Karolinska Institutet

Thesis for doctoral degree (Ph.D.) 2008

Statistical methodology for testing genetic association in family based studies

Gudrun Jonasdottir

Thesis for doctoral degree (Ph.D.)
2008

Statistical methodology for testing genetic association in family based studies

```
1 1 0 1 1 0 1 1 1 1 1 0 1 1 1 1 0 1 0 1 0 1 1 1
1 0 0 0 1 1 1 1 0 1 1 1 1 1 0 1 0 0 0 1 0 0 1 1
0 1 1 0 1 1 1 1 1 1 1 0 1 0 0 0 1 0 0 1 1 0 1 1
0 0 1 1 1 1 1 1 1 0 0 1 1 1 0 1 0 0 1 1 1 0 1 1
1 1 1 1 0 1 1 1 1 0 1 0 0 1 1 0 1 1 1 1 0 0 1 1
1 0 1 1 1 1 1 1 1 0 0 1 1 1 1 1 1 0 1 1 1 1 1 0
0 1 0 1 1 0 0 1 1 1 1 1 1 1 0 1 0 1 1 0 1 0 1 1
1 1 0 1 0 1 0 0 1 0 0 1 1 0 1 1 1 1 0 0 1 1 1 0
1 0 1 1 1 1 1 1 1 1 0 0 1 0 0 1 1 0 1 1 0 1 0 1
0 1 1 1 1 0 0 0 1 1 1 1 0 1 1 1 1 0 1 1 0 1 1 1
1 1 0 0 1 0 1 0 0 1 1 0 1 1 0 0 1 1 1 1 1 1 0 1
0 0 1 1 0 1 0 1 0 0 1 0 0 1 1 0 1 1 1 1 0 0 0 1
1 1 0 1 0 1 1 1 0 1 0 0 1 1 1 1 0 0 0 1 1 1 1 0
1 0   0 1     1   1 1 0   0 0     1   1 1     0
1 0   1       1   1   1   0 0     1   0       1
  1           1   0   0   1       0
```

Gudrun Jonasdottir

Karolinska Institutet

From The Department of Medical Epidemiology and
Biostatistics, Karolinska Institutet, Stockholm, Sweden

Statistical methodology for testing genetic association in family-based studies

Gudrun Jonasdottir

Stockholm 2008

All previously published papers were reproduced with permission from the publisher.
Published by Karolinska Institutet.

ISBN 978-91-7357-540-9

Printed by
REPROPRINT AB
Stockholm 2008
www.reproprint.se
Gårdsvägen 4, 169 70 Solna

ABSTRACT

This thesis is concerned with family-based studies of association between genetic markers and binary traits. A special point of interest in family-based association studies is to separate the *within family* correlation and the genetic effect common among all families in the study. In family-based studies of *quantitative traits* within family trait correlation is explicitly modeled in terms of both alleles shared identical by descent and common environment. We have extended this notion to a binary trait setting, and formulate a generalized linear mixed model based on a log-log link and gamma distributed random effects capturing the within family correlation induced by linkage. The genetic effect common among all families is captured in the linear predictor of the model.

We show that the model can be used to construct tests for a variety of situations; for testing association between single markers and a trait, for testing association between multiple markers (jointly) and a trait, and for testing association between a single marker and two diseases jointly. We have evaluated the model in four papers and show that the power of the test is up to double that of the gold standard for testing association in the presence of linkage - the *Family-Based Association Test* (FBAT).

LIST OF PUBLICATIONS

I. Jonasdottir G, Palmgren J, Humphreys K. Analysis of binary traits: testing association in the presence of linkage. BMC Genetics 6(Suppl 1):S92, 2007.

II. Jonasdottir G, Humphreys K, Palmgren J. Testing association in the presence of linkage - a powerful score for binary traits. Genetic Epidemiology 31(6): 528-540, 2007.

III. Jonasdottir G, Becker T, Humphreys K, Palmgren J. Testing association in the presence of linkage using the GRE and multiple markers. To appear in Genetic Epidemiology.

IV. Jonasdottir G, Humphreys K, Palmgren J. Testing association in family-based studies of bivariate (co-morbid) traits. Submitted.

Contents

LIST OF ABBREVIATIONS

APL Association in the Presence of Linkage

COGA Collaborative Study on the Genetics of Alcoholism

DNA Deoxyribonucleic Acid

DS Disease Susceptibility

FBAT Family-Based Association Test

GAW Genetic Association Workshop

GAW14 14th Genetic Association Workshop

GAW15 15th Genetic Association Workshop

GHRR Genotype-based Haplotype Relative Risk

GLMM Generalized Linear Mixed Model

GRE Gamma Random Effects

GWA Genome-Wide Association

HHRR Haplotype-based Haplotype Relative Risk

IBD Identity-By-Descent

IBS Identity-By-State

IWLS Iterative Re-Weighed Least Squares

LD Linkage Disequilibrium

LRT Likelihood Ratio Test

MGRR Matched Genotype-based Relative Risk

ML Maximum Likelihood

RL Rabinowitz-Laird

SNP Single Nucleotide Polymorphism

TDT Transmission-Disequilibrium Test

VCM Variance Components Model

Part I

BACKGROUND MATERIAL

Chapter 1

GENETIC PRIMER

1.1 Some Important Concepts

Humans carry genetic information in double helix strings of *nucleotides* called DNA (*deoxyribonucleic acid*). There are four nucleotides which form complementary pairs; Adenine and Cytosine, Guanine and Thymine; see Figure 1.1.1. 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 then the complimentary location on the other chromosome in the pair also carries information about the same trait.

Mitosis is the process by which a cell copies it self to form a new identical cell; both cells contain the same set of chromosome pairs. *Meiosis* is the process which in humans forms the *gametes*, *i.e.* the egg in females and the sperm in males. A gametes only contains half a set of chromosomes and when an egg and a sperm merge they form a new cell containing a full set of chromosomes. During the meiosis the chromosomes in a pair *recombine* at random locations, and form new chromosomes. The probability of a recombination occurring between two loci depends on the distance between the loci, chromosome type and sex. The *recombination fraction* is defined, such that a recombination fraction of 1/2 means that the probability of recombination between two loci is 0.5, *i.e.* random assortment, whilst a recombination fraction of 0 means that the probability of recombination between two loci is zero. In conclusion, a child will have half of its DNA from its mother and half of its DNA from its father, but the child does not inherit whole chromosomes from its parents. In the context of this thesis we formalize the process of meiosis and say that

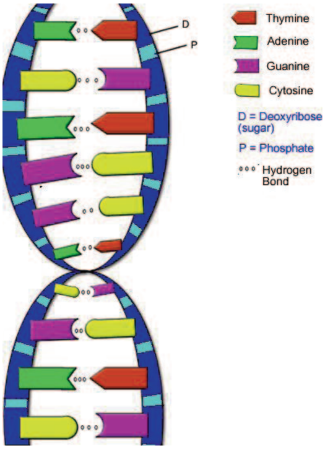


Figure 1.1.1: **The DNA double helix.** The picture has been obtained from http://www.biologycorner.com.

when two individuals mate, chromosomes (perhaps recombined), or parts of chromosomes, are *transmitted* from the parents to the offspring.

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*. One type of biallelic locus is defined on a single nucleotide and if the rare allele has a frequency of at least 1 % in the population, the variant is called a *Single Nucleotide Polymorphism* (SNP). When more than two alleles exists we refer to the locus as *multi-allelic*.

All humans carry two alleles of a given variant; one at each chromosome. We refer to individuals carrying the same allelic variant on both chromosomes as *homozygote* and individuals carrying different variants as *heterozygote*. 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. Consider for example an individual homozygote at one locus and heterozygote on another locus (both on the same chromosome), *e.g.* AA and Dd. The genotype may be written without order, $\{AA.Dd\}$ (our notation - ”.” is a separator between locus genotypes), whereas the haplotypes are ordered by chromosome, AD/Ad (our notation - ”/” is a separator between haplotypes).

In genetic association studies interest lies in finding a locus involved in a disease or trait. Such a locus is called a *Disease Susceptibility* (DS) locus. 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*. Our hope is that at least one of the markers either is the DS locus, or is in close proximity to the DS locus. If the marker and the DS locus are close enough, with little (or no) recombination between them, we will see a *co-transmission* of the two.

1.1.1 Identity-by-descent and inheritance vectors

When studying transmission of alleles to family members it is common to refer to allele similarities between relatives in terms of *Identity-By-State* (IBS) and *Identical-By-Descent* (IBD). If, *e.g.* two siblings share an allele IBS this means that they both have the same allele type, but possibly from different parental chromosomes. If the siblings share an allele IBD then they share the same allele from the same parental chromosome. Siblings can share either 0,1 or 2 alleles IBD. Consider a family where both parents are heterozygous at a biallelic marker; *e.g.* both parents have genotype Aa. Now consider a sib pair with genotypes Aa and aa. The sibs share one a allele IBS, and we can deduce that they have to share one a allele IBD.

Co-transmission can also be described using the *inheritance vector* [29]. Consider a family with two offspring, and a multi- allelic scenario where the father carries alleles a and b and the mother carries alleles c and d. There can be no ambiguity about IBD sharing in this example. The inheritance vector in the two offspring case is a vector with four indicators, where the 1st and the 2nd value indicate which paternal and maternal allele offspring 1 carries, and the 3rd and the 4th value indicates which paternal and maternal alleles offspring 2 carries. For example, the first value in the vector may indicate whether the offspring received allele a (indicator $= 1$) or b (indicator $= 0$), and the second value may indicate whether the offspring received allele c ($= 1$) or d ($= 0$). If the offspring carries genotypes ac and bd, respectively, the inheritance vector is (1100). Offspring genotypes ac and ad correspond to inheritance vector (1110) and genotypes ac and ac correspond to (1111). The inheritance vector is thus a $(2J)$-vector, where J is the number of offspring in the family, containing full information on how the 4 parental alleles have been transmitted to the offspring.

Generally, some inheritance vectors will contain the same amount of information due to the symmetry in enumerating the parental alleles. In the example with two offspring: (0011), (1100), (0110) and (1001) form a group, (0001), (0010), (1000), (0100), (0111), (1011), (1101) and (1110) form a group, and (0000), (0101), (1111) and (1010) form a group. The three groups of inheritance vectors correspond to the offspring sharing 0, 1 and 2 alleles IBD, respectively. Thus, the information contained in pairwise IBD sharing is identical to the information contained in the inheritance vector in the two offspring case. It can easily be shown, however, that when there are more than three sibs, information on pairwise IBD sharing is less informative than inheritance vectors.

The mode of transmission of the parental alleles is not always known. For example, if the parents and offspring are all heterozygotes, all with genotype Aa (say), then all inheritance vectors are equally likely, or equivalently the probabilities of sharing 0, 1 and 2 alleles IBD are 0.25, 0.5 and 0.25, respectively. In contrast, if both offspring are aa homozygotes (parents still heterozygote Aa) then only one inheritance vector is possible, *i.e.* (0000), and 2 alleles are shared IBD.

1.2 Linkage Disequilibrium and Linkage

The concepts of *Linkage Disequilibrium* (LD) and *Linkage*, which are key to the subject of this work, are closely related.

Linkage is defined in the context of transmission of genetic material 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 is less than 0.5.

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 and p_D be population frequencies of haplotype AD, and alleles A and D, respectively. One measure of LD is the *correlation coefficient*, $r = (p_{AD} - p_A p_D)/\sqrt{p_A p_a p_D p_d}$. Another measure of LD, more commonly used by geneticists is $D' = (p_{AD} -$

$p_A p_D)/D_{max}$, where D_{max} is the maximum of $p_A \cdot p_d$ and $p_a \cdot p_D$ if $(p_{AD} - p_A p_D) \leqslant 0$, and the minimum of $p_D \cdot p_d$ and $p_A \cdot p_a$, otherwise.

The level of LD between two markers decreases over generations with the rate of the decay in LD being dependent on the degree of linkage [56]. That is, linkage preserves LD in the population. Another way that LD can be preserved is through non-random mating.

Chapter 2

STATISTICS PRIMER

The test developed in this thesis, the *Gamma Random Effects* (GRE) test, as well as many of the tests presented in Section 3, are based on *likelihood inference*. We present some key concepts, starting with a general formulation of the *Generalized Linear Mixed Model* (GLMM), see *e.g.* [42], in Section 2.1. Not all likelihood based genetic association tests are based on a GLMM, but the GRE test and a test for time-to-event data [67] are. These particular two tests are based on GLMMs with a closed form likelihood solution. For the GRE, the closed-form likelihood solution originates from a result in Conaway [15]. We describe these GLMMs briefly in Section 2.2 and more detailed in Sections 3.1.3 and 5.1. In Section 2.3 we describe two types of tests in likelihood based analysis; the likelihood ratio test and the score test.

2.1 Generalized linear mixed models

Let Y_{ij} be a random variable taking observed value y_{ij}, $i = 1, \ldots, n$ and $j = 1, \ldots, J_i$. Let $\boldsymbol{X}_{ij}$ be a k-vector of predictors. The GLMM can be viewed as an extension of the *Generalized Linear Model* (GLM), described in McCullagh & Nelder [42]. The GLMM allows for dependencies among the Y_{ij}, $j = 1, \ldots, J_i$. The GLMM can be defined, similarly to the GLM, in steps:

1. Let μ_{ij} be the conditional mean of the response Y_{ij}, $E(Y_{ij}|\boldsymbol{b}, \boldsymbol{\beta}, \boldsymbol{X}_{ij}, \boldsymbol{Z}_i)$, and let $h(\cdot)$ be a twice differentiable, continuous function. The conditional mean can then be expressed as

$$h(\mu_{ij}) = \gamma_{ij} = \boldsymbol{X}_{ij}\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b} \ ,$$

where $\boldsymbol{\beta}$ are *fixed effects* and $\mathbf{b}$ are *random effects.* $\boldsymbol{X}_{ij}$ and $\boldsymbol{Z}_i$ are design matrices for the fixed effects and random effects respectively. The function $h(\cdot)$ maps the mean of Y_{ij} to the *linear predictor* γ_i and is called the *link function.*

2. Conditional on the random effects $\boldsymbol{b}$, the fixed effect $\boldsymbol{\beta}$, and the design matrices, $\boldsymbol{X}_{ij}$ and $\boldsymbol{Z}_i$, $\boldsymbol{Y}_i$ is a random variable following a probability, $P(\boldsymbol{Y}_i = \boldsymbol{y}_i|\boldsymbol{b}, \boldsymbol{\beta}, \boldsymbol{X}_{ij}, \boldsymbol{Z}_i)$, with mean $\boldsymbol{\mu}_i$ and variance $\boldsymbol{V}$, determined by the distribution.

3. The random effects $\boldsymbol{b}$ follow a distribution $P(\boldsymbol{b}|\boldsymbol{\theta})$ with parameter $\boldsymbol{\theta}$, calibrated to have zero mean and variance $\boldsymbol{D}$.

The observed likelihood can be formulated in terms of the conditional distribution of $\boldsymbol{Y} = \{Y_{ij}, i = 1, \ldots, n \text{ and } j = 1, \ldots, J_i\}$ and the distribution of the random parameter $\boldsymbol{b}$, integrated over $\boldsymbol{b}$,

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}|\boldsymbol{Y}, \boldsymbol{X}) = \prod_{i=1}^{n} \int_{\boldsymbol{b}} P(\boldsymbol{Y}_i = \boldsymbol{y}_i|\boldsymbol{b}, \boldsymbol{\beta}, \boldsymbol{X}_{ij}, \boldsymbol{Z}_i) \ P(\boldsymbol{b}|\boldsymbol{\theta}) \ \partial \boldsymbol{b} \ . \tag{2.1.1}$$

We maximize the likelihood in Equation (2.1.1) with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. The primary interest is typically estimation or testing of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ is then considered to be a *nuisance* parameter. The values of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ which maximizes the likelihood in Equation (2.1.1) are called the *Maximum Likelihood* (ML) estimates, and are denoted $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$, respectively. A problem with the likelihood in Equation (2.1.1) is that it in general it has no closed form solution and its evaluation requires computationally intensive numerical integration. The GRE is a special case where a closed form solution exists; see Sections 2.2 and 5.1.

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

$$P(\boldsymbol{Y}_i = \boldsymbol{y}_i|\gamma_i, \boldsymbol{b}) = c(\boldsymbol{y}_i, \psi) \exp\left(\frac{S(\boldsymbol{y}_i)\gamma_i - a(\gamma_i)}{\psi}\right) \ ,$$

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 conditional mean can be expressed in terms of the derivative of the *canonical term*, $a(\boldsymbol{\gamma_i})$,

$$E(\boldsymbol{Y}_i|\boldsymbol{\gamma}_i, \boldsymbol{b}) = \boldsymbol{\mu}_i = a'(\boldsymbol{\gamma}_i) \ , \tag{2.1.2}$$

and the variance can be expressed in terms of the second derivative of $a(\boldsymbol{\gamma}_i)$

$$\mathrm{Var}(\boldsymbol{Y}_i|\boldsymbol{\gamma}_i, \boldsymbol{b}) = v(\boldsymbol{\mu}_i) = a''(\boldsymbol{\gamma}_i) \ .$$

The link function h is the inverse function of a' in Equation (2.1.2). Examples of distributions belonging to the exponential family are the Normal distribution, for continuous outcomes, and the Bernoulli distribution for binary data.

2.2 A closed form solution to the likelihood of a generalized linear mixed model for binary outcomes

The likelihood in Equation (2.1.1) is generally difficult to optimize; numerical integration is necessary, except for some special GLMMs. One of these special cases is for a binary response, $Y_{ij} = 0$ or 1, with the log(-log) link and log-gamma distributed random effects, b_i, with scale λ and shape α. Consider the *local independence model*,

$$\log\left(-\log\left(\ P(Y_{ij} = 1|b_i, \beta_j)\ \right)\right) = b_i + \beta_j \ ,$$

where β_j are fixed effects, and $i = 1, \ldots, n$ and $j = 1, \ldots, J_i$. Dropping the conditioning on b_i and β_j we introduce the following notation:

$$\boldsymbol{p}^* = \begin{pmatrix} p^*_{\{\emptyset\}} \\ p^*_{\{1\}} \\ p^*_{\{2\}} \\ p^*_{\{1,2\}} \end{pmatrix} = \begin{pmatrix} 1 \\ P(Y_{i1} = 1) \\ P(Y_{i2} = 1) \\ P(Y_{i1} = 1,\ Y_{i2} = 1) \end{pmatrix} \ ,$$

and

$$\boldsymbol{p} = \begin{pmatrix} p_{11} \\ p_{01} \\ p_{10} \\ p_{00} \end{pmatrix} = \begin{pmatrix} P(Y_{i1} = 1,\ Y_{i2} = 1) \\ P(Y_{i1} = 0,\ Y_{i2} = 1) \\ P(Y_{i1} = 1,\ Y_{i2} = 0) \\ P(Y_{i1} = 0,\ Y_{i2} = 0) \end{pmatrix} .$$

Conaway [15] shows that the joint probability of observable outcomes, $\boldsymbol{p}$, can be written as an equation system of $\boldsymbol{p}^*$. We can write $\boldsymbol{p} = \boldsymbol{B}\boldsymbol{p}^*$, where

$$\boldsymbol{B} = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & -1 \\ 0 & 1 & 0 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix} .$$

We may thus formulate $P(Y_{i1} = y_{i1}, Y_{i2} = y_{i2} | b_i, \beta_j)$ in terms of the rows in $\boldsymbol{p} = \boldsymbol{B}\boldsymbol{p}^*$. For a specific outcome we may write $P(Y_{i1} = y_{i1}, Y_{i2} = y_{i2} | b_i, \beta_j) = \sum_{T \in \Psi} c_T^{\boldsymbol{y}_i} P(Y_{ij} = 1, \forall j \in T | b_i, \beta_j)$, where $\Psi = \{\{\emptyset\}, \{1\}, \{2\}, \{1,2\}\}$ are the indices of $\boldsymbol{p}^*$ and where $c_T^{\boldsymbol{y}_i}$ is the row in $\boldsymbol{B}$ corresponding to the observed outcome, as indexed in $\boldsymbol{p}$. This argument can be generalized to larger sibships as outlined in Conaway [15]. Note that the inverse of matrix $\boldsymbol{B}$ (denoted $\boldsymbol{Z}$ in paper I and $\boldsymbol{A}$ in papers II-IV) is the *J_i-factorial design matrix.*

The above result allows us to consider the integration over the probabilities in $\boldsymbol{p}^*$, instead of $\boldsymbol{p}$. Conaway [15] shows that,

$$P(Y_{ij} = 1,\ \forall j \in T | \beta_j, \lambda, \alpha) = \int_{b_i} P(Y_{ij} = 1,\ \forall j \in T | b_i, \beta_j) P(b_i | \lambda, \alpha) \partial b_i$$

$$= \int_{b_i} \prod_{j \in T} \exp(-\exp(b_i + \beta_j)) P(b_i | \lambda, \alpha) \partial b_i$$

$$= \int_{b_i} \exp\left(-\exp(b_i) \cdot \sum_{j \in T} \exp(\beta_j) \right) P(b_i | \lambda, \alpha) \partial b_i$$

$$= \left(\frac{\lambda}{\lambda + \sum_{j \in T} \exp(\beta_j)} \right)^{\alpha} , \tag{2.2.1}$$

where $P(b_i | \alpha, \lambda)$ is the log-gamma distribution of b_i. From Equation (2.2.1) we see that a closed form solution of the likelihood can be obtained.

2.3 The score test and the Likelihood ratio test

Based on the likelihood in Equation (2.1.1) we formulate two types of tests; the *score test* and the *likelihood ratio test* (LRT). Assume that the parameter we want to test is $\boldsymbol{\beta}$ and that $\boldsymbol{\theta}$ is a nuisance parameter. The score test is based on the derivative of the log of the likelihood with respect to the parameter of interest, in this case $\boldsymbol{\beta}$. The score test is written,

$$S(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \sum_{i=1}^{n} \log L(\boldsymbol{\beta}, \hat{\boldsymbol{\theta}} | \mathbf{Y}_i, \boldsymbol{X}_i) \ , \tag{2.3.1}$$

where $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate of $\boldsymbol{\theta}$, evaluated at $\boldsymbol{\beta}$ (fixed). The score test in Equation (2.3.1) does not take the variability of $\hat{\boldsymbol{\theta}}$ into account; taking it into account requires derivation and computation of the second derivative with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. We test the null hypothesis that $\boldsymbol{\beta}$ is zero by computing the score in Equation (2.3.1) at $\boldsymbol{\beta} = 0$, $S(0)$. The square of the score, $S(0)^2$, is asymptotically *chi-squared* with the *degrees of freedom* equal to the length of vector $\boldsymbol{\beta}$ [42].

The likelihood ratio test compares the likelihood under the null with the likelihood under the alternative. The likelihood under the null is obtained by evaluating the likelihood with $\boldsymbol{\beta} = 0$ (fixed) and $\hat{\boldsymbol{\theta}}$ (the ML estimate of $\boldsymbol{\theta}$ at $\boldsymbol{\beta} = 0$). The likelihood under the alternative is obtained by evaluating the likelihood at the ML estimates of both $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. We write,

$$LRT = -2 \cdot \Big(\log L(\boldsymbol{\beta} = 0, \hat{\boldsymbol{\theta}} | Y, X) - \log L(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}} | Y, X) \Big) \ .$$

The likelihood ratio statistic is, like the score test, chi-squared with degrees of freedom equal to the number of parameters in β [42].

The LRT test requires an extra evaluation of the likelihood at the alternative and it requires estimation of $\boldsymbol{\beta}$, which makes it more cumbersome to evaluate than the score test. The score test requires that the derivative of the likelihood is calculated, either numerically or exactly. Finding the derivative, numerically or exactly, is also required in the estimation of β and is thus also required in the LRT.

Chapter 3

LINKAGE AND ASSOCIATION STUDIES

Burton *et al.* [10] define six steps for discovering and characterizing genes involved in binary and continuous traits. Genetic linkage and association studies constitute steps four and five. The first step is finding evidence that the disease aggregates in families. The last step is to study how the DNA variation in question affects the function of the cell.

In *population-based association studies* we test for association between marker loci and disease by comparing the distribution of marker alleles in a group of unrelated individuals with the disease versus a group of unrelated individuals without the disease. The hope is that one, or several of the markers are either, the DS loci, or in LD with DS loci. An increasingly common type of association study is the *Genome-Wide Association* (GWA) study, see *e.g.* Carlson *et al.* [11] and Kruglyak [28]. In the GWA study the full genome is scanned using a dense distribution of markers. Illumina$^{®}$ and Affymetrics$^{®}$ are two companies providing genotyping arrays for GWA studies. The technology is moving fast and both manufacturers have today the technology to cover about 1 million SNPs in ”one go”.

In family-based studies of linkage, the transmission of genes to cases is typically compared to the expected transmission probabilities; the *Mendelian* transmission probabilities. Linkage analysis/testing and association analysis/testing have historically been viewed as separate entities, both however, rely on ancestral recombination to define ”closeness” between DS loci and marker loci. In association studies the population can be viewed as one big family in which relatives are distantly related and no knowledge of specific relationships exists.

It is however possible in family-based studies to test both association and linkage. In family-based studies of association, cases are compared to their healthy relatives, typically their healthy siblings. An advantage with choosing a family member as control is that they are ethnically matched and they can also share environment and lifestyle. A drawback is that their genetic similarity reduces power.

Linkage between the marker and DS locus will induce a within family trait correlation. For this reason, family-based studies (of genetic association) tests either association and linkage jointly, or association controlling for linkage. The null hypothesis, in the joint test of association and linkage, is that neither linkage nor association exists between the marker and trait. The null hypothesis when testing for association while controlling for linkage, is that linkage, but not association, exists between the marker and trait. The first null hypothesis is referred to as the *type-I hypothesis* and the latter is referred to as the *type-II hypothesis* [31]. Testing the type-II hypothesis is more commonly referred to as testing *Association in the Presence of Linkage* (APL).

A confounding factor which may induce false association between markers and disease is *population stratification* or *population admixture*. Population stratification/admixture refers to a situation where several sub-populations with different genetic background exists in the study population, which may be the case in ethnically mixed populations. Methods for handling population stratification have been developed for both population-based and family-based studies. However, family-based studies have the advantage of offering natural matching of genetic background; siblings, for example, share the exact same genetic background; at the expense of reduced power.

3.1 Family-Based Association and Linkage Studies

3.1.1 The conditional prospective and retrospective likelihoods

Several different likelihoods have been proposed and used for the analysis of data from family-based genetic studies. Kraft & Thomas [27] discuss likelihoods which do not condition on parental marker genotypes. In contrast, the likelihoods presented in this section condition on parental marker genotypes. By conditioning on the parental marker genotype we avoid having to estimate the population allele frequencies [50].

Ascertainment of families for mapping is often based on one or several family members having the disease. The family are thus ascertained based on trait values and statistical analyzes need to acknowledge the sampling procedure. Most tests of genetic association in family-based studies are for this reason based on the retrospective likelihood. Another type of study is a cohort study where individuals and their families are followed prospectively in time. In such studies it is appropriate to base inference on the *prospective likelihood*,

$$L_p = \prod_{i=1}^{n} P(\boldsymbol{Y}_i = \boldsymbol{y}_i | \boldsymbol{G}_i, \boldsymbol{g}_i) \ ,$$

where $\boldsymbol{Y}_i$ denotes the trait vector for family i, $\boldsymbol{G}_i$ and $\boldsymbol{g}_i$ denote the offspring and parental genotypes, respectively, and where n is the number of families in the study. The prospective likelihood has been used extensively in searching for loci involved in quantitative traits, see *e.g.* Fulker *et al.* [20] and Sham *et al.* [57].

We present the *conditional* (on parental genotypes) *retrospective* (with respect to trait) likelihood, written,

$$L_r = \prod_{i=1}^{n} P(\boldsymbol{G}_i | \boldsymbol{g}_i, \boldsymbol{Y}_i) \ .$$

We will use the fact that the retrospective likelihood can be written in terms of the prospective probability of trait, $P(\boldsymbol{Y}_i = \boldsymbol{y}_i | \boldsymbol{G}_i, \boldsymbol{g}_i)$. Using Bayes Theorem,

$$L_r = \prod_{i=1}^{n} \frac{P(\boldsymbol{Y}_i = \boldsymbol{y}_i | \boldsymbol{G}_i, \boldsymbol{g}_i) P(\boldsymbol{G}_i | \boldsymbol{g}_i)}{\sum_{\boldsymbol{G} \in \boldsymbol{G}_i^*} P(\boldsymbol{Y}_i = \boldsymbol{y}_i | \boldsymbol{G}, \boldsymbol{g}_i) P(\boldsymbol{G} | \boldsymbol{g}_i)} \ ,$$

where the summation in the nominator is over all offspring genotype configurations, consistent with the observed parental genotypes (further discussed in Section 6). The conditional retrospective likelihood has been proposed in the context of family-based association studies by several authors, including Clayton [13], Shih & Whittemore [59] and Zhong & Li [67].

3.1.2 Testing and Estimating Association and Linkage Jointly

The main topic of the present thesis is testing and estimation of APL. Many existing family-based tests are, however, appropriate for testing association and linkage jointly. We provide a short review of these tests with the purpose of providing a general background of statistical methods for analyzing family-based association and linkage studies.

Most of the early family-based studies are based on *trios*, *i.e.* studies which collect genetic information from one affected offspring and its parents. Here we summarize four tests which, in their original form, were designed to test association and linkage from trios data; the *Matched Genotype-based Relative Risk (MGRR)* test [52], the *Genotype-based Haplotype Relative Risk* (GHRR) test [19], the *Haplotype-based Haplotype Relative Risk* (HHRR) test [46] and the *Transmission Disequilibrium Test* (TDT) [61]. We deal with a binary trait and a biallelic marker locus with alleles A and a. Let n be the number of families which are in the study.

Rubenstein *et al.* [52] described a test based on the transmission of genotypes from parents to offspring. Let index i denote the genotype transmitted to the offspring and let index j denote genotype not transmitted to the offspring. Rubenstein *et al.* [52] let i and j take values 1 for genotypes AA and Aa and 2 for genotype aa. T_{ij} denotes the number of transmitted (i) and non-transmitted (j) genotypes. Thus, $\sum_{i,j} T_{ij} = n$. Rubenstein *et al.* [52] suggest treating the two genotypes; the transmitted and the non-transmitted; as being dependent (matched). The test suggested by Rubenstein *et al.* [52] is,

$$MGRR = \frac{(T_{12} - T_{21})^2}{T_{12} + T_{21}} \ . \tag{3.1.1}$$

Falk & Rubenstein [19] suggest breaking the matching and instead of looking at the pairs of genotypes, they separate the transmitted and non-transmitted genotypes. Falk & Rubenstein [19] propose using the test

$$GHRR = \frac{(T_{12} - T_{21})^2}{(2T_{11} + T_{12} + T_{21})(T_{12} + T_{21} + 2T_{22})/2n} \ . \tag{3.1.2}$$

The difference between the MGRR and the GHRR tests lies in the variance estimator of $(T_{12} - T_{21})^2$, i.e. in the denominators of the test statistics in

Equations (3.1.1) and (3.1.2). Both test statistics are used for jointly testing linkage and association.

It is also possible to test for association and linkage based on the transmission of alleles, rather than genotypes. Ott [46] and Terwilliger & Ott [61] propose a matched analysis. Let t_{ij} denote the number of pairs of alleles, where the index i denotes the allele transmitted to the affected offspring and where index j denotes the allele which was not transmitted. Hence, $\sum_{ij} t_{ij} = 2n$. Ott [46] and Terwilliger & Ott [61] proposed a test statistic similar in spirit to the test statistic in Equation (3.1.2),

$$HHRR = \frac{(t_{12} - t_{21})^2}{(2t_{11} + t_{21} + t_{12})(t_{12} + t_{21} + 2t_{22})/4n} \ . \tag{3.1.3}$$

The TDT considers non-matched alleles and is written,

$$TDT = \frac{(t_{12} - t_{21})^2}{(t_{12} + t_{21})} \ . \tag{3.1.4}$$

Note the similarity between the TDT in Equation (3.1.4) and the MGRR in Equation (3.1.1). All tests presented in this section are in their original form χ^2 distributed with one degree of freedom.

Ott [46] derived the expected value of the HHRR and the TDT. Consider a biallelic DS locus with alleles D and d, and a marker locus with alleles A and a. Let p_{AD}, p_{Ad}, p_{aD} and p_{ad} denote the frequencies of haplotypes AD, Ad, aD and ad, respectively. Let also p denote the prevalence of the DS allele, D. Let θ denote the recombination fraction and let $\delta = p_{Ad}p_{aD} - p_{ad}p_{AD}$ be a measure of LD. The expected squared difference between t_{12} and t_{21} can be expressed in terms of θ, δ and the population frequency of the DS allele, p. That is, the expected value of the numerator of the HHRR and the TDT, $(t_{12} - t_{21})^2$, equals,

$$\left(\frac{\delta}{p}(1 - 2\theta)\right)^2 \ . \tag{3.1.5}$$

Equation (3.1.5) equals zero if and only if θ equals 0.5 or δ equals zero, i.e. only if there is no linkage ($\theta = 0.5$), or if there is no LD between the marker and the DS locus ($\delta = 0$). Alternatively, Equation (3.1.5) differs from zero only if there is *both* linkage and LD between the marker and the DS locus.

The HHRR and the TDT are thus joint tests of linkage and association. Note that trait values are not included in the expected value in Equation (3.1.5); they cancel out in the derivations of the transmission probabilities [46].

Many extensions of the original TDT have been proposed; for multiple affected sibs [60], multiple markers [58, 55, 53], for general pedigrees [41], allowing for missing parental genotypes [58], for haplotypes phase and missing parental genotypes [13, 12], and for continuous traits [3, 49, 1, 2]. Another family of tests of association and linkage is the (original) *Family-Based Association Test* (FBAT) [50], which we describe in more detail in Section 3.1.3.

3.1.3 Testing and Estimating Association in the Presence of linkage

In this section we describe three general lines of methodological development for testing APL; the *Variance Components Model* (VCM) [20] for continuous traits, the Family Based Association Test (FBAT) [49] for binary and continuous traits, and a score test for time-to-event data [67].

We deal here with n independent nuclear families, consisting of parents and their offspring. Let i denote family ($i = 1, 2, ..., n$), j denote offspring within a family i ($j = 1, 2, ..., J_i$) and let $\boldsymbol{g}_i$ and $\boldsymbol{G}_i$ denote parental and offspring genotypes, respectively. Let $X(G_{ij})$ denote some genotype score (possibly a vector) of the offspring genotype, G_{ij}. For example, in the biallelic setting with alleles A and a, $X(G_{ij})$ may be equal to the number of A alleles in genotype G_{ij}. The trait, $\boldsymbol{Y}_i$, of the offspring in family 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).

The Variance Components Model

The Variance Components Model (VCM) has a long history of quantifying the relative importance of the genetic component of quantitative traits (without genotype data), for example in twin-studies [45]. Almasy & Blangero [4] extended Variance Components methodology to asses linkage between a marker and a quantitative trait.

Fulker *et al.* [20] were the first to propose the use of the VCM to analyze association and linkage jointly. The VCM that Fulker *et al.* [20] describe is a Generalized Linear Mixed Model (GLMM) with an identity link function, normally distributed trait and normally distributed random effects. For trait, Y_{ij}, fixed effect, μ_{ij}, and random effects, a_{ij}, s_{ij} and e_{ij}, the model is written,

$$Y_{ij} = \mu_{ij} + a_{ij} + s_{ij} + e_{ij} \ ,$$

where e_{ij} denotes a non-shared random effect, s_{ij} denotes a shared random effect, and a_{ij} denotes an additive genetic random effect, all normally distributed with zero means and variances σ^2_N, σ^2_S and σ^2_A, respectively. The mean of Y_{ij}, can be written as $\mu_{ij} = \beta_0 + \beta_1 X(G_{ij})$, where β_0 denotes a common mean and β_1 denotes an additive genetic effect. The trait vector of the offspring in family i, $\boldsymbol{Y}_i$, is multivariate normal with mean $(\beta_0 + \beta_1 X(\boldsymbol{G}_i))$ and covariance matrix Σ_i. Consider for simplicity of exposition a sib-pair. The model-postulated covariance matrix is then written,

$$\Sigma_i = \begin{pmatrix} \sigma^2_N + \sigma^2_S + \sigma^2_A & \sigma^2_S + \pi\sigma^2_A \\ \sigma^2_S + \pi\sigma^2_A & \sigma^2_N + \sigma^2_S + \sigma^2_A \end{pmatrix} \ ,$$

where π is the expected proportion of alleles shared IBD, *i.e.* $\pi = 0.5$ for sib-pairs (except monozygotic twins for which $\pi = 1$). The correlation structure expands straightforwardly to other sibship sizes. Population stratification can be handled by allowing β_1 to be partitioned into a between-family, and a within-family effect [20]. Fulker *et al.* [20] consider inference based on the prospective likelihood (Section 3.1.1). We can test for APL using a likelihood ratio test or a score test of $\beta_1 = 0$, and estimate σ_A, σ_N and σ_S as nuisance parameters (Section 2.3). Note that it is possible to test several null hypothesis using the VCM, including *linkage only* and *association and linkage jointly* [20].

The Family-Based Association Tests

Here we describe the original FBAT statistic [50] for testing association and linkage jointly, and the extension for testing APL [31]. Let $T(Y_{ij})$ be a function of the trait, for example equal to $Y_{ij} - o$, where o is an offset. For binary traits, setting $o \neq 0$ makes T_{ij} non-zero for both affected and unaffected individuals. It is shown in Lange & Laird [34] and Lange & Laird [33] that an optimal choice of o, in terms of power, is the sample mean of the trait. For simplicity of notation, we will write T_{ij} and X_{ij} in place of $T(Y_{ij})$ and $X(G_{ij})$, respectively. Rabinowitz & Laird [50] propose the following score statistic for testing association,

$$S = \sum_i S_i = \sum_{i=1}^{n} \sum_{j=1}^{J_i} T_{ij} X_{ij} \ . \tag{3.1.6}$$

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

Rabinowitz & Laird [50] propose calculating the expected value of the family score, S_i in Equation (3.1.6), by conditioning on the sufficient statistic of the parental genotypes, $\boldsymbol{g}_i$, and the trait, $\boldsymbol{Y}_i$. By this conditioning, Rabinowitz & Laird [50] design a valid test for association, regardless of genetic model and population admixture or stratification. Rabinowitz & Laird [50] present an algorithm for finding the sufficient statistic of the parental genotypes, and for calculating the conditional probabilities of the possible sibship genotype vector, given the sufficient statistic for parental genotype. The *Rabinowitz-Laird* (RL) 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 minimal 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.* [31] show that $S_L = \sum_{i=1}^{n} (S_i - E(S_i|\phi))$ is a valid test statistic for testing association in the presence of linkage. However, the covariance of the statistic will not be the same, so instead Lake *et al.* [31] propose using a robust covariance estimator [66, 38],

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

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

$$Z_L = \Sigma_L^{-1} S_L \ .$$

The Lake extension of FBAT is valid under any genetic model and population stratification / admixture [50]. It also deals with missing marker data, through conditioning on $\xi(\boldsymbol{g}_i)$.

Several extensions of the FBAT have been proposed. Notable in the context of this thesis are the extensions which handle testing haplotypes [22] and multivariate traits [35]. We have formulated the five steps in the RL-algorithm such that they include haplotypes (by adding *phase* in several of the steps) as well as single markers. The multivariate FBAT [35] is based on a GEE [38] formulation of the trait specific FBAT scores, S_i.

Note that the FBAT is based on assuming that the trait is fixed. The FBAT is thus conditional on trait and is therefore, in general terms, a retrospective test, see Section 3.1.1.

A Score Test for Time-to-Event Data

Li & Zhong [36] propose a so called *frailty model* [64] for *survival data.* The model represents an extension of the *Cox proportional hazards model* [16] that allows for random individual *risks* of having an *event.* The *hazard* at a specific time point (or age) is a measure of the risk which a person, healthy up to that time point, has of developing the disease at that time. The model allows for right censoring, acknowledging that some individuals are not observed to develop the disease, due to death or loss to follow up for other reasons (which ever came first). Survival methodology is not within the scope of this thesis. However, the GRE is an adaption of the methodology presented in Zhong & Li [67] to binary traits. We therefore shortly outline the starting point of the Zhong & Li [67] model, here.

For simplicity of exposition, consider families with two offspring. Let t_{ij} be the age at which offspring j in family i develops the disease. The hazard takes the form,

$$\lambda_{ij}(t_{ij}|Z_{ij}) = \lambda_0(t_{ij}) \exp(X(G_{ij})\beta) Z_{ij} \ , \tag{3.1.7}$$

where $\lambda_0(t_{ij})$ is some unknown baseline hazard at age t_{ij} and β is a parameter measuring the common genetic effect. The vector of random effects, Z_{ij}, is defined in terms of the inheritance vector. Assuming that the mode of inheritance is known Zhong & Li [67] write,

$$Z_{ij} = \epsilon_{i,v_{2j-1}} + \epsilon_{i,v_{2j}} + \epsilon^s_{ij} \ , \tag{3.1.8}$$

where v_{2j-1} and v_{2j} are the two components of the inheritance vector attributable to offspring j. The random effects, $\epsilon_{i,v_{2j-1}}$, $\epsilon_{i,v_{2j}}$ and ϵ^s_{ij}, are assumed to be gamma distributed. The parameters of the gamma distributions are restricted so that the sum, Z_{ij}, is gamma distributed with mean 1 and variance $1/\lambda$. The restriction assures identifiability of the baseline hazard, $\lambda_0(t)$, and has the advantage of preventing arbitrary scaling of the model in Equation (3.1.7) [67].

Zhong & Li [67] chose the model out of mathematical convenience; the hazard in Equation (3.1.7) and random effects formulation in Equation (3.1.8) yields a closed form expression of the likelihood. The model is in fact similar to the GLMM for binary traits, presented in Section 2.2. Based on the conditional retrospective likelihood, Zhong & Li [67] propose a score test for testing association in the presence of linkage between a marker and time-to-event data.

Part II

THESIS MATERIAL

Chapter 4

AIMS OF THE THESIS

The Variance Components Model [20], described in Section 3.1.3, has been used widely in testing for genetic association in family-based studies of quantitative traits; see for example [57]. The VCM includes a fixed effect, which can be estimated to assess genotype-trait association, and it also contains random effects which accounts for correlation in trait values between related individuals, based on knowledge of IBD sharing and the extent to which related individuals share environmental factors. For survival data the idea of testing for a fixed effect (genotype association) while accounting for the trait correlations between related individuals, using a frailty model, was introduced by Li & Zhong [36]; Section 3.1.3.

The broad aim of this thesis is to extend the methodology of Fulker *et al.* [20] and Zhong & Li [67] to the binary outcome setting. The VCM is a GLMM (Section 2.1) and the Zhong & Li model for the time-to-event data is based on a random effects formulation. Most GLMMs for binary outcomes have a likelihood whose evaluation requires numerical integration over the random effect distribution. Consideration of computational simplicity and feasibility is thus a central issue in this thesis. Specific aims:

(i) To adapt the method of Zhong & Li [67] to a simple binary trait setting with single markers and no missing parental genotypes, using the methodology of Conaway [15]. To study the validity and power of a test, based on the 'novel' model under a conditional retrospective likelihood (Section 3.1.1), for testing APL while protecting for population stratification and admixture.

(ii) To extend the model in (i) in such a way as to derive a test of APL for multiple markers, which, as a result, handles families with missing parental genotypes. To develop a haplotype-based test of APL.

(iii) To extend the model in (i) to a bivariate binary trait setting, for use in genetic association studies of comorbidity.

Chapter 5

THE STRUCTURE OF THE THESIS

The tests developed in this thesis are based on a prospective GRE model for the probability of a (binary) trait, given an observed set of offspring genotypes and (inferred) knowledge of IBD sharing between sibs. Inferences are based on a likelihood, which is chosen according to study design. For example, when families are ascertained based on trait values, we use a likelihood which conditions on trait values, relying on Bayes theorem to formulate the likelihood in terms of prospective probabilities.

All four papers included in this thesis are based on the same general model, under different restrictions and modifications. We begin in this Chapter by describing the general GRE model and then proceed in Chapter 6 by describing the specific applications included in each paper. From paper I through to paper IV the GRE tests are developed to become successively more general. Developments for missing genotype information, multi-marker tests and bivariate traits are described.

5.1 The Gamma Random Effects Model

We use index i to denote families, $i = 1, \ldots, n$, and index j to denote individuals within families i, $j = 1, \ldots, J_i$. Offspring trait values (= 0 or 1) are denoted by Y_{ij}. In paper IV, Y_{ij} is allowed to be a a vector of trait values for each offspring. We use G_{ij} to denote the genotype of offspring j and $\boldsymbol{g}_i$ to denote parental genotypes. Genotypes may be single-marker or multi-marker. In what follows, 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 $\boldsymbol{Y}_i$.

In this thesis we mainly consider cases where the region of interest is known to be in linkage with a DS locus. Under the alternative hypothesis of association and linkage the joint probability of trait, $P(\boldsymbol{Y_i} = \boldsymbol{y}_i|\boldsymbol{G}_i, \boldsymbol{g}_i)$, depends on the observed genotypes in the sibship and on the pattern of IBD sharing. Many of the existing methods for testing association in the presence of linkage specify only pairwise IBD proportions, see *e.g.* [20, 57, 40]. The GRE is, however, based on full specification of the pattern of IBD sharing in a family, using the *inheritance vector* [29]; see Section 1.1.1. We specify the GRE working model as,

$$\log\left(-\log\left(q_{ij}\right)\right) = \log(\boldsymbol{a}_{ij} \cdot \boldsymbol{\epsilon}_i + \epsilon_i^s) + \beta_0 + X(G_{ij}) \cdot \beta_1 \ , \tag{5.1.1}$$

where $q_{ij} = P(Y_{ij} = 1|X(G_{ij}), \boldsymbol{v}_i, \boldsymbol{\epsilon_i}, \epsilon_i^s, \beta_0, \beta_1)$, where $\boldsymbol{v}_i$ is the inheritance vector for family i. The vector $\boldsymbol{a}_{ij}$ contains the same information as the inheritance vector. To simplify notation we write $a_{ij} = [v_{i,2j-1}, \overline{v}_{i,2j-1}, v_{i,2j}, \overline{v}_{i,2j}]$, where $\overline{v}_{i,k}$ takes the value of 1 if $v_{i,k} = 0$, and 0 otherwise ($k = 2j - 1$ or $2j$). Here $\boldsymbol{\epsilon_i} = (\epsilon_{i1}, \epsilon_{i2}, \epsilon_{i3}, \epsilon_{i4})$ denotes family specific transmission effects, one for each of the four parental alleles and $\boldsymbol{a}_{ij} \cdot \boldsymbol{\epsilon}_i$ is the sum of the two transmission effects corresponding to the parental alleles transmitted to offspring j. We also allow for a shared environmental effect within family, ϵ_i^s. We use a flexible distribution for the transmission effects, $\epsilon_1, \dots \epsilon_4$; the gamma distribution, with scale λ and shape α. The shared effect, ϵ_i^s is similarly assumed to be gamma distributed with scale δ and shape η. $X(G_{ij})$ is the marker genotype G_{ij} score for offspring j in family i. The marker genotype score can be formulated in several ways, but we restrict ourselves to the biallelic case and let $X(G_{ij})$ take values 0, 1 or 2 for genotypes AA, Aa and aa, respectively. Thus, the parameter β_1 measures the additive genotype effect of marker genotype G_{ij}, and β_0 is a baseline parameter.

The model in Equation (5.1.1) is a GLMM, on the log(-log) scale, with a fixed genetic effect capturing the population level marker association, and a random transmission effect capturing the within family effect inherent to linkage between the marker and a DS locus. The next step in the model formulation involves the joint probability of the traits in the sibship and integrating over the random effects. Exactly how this is done depends on the design and on the marker genotype information, and whether or not the mode of inheritance is known and we thus present these details for each paper separately.

In paper I we start to address Aim (i) in a first development of the GRE model, based on a prospective likelihood. A LRT for single marker data is described. The LRT ignores the information from the parental genotypes in inferring the pattern of transmission (inheritance vector) and we do not account for population stratification. Missing parental genotypes are not accounted for. See Section 6.1.

In paper II we continue to address Aim (i), and develop a GRE score test for a single marker, based on the retrospective likelihood. The score test accounts for population stratification by conditioning the probability of offspring marker genotypes on the parental marker genotypes. The properties of the GRE are studied in terms of empirical power and type-I-error, using simulated data. See Section 6.1.

In paper III we generalize the single-marker GRE score for multiple markers (Aim (iii)). Based on the retrospective likelihood we develop a multi-marker score which accounts for missing parental genotypes. A haplotype test, as well as a single-marker test which, uses information from multiple markers to infer the pattern of transmission, is studied, using simulated data. See Section 6.2

In paper IV we extend the single-marker GRE to a bivariate trait setting. Focus is on the prospective likelihood. We describe a LRT and demonstrate that it protects against population stratification and is valid and powerful in testing single-marker trait association. See Section 6.3.

Chapter 6

THE PAPERS

We consider only sib-pairs in all four papers. The methods are however directly extendable to any family data. We noted in Section 1.1.1 that the information in the inheritance vector collapses to information on IBD sharing when considering sib-pairs. As a consequence, we can express the likelihoods (based on the GRE model) in terms of IBD sharing.

Note that we have adopted the notation in Section 5.1 in what follows. Deviations from the notation will be described in each Section.

6.1 Papers I and II - A Powerful GRE Score Test for a Single Trait and a single marker

In papers I and II we focus on the formulation of the transmission effects in Equation (5.1.1). The shared environmental random effect, ϵ_i^s is left out of the model. We reformulate the model in Equation (5.1.1) as

$$\log\left(-\log\left(q_{ij}\right)\right) = \log(\boldsymbol{a}_{ij} \cdot \boldsymbol{\epsilon}_i) + \beta_0 + X(G_{ij}) \cdot \beta_1 \ . \qquad (6.1.1)$$

If we assume that the traits of offspring in family i are independent, conditional on the random transmission effects and $\boldsymbol{G}_i$, then based on the model Equation (6.1.1) we can write the joint probability of trait in the sib pair as,

$$P(Y_{i1} = y_{i1}, Y_{i2} = y_{i2} | \boldsymbol{G}_i, \boldsymbol{\epsilon}_i) = P(\boldsymbol{Y}_i = \boldsymbol{y}_i | \boldsymbol{G}_i, \boldsymbol{\epsilon}_i)$$

$$= \prod_{j=1}^{2} q_{ij}^{y_{ij}} (1 - q_{ij})^{1-y_{ij}} \ .$$

We want to evaluate the prospective probability of trait, given the observed marker genotype data, $P(\boldsymbol{Y}_i|\boldsymbol{G}_i, \boldsymbol{g}_i)$ and write,

$$P(\boldsymbol{Y}_i = \boldsymbol{y}_i|\boldsymbol{G}_i, \boldsymbol{g}_i) = \sum_{\boldsymbol{v} \in \boldsymbol{v}_i^*} P(\boldsymbol{Y}_i = \boldsymbol{y}_i|\boldsymbol{G}_i, \boldsymbol{v}) P(\boldsymbol{v}|\boldsymbol{G}_i, \boldsymbol{g}_i) \ ,$$

where the summation is over all possible inheritance vectors, $\boldsymbol{v} \in \boldsymbol{v}_i^*$, given the observed genotype data, $\boldsymbol{g}_i$ and $\boldsymbol{G}_i$. We restrict the parameters of the random effects, $\lambda = \alpha/2$, so that the mean of $\log(\boldsymbol{a}_{ij} \cdot \boldsymbol{\epsilon}_i)$ is approximately zero. Following Conaway [15] (see Equation (2.2.1) in Section 2.2) we write,

$$P(\boldsymbol{Y}_i = \boldsymbol{y}_i|\boldsymbol{G}_i, \boldsymbol{v}_i) = \sum_{T \in \psi} c_T^{\boldsymbol{y}_i} P(Y_{ij} = 1, \ \forall j \in T|\boldsymbol{G}_i, \boldsymbol{v}_i)$$

$$= \sum_{T \in \psi} c_T^{\boldsymbol{y}_i} \int_{\boldsymbol{\epsilon}_i} P(\boldsymbol{Y}_i = 1, \ \forall j \in T|G_{ij}, \boldsymbol{\epsilon_i}) P(\epsilon_i) \partial \boldsymbol{\epsilon}_i$$

$$= \left(\frac{\lambda}{\lambda + \sum_{j \in T} \exp(\beta_0 + \beta_1 X_{ij})} \right)^{\pi\lambda} \cdot$$

$$\prod_{j=1}^{2} \left(\frac{\lambda}{\lambda + \exp(\beta_0 + \beta_1 X_{ij})} \right)^{(1-\pi)\lambda} \ , \qquad (6.1.2)$$

where π is the proportion of alleles shared IBD in the sib-pair, *i.e.* 0, 0.5 or 1, and where $P(\epsilon_i)$ is the gamma distribution function of ϵ_i.

6.1.1 The Application in Paper I

In paper I we studied a first version of the GRE, built on some simplifying assumptions. It is assumed that all inheritance vectors are equally likely, *i.e.* we ignore the information from the parental genotypes and set $P(\boldsymbol{v}|\boldsymbol{G}_i, \boldsymbol{g}_i) = P(\boldsymbol{v}|\boldsymbol{G}_i) = 1/16$. We based analysis on the prospective probability of trait, $P(\boldsymbol{Y}_i = \boldsymbol{y}_i|\boldsymbol{G}_i, \boldsymbol{g}_i)$ and formulate a likelihood ratio test. This will not be valid when families are ascertained on trait. Based on these assumptions, we formulated the prospective likelihood,

$$L(\beta_0, \beta_1, \lambda) = \prod_{i=1}^{n} P(\boldsymbol{Y}_i = \boldsymbol{y}_i | \boldsymbol{G}_i, \boldsymbol{g}_i)$$

$$= \prod_{i=1}^{n} \sum_{\boldsymbol{v}_i \in \boldsymbol{v}_i^*} \frac{1}{16} P(\boldsymbol{Y}_i = \boldsymbol{y}_i | \boldsymbol{G}_i, \boldsymbol{v}_i)$$

$$= \prod_{i=1}^{n} \sum_{\boldsymbol{v}_i \in \boldsymbol{v}_i^*} \frac{1}{16} \sum_{T \in \Psi} c_T^{y_{ij}} P(\boldsymbol{Y}_i = 1 | \boldsymbol{G}_i, \boldsymbol{v}_i) \ , \qquad (6.1.3)$$

where $\boldsymbol{G}_i$ and $\boldsymbol{g}_i$ refers to single biallelic marker data from the offspring and parents in family i, and where $\boldsymbol{v}_i^*$ refers to all (16) possible inheritance vectors.

Testing the null hypothesis of no association in the presence of linkage corresponds to testing $\beta_1 = 0$ in the likelihood represented in Equation (6.1.3). In Jonasdottir *et al.* [26] we propose to use a likelihood ratio test,

$$LRT = -2 \left(\log L(\hat{\beta}_0, \beta_1 = 0, \hat{\lambda}) - \log L(\hat{\beta}_0, \hat{\beta}_1, \hat{\lambda}) \right) \ , \qquad (6.1.4)$$

where $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\lambda}$ are maximum likelihood estimates of the GRE model parameters. The likelihood ratio test in Equation (6.1.4) is chi-squared distributed with one degree of freedom.

Using the LRT we analyze data simulated for the use of the participants at the *14th Genetic Association Workshop* (GAW14). See Section 6.1.3 for results.

6.1.2 The Application in Paper II

In Jonasdottir *et al.* [24] we presented a score test based on the conditional prospective likelihood. Using Bayes rule the retrospective probability of trait can be written in terms of prospective probabilities,

$$L = \prod_{i=1}^{n} \frac{P(\boldsymbol{Y}_i = \boldsymbol{y}_i | \boldsymbol{G}_i, \boldsymbol{g}_i) P(\boldsymbol{G}_i | \boldsymbol{g}_i)}{\sum_{\boldsymbol{G} \in \boldsymbol{G}_i^*} P(\boldsymbol{Y}_i = \boldsymbol{y}_i | \boldsymbol{G}, \boldsymbol{g}_i) P(\boldsymbol{G} | \boldsymbol{g}_i)} \ , \qquad (6.1.5)$$

where $\boldsymbol{Y}_i$ refers to the trait vector in family i of a single trait. We focused on the biallelic case; $\boldsymbol{G}_i$ and $\boldsymbol{g}_i$ refers to (biallelic) genotype information in the offspring and parents, respectively. In the single marker case the probability $P(\boldsymbol{G}|\boldsymbol{g}_i)$ is simply derived by enumerating all possible offspring genotypes given the observed parental genotypes, and the set of all such genotypes is referred to as $\boldsymbol{G}_i^*$.

The conditional retrospective likelihood in Equation (6.1.5) accounts for population stratification by conditioning on parental genotypes and it accounts for non-random ascertainment by conditioning on disease status. We derived a score test based on this likelihood. Both Zhong & Li [67] and Jonasdottir *et al.* [24] note that, in order to get consistent estimation of λ, using the conditional retrospective likelihood, β_0 has to be calculated from external data. Jonasdottir *et al.* [24] estimate β_0 by taking the log(-log) of the population prevalence of disease and demonstrate the validity of this choice. Under the null hypothesis β_1 is zero, so an appropriate score is given by,

$$S = \sum_{i=1}^{n} \frac{\partial}{\partial \beta_1} \log \left(L_i(\widetilde{\beta}_0, 0, \hat{\lambda}) \right)$$

$$= \sum_{i=1}^{n} \left(\frac{\sum_{\boldsymbol{v} \in \boldsymbol{v}_i^*} \frac{\partial}{\partial \beta_1} \left(P(\boldsymbol{Y}_i = \boldsymbol{y}_i | \boldsymbol{G}_i, \boldsymbol{v}) \right) P(\boldsymbol{v}|\boldsymbol{G}_i, \boldsymbol{g}_i) P(\boldsymbol{G}_i|\boldsymbol{g}_i)}{\sum_{\boldsymbol{v} \in \boldsymbol{v}_i^*} P(\boldsymbol{Y}_i = \boldsymbol{y}_i | \boldsymbol{G}_i, \boldsymbol{v}) P(\boldsymbol{v}|\boldsymbol{G}_i, \boldsymbol{g}_i) P(\boldsymbol{G}_i|\boldsymbol{g}_i)} - \right.$$

$$\left. - \frac{\sum_{\boldsymbol{G} \in \boldsymbol{G}_i^*} \sum_{\boldsymbol{v} \in \boldsymbol{v}_i^*} \frac{\partial}{\partial \beta_1} \left(P(\boldsymbol{Y}_i = \boldsymbol{y}_i | \boldsymbol{G}, \boldsymbol{v}) \right) P(\boldsymbol{v}|\boldsymbol{G}, \boldsymbol{g}_i) P(\boldsymbol{G}|\boldsymbol{g}_i)}{\sum_{\boldsymbol{G} \in \boldsymbol{G}_i^*} \sum_{\boldsymbol{v} \in \boldsymbol{v}_i^*} P(\boldsymbol{Y}_i = \boldsymbol{y}_i | \boldsymbol{G}, \boldsymbol{v}) P(\boldsymbol{v}|\boldsymbol{G}, \boldsymbol{g}_i) P(\boldsymbol{G}|\boldsymbol{g}_i)} \right) ,$$

where $\widetilde{\beta}_0$ is equal to log(-log) of the population prevalence of the trait and $\hat{\lambda}$ is the maximum likelihood estimate of λ. The probability $P(\boldsymbol{Y}_i = \boldsymbol{y}_i|\boldsymbol{G}, \boldsymbol{v})$ is given by Equation (6.1.2). Using this score we tested for APL using simulated data as well as data from the Collaborative study on the Genetics of Alcoholism (COGA). See Section 6.1.3 for results.

6.1.3 Results

Data Simulated for the 14th Genetic Association Workshop

In paper I we analyzed simulated data from the GAW14 [21]. We analyzed one trait in a data set containing 10 (out of 100) replicates from a specific

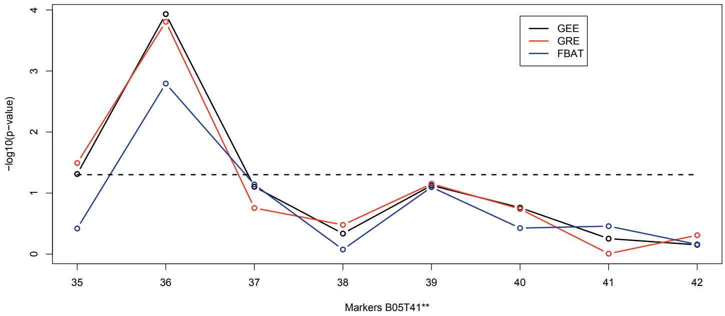


Figure 6.1.1: **Analysis of Region D2 in data simulated for the GAW14.** This region is simulated not to contain any DS loci. The dotted verticle line represents the p-value of 0.05.

subset of the GAW14 simulated data. The trait was labeled *Trait A* and was simulated to be associated with haplotypes in a specific region, entitled Region D3. We analyzed 20 markers from that region (B03T30**: 48 - 67), as well as 8 markers from another region, Region D2 (B05T41**: 35 - 42), not associated with *Trait A*. We compared the results of the first GRE presented in Section 6.1.1, with a GEE analysis, based on a logistic model, and a FBAT analysis, using the optimal offset and a robust variance estimator [31].

In our analysis of SNPs in the non-associated Region D2 the GRE followed the FBAT and the GEE test reasonably closely, see Figure 6.1.1. All tests falsely detected marker B05T4136 as being associated with *Trait A*; the GEE and the GRE with a p-value approximately 10 times smaller than the FBAT.

Analysis of the associated Region D3 shows that the GRE outperforms the GEE and the FBAT in pinpointing certain markers (Figure 6.1.2); see for example markers B03T3056 and B03T3059.

We note that the GRE used in paper I did not properly account for the fact that only families with at least one affected sib had been selected, and it did not protect against population stratification. Also the test assumed that all inheritance vectors are equally likely which is not optimal in terms of power. These issues are addressed in paper II.

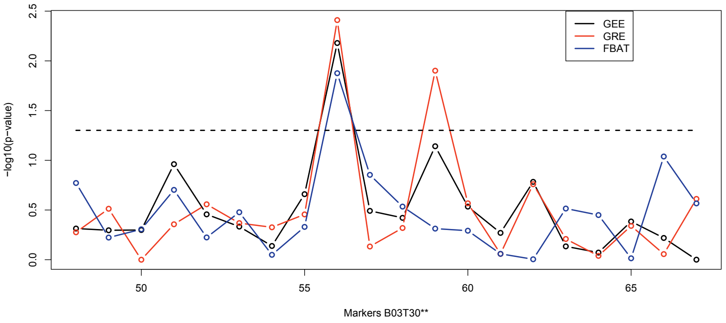


Figure 6.1.2: **Analysis of region D3 in data simulated for the GAW14.** This region is simulated to contain a haplotype DS locus. The dotted verticle line represents the p-value of 0.05.

Simulated Data using SIMLA

In paper II, we studied the properties of the GRE score test, presented in Section 6.1.2, using simulated genetic family data obtained by using the SIMLA software [6]. We ascertained families with sib pairs discordant for trait. The DS locus, as well as the observed marker locus was assumed to be biallelic. Let f_k denote the probability of disease, *the penetrance*, given k ($=$ 0, 1, or 2) DS alleles. The DS locus was assumed to have either a co-dominant effect on disease, with $f_0 = 0.004$, $f_1 = 0.008$ and $f_2 = 0.016$, or a dominant effect on disease, with $f_0 = 0.004$, $f_1 = 0.008$ and $f_2 = 0.008$. The frequency of the DS alleles and marker alleles were assumed equal. We varied two parameters in our simulations: (i) the disease and marker allele frequencies and (ii) the strength of LD between the marker locus and the DS locus. We assumed zero recombination between the marker and the DS locus, and the number of families, in each replicate, was set to 1000.

For each scenario, defined by (i) and (ii) we calculated the GRE score and the FBAT [50] score, for the marker locus. The GRE is calculated using log(-log) of the simulated population prevalence for the β_0 parameter. FBAT is calculated using an optimal offset of 0.5 and using a robust variance estimator [31]. The simulation of families and calculation of the scores was replicated 1,000 times and the proportion of score statistics exceeding the nominal 0.05 percentile of the chi-squared distribution (with 1 degree of freedom) was recorded.

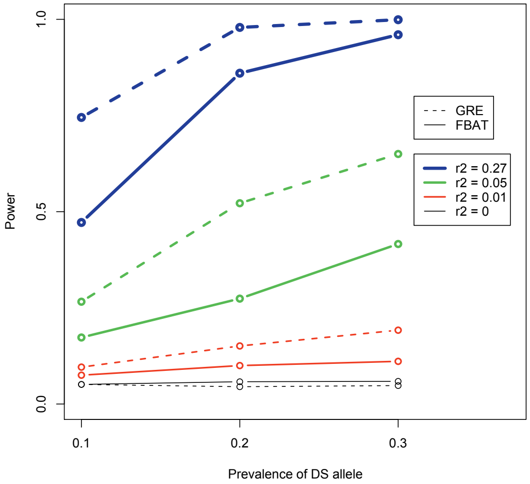


Figure 6.1.3: **Co-dominant Model:** Studying the validity and power of the GRE and the FBAT. Number of families=1000, number of simulations=1000, $\theta = 0$.

Figures 6.1.3 and 6.1.4 show results for the co-dominant model and the dominant model, respectively. In Figure 6.1.3 we plot the proportion of rejected null hypotheses based on the co-dominant model for various values of LD (in terms of the squared correlation coefficient r^2), ranging from 0 to 0.27. For no association ($r^2 = 0$) both the GRE and FBAT tests are valid. However, for all non-zero values of r^2, the GRE is more powerful than the FBAT. We see a similar result in Figure 6.1.4. The power is up to 50 % more powerful for the GRE than the FBAT.

We also tried varying the recombination fraction and the number of families, but these results were left out of the paper and are also left out here; they show as expected that power goes down when the recombination fraction is increased and that the power is approximately linearly related to the number of families in the data.

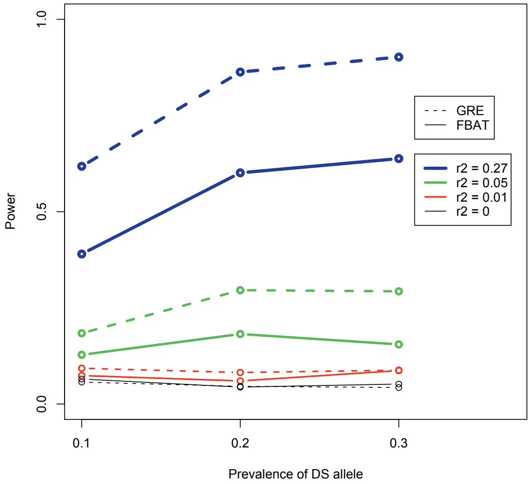


Figure 6.1.4: **Dominant Model:** Studying the validity and power of the GRE and the FBAT. Number of families=1000, number of simulations=1000, $\theta = 0$.

COGA

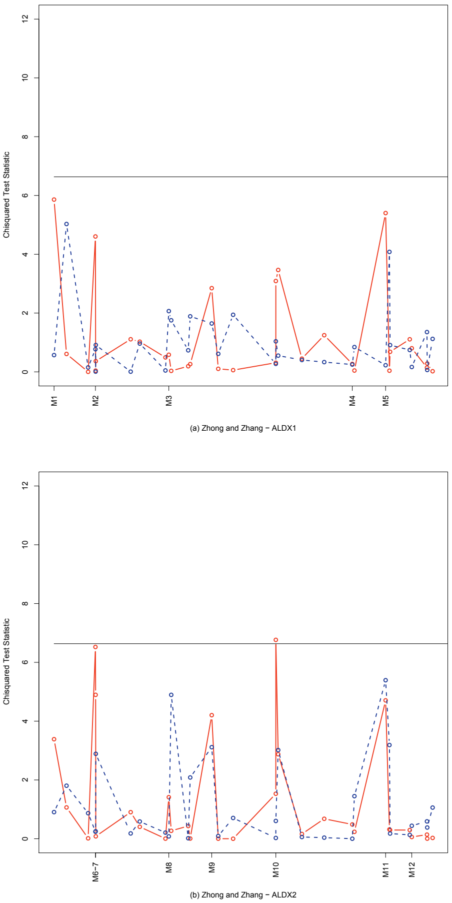
We analyzed data from the Collaborative Study on the Genetics of Alcoholism (COGA) which was made available to the participants of the GAW14 [17]. The general aim of COGA is to identify and characterize genes that affect the susceptibility to develop alcohol dependence. We focused on two clinical measures of alcoholism; ALDX1 that is based on the DSM-III-R (Diagnostic and Statistical Manual of the American Psychiatric Association-Revised) criteria for alcohol dependence and the Feigner criteria for alcoholism, ALDX2, based on the Diagnostic and Statistical Manual of the American Psychiatric Association-IV criteria. ALDX1 and ALDX2 have five categories: No Info, Pure Unaffected, Never Drank, Unaffected with some symptoms and Affected. We dichotomized ALDX1 and ALDX2 to focus on individuals with a clear diagnosis, and assigned the value of 1 to Affected individuals, and the value of 0 to all others, except "No info" which we treated as missing. As we note in paper II, one could argue against the relevance of this dichotomization; we simply made a pragmatic choice for illustration of the GRE. We investigated the association between the dichotomized ALDX1 and ALDX2 with markers which had been highlighted in two published analyzes of the GAW14 COGA data [68, 69].

The data consists of 143 pedigrees of varying size, in total including 1,614 individuals. As many as 136 families have four or more children. We selected one sib pair (with at lease one affected sib) per pedigree, at random, for subsequent analysis. A total of 304 individuals had no marker information and, overall, 22 % of the marker information was missing. Families with missing genotypes were not selected for analysis.

The GRE scores were compared with FBAT [50] scores, for each marker, using an optimal offset and an empirical variance estimator [31]. The p-values are not adjusted for multiplicity. Scores with a p-value smaller than 0.01 are considered significant.

For the GRE we obtained comparable results to the FBAT. The most notable result was found in the analysis of the Zhu *et al.* [69] data, where we obtained a strong signal at marker tsc0594280 (score = 12.305, p-value = 0.0048). One of the more important findings in Zhu *et al.*, marker tsc0593964, was not found association by FBAT or the GRE.

The discrepancies between the markers found (significantly associated with trait) in our analyzes and the markers found in the analyzes of Zhong & Zhang [68] and Zhu *et al.* [69] are partly attributable to differences in methodology and data analyzed. One major disadvantage of our analysis was that we could not include families with missing parental genotypes. This shortcoming was however addressed in paper III, and the COGA data were reanalyzed.



(a) Zhong and Zhang − ALDX1

(b) Zhong and Zhang − ALDX2

Figure 6.1.5: **Results from analysis of Zhong and Zhang [68] data using the GRE.** Each point represents a test of association between a specific marker and ALDX1/2. The full red line represents the GRE score and the dotted blue line represents the FBAT score. The straight horizontal line corresponds to the nominal 5 per cent chi-squared quantile (=3.84). The distance from the left most marker to the right most marker is approximately 13 mb. In figure (a) markers, M1-M5 refers to rs889826, rs1559534, rs273954, rs727714 and rs2056553, respectively. In figure (b) markers M6-M12 refers to rs1559534, rs2059367, rs273954, rs13068, rs768055, rs2056553 and rs700273, respectively.

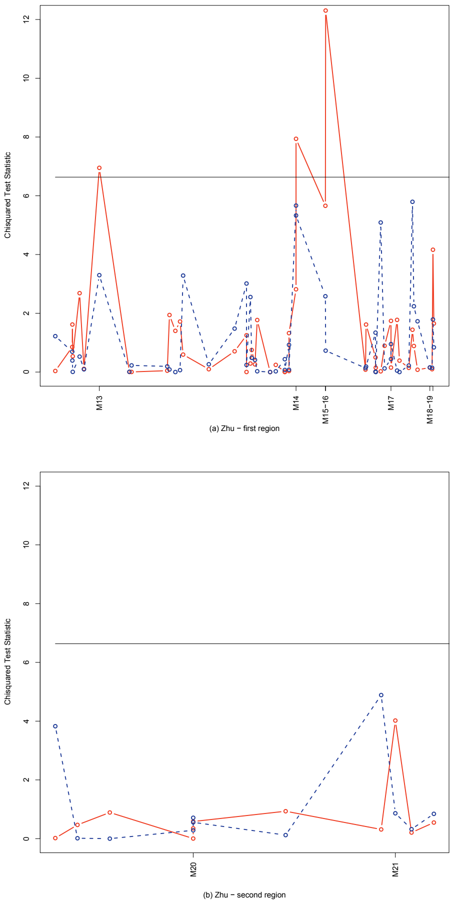


Figure 6.1.6: **Results from analysis of Zhu *et al.* [69] data using the GRE.** Each point represents a test of association between a specific marker and the ALDX1 measure of alcoholism. The full red line represents the GRE score and the dotted blue line represents the FBAT score. The straight horizontal line corresponds to the nominal 5 per cent chi-squared quantile (=3.84). The distance from the left most marker to the right most marker is approximately 11 mb in (a) and 320 kb in (b). In figure (a) markers M13- M19 refers to tsc0331830, tsc0018712, tsc0593964, tsc0594280, tsc0042959, tsc0051325 and tsc0505383, respectively. In figure (b) M20 and M21 refers to tsc0082737 and tsc0109702, respectively.

6.2 Paper III - A Multi-Marker Extension of the GRE

One disadvantage of the GRE test presented in paper II [24] is that only information from one marker is considered in the test. In paper III [23] we considered two ways to use information from multiple, m markers. We proposed a GRE for testing association between disease status and m-marker haplotypes, and a single marker GRE test which uses surrounding markers to infer the inheritance vector. This approach naturally incorporates efficient resolution of missing parental genotype information.

The likelihood in Equation (6.1.5) and the model in Equation (5.1.1) is modified to allow adjacent markers to give additional information to the distribution of inheritance vectors. The test, which we refer to as the I-GRE, is a single marker association test, although it uses multi-point information in the specification of correlation structure (linkage). For simplicity we consider nuclear families with two offspring, genotyped for a set of m markers, which are close enough to assume a zero recombination fraction between the markers.

The notation previously introduced in Section 5.1 needs to be adapted to the multi-marker case. To distinguish single marker and multi-marker genotypes we add a superscript to $\boldsymbol{g}_i$ and $\boldsymbol{G}_i$, containing the index numbers of the markers considered. To simplify further we add the superscript *all* when all observable markers are considered. The notation also needs to distinguish the unphased genotype data from the phased genotype, or haplotype, data; the vector of offspring haplotype pairs is denoted by $\boldsymbol{H}_i = \{\boldsymbol{H}_{i1}, \boldsymbol{H}_{i2}, \ldots \boldsymbol{H}_{iJ_i}\}$ and parental haplotypes are denoted by $\boldsymbol{h}_i = \{\boldsymbol{h}_{i1} \times \boldsymbol{h}_{i2}\}$, where the cross sign indicates a *mating type*. As with the genotypes, a superscript is added, when needed, to index the markers considered in the haplotype.

6.2.1 Phase Uncertainty and the FAMHAP Algorithm

The denominator in Equation (6.1.5) is a sum over all genotypes $\boldsymbol{G}_i^*$ that could have been transmitted from parent to offspring. It is simple to enumerate the genotypes of $\boldsymbol{G}_i^*$ in the single-marker situation, but it is more complicated when haplotypes are considered, since phase needs to be considered.

Uncertain phase in the offspring genotypes: Consider two markers and a nuclear family; parents with genotypes $\{AA.BB\}$ and $\{Aa.bb\}$, and offspring with genotypes $\{Aa.Bb\}$ and $\{AA.Bb\}$. With help from the parental genotypes (assuming no recombination between the two loci) we can infer the phase of the first offspring genotype. The first offspring carries haplotypes AB and ab, and the second one carries haplotypes AB and Ab. We can infer that the offspring share one haplotype IBD; the AB haplotype from the first parent. If we only consider the genotype from locus $\mathbf{A}$ (with alleles A and a), we can only infer that offspring share 0 or 1 alleles IBD, with equal probability.

Uncertain phase in the parental genotypes, as well as the offspring genotypes: Consider an example with three biallelic markers. If parents have genotypes $\{AA.bb.Cc\}$ and $\{AA.Bb.Cc\}$, and offspring have genotypes $\{AA.Bb.Cc\}$ and $\{AA.bb.Cc\}$, then the phase of the second offspring can be determined unambiguously, *i.e.* $\{AbC/Abc\}$ whereas the first offspring can carry either $\{ABC/Abc\}$ or $\{AbC/ABc\}$. The phase from the first parent can be inferred unambiguously, $\{AbC/Abc\}$, whereas the genotype of the second parent has two possible phases. That is, two possible mating types, or *haplotype explanations* exist [8]; $\{AbC/Abc\} \times \{ABC/Abc\}$ or $\{AbC/Abc\} \times \{ABc/AbC\}$. Given either haplotype explanation, we can infer the phase of the first offspring and that the offspring share no alleles IBD. If, on the other hand, we consider only the alleles of the 2nd marker, then we can only infer that the offspring share either one or none allele IBD.

Determining which genotypes ($\boldsymbol{G}_i^*$) to include in the summation: Consider the scenario depicted in Figure 6.2.1 with two biallelic SNPs with alleles A/a and B/b, respectively. Using the information on the child's haplotypes, both parents have phased genotype $\{AB/ab\}$. Given this knowledge, the genotypes that could have been transmitted from the parents to an offspring, including the observed genotype, are $\{\{AA.BB\}, \{Aa.Bb\}, \{aa.bb\}\}$. However, the transmission leading to offspring genotype $\{Aa.Bb\}$ will not be allowed in $\boldsymbol{G}_i^*$; the reason is that it would not have been possible to infer the same haplotype phase of the parents and the child as from the observed data. In summary, we allow only $\{ab/ab\}$ as additional genotype in the summation in the denominator of Equation (6.1.5), as it is the only offspring genotype that leads to the same set of inferred phased genotypes as the actually observed child genotype $\{AB/AB\}$. Generally, in the denominator of the likelihood in Equation (6.1.5),

(i) we allow only such offspring genotypes that lead to the same set of possible,

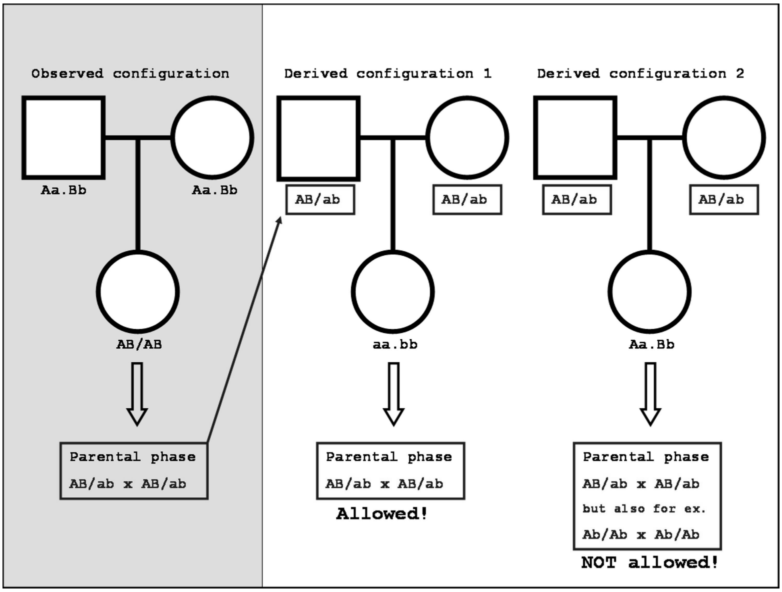


Figure 6.2.1: An example of the algorithm which determines the genotypes in set $\boldsymbol{G}_i^*$. See text for further description.

phased genotypes as the actually observed offspring genotypes.

(ii) we require that for each phased multi-marker genotype configuration of the parents all possible un-phased offspring genotype configurations have the same transmission probability.

$\boldsymbol{G}_i^*$ is thus the set of offspring genotypes that fulfill (i) and (ii), given the parental genotypes. Note that our condition (i) is equivalent to steps 1 and 2, and that (*ii*) is equivalent to step 4, in the haplotype extension [22] of the RL-algorithm [50], described in Section 3.1.3.

We have generalized the procedure described above to a systematic approach that uses an extension of FAMHAP [8]. FAMHAP computes ML haplotype frequency estimates and uses them to obtain, for each nuclear family, a list of likelihood weighted haplotype explanations for the parents, together with the possible transmission patterns to all of the offspring. In particular, these lists contain the maximal information on IBD status that can be obtained from the joint distribution of the m markers. The probabilities for the different IBD values derived from the haplotype distribution can then be used in the calculation of the GRE test.

6.2.2 The I-GRE and H-GRE model

One aim of our work was to extend the single-marker GRE model to be able to test for multi-marker, haplotype, association and to incorporate information from multiple marker into the determination of inheritance patterns. Given the phase of the parental genotypes and the inheritance vector, we may rewrite model in Equation (6.1.1) as,

$$\log(-log(q_{ij})) = \log(\boldsymbol{a}_{ij}^{\{K_1\}} \cdot \boldsymbol{\epsilon}_i) + \beta_0 + X_{ij}^{\{K_2\}} \cdot \beta_1 \ , \tag{6.2.1}$$

where $\boldsymbol{a}_{ij}^{\{K_1\}}$ now refers to a possibly multi-point patterns of inheritance, for the $\{K_1\}$ set of markers, derived using the FAMHAP[8] algorithm described in Section 6.2.1. That is to say, we use the information from all markers in the inference about IBD sharing. $X_{ij}^{\{K_2\}}$ refers to either the single-marker genotype score or a haplotype score for the $\{K_2\}$ set of markers.

In I-GRE, all (or a subset of all) markers contribute to the information in the IBD sharing, $\boldsymbol{a}_{ij}^{\{K_1\}} = \boldsymbol{a}_{ij}^{\{all\}}$, whereas the fixed effect is specified in terms of a single-marker allelic count $X_{ij}^{\{K_2\}} = X(G_{ij}^{\{k\}})$, where k represents the marker being tested ($k = 1, \ldots, m$). In the H-GRE, in addition to using all markers in the specification of the IBD sharing, fixed effects are specified for phased multi-marker haplotypes, $X_{ij}^{\{K_2\}} = X(H_{ij}^{\{all\}})$. For the haplotype test we compare the haplotype of interest against all other haplotypes, *i.e.* we let $X(H_{ij})$ count the number of occurrences of a specific haplotype, The FBAT haplotype test, the HBAT, also tests one haplotype versus all other haplotypes. Comparisons of H-GRE versus HBAT are therefore meaningful.

6.2.3 The Multi-Marker Likelihood and Score

Consider the prospective probability of the trait, given the observed m marker genotypes, $P(\boldsymbol{Y}_i = \boldsymbol{y}_i | \boldsymbol{G}_i, \boldsymbol{g}_i)$. For the observed genotypes, given haplotype frequency estimates from FAMHAP [8], we can infer the possible sets of parental haplotype configurations and their corresponding probabilities. Let $\boldsymbol{h}$ denote a specific parental haplotype configuration and let $\boldsymbol{h}_i^*$ denote the set of possible parental haplotype configurations, given the observed, m marker genotypes.

For simplicity of exposition we drop the superscripts identifying the markers tested. We write,

$$
\begin{aligned}
P(\boldsymbol{Y}_i = \boldsymbol{y}_i | \boldsymbol{G}_i, \boldsymbol{g}_i) &= \sum_{\boldsymbol{h} \in \boldsymbol{h}_i^*} P(\boldsymbol{Y}_i = \boldsymbol{y}_i | \boldsymbol{G}_i, \boldsymbol{g}_i, \boldsymbol{h}) \cdot P(\boldsymbol{h} | \boldsymbol{G}_i, \boldsymbol{g}_i) \\
&= \sum_{\boldsymbol{h} \in \boldsymbol{h}_i^*} P(\boldsymbol{h} | \boldsymbol{G}_i, \boldsymbol{g}_i) \cdot \left(\sum_{\boldsymbol{v} \in \boldsymbol{v}_i^*} P(\boldsymbol{Y}_i = \boldsymbol{y}_i | \boldsymbol{G}_i, \boldsymbol{v}) \cdot P(\boldsymbol{v} | \boldsymbol{G}_i, \boldsymbol{h}) \right) \ ,
\end{aligned}
$$

and

$$
\begin{aligned}
&\frac{\partial}{\partial \beta_1} P(\boldsymbol{Y}_i = \boldsymbol{y}_i | \boldsymbol{G}_i, \boldsymbol{g}_i) \\
&= \sum_{\boldsymbol{h} \in \boldsymbol{h}_i^*} P(\boldsymbol{h} | \boldsymbol{G}_i, \boldsymbol{g}_i) \cdot \left(\sum_{\boldsymbol{v} \in \boldsymbol{v}_i} \frac{\partial}{\partial \beta_1} \left(P(\boldsymbol{Y}_i = \boldsymbol{y}_i | \boldsymbol{G}_i, \boldsymbol{v}) \right) \cdot P(\boldsymbol{v} | \boldsymbol{G}_i, \boldsymbol{h}) \right) \ ,
\end{aligned}
$$

were $P(\boldsymbol{Y}_i = \boldsymbol{y}_i | \boldsymbol{G}_i, \boldsymbol{v}) = \sum_{T \in \Psi} c_T^{y_{ij}} P(Y_{ij} = 1, \ \forall j \in T | \boldsymbol{G}_i, \boldsymbol{v}_i)$, with $P(Y_{ij} = 1, \ \forall j \in T | \boldsymbol{G}_i, \boldsymbol{v}_i)$ derived from the model in Equation (6.2.1), leading to the same form as the probability in Equation (6.1.2), the only difference being the form of the genotype score (X_{ij} in Equation (6.1.2)). We continue by formulating a score test in the same way as for the original GRE leading to a 1 degree of freedom chi-squared test.

6.2.4 Results

Simulated data

We generated data containing one DS locus and three biallelic marker loci. Let p_D denote the frequency of the DS allele D and let f_0, f_1 and f_2 denote the probabilities (the *penetrance*) of being affected when carrying 0, 1 or 2 D alleles. In the data generation we have varied $p_D = \{0.1, 0.2, 0.3\}$ and we have considered two penetrance scenarios; a co-dominant disease model with $f_0 = 0.004$, $f_1 = 0.008$ and $f_2 = 0.016$, and a dominant disease model with $f_0 = 0.004$ and $f_1 = f_2 = 0.008$. Families with one affected and one unaffected sibling were ascertained. The DS locus was treated as unobserved and we generated three biallelic markers, all with alleles denoted 1 and 2, completely linked to the DS locus. We have generated data sets of nuclear families under a single-marker model (I) and under a haplotype model (II):

(I) Single-Marker Model: Marker 2 was set to have the strongest LD with the DS locus. We set the frequency of allele 1 (at Marker 2) equal to p_D and considered r^2 values of $\{0, 0.01, 0.05, 0.27\}$ between Marker 2 and the DS locus. The distribution of markers 1 and 3 was set up so that the r^2 between the markers and the DS locus is zero, which ensures that the markers are maximally informative for the inheritance vectors.

(II) Haplotype Model: Haplotype 111 was set to have an increased risk of disease, mimicking a cis-acting effect. Only four other alleles were allowed; 112, 121 and 211 (the $\overline{111}$ haplotypes). We define r^2 in terms of the dichotomized distribution of haplotype 111 versus $\overline{111}$ against the alleles at the DS locus, and let r^2 take the same values as under model (I). This haplotype model is chosen to mimic the situation of a haplotype block that has not been subject to historical recombination events.

We simulated (100 or 1000) data sets and calculate the proportion exceeding the nominal 0.05 percentile of the chi-squared distribution (in this case 3.84). We let this proportion estimate power when the data has been simulated under $r^2 > 0$, and type-I-error when $r^2 = 0$.

We use the I-GRE to analyze the data simulated under model (I) and compare the power and Type-I-error of the I-GRE with the original GRE, and with the FBAT. The FBAT analysis was performed using an optimal offset of 0.5 and an empirical variance estimator [31]. Results from simulation of the co-dominant disease model can be found in Table 6.2.1 and results from for the dominant disease model can be found in Table 6.2.2. As in the similar simulation study in paper I (Figures 6.1.3 and 6.1.4), the I-GRE outperforms FBAT for all scenarios $r^2 > 0$. The increase is found to be even more pronounced, at $p_D = 0.1$ and $r^2 = 0.05$, where the power of the GRE compared to the FBAT is doubled (Table 6.2.1). We find, however, no consistent evidence of I-GRE being an improvement over the original GRE. We failed to see an improvement of power of the I-GRE compared to the GRE, even after increasing the number of replicates to 1000 (marked by * in Table 6.2.1). See Table 6.2.1 for more results. The results in Table 6.2.2 suggest that the power of the I-GRE is well maintained even under misspecification of the genotype score $X(\cdot)$. This is also supported by the results for the single-marker GRE, see Figure 6.1.4. However, the type-I-errors of the I-GRE under the dominant model, presented in Table 6.2.2, also possibly suggest that the type-I-error of the I-GRE (and the GRE) is too small.

p_D	r^2	FBAT	GRE	I-GRE
0.1	0	0.061*	0.049*	0.044*
0.1	0.01	0.07	0.09	0.09
0.1	0.05	0.17	0.36	0.34
0.1	0.27	0.833*	0.888*	0.882*
0.2	0	0.065*	0.045*	0.040
0.2	0.01	0.125*	0.145*	0.156*
0.2	0.05	0.44	0.66	0.63
0.2	0.27	0.963*	0.968*	0.964*
0.3	0	0.055*	0.057*	0.038*
0.3	0.01	0.11	0.17	0.16
0.3	0.05	0.42	0.67	0.71
0.3	0.27	0.98	0.99	0.99

Table 6.2.1: **Empirical power and type-I-error estimates for data simulated under model I.** The penetrance values are 0.004/0.0/0.016 for 0/1/2 copies of the DS locus allele. (*) indicates that 1000 replicates have been used, all other simulation results based on 100 replicates.

p_D	r^2	FBAT	GRE	I-GRE
0.1	0	0.050*	0.051*	0.038*
0.1	0.05	0.20	0.33	0.34
0.2	0	0.039*	0.032*	0.029*
0.2	0.05	0.24	0.61	0.65
0.3	0.	0.048*	0.041*	0.041*
0.3	0.05	0.18	0.71	0.70

Table 6.2.2: **Empirical power and type-I-error estimates for data simulated under model I.** The penetrance values are 0.004/0.008/0.008 for 0/1/2 copies of the DS locus allele. (*) indicates that 1000 replicates have been used, all other simulation results based on 100 replicates.

Only the co-dominant disease model was used in the simulation under the haplotype model (II). We analyzed the data with the H-GRE and the HBAT using an optimal offset of 0.5 and an empirical variance estimator. The H-GRE

p_D	r^2	HBAT	H-GRE
0.1	0	0.04	0.04
0.1	0.01	0.08	0.14
0.1	0.05	0.36	0.41
0.1	0.27	0.84	0.98
0.2	0	0.05	0.02
0.2	0.01	0.09	0.13
0.2	0.05	0.32	0.62
0.2	0.27	0.98	0.98
0.3	0	0.02	0.04
0.3	0.01	0.11	0.19
0.3	0.05	0.40	0.61
0.3	0.27	0.96	0.99

Table 6.2.3: **Empirical power and type-I-error estimates for data simulated under model II.** All simulation results based on 100 replicates.

is consistently more powerful than the HBAT; up to two times as powerful ($p_D = 0.2$ and $r^2 = 0.05$). See Table 6.2.3 for more results.

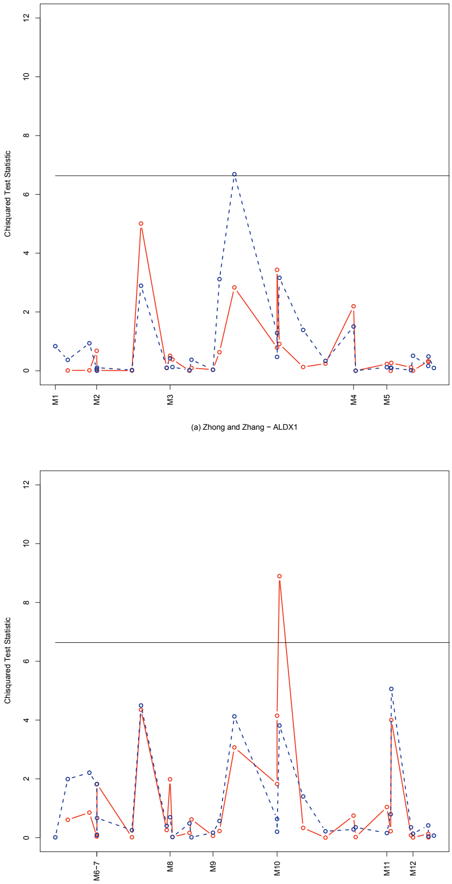
COGA data

We reanalyzed the GAW14 COGA data analyzed in paper II [24] using the I-GRE, this time including families with missing parental genotypes. The same markers were analyzed as in the original analysis, and one marker at each side of the tested marker was used to increase information about allele sharing and missing parental genotypes. We chose not to test the markers at the end of the regions, since they have only one adjacent marker. We also excluded two markers which were too far apart to assume zero recombination; an assumption in the formulation of the I-GRE is that no recombination occurs between the markers included in the analysis. This assumption is also made in the HBAT [22] test.

In Zhong & Zhang [68] association to age-at-onset of ALDX1 and ALDX2 is tested using the Zhong & Li score test [67] presented in Section 3.1.3. We were not able to replicate any of the results from Zhong & Zhang [68]; see Figure

6.2.2. A plausible reason for this inconsistency is the difference in choice of phenotype in our analysis and in the analysis of Zhong & Zhang [68].

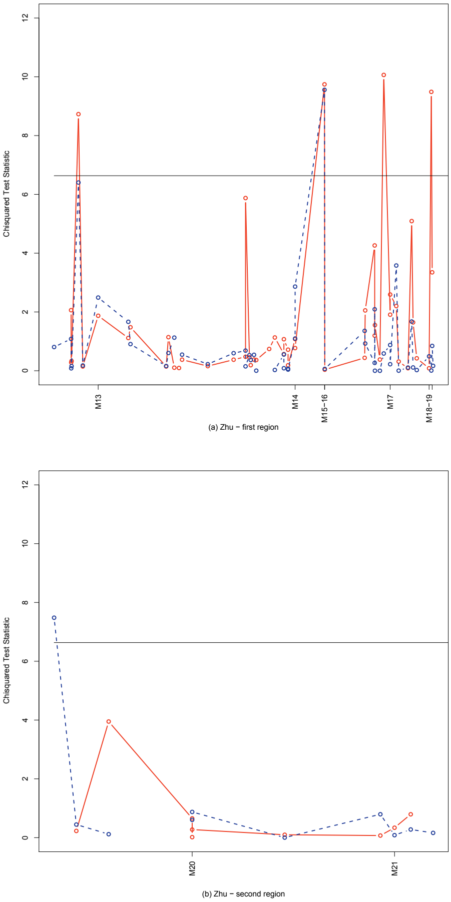
In Zhu *et al.* [69] FBAT was used to test the association between a dichotomized ALDX1 and markers in large families. We were able to replicate the major finding in Zhu *et al.* [69]; tsc0593964 is associated with ALDX1 in both the FBAT ($\chi^2 = 9.56$, p-value = 0.0020) and in the I-GRE ($\chi^2 = 9.74$, p-value = 0.0018) test, compared with the p-value of 0.00328 in Zhu *et al.* [69]. The GRE test in Jonasdottir *et al.* [24] found the neighboring marker tsc0594280 significantly associated, not tsc0593964; see Figure 6.1.6. The I-GRE test of marker tsc0229629, not tested by Zhu *et al.* [69], is also highly significant ($\chi^2 = 10.06$ p-value = 0.0015). This result was not duplicated by FBAT. See Figure 6.2.3.



(a) Zhong and Zhang − ALDX1

(b) Zhong and Zhang − ALDX2

Figure 6.2.2: **Results from analysis of Zhong and Zhang [68] data using the I-GRE.** Each point represents a test of association between a specific marker and ALDX1/2. The full red line represents the GRE score and the dotted blue line represents the FBAT score. The straight horizontal line corresponds to the nominal 5 per cent chi-squared quantile (=3.84). The distance from the left most marker to the right most marker is approximately 13 mb. In figure (a) markers, M1-M5 refers to rs889826, rs1559534, rs273954, rs727714 and rs2056553, respectively. In figure (b) markers M6-M12 refers to rs1559534, rs2059367, rs273954, rs13068, rs768055, rs2056553 and rs700273, respectively.



(a) Zhu − first region

(b) Zhu − second region

Figure 6.2.3: **Results from analysis of Zhu *et al.* [69] data using the I-GRE.** Each point represents a test of association between a specific marker and the ALDX1 measure of alcoholism. The full red line represents the GRE score and the dotted blue line represents the FBAT score. The straight horizontal line corresponds to the nominal 5 per cent chi-squared quantile (=3.84). The distance from the left most marker to the right most marker is approximately 11 mb in (a) and 320 kb in (b). In figure (a) markers M13- M19 refers to tsc0331830, tsc0018712, tsc0593964, tsc0594280, tsc0042959, tsc0051325 and tsc0505383, respectively. In figure (b) M20 and M21 refers to tsc0082737 and tsc0109702, respectively.

6.3 Paper IV - A Bivariate GRE Test

Paper IV is concerned with correlated traits. One example of correlated traits is blond hair and blue eyes. A somewhat less trivial example is depression and social phobia; depression is more common among individuals with social phobia than among those without social phobia [54, 51, 39].

In Jonasdottir *et al.* [25] we use the term co-morbidity to describe correlated traits sharing genetic and/or environmental pathways. Consider two diseases with a (partial) common genetic pathway. In Figure 6.3.1 (a) one DS locus has a direct causal effect on both diseases. In Figure 6.3.1 (b) the two diseases are causally affected by two different DS loci in close proximity to each other in terms of Linkage Disequilibrium. Given this model, the degree of correlation will depend on the allele frequencies of the DS loci, penetrance values and the degree of linkage and LD between the DS loci [25].

Three reasons can be identified for using a multivariate model instead of separate univariate models for each disease,

1. If the diseases share genetic background, then multivariate association testing [32] has increased power compared to the corresponding univariate tests
2. when ascertainment is on disease, univariate tests may not be valid in the presence of co-morbidity [51]
3. if a marker is associated and linked with multiple diseases, multivariate modeling might help in understanding functionality.

In Jonasdottir *et al.* [25] we focus on the scenario in Figure 6.3.1 (a) and extend the GRE to a test for association between two correlated traits and one marker. We will briefly address the scenario in Figure 6.3.1 (b) in the discussion.

6.3.1 The Bivariate extension of the GRE

In extension to the notation introduced in Section 5.1 we need to introduce notation referring to two traits. We denote disease status for offspring j in

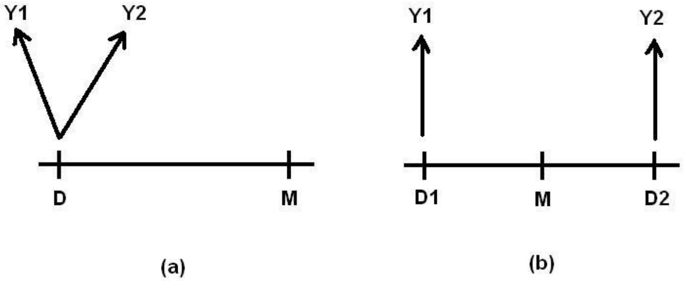


Figure 6.3.1: **Two scenarios of genetically co-morbid diseases:** In scenario (a) one DS locus, marker D, affects both diseases, Y_1 and Y_2. In scenario (b) separate DS loci, D_1 and D_2, affect the diseases, Y_1 and Y_2. Scenario (a) is a special case of scenario (b), where the distance between the two DS loci, D_1 and D_2, is zero. The marker, M is the marker tested and is assumed to be in LD with the DS loci.

family i by Y_{ijk}, where the index k (= 1 or 2) is used to refer to the two diseases.

In Jonasdottir *et al.* [24] and Jonasdottir *et al.* [23] it was noted that external data was needed to estimate the baseline parameter β_0. This problem may become even more complicated in a two disease scenario. We have therefore chosen to focus on the prospective scenario, with analysis based on the prospective likelihood. In papers I-III the shared environmental effect, ϵ_i^s, is removed. In paper IV we deal with both random transmission effects and shared environmental random effects.

The tests in papers II and III were based on the retrospective likelihood where population stratification was accounted for by conditioning the probability of offspring marker genotypes on parental marker genotypes. Baksh *et al.* [5] show that it is possible to control for population stratification in the prospective setting by allowing for mating type specific baseline parameters. In the biallelic case with genotypes denoted 0, 1, 2, there are 6 mating types; 0x0, 0x1, 0x2, 1x1, 1x2 and 2x2. Let $\beta_{0k}^{\{g\}}$ denote mating type specific baseline parameters, for disease k and mating type g. In the specific bivariate, biallelic, setting considered here, there are 12 baseline parameters (6 per disease). We write,

$$\log\left(-\log q_{ijk}\right) = \log(\boldsymbol{a}_{ij}\cdot\boldsymbol{\epsilon}_i + \epsilon_i^s) + \beta_{0k}^{\{g_i\}} + X(G_{ij})\beta_{1k} \ , \tag{6.3.1}$$

where $q_{ijk} = P(Y_{ijk} = 1|G_{ij}, \boldsymbol{g}_i, \boldsymbol{v}_i, \epsilon_i)$. Given the mating-types, the marker genotypes and the random effects, the offspring trait values are independent. Following the line of derivation as in paper I-III we obtain,

$$\begin{aligned}
&P(Y_{i11} = y_{i11}, Y_{i12} = y_{i12}, Y_{i21} = y_{i21}, Y_{i22} = y_{i22}|\boldsymbol{G}_i, \boldsymbol{v}_i) = \\
&\sum_{\{T_1,T_2\}\in\Psi} C_{T_1,T_2}^{\boldsymbol{y}_i}\cdot\left(\frac{\delta}{\delta+\sum_{k=1}^{2}\sum_{j\in T_k}\exp(\beta_{0k}^{\{g_i\}}+\beta_{1k}X_j)}\right)^{\eta} \\
&\quad\cdot\left(\frac{\lambda}{\lambda+\sum_{k=1}^{2}\sum_{j\in T_k}\exp(\beta_{0k}^{\{g_i\}}+\beta_{1k}X_j)}\right)^{\pi\alpha} \\
&\quad\cdot\prod_{k=1}^{2}\prod_{j\in T_k}\left(\frac{\lambda}{\lambda+\exp(\beta_{0k}^{\{g_i\}}+\beta_{1k}X_j)}\right)^{(1-\pi)\alpha} \quad ,
\end{aligned} \tag{6.3.2}$$

where T_1 and T_2 are the same sets of indices $T \in \psi$ (one per disease) as described in Section 2.2. The constants can be obtained by taking the croenecker product of two matrices, $\boldsymbol{B}$ (Section 2.2). The constants, for the sib pair case, are presented in Table 6.3.1. The first row of Equation (6.3.2) corresponds to the family specific shared effect and rows 2 and 3 correspond to the transmission specific effects, where π is proportion of alleles shared IBD; 0, 0.5 or 1 in the present sib pair case.

We wish to test parameters β_{11} and β_{12}. In this context $\delta, \eta, \lambda, \alpha$ and the baseline parameters are nuisance parameters and we denote them with Φ. The prospective likelihood is written,

$$L(\beta_{11}, \beta_{12}, \hat{\Phi}) = \prod_{i=1}^{n} P(\boldsymbol{Y}_{i1} = \boldsymbol{y}_{i1}, \boldsymbol{Y}_{i2} = \boldsymbol{y}_{i2}|\boldsymbol{G}_i, \boldsymbol{g}_i) \ .$$

We test the null hypothesis that both β_{11} and β_{12} are zero, versus that at least one is non-zero, by using a LRT. Two null hypotheses are considered; the type-I-hypothesis *no association or linkage* and the type-II-hypothesis *association in the presence of linkage.* The alternative hypothesis is *association and linkage.* Under the null, we estimate values of the nuisance parameters, $\hat{\Phi}$, assuming that β_{11} and β_{12} are zero. Under the alternative, we estimate β_{11} and β_{12}, as well as the nuisance parameters. The LRT is written,

Affection status				Constants $c^{\boldsymbol{y}_i}_{T_1,T_2}$ for Indexes Ψ: T_1 row 1; T_2, row 2															
				.	.	.	.	1	1	1	1	2	2	2	2	1,2	1,2	1,2	1,2
Sib 1		Sib 2		.	1	2	1,2	.	1	2	1,2	.	1	2	1,2	.	1	2	1,2
1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	-1
1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	-1
1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	-1	-1	1
0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	-1
0	1	0	1	0	0	0	0	0	0	0	0	0	0	1	-1	0	0	-1	1
0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	-1	0	-1	0	1
0	1	0	0	0	0	0	0	0	0	0	0	1	-1	-1	1	-1	1	1	-1
1	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	-1
1	0	0	1	0	0	0	0	0	0	1	-1	0	0	0	0	0	0	-1	1
1	0	1	0	0	0	0	0	0	1	0	-1	0	0	0	0	0	-1	0	1
1	0	0	0	0	0	0	0	1	-1	-1	1	0	0	0	0	-1	1	1	-1
0	0	1	1	0	0	0	1	0	0	0	-1	0	0	0	-1	0	0	0	1
0	0	0	1	0	0	1	-1	0	0	-1	1	0	0	-1	1	0	0	1	-1
0	0	1	0	0	1	0	-1	0	-1	0	1	0	-1	0	1	0	1	0	-1
0	0	0	0	1	-1	-1	1	-1	1	1	-1	-1	1	1	-1	1	-1	-1	1

Table 6.3.1: **The constants and the indexes needed in the summation in probability of trait, given the observed offspring and parental genotypes.**

$$LRT = -2 \cdot \left(\log L(0, 0, \hat{\Phi}) - \log L(\hat{\beta}_{11}, \hat{\beta_{12}}, \hat{\Phi}) \right) \ ,$$

which, under the null, is a chi-squared distributed with 2 degrees of freedom.

6.3.2 Properties of the bivariate GRE

We study the properties of the test under two scenarios: (i) population stratification and, (ii) linkage.

(1) Accounting for population stratification:

For the purpose of studying the properties of the GRE under population stratification we simplify the model in Equation (6.3.1) and remove the transmission effects (ϵ_i) from the model. The test will thus be a joint test of association and linkage. We set the shape and scale parameters of the shared random effect equal, $\delta = \eta$, so as to set the mean of $\log(\epsilon_i^*)$ to zero. We have the following simplified form from Equation (6.3.2),

$$P(\boldsymbol{Y}_{i1} = \boldsymbol{y}_{i1}, \boldsymbol{Y}_{i2} = \boldsymbol{y}_{i2} | \boldsymbol{G}_i, \boldsymbol{g}_i)$$
$$= \sum_{\{T_1,T_2\}\in\Psi} c^{\boldsymbol{y}_i}_{T_1,T_2} \cdot \left(\frac{\delta}{\delta + \sum_{k=1}^{2} \sum_{j\in T_k} \exp(\beta^{\{g_i\}}_{0k} + \beta_{1k} X_j)} \right)^{\delta} . \qquad (6.3.3)$$

(2) Accounting for linkage:

To study the properties of the bivariate GRE as a test of association in the presence of linkage we again simplify the model in Equation (6.3.1) by removing the shared environmental effect ϵ^s_i. We set $\lambda = \alpha/2$, so as to set the mean of $\log(\boldsymbol{a}_{ij} \cdot \boldsymbol{\epsilon}_i)$ to zero.

$$P(\boldsymbol{Y}_{i1} = \boldsymbol{y}_{i1}, \boldsymbol{Y}_{i2} = \boldsymbol{y}_{i2} | \boldsymbol{G}_i, \boldsymbol{g}_i)$$
$$= \sum_{\boldsymbol{v}_i \in \boldsymbol{v}^*_i} \left\{ \sum_{\{T_1,T_2\}\in\Psi} c^{\boldsymbol{y}_i}_{T_1,T_2} \cdot \left(\frac{\lambda}{\lambda + \sum_{k=1}^{2} \sum_{j\in T_k} \exp(\beta^{\{g_i\}}_{0k} + \beta_{1k} X_j)} \right)^{\pi\lambda} \cdot \right.$$
$$\left. \prod_{k=1}^{2} \prod_{j\in T_k} \left(\frac{\lambda}{\lambda + \sum_{j\in T_k} \exp(\beta^{\{g_i\}}_{0k} + \beta_{1k} X_j)} \right)^{(1-\pi)\lambda} \right\} P(\boldsymbol{v}_i | \boldsymbol{G}_i, \boldsymbol{g}_i) . \qquad (6.3.4)$$

Note that the GRE without mating-type specific baselines is obtained by assuming that, for disease k, all $\beta^{\{g\}}_{0k}$ are equal, thus only requiring estimation of 2 baseline parameters. We evaluate the test under both scenarios, including and excluding the mating type specific baseline parameters. In summary, we use four types of test:

(A) GRE for testing association and linkage (Equation (6.3.3)) *with* mating-type specific baseline parameters.

(B) GRE for testing association and linkage (Equation (6.3.3)) *without* mating-type specific baseline parameters.

(C) GRE for testing APL (Equation (6.3.4)) *with* mating-type specific baseline parameters.

(D) GRE for testing APL (Equation (6.3.4)) *without* mating-type specific baseline parameters.

6.3.3 Results

Population stratification and power

Population stratification: We simulated two population subgroups, assuming that they have different DS allele frequencies and different penetrance values. The DS allele frequency in population 1 and 2 are denoted by p_1 and p_2, respectively. We set $p_2 = 1 - p_1$ and let p_1 take values 0.9, 0.8, 0.7, 0.6 and 0.5. In population 2 we set the prevalence of the diseases to 1 minus the prevalences in population 1. Note that although we impose a severe difference between the prevalences of disease, the degree of population stratification is determined by the difference in allele frequencies between the populations (p_1 and p_2). That is, we simulate scenarios going from severe population stratification ($p_1 = 0.9$, $p_2 = 0.1$) to no population stratification ($p_1 = 0.5$, $p_2 = 0.5$).

No population stratification: The DS locus is assumed bi-allelic, with equal frequencies of the alleles ($= 0.5$). The DS locus is set to act additively on the probability of disease; the penetrance of disease 1 is $0.04 \cdot (1, 1+\gamma, 1+2\gamma)$ for 0, 1 and 2 a alleles, where γ take values 0, 0.1, 0.2, 0.3, 0.4, 0.5 and 1. The penetrance of disease 2 is $0.08 \cdot (1, 1+\gamma, 1+2\gamma)$, where the same values of γ are used.

The simulation scenarios are repeated 100 times and for or each replicated data set, we calculate the (A) and (B) GRE tests. The rejection rates (the proportion exceeding the nominal chi-squared quantile with 2 degrees of freedom, 5.99) recorded for each scenario. In null scenarios, the rejection rate is a measure of validity (type-I-error), and in alternative scenarios, the rejection rate is a measure of power.

With the mating type specific baseline effects the GRE remains valid through different levels of population stratification, ranging from mild to severe; see Table 6.3.2. Omitting the mating type specific baseline parameters leads to an increasing level of bias as the level of stratification increases, in terms of p_1 versus p_2 (Table 6.3.2). However, accounting for population stratification where there is none decreases the power of the test and the difference in power is consistent over all levels of association; see Table 6.3.3. The difference in power, between the test without mating-type specific baselines and the test with mating-type specific baselines, is at most twofold; Table 6.3.3.

Population stratification, under H_0:

p_1/p_2	*With*	*Without*
0.5/0.5	0.05	0.02
0.6/0.4	0.04	0.33
0.7/0.3	0.02	0.73
0.8/0.2	0.03	0.93
0.9/0.1	0.03	1.00

Table 6.3.2: **Validity of the GRE association test ((A) and (B)) under population stratification:** GRE results based on 100 replicates, *with* (see (A) in main text) and *without* (see (B) in main text) mating type specific baseline effects. p_1 and p_2 refers to the A allele frequencies in population 1 and 2, respectively. The a allele frequency is 1 minus the A allele frequency. In population 1, the penetrance is 0.04 for disease 1 and 0.08 for disease 2. The penetrance in population 2 is 1 minus the penetrance in population 1. The disease and markers are not associated.

No population stratification, under H_1:

γ	*With*	*Without*
0	0.04	0.03
0.1	0.11	0.20
0.2	0.36	0.50
0.3	0.49	0.77
0.4	0.65	0.96
0.5	0.85	0.99
1.0	1.00	1.00

Table 6.3.3: **Power of the GRE association test ((A) and (B)):** GRE results based on 100 replicates, *with* (see (A) in main text) and *without* (see (B) in main text) mating type specific baseline effects. p_1 and p_2 refers to the A allele frequencies in population 1 and 2, respectively. The a allele frequency is 1 minus the A allele frequency. The penetrance for disease 1 is $0.04/(0.04+\gamma)/(0.04+2\gamma)$ for 0/1/2 A alleles. The penetrance for disease 2 is $0.08/(0.08+\gamma)/(0.08+2\gamma)$ for 0/1/2 A alleles.

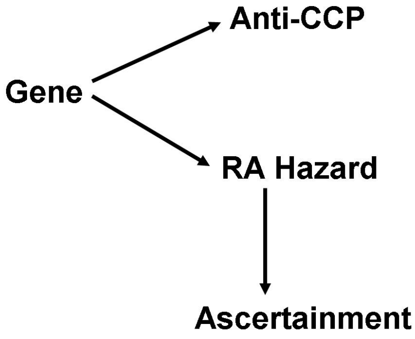


Figure 6.3.2: **The GAW15 simulated disease scenario.** The gene (locus C) affects both anti-CCP and the hazard of Rheumatoid Arthritis (RA), which in turn affects the age-at-onset of RA. The ascertainment is then on RA status.

Testing association in the presence of linkage

To evaluate the GRE APL test (C) and (D) we analyze data simulated to mimic a *Rheumatoid Arthritis* (RA) study, supplied to participants at the *15th Genetic Association Workshop* (GAW15) [43]. We focus on one of the simulated DS loci which was simulated to have a strong effect on the risk of RA; locus C on chromosome 6. Locus C is was simulated to be in complete LD (D' = 1) with another simulated DS locus, the HLA-DRB1 locus. HLA-DRB1 has a direct effect on anti-CCP (a diagnostic marker for RA). Thus, locus C is a common genetic marker for both RA risk and anti-CCP level.

A complicating feature of the data is that only sib pairs, where both are affected with RA, are ascertained. We have defined age-at-onset for RA and anti-CCP as the two traits of interest. The simulation model is graphically represented in Figure 6.3.2. Since families are ascertained on trait, estimation of genetic effects based on fitting the prospective GRE will be biased, although testing based on the GRE model will be valid [51].

Age-at-onset of RA and anti-CCP are observed to be correlated within individuals ($r = -0.15$, CI: [-0.18,-0.11]). We investigate the power of the bivariate GRE test using this example of co-morbidity. We dichotomize anti-CCP (low/high), based on a cut-off of 150, and age-at-onset (young/old) based on a cut-off at age 30. In the simulated data 72.5 % of the subjects are older than 30 and 49.4 % have an anti-CCP level higher than 150. A published estimate of the mean anti-CCP level in the population is 20 [18].

Each of the 100 simulated GAW15 replicates had 1500 sib pairs. We selected the first 400 from each replicate for analysis. We base our analysis on six marker loci (SNPs 150-155) from the *linkage SNP data* [43]. Locus C is located in the middle of this region, between SNPs 152 and 153. The GRE APL tests (C) and (D) were carried out for each replicate and the rejection rates on the nominal 5.99 level (chi-squared with 2 degrees of freedom) were recorded. We also calculate the level of LD (in terms of r^2) between the markers tested and markers 152 and 153, which flank locus.

The distance between the studied markers and locus C is small, which means that the region is tightly linked. Locus 152 and locus 153 are strongly associated with both age-at-onset and anti-CCP, see Table 6.3.4. From Table 6.3.4 we see that the power is high for the markers close to the true DS locus; markers 152 to 155. The other markers (150 and 151) are far from locus C and are not in LD with either marker 152 or marker 153 (all r^2 values are less than 0.05). The GRE results for these markers is close to the nominal 0.05 level, as should be expected. See Table 6.3.4 for more results.

We have also computed bivariate FBAT [35] scores for markers 150-155, using an optimal offset and a robust variance estimator. These results show that the power of the GRE score is higher than the power of the FBAT score on the markers close to to the DS locus (markers 152-155). It has however been noted that a comparison between the GRE LRT (based on the prospective likelihood) and the retrospective FBAT, is unfair. The results are therefor not presented here or in the paper.

	r^2		Distance	GRE	
SNP	152	153	to C	*With*	*Without*
155	0.010	0.100	287.555	0.62	0.63
154	0.380	0.588	366.290	1.00	1.00
153	0.222	-	14.817	1.00	1.00
152	-	0.222	-37.499	1.00	0.99
151	0.004	0.001	-948.504	0.10	0.06
150	0.020	0.003	-1390.619	0.06	0.06

Table 6.3.4: **Power of the GRE APL test ((C) and (D)):** GRE results, with and without mating-type specific baselines. In this table we present results from an empirical power study, based on data from 100 simulated data sets. For each simulated data set, we use the GRE APL bivariate test (anti-CCP and age-at-onset of RA) and marker SNPs 155 to SNP 150. A disease susceptibility locus (C) is located between SNPs 152 and 153. Columns 2 and 3 present the pair wise LD, as measured by the squared correlation coefficient (r2), between each SNP and SNP 152 and153, respectively. Column 4 lists the distances, in kilo base-pairs (kb), to locus C.

Chapter 7

Discussion

7.1 On Some Properties of the GRE

In this thesis we have presented statistical methods for testing genetic association in family-based studies. We have focused on testing association in the presence of linkage, but have also used an adaption of the GRE to test association and linkage jointly (paper IV). Both prospective (paper I and IV) and retrospective scenarios (paper II and III) have been considered and we have dealt with multiple markers (paper III), missing parental genotypes (paper III) and bivariate, comorbid, diseases (paper IV). We have used two different methods to deal with population stratification in the context of the GRE model, for the prospective and the retrospective study setting, respectively.

We have shown that the method is powerful and that it is *reasonably* valid; the empirical type I error of the GRE tends to be conservative (i.e. lower than the nominal level) for data simulated under the null hypothesis of no association; see *e.g.* Table 6.2.1. The GRE was, however, shown to be substantially more powerful than the FBAT. We have also demonstrated that population stratification can be accounted for properly in the prospective study setting, although (as expected) at the cost of a loss in power; see *e.g* Table 6.3.3.

Although not shown here, the GRE should be able to handle missing offspring genotypes when multiple markers are considered. An assumption underlying the I-GRE and the H-GRE is that the markers are tightly linked so that no recombination occurs between the markers. Using the surrounding markers it should be possible to infer the set of possible missing offspring genotypes. Given this information we could either impute the mean genotype score, $E(X(G_{ij}))$, or add a summation over possible offspring configurations. How this affects the power compared to removing the offspring with missing data from the data analyzed remains an open question.

In paper IV, where we considered a GRE to test genetic comorbidity, we focused on a scenario where one marker affects two diseases; see Figure 6.3.1 (a). Another plausible scenario when considering two diseases with a (partially) common genetic background is that two markers, in close proximity of each other, affects the diseases separately as depicted in Figure 6.3.1 (b). The GRE could be extended to include disease specific, correlated, transmission effects to account for the effect of linked DS loci. The bivariate gamma distribution could be considered for this purpose.

The GRE in the present implementation in the free statistical software **R** [48] is, however, very computer intensive. An implementation in another software may make the implementation of the GRE quicker. In the present form, FBAT wins by far in terms of computational run time.

7.2 Family-based versus population-based association testing/estimation

There are several *pros* and *cons* of family-based association testing/estimation, compared to population based association testing/estimation. Several authors have discussed the future role of family-based association testing/estimation; see *e.g.* Laird & Lange [30], Clerget-Darpoux & Elston [14] and Bourgain *et al.* [9].

Pros:

- Family-based association tests offers protection against population stratification; as was described in *e.g.* paper III and IV.
- If a marker locus is found associated in a family-based study that means that the marker locus is both linked and in LD with the DS locus.
- Family-based designs offer informative imputation of missing genotypes by use of existing data and information on family structures, as was shown in paper III.

- Family-based studies offer possibilities to separate genetic and environmental effect on disease, for example in twin-studies [45].
- It is possible to detect parent-of-origin effects in family-based studies, see *e.g.* Becker *et al.* [7].

Cons:

- In studies of late onset diseases it may be difficult to collect parental genotype information.
- The power of family-based association testing is lower than that of population -based association testing, in terms of power per genotyped individual, and is typically more time consuming.
- Many population-based methods for testing association, such a the logistic model for case-control data, are readily available in standard statistical software. Family-based association tests often require *home made* software.
- Family-based association tests are often more computer intensive than population-based association tests. This is often a minor issue, however.

7.3 Some other important contributions to the field

The arrival of genome-wide association studies, in which hundreds of thousands of SNPs are typed in study subjects has lead to the discovery of new DS loci. Replication of genetic association are difficult, but major breakthroughs in 2007 have confirmed the common disease common variant hypothesis [47].

Simple multiplicity corrections, such as the Bonferroni correction, suffer from assuming independence between tests; tests based on neighboring markers are not independent. Two stage strategies for selecting the most promising loci are one way in which the number of tested markers can be reduced, thereby lowering the number of tests that need to be taken into account in the overall p-value. One of the more recent such developments is the two independent stage strategy (using the same data) by Van Steen *et al.* [63]. Since the steps are independent one only needs to correct the p-value for the number of steps in the second step.

Finally, we want to mention some family-based association tests not discussed in the Background of the thesis:

Martin *et al.* [40] propose a fully parametric model which extends the retrospective probability of offspring marker genotypes in terms of identity-by-descent (IBD) probabilities.

Millstein *et al.* [44] present a likelihood upon which they base a test for both association and linkage using fixed effects.

Wheeler *et al.* [65] propose a prospective based approach; the Quantitative 'Conditioning on Parental Genotype' QCPG; approach. for continuous traits with parameters accounting for non-Mendelism (*i.e.* non-random transmission) and population stratification.

Li *et al.* [37] present a test which, like the I-GRE, uses neighboring markers to infer linkage, but it is not based on a formulation which conditions on genotypes.

Tzeng & Zhang [62] have proposed a general framework for testing haplotype effects starting from a VCM model for quantitative traits and show that it can be generalized to other types of trait.

Acknowledgements

I had the fortune of meeting Juni Palmgren, my supervisor, way back in the year of 2000. I wanted two leave my undergraduate studies at the Department of Mathematics at Stockholm University to study abroad and Juni suggested that I should apply to the Master of Science program at Oxford University. Much to my surprise they accepted my application. My studies at Oxford actually became the starting point of this thesis; it was at the University of Oxford that I discovered genetic epidemiology.

I am very grateful for the the opportunities Juni has given me; by taking me on as a Ph.D. student, letting me travel to conferences to meet people and present my research, and letting me study a semester at Harvard School of Public Health. I believe that I have grown, not only towards becoming a researcher, but also as a human being. For that, Juni, I am forever grateful.

My warmest thanks also goes to my co-supervisor, Keith Humphreys. It seems like every time I ask Keith a question he looks through the piles on his desk and finds some notes that he scribbled down months ago. Thank you, Keith, for all your support and help during these years. It has been invaluable for me to have you as my co-supervisor.

I would also like to acknowledge the people that I have collaborated with during the years at Karolinska Institutet and especially the group studying the genetics of Multiple Sclerosis at the Department of Applied Neuroscience; Thomas Masterman, Helena Modin, Kerstin Imrell, Virginija Danylaité Karrenbauer, Leszek Stawiarz, Izaura Roos and Eva Åkesson. From this group I want to especially thank Jan Hillert, Boel Brynedal, Kristina Duvefelt and Frida Lundmark - I have learned so much from you all. We have had both intense discussions and some laughs.

My warmest regards also goes to my co-author on paper III, Tim Becker. We basically wrote the paper during the summer of 2007. It was an intense working summer, but we also found time for some laughs. I truly enjoyed working you, Tim.

I would also like to thank Kristel Van Steen for inspiring discussions.

Many thanks to my friends at MEB. This could be a long list and I will only mention a few names. Lots of thanks to the wonderful IT-crew for all their help, especially Denny Rönngren, Gastón Nuñez , Torbjörn Gustafsson and Zack Yusof. My warmest thanks also goes to the education administration; Monica Rundgren, Camilla Ahlqvist and Marie Dokken. I would also like to thank Ulrika Eriksson, Elinor Fondell and Junmei Miao Jonasson with whom I arranged a doctoral retreat in 2007. Thanks also to Cat Halthur, Kristjana Einarsdottir and Anthony Gunell for their friendship.

To my friends in the Biostat group at MEB: Thanks for the open research environment, for the enjoyable Thursday *Fika*, for the fun and laughter going out to eat and watching international movies together. Thank you all; Rino Bellocco, Paul Dickman, Jonas Hällgren, Paul Lambert, Yudi Pawitan, Alex Ploner, Samuli Ripatti, Sven Sandin, Lucy Smith, Fredrik Wiklund, Li Yin, Barbara Mascialino, Davide Valentini, Cecilia Lundholm, Therese Andersson, Anna Johansson and Sandra Eloranta. Many thanks also to Marie Jansson, *the wonder woman*, who seems to know everything about anything. A special thanks to my (former and present) doctoral colleagues; Arief Gusnanto, Monica Leu, Ben Yip, Chuen Seng Tan and Hatef Darabi (future (?) Ph.D. student). A special *thank you* goes to Arvid Sjölander, with whom I have shared office with for the last couple of years; thanks for being a good colleague, a considerate room mate and a good friend.

I would also like to thank my (former) colleagues in the Matstat group at Stockholm University; especially my doctoral colleagues Patricia Geli, Elinor Gustafsson, Mathias Lindholm, Andreas Nordvall and Annica Dominicus. I also want to give my warm thanks Louise af Klintberg for her support during my undergraduate studies at the department. Many thanks also to Barbro Fernström, the perhaps most helpful librarian in the world.

A very warm *thank you* to all my friends who have continued being my friends even in my mental absence (all these deadlines!); Ida Lundgren, my dearest and oldest friend - you mean the world to me; Johanna Mannung - for always listening and being so wise; Maria Grünewald, my friend and colleague - you have listened and supported me so many times that I think I owe my sanity to you; Catherine Mortimer-Hawkins - thank you for your support and friendship from the beginning to the end of this process; Ulla and Elaine Bryntesson, Karin Enblom, Calle Forsmark, John Windå and Per Hofgren - thanks for all the fun we have had!

I would not have made it this far without the love and support of my family. I want to thank my parents; my father Jonas, and his wife Ragnhildur, and my mother Hrafnhildur, for supporting and believing in me. I would also like to thank my *grandmother-in-law*, Ingrid Bergman; a very wise and good hearted woman. Last, but not least, I would like to thank my fiancé, Sophia Bergman. Thank you, Sophia, for your love and support. *You have lightened up my life*!

Bibliography

[1] G R Abecasis, L R Cardon, and W O Cookson. A general test of association for quantitative traits in nuclear families. *American Journal of Human Genetics*, 66:279–292, 2000.

[2] G R Abecasis, W O Cookson, and L R Cardon. Pedigree tests of transmission disequilibrium. *European Journal of Human Genetics*, 8:545–551, 2000.

[3] D B Allison. Transmission-disequilibrium tests for quantitative traits. *American Journal of Human Genetics*, 60:676–690, 1997.

[4] L Almasy and J Blangero. Multipoint quantitative-trait linkage analysis in general pedigrees. *American Journal of Human Genetics*, 62:1198–1211, 1998.

[5] M F Baksh, D J Balding, T J Vyse, and JC Whittaker. A likelihood ratio approach to family-based association studies with covariates. *Annals of Human Genetics*, 70:131–139, 2006.

[6] M Bass, E Martin, and E Hauser. Pedigree generation for analysis of genetic linkage and association. *Pacific Symposium on Biocomputing*, 9:93–103, 2004.

[7] T Becker, M P Baur, and M Knapp. Detection of parent-of-origin effects in nuclear families using haplotype analysis. *Human Heredity*, 62:64–76, 2006.

[8] T Becker and M Knapp. A powerful strategy to account for multiple testing in the context of haplotype analysis. *American Journal of Human Genetics*, 75:561–570, 2004.

[9] C Bourgain, E Génin, N Cox, and F Clerget-Darpoux. Are genome-wide association studies all that we need to dissect the genetic component of

complex human diseases? *European Journal of Human Genetics*, 15:260–263, 2007.

[10] P R Burton, M D Tobin, and J L Hopper. Key concepts in genetic epidemiology. *The Lancet*, 366:941–951, 2005.

[11] C S Carlson, M A Eberle, L Kruglyak, and D A Nickerson. Mapping complex disease loci in whole-genome association studies. *Nature*, 429:446–452, 2004.

[12] D Clayton and H Jones. Transmission/disequilibrium tests for extended marker haplotypes. *American Journal of Human Genetics*, 65:1161–1169, 1999.

[13] D G Clayton. A generalization of the transmission/disequilibrium test for uncertain haplotype transmission. *American Journal of Human Genetics*, 65:1170–1177, 1999.

[14] F Clerget-Darpoux and R C Elston. Are linkage analysis and the collection of family data dead? prospects for family studies in the age of genome-wide association. *Human Heredity*, 64:91–96, 2007.

[15] M R Conaway. A random effects model for binary data. *Biometrics*, 46:317–328, 1990.

[16] D R Cox. Regression models and life tables. *Journal of the Royal Statistical Society Series B*, pages 187–220, 1972.

[17] H J Edenberg, L J Bierut, P Boyce, M Cao, S Cawley, R Chiles, K F Doheny, M Hansen, T Hinrichs, and K Jones et al. Description of the data from the collaborative study on the genetics of alcoholism (coga) and single-nucleotide polymorphism genotyping for genetic analysis workshop 14. *BMC Genetics*, 6(Suppl 1)(S2):S2, 2005.

[18] O Elkayam, R Segal, M Lidgi, and D Caspi. Positive anti-cyclic citrullinated proteins and rheumatoid factor during active lun tuberculosis. *Annals of the Rheumatic Diseases*, 65:1110–1112, 2005.

[19] C T Falk and P Rubinstein. Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculation. *Annals of Human Genetics*, 51:227–233, 1987.

[20] D W Fulker, S S Cherny, P C Sham, and J K Hewitt. Combined linkage and association sib-pairs analysis for quantitative traits. *American Journal of Human Genetics*, 64:259–267, 1999.

[21] D A Greenberg, J Zhang, D Shmulewitz, LJ Strug, R Zimmerman, V Singh, and S Marathe. Construction of the model for the genetic analysis workshop 14 simulated data: genotype-phenotype relationships, gene interaction, linkage, association, disequilibrium, and ascertainment effects for a complex phenotype. *BMC Genetics*, 6(Suppl 1):S3, 2005.

[22] S Horvath, X Xu, S Lake, E Silverman, S Weiss, and N Laird. Tests for associating haplotypes with general phenotype data: Application to asthma genetics. *Genetic Epidemiology*, 26:61–69, 2004.

[23] G Jonasdottir, T Becker, K Humphreys, and J Palmgren. Testing association in the presence of linkage using the gre and multiple markers. *To appear in Genetic Epidemiology*, 2008.

[24] G Jonasdottir, K Humphreys, and J Palmgren. Testing association in the presence of linkage - a powerful score for binary traits. *Genetic Epidemiology*, 31:528–540, 2007.

[25] G Jonasdottir, K Humphreys, and J Palmgren. Testing association in family-based studies of bivariate (co-morbid) traits. *Submitted*, 2008.

[26] G Jonasdottir, J Palmgren, and K Humphreys. Analysis of binary traits: testing association in the presence of linkage. *BMC Genetics*, 6(Suppl 1)(S92), 2005.

[27] P Kraft and D C Thomas. Bias and efficiency in family-based gene-characterization studies: Conditional, prospective, retrospective, and joint likelihoods. *American Journal of Human Genetics*, 66:1119–1131, 2000.

[28] L Kruglyak. The road to genome-wide association studies. *Nature Reviews Genetics*, [Epub ahead of print], 2008.

[29] L Kruglyak and E S Lander. Complete multipoint sib-pair analysis of qualitative and quantitative traits. *American Journal of Human Genetics*, 57:439–454, 1995.

[30] N M Laird and C Lange. Family-based design in the age of large-scale gene association studies. *Nature Reviews Genetics*, 7:385–394, 2006.

[31] S L Lake, D Blacker, and N M Laird. Family-based tests of association in the presence of linkage. *American Journal of Human Genetics*, 67:1515–1525, 2000.

[32] C Lange, D L DeMeo, and N M Laird. Power and design considerations for a general class of family-based association tests: Quantitative traits. *American Journal of Human Genetics*, 71:1330–1341, 2002.

[33] C Lange and N M Laird. Power and design considerations for a general class of family-based association tests: Quantitative traits. *American Journal of Human Genetics*, 71:1330–1341, 2002.

[34] C Lange and N M Laird. Power calculations for a general class of family-based association tests: Dichotomous traits. *American Journal of Human Genetics*, 71:575–584, 2002.

[35] C Lange, E K Silverman, X Xu, S T Weiss, and N M Laird. A multivariate family-based association test using generalized estimating equations: Fbat-gee. *Biostatistics*, 4(2):195–206, 2003.

[36] H Li and X Zhong. Multivariate survival models induced by genetic frailties, with application to linkage analysis. *Biostatistics*, 3(1):57–75, 2002.

[37] M Li, M Boehnke, and G R Abecasis. Joint modeling of linkage and association: Identifying snps responsible for a linkage signal. *American Journal of Human Genetics*, 76:934–949, 2005.

[38] K Y Liang and S L Zeger. Longitudinal data analysis using generalized estimating equations. *Biometrika*, 73:13–22, 1986.

[39] W Magee, W Eaton, H-U Wittchen, K McGonagle, and R Kessler. Agoraphobia, simple phobia, and social phobia in the national comorbidity survey. *Archives of General Psychiatry*, 53:159–68, 1996.

[40] E R Martin, M P Bass, E R Hause, and N L Kaplan. Accounting for linkage in family-based tests of association with missing parental genotypes. *American Journal in Human Genetics*, 73:1016–1026, 2003.

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

[42] P McCullagh and J A Nelder. *Generalized Linear Models*, volume 37 of *Monographs on Statistics and Applied Probability.* CRC Press, second edition, 1989.

[43] M B Miller, G R Lind, N Li, and S-Y Jang. Genetic analysis workshop 15: simulation of a complex genetic model for rheumatoid arthritis in nuclear families including a dense snp map with linkage disequilibrium between marker loci and trait loci. *BMC Proceedings*, 1(Suppl 1):S4, 2007.

[44] J Millstein, K D Siegmund, D V Conti, and W J Gauderman. Testing association and linkage using affected-sib-parent study designs. *Genetic Epidemiology*, 29:225–233, 2005.

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

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

[47] Elizabeth Pennisi. Breakthrough of the year: Human genetic variation. *Science*, 318:1842–1843, 2007.

[48] R Development Core Team. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria, 2005. ISBN 3-900051-07-0 http://www.R-project.org.

[49] D Rabinowitz. A transmission disequilibrium test for quantitative trait loci. *Human Heredity*, 47:342–350, 1997.

[50] D Rabinowitz and N Laird. A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Human Heredity*, 50(4):211–223, 2000.

[51] J M Robins, J W Smoller, and K L Lunetta. On the validity of the tdt test in the presence of comorbidity and ascertainment bias. *Genetic Epidemiology*, 21:326–336, 2001.

[52] P Rubinstein, M Walker, C Carpenter, C Carrier, J Krassner, C Falk, and F Ginsberg. 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, 1981.

[53] D J Schaid and S S Sommer. General score tests for associations of genetic markers with disease using cases and their parents. *Genetic Epidemiology*, 13:423–449, 1996.

[54] F R Schneier, J Johnson, C D Hornig, M R Liebowitz, and M M Weissman. Social phobia: comorbidity and morbidity in an epidemiologic sample. *Archives of General Psychiatry*, 49:282–8, 1992.

[55] S G Self, G Longton, K J Kopecky, and K-Y Liang. On estimating hla/disease association with application to a study of aplastic anemia. *Biometrics*, 47:53–61, 1991.

[56] P C Sham. *Statistics in Human Genetics.* John Wiley & Sons Inc., New York, 1998.

[57] P C Sham, S S Cherny, S Purcell, and 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.

[58] P C Sham and D Curtis. An extended transmission/equilibrium test (tdt) for multi-allele marker loci. *Annals of Human Genetics*, 59:323–336, 1995.

[59] M-C Shih and A S Whittemore. Tests for genetic association using family data. *Genetic Epidemiology*, 22:128–145, 2002.

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

[61] J Terwilliger and J Ott. A haplotype-based 'haplotype relative risk' approach to detecting allelic associations. *Human Heredity*, 42:337–346, 1992.

[62] Jung-Ying Tzeng and Daowen Zhang. Haplotype-based association via variance-components score test. *American Journal of Human Genetics*, 81:927–938, 2007.

[63] K Van Steen, M B McQueen, A Herbert, B Raby, H Lyon, D L DeMeo, A Murphy, J Su, S Datta, C Rosenow, M Christman, E K Silverman, N M Laird, S T Weiss, and C Lange. Genomic screening and replication using the same data set in family-based studies. *Nature Genetics*, 37:683–691, 2005.

[64] J W Vaupel, K G Manton, and E Stallard. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16:439–454, 1979.

[65] E Wheeler and H J Cordell. Quantitative trait association in parent offspring trios: Extension of case/pseudocontrol method and comparison of prospective and retrospective approaches. *Genetic Epidemiology*, 31:813–833, 2007.

[66] H White. A heteroscedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity. *Econometrica*, 48:817–838, 1980.

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

[68] X Zhong and H Zhang. Linkage analysis and association analysis in the presence of linkage using age at onset of coga alcoholism data. *BMC Genetics*, 6(Suppl 1)(S31), 2005.

[69] X Zhu, R Cooper, D Kan, G Cao, and X Wu. A genome-wide linkage and association study using coga data. *BMC Genetics*, 6(Suppl 1)(S128), 2005.