}$)
Aa	Aa	(0,0)	(0,0)	(0,0)
Aa	aa	(0,-1)	(-$\frac{1}{2}$,-$\frac{1}{2}$)	($\frac{1}{2}$,-$\frac{1}{2}$)
aa	AA	(-1,1)	(0,0)	(-1,1)
aa	Aa	(-1,0)	(-$\frac{1}{2}$,-$\frac{1}{2}$)	(-$\frac{1}{2}$,$\frac{1}{2}$)
aa	aa	(-1,-1)	(-1,-1)	(0,0)

Tab. 1: The elements of $\boldsymbol{X}$, $\boldsymbol{X}_b$ and $\boldsymbol{X}_w$ for all possible sib pairs. Here $\boldsymbol{X}$ is the vector of X_j's for the two sibs. The elements of $\boldsymbol{X}_b$ are the mean of the elements in $\boldsymbol{X}$ and $\boldsymbol{X}_w = \boldsymbol{X} - \boldsymbol{X}_b$. Note that all indexes i have been omitted and that T denotes transposition of the vectors.

Let ψ denote the vector of parameters, $(\mu, a_b, a_w, \sigma_N^2, \sigma_S^2, \sigma_A^2)$. Using the likelihood in (2.3), the Score function, $S(\psi)$ and the Fisher Information matrix, $I(\psi)$, we can construct a test of association in the presence of linkage.

Maximum Likelihood estimates of the parameters in ψ can be obtained using the Newton-Raphson algorithm. We get an updated estimate $\hat{\psi}^{i+1}$ by,

$$\hat{\psi}^i + \left(I(\psi^i)\right)^{-1} S(\psi^i) \ .$$

The iteration stops when $|\hat{\psi}^{i+1} - \hat{\psi}^{i+1}|$ is smaller than some predefined small δ. A Likelihood Ratio test of the null hypothesis is given by,

$$\text{LRT} = \frac{L(\hat{\mu}, \hat{a}_b, a_w = 0, \hat{\sigma}_A^2, \hat{\sigma}_N^2, \hat{\sigma}_S^2)}{L(\hat{\psi})} \ ,$$

where $\hat{\mu}$ $\hat{a}_b$ $\hat{\sigma}_A^2$, $\hat{\sigma}_N^2$, $\hat{\sigma}_S^2$, and $\hat{\psi}$ are maximum likelihood estimates of μ a_b σ_A^2, σ_N^2, σ_S^2, and $\hat{\psi}$, respectively. The test statistic LRT is χ^2 distributed with 1 degree of freedom.

3 The Simulations

The aim of the simulations is to evaluate the VCM under different scenarios of population stratification, and different degrees of linkage disequilibrium (LD) and linkage (as measured by the recombination fraction θ) between the marker locus and the QTL. Specifically, we want to study the properties of the mean parameters a_b and a_w, and test whether the parameter a_w is robust against population stratification. Another question we wish to address is how efficient the mean model is when there is no population stratification. In that case $a_b = a_w$, which means that the mean model is over-parameterised.

We simulate nuclear families with two sibs. The marker alleles are denoted A/a and the QTL alleles are denoted D/d. The simulation can be divided into three steps;

1. Simulate parents: The parental haplotypes are simulated according to the LD structure between the marker and the QTL.

2. Simulate the offspring: Firstly, the parental haplotypes undergo recombination according to the recombination fraction θ. Secondly, the recombined haplotypes are transmitted, assuming random mating.

3. Simulate the traits: the trait for the offspring in a family is simulated according to model (1), given parameters a_b, a_w, σ^2_N, σ^2_S and σ^2_A, and conditional on the offspring IBD sharing at the QTL.

For all simulations, we fix the model parameters:

- The overall mean, $\mu = 0$.
- The between-family effect, $a_b = 0.5$.
- The within-family effect, $a_w = 0.5$.
- The QTL variance, $\sigma^2_A = 0.2$.

- The non-shared variance, $\sigma_N^2 = 0.6$.

- The shared variance, $\sigma_S^2 = 0.2$.

As explained in Section 2.1 we assume that the polygenic effect gets absorbed into the random environmental and shared random effects and ignore simulating it.

We can describe the LD structure in our simulations using Table (2). This general form of LD structure allows for specifying different degrees of population stratification (population 1 and 2) and LD. Let q_1 and q_2 be the probability of allele A in population 1 and 2, respectively. Further, let r_1 and r_2 be the probability of allele D in population 1 and 2, respectively. We simulate population stratification by varying the marginals q_1 and r_1 for population 1, and q_2 and r_2 for population 2. No population stratification corresponds to setting $q_1 = q_2$ and $r_1 = r_2$. Population k $(= 1, 2)$ is in LD if $p_k \neq q_k \cdot r_k$. Given that we have fixed the marginals, we can vary the degree of LD in the populations by varying p_1 and p_2. We calculate r^2 as,

$$\frac{(p_k - q_k \cdot r_k)^2}{\sqrt{q_k \cdot r_k \cdot (1 - q_k) \cdot (1 - r_k)}} \,,$$

for $k = 1$ or 2. We will use r^2 to quantify the degree of LD between the marker and the QTL. A r^2 of zero corresponds to linkage equilibrium (ie, no LD), and a r^2 of 1 corresponds to complete LD between the marker and the QTL. If $r^2 > 0$ and $r^2 < 1$, we will say that the marker and the QTL are in LD.

	Population 1				Population 2		
	D	d			D	d	
A	p_1	$q_1 - p_1$	q_1	A	p_2	$p_2 - q_2$	q_2
a	$r_1 - p_1$	$1 - q_1 - p_1 - r_1$	$1 - q_1$	a	$r_2 - p_2$	$1 - p_2 - q_2 - r_2$	$1 - q_2$
	r_1	$1 - r_1$	1		r_2	$1 - p_2 - q_2 - r_2$	1

Tab. 2: A general form of LD-structure. We can allow for population stratification by setting $q_1 \neq q_2$ and $r_1 \neq r_2$. The populations $k = 1, 2$ are in LD when $p_k \neq q_k \cdot r_k$.

We set the number of simulations to 20, but in future studies we plan to increase the number to obtain better precision. In each simulation, we estimate all parameters

and test the null using the LRT statistic. We take the mean over all 20 simulation for the mean estimates, $\overline{a}_b$ and $\overline{a}_w$. In future simulations, we will calculate the Fisher Information matrix for each step of the simulation, and from that get the estimated model variance for a_b and a_w, $\overline{\sigma}_b$ and $\overline{\sigma}_w$ respectively.

3.1 Simulation 1

The aim in this simulation is to assess how well the VCM estimates the fixed parameters, varying θ and r^2. In a future study, this simulation setup can be used to address the efficiency of the model when there is no population stratification.

Redefining Table (2), we set $q_1 = q_2 = r_1 = r_2 = 0.5$ and set $p_1 = p_2 = p$. Hence, no population stratification is created in this simulation. We let θ take values 0 and 0.2. We vary the degree of LD by letting p take values 0.25, 0.30 0.35, 0.40 and 0.45, corresponding to a r^2 of 0, 0.04, 0.16, 0.36 and 0.64, respectively.

	D	d	
A	p	$0.5-p$	0.5
a	$0.5-p$	p	0.5
	0.5	0.5	1

Tab. 3: LD-structure for Simulation 1. The population is in LD when $p \neq 0.25$.

3.2 Simulation 2

The aim of this simulation is to assess whether or not a_w is robust against stratification. It will also address how well the VCM estimates the parameters, varying θ and r^2.

Redefining Table 2, we set $q_1 = r_1 = q$ and $q_2 = r_2 = 1 - q$. We let q be either 0.1 or 0.3. Hence, we create population stratification in this simulation. We can quantify the degree of stratification with the fraction $q/(1-q)$, i.e. 1/9 or 3/7. To vary the degree of LD in population 1 and 2, we set $p_1 = p$ and $p_2 = 1 - 2q + p$, and we let p take values q^2 (no LD, $r^2 = 0$) or q (full LD, $r^2 = 1$). With this relation between population 1 and 2, both populations have the same r^2, but are as different as they can be, given the restriction that $q_k = r_k$ $(k = 1, 2)$. We let

r_T^2 denote the LD for population 1 and 2 combined. A false association due to population stratification will be indicated by $r_T^2 > r^2$

	Population 1				Population 2		
	D	d			D	d	
A	p	$q-p$	q	A	$1-2q+p$	$q-p$	$1-q$
a	$q-p$	$1-2q+p$	$1-q$	a	$q-p$	p	q
	q	$1-q$	1		$1-q$	q	1

Tab. 4: LD-structure for Simulation 2. There is some degree of population stratification as long as $q \neq 0.5$. The two populations are in LD when $p \neq q^2$.

4 Results

All programming have been made in the statistical program **R** (version 1.9.0). We have not been able to obtain all results as of yet. Missing results will be denoted by "-".

From simulation 1 (no population stratification), we see that under H_0 ($r^2 = 0$), both additive parameters, the between family parameter a_b and the within family parameter a_w, are estimated close to zero, regardless of θ (see Table (5)). This is consistent with what we should expect. It is, however, curious that a_b is estimated higher than a_w, for all values of θ and r^2. We also see that, under H_1 ($r^2 > 0$), the parameter estimates depend both on the degree of linkage and LD (see Figure (1) and Table (5)). When there is complete linkage ($\theta = 0$), both parameter estimates are approximately equal, regardless of the degree of LD, which is consistent with what we expect to see (Figure (1)). However, as θ becomes larger, the difference between a_b and a_w increase (Figure (1)).

From simulation 2 (population stratification), we see that, under H_0, a_b and a_w differ (see Table (6) and Table (7)). The estimates of a_w are, regardless of θ, close to zero, while a_b is not (Table (6) and Table (7)). This is consistent with what we expect to see. We also see that with a smaller degree of stratification, the difference between a_b and a_w gets smaller, which is also what we would expect to see (Table (6) and Table (7)).

θ	r^2	(p)	Mean (True) $\overline{a}_b$	$\overline{\sigma}_b$	$\overline{a}_w$	$\overline{\sigma}_w$
0	0	(0.25)	-0.0055 (0.5)	- (-)	-0.0184 (0.5)	- (-)
0	0.04	(0.30)	0.0961 (0.5)	- (-)	0.0800 (0.5)	- (-)
0	0.16	(0.35)	0.1955 (0.5)	- (-)	0.1823 (0.5)	- (-)
0	0.36	(0.40)	0.2960 (0.5)	- (-)	0.2843 (0.5)	- (-)
0	0.64	(0.45)	0.3956 (0.5)	- (-)	0.3830 (0.5)	- (-)
0.2	0	(0.25)	-0.0052 (0.5)	- (-)	-0.0191 (0.5)	- (-)
0.2	0.04	(0.30)	0.0822 (0.5)	- (-)	0.0410 (0.5)	- (-)
0.2	0.16	(0.35)	0.1684 (0.5)	- (-)	0.1030 (0.5)	- (-)
0.2	0.36	(0.40)	0.2551 (0.5)	- (-)	0.1635 (0.5)	- (-)
0.2	0.64	(0.45)	0.3420 (0.5)	- (-)	0.2224 (0.5)	- (-)

Tab. 5: Simulation 1: No population stratification. Varying degree of linkage and LD.

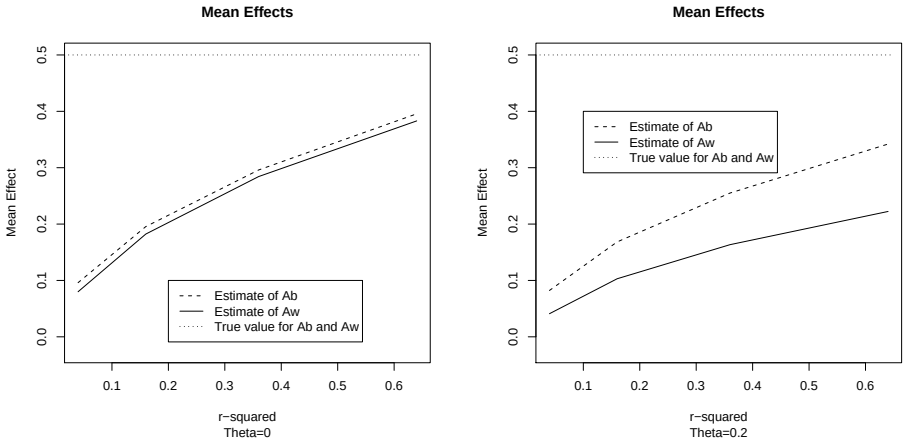

Fig. 1: Simulation 1: Plot of the mean estimates from VCM, under H_1: linkage and association.

5 Discussion

We have simulated scenarios with varying values of r^2 and θ. Some of the values of θ are unrealistically high, corresponding to low linkage. We assume linkage in both

			Mean				True			
θ	r^2	(p, r_T^2)	$\overline{a}_b$	$\overline{\sigma}_b$	$\overline{a}_w$	$\overline{\sigma}_w$	a_b	σ_b	a_w	σ_w
0	0	(0.01, 0.41)	0.4684	-	0.0968	-	0.5	-	0.5	-
0.2	0	(0.01, 0,41)	0.4344	-	0.0138	-	0.5	-	0.5	-
0	1	(0.1, 1)	0.5276	-	0.6094	-	0.5	-	0.5	-
0.2	1	(0,1, 1)	0.4861	-	0.3742	-	0.5	-	0.5	-

Tab. 6: Simulation 2: Population stratification. Varying degree of linkage and LD. $q/(1-q) = 1/9$

			Mean				True			
θ	r^2	(p, r_T^2)	$\overline{a}_b$	$\overline{\sigma}_b$	$\overline{a}_w$	$\overline{\sigma}_w$	a_b	σ_b	a_w	σ_w
0	0	(0.09, 0,03)	0.1879	-	0.0133	-	0.5	-	0.5	-
0.2	0	(0,09, 0,03)	0.1524	-	-0.0824	-	0.5	-	0.5	-
0	1	(0.3, 1)	0.4856	-	0.5166	-	0.5	-	0.5	-
0.2	1	(0.3, 1)	0.4778	-	0.3149	-	0.5	-	0.5	-

Tab. 7: Simulation 2: Population stratification. Varying degree of linkage and LD. $q/(1-q) = 3/7$

the null and the alternative hypothesis, and a large θ is a violation to that assumption. For example, a θ of 0.2 corresponds to a genetic distance of 25.5 cM (with Haldanes mapping function). However, they still serve as an illustration how the estimates of the association parameters are affected as θ increases. Further simulations should be carried out, using more realistic linkage and LD scenarios, and using more simulations. These simulations should be illustrated with confidence intervals for the estimated parameters, giving more interpretable results.

In Figure 1 we see an increasing difference between $\overline{a}_b$ and $\overline{a}_w$. The reason for this is that recombination breaks the LD-structure. Remember that the LD-structure is simulated in the parental generation, and as recombination occur in the second step of the simulation (transmission of haplotypes to offspring), the LD-structure change. So, if there is strong LD in the parental generation and if there is high recombination in the transmission from parents to offspring, we will see a completely different LD-pattern in the offspring generation. This may not be a realistic scenario, but again, it serves as an illustration to what happens on a smaller scale when we change the recombination fraction less dramatically.

As may be noted, some parts of the study is left unfinished to be continued in

a near future. One may also note that the expected values of σ_b and σ_w have not been derived. That is also a task for the near future. This paper is still an unfinished manuscript. A topic for future studies is the error induced by non-random ascertainment. Another interesting question to address is to study how the VCM performs when only a small number of QTL are associated with the trait.

References

[1] D. W. Fulker, S. S. Cherny, P. C. Sham and J. K. Hewitt; Combined linkage and association sib-pairs analysis for quantitative traits; Am. J. Hum. Genet. 64:259-267, 1999.

[2] M. C. Neal, L. R. Cardon; Methodology for genetic studies of twins and families; Kluwer Academic Publishers: Dordrecht, 1992.

[3] P.C. Sham, S.S. Cherny, S. Purcell, J.K. Hewitt; Power of Linkage versus Association Analysis of Quantitative Traits, by Use of Variance-Components Models, for Sibship Data; American Journal of Human Genetics 66:1616–1630, 2000.