Linkage Mapping for Complex Traits

A Regression-based Approach

This work was carried out as part of the GENOMEUTWIN project which is supported by the European Union Contract No. QLG2-CT-2002-01254. The publication of this thesis was supported by the Fonds Medische Statistiek.

Cover: Pierre Darcel

ISBN-10: 90-9021504-2
ISBN-13: 978-90-9021504-4

Linkage Mapping for Complex Traits

A Regression-based Approach

Proefschrift

ter verkrijging van de graad van Doctor
aan de Universiteit Leiden,
op gezag van de Rector Magnificus prof.mr.dr. P.F. van der Heijden,
volgens besluit van het College voor Promoties
te verdedigen op woensdag 21 februari 2007
te klokke 13.45 uur

door

Jérémie Jacques Paul Lebrec

geboren te Rennes (Frankrijk) in 1974

Promotiecommissie

Promotor:	Prof. dr. J. C. van Houwelingen
Co-promotor:	Dr. H. Putter
Referent:	Prof. dr. D. O. Siegmund · *Stanford University*
Overige leden:	Prof. dr. P. Slagboom
	Prof. dr. A. W. van der Vaart · *Vrije Universiteit Amsterdam*

Contents

Chapter 1

Introduction

Once the heritable character of a trait has been established, the strategies available for gene mapping may be split into two classes. In the first 'candidate gene' approach, prior biological knowledge is available about the function of one or several genes, the scientific question to be tested is whether this limited number of pre-identified genes influences the trait of interest. Subsequently, researchers are usually interested in quantifying those effects. Although the field of genetics offers some peculiarities, well known epidemiological methods are suited to answer this type of questions. The second 'positional mapping' approach requires, in principle, no prior biological knowledge but its purpose is perhaps less ambitious: it aims at identifying chromosomal regions which contain genes influencing a trait. As far as the search for genes is concerned, the first approach therefore is an hypotheses-testing exercise while the second approach generates hypotheses. linkage as well as association studies fall into the positional mapping category. The former relies on the biological process of recombination (see 1.1) and the latter on the presence of linkage disequilibrium (see also 1.1) in populations. In the traditional gene-mapping paradigm, positional mapping precedes candidate gene-mapping but the frontiers between the two categories are sometimes fuzzy. Indeed nowadays, association scans often attempt to combine the two steps together. This thesis only deals with issues related to linkage mapping.

1.1 Some basics in genetics

This section introduces some basic concepts of genetics that are a pre-requisite to the understanding of the problem of linkage.

A gene is defined as a sequence of desoxyribonucleic acid (DNA) that codes a protein; most of our DNA is non-coding. Despite this formal definition, the term gene

is often loosely used to refer to a piece of DNA or genetic material, whether coding or not. This imprecision in terminology is often a hurdle for statisticians willing to enter the realm of genetics. Nevertheless, I will adhere to this practice. The genetic material of human beings is stored in 23 pairs of chromosomes, 22 pairs of autosomes and 1 pair of sex chromosomes. The transmission of this material from parents to offspring occurs independently at each chromosome: each parent contributes one copy of his/her two genes at random to an offspring via their gametes, this is known as the law of segregation or Mendel's first law. Parents, however, rarely transmit an entire copy of one of their two chromosomes (termed grand-paternal and grand-maternal). Instead, their transmitted chromosome is made up of alternating segments from the grand-paternal and grand-maternal chromosomes. This exchange of genes between the grand-paternal and grand-maternal chromosomes occurs during the formation of gametes or meiosis at points called crossovers, as a result chromosomes in gametes and resulting offspring are made up of recombinant chromosomes (see Fig.1).

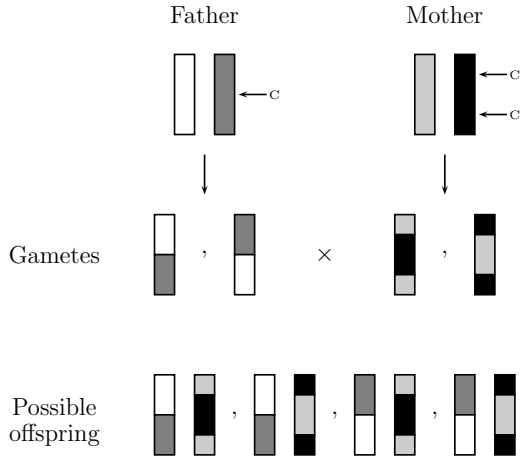

Figure 1.1: Chromosomes in gametes and offspring after recombinations - C indicates a crossover event

This recombination process ensures genetic diversity, it is also the phenomenon that makes linkage analysis possible because it introduces variation in genetic similarity between relatives across one single chromosome. A recombination event between two chromosomal positions or loci is equivalent to an odd number of crossovers

between those two loci in one meiosis, this happens at a certain rate called the recombination fraction θ. The recombination fraction increases with physical distance, however the relation between the two varies across the genome. If two loci are close together on the same chromosome, they are said to be linked; if they are very far apart, on the same chromosome or on different chromosomes, they are unlinked and the law of segregation implies that $\theta = 0.5$. The genetic distance d_{AB} (unit=Morgan) between two loci A and B is defined as the average number of crossovers between them per meiosis, by linearity of the expectation $d_{\mathrm{AC}} = d_{\mathrm{AB}} + d_{\mathrm{BC}}$ (if B lies between A and C). This additive property of the genetic distance scale is extremely convenient but obviously does not apply to recombination fractions although this is the probabilistic quantity needed for computations in linkage testing. Mapping functions that convert recombination fraction θ into genetic distance m, or conversely, are therefore available. One slightly simplistic but practically important such function is given by Haldane's function $\theta = \frac{1}{2}(1 - \mathrm{e}^{-2m})$ which is obtained by assuming that the number of crossovers between two loci follows a Poisson distribution with mean proportional to the genetic distance between loci.

Since the genetic similarity between relatives extends over relatively large chromosomal segments, it would be far too costly and inefficient to sequence the whole genome of each individual. Geneticists have identified DNA polymorphisms (so called markers) which can be seen as genes (in the loose sense) whose alleles (the different forms that a gene can take) can easily be identified by modern molecular biology techniques. It must be stressed that this technology can only determine the unordered pair of alleles (or genotype) at each marker for the two paired chromosomes of an individual. Classically, a few hundreds highly polymorphic genetic markers known as micro-satellites are scattered more or less evenly across all chromosomes. Since they have many and therefore relatively rare alleles, those markers allow one to tell whether relatives share the same genes at that location with little uncertainty. Those markers are usually taken in non-coding regions of the genome and are therefore believed, due to lack of selective pressure, to be neither related with each other nor with the potentially causing genes, in the overall population. In genetic jargon, the markers are said to be in linkage equilibrium with each other and with the genes [1]. Another

[1] In statistical terms, considering the one-allele genotypes of gametes at different loci as random

type of (bi-allelic) markers known as single nucleotide polymorphisms (SNP) is now routinely used in gene-association studies, these markers are more densely available across the genome and they can be cheaply typed in chips called SNP-arrays. They are now being used in linkage analysis too although their use is more problematic due to linkage disequilibrium between them. Despite the intensive computations involved in their use in linkage analysis, they offer the promise of a cheap and evenly distributed linkage information map across the genome.

1.2 Overview of linkage methods

The first traits to be mapped by linkage methods were Mendelian i.e. they were rare and determined in an almost one-to-one relation by the genotype at a single location. With such strong genetic effects, the actual mode of inheritance (i.e. genetic model) was fairly well known via segregation analysis (which only requires phenotypic data in families). This type of traits lent itself very well to the so-called parametric linkage methods. In its simplest version, this methodology postulates a genetic model for the trait values Y given the genotype at the causing locus with genotype G via a penetrance function $\mathbf{P}(Y \,|\, G)$. The likelihood $L(M \,|\, Y; \theta)$ of the data at a marker M given the recombination fraction θ between marker M and true locus can be computed and the corresponding likelihood ratio test $\sup_\theta \frac{L(M \,|\, Y; \theta)}{L(M \,|\, Y \;;\; \theta=0.5)}$ provides a test for linkage.

This model for linkage was appealing for Mendelian traits and did yield an unprecedented harvest of genes for those rare diseases but it is much less suited for the analysis of complex traits. The methodological emphasis has long switched to biometrical models and to the so-called non-parametric linkage methods. This other branch of methods is essentially based on identifying chromosomal regions where phenotypic similarity coincides with genotypic similarity. The concept of identity-by-descent (IBD) formalizes the idea of genetic similarity between relatives: two genes are said to be IBD if they are copies of the same ancestral gene. The IBD configuration at different loci in a pedigree is not observable directly but it can be conceived of as

variables (a haplotype is a possible value of the resulting multivariate random variable), two loci are said to be in linkage equilibrium if the genotypes at those two loci are independently distributed, if not they are said to be in linkage disequilibrium

a hidden Markov process whose transition probabilities depend upon the recombination fractions [Lander and Botstein, 1989] between loci. The observations at the markers are used to calculate the IBD distribution at any arbitrary position on the chromosome [Kruglyak et al., 1996; Abecasis et al., 2002].

Continuous traits

For a quantitative trait, a Gaussian distribution naturally arises from the view that many factors, whether environmental or genetic, with equally small individual effects contribute to the trait. By further assuming a random mating population, one obtains the so-called variance components model [Lange et al., 1976; Amos, 1994; Almasy and Blangero, 1998]. In a simple additive version of the model, the total trait variance is decomposed into three sources: familial or common environment, additive genetic and measurement error or unique environment. The covariance of two relatives turns out to be the sum of the common environment variance and the additive genetic variance times a kinship coefficient which is proportional to the average proportion of genes that the relatives share. The model is often used in heritability and segregation analysis where the purpose is to establish the genetic character of a trait and to further characterize its mode of inheritance. Monozygotic twins have the same genes while dizygotic twins share only half of them but the degree to which the environment is shared by individuals in the two types of twinships is identical. Twin studies therefore provide a simple design for testing for a purely genetic component.

If IBD was measured exactly at a causative additive gene, the covariance for two relatives in the variance components model would include a term equal to the product of kinship coefficient by the gene attributable variance σ_q^2 times the IBD sharing. The test for linkage at any putative position is therefore based on rejecting the null hypothesis that $\sigma_q^2 = 0$ in favor of the alternative $\sigma_q^2 > 0$. In unselected families, this is traditionally done using a likelihood ratio test statistic. In practice, IBD is measured at locations nearby the causing gene(s) and the estimated attributable variance will be a deteriorated version of σ_q^2, nevertheless the test statistic will tend to be maximal at positions closest to the true gene location. The popularity of the variance components model in quantitative trait locus (QTL) mapping is undoubtedly due to its extreme flexibility: variance components corresponding to non-additive

(dominant) gene effects, gene-gene interactions, gene by covariate interactions can be accommodated, the model mean can be corrected for important covariate effects, multivariate phenotypes can be conjunctly analyzed, the method can be adapted for analysis of the sex-chromosomes [Ekstrøm, 2004] and mixtures of variance components models can be used to face the problem of locus heterogeneity (see 1.3) [Ekstrøm and Dalgaard, 2003]; these extensions are only hindered by the computations required for fitting the corresponding models.

The much less computationally greedy regression-based methods for linkage analysis stem back to the work of Haseman and Elston [1972] who proposed to regress the squared difference in phenotypic values of siblings on their IBD sharing. In 30 years, many variations have appeared on the theme and they are all based on the regression of some form of phenotypic similarity statistic on the IBD sharing. It is only recently that light has been shed on the relation between Haseman-Elston regressions and the score test of the linkage parameter $\sigma_q^2 = 0$ in the variance components model [Tang and Siegmund, 2001; Putter et al., 2002; Wang and Huang, 2002a]: some optimal form of Haseman-Elston regression happens to coincide with such a score test in an additive variance components model for sibling pairs. The conceptualization of those regression methods as score tests in the flexible variance components model frameworks has opened the way to fruitful generalizations of the regression-based methods e.g. to arbitrary pedigrees. In addition to their light computational burden, regression-based or score test based methods are appealing because of their potential robustness (in terms of false positive rate) to normality and to outliers. Finally, by inverting the regression i.e. IBD is regressed on a function of phenotypic similarity, the method can in principle be used to make valid inference in families sampled using their trait values [Sham et al., 2002].

Qualitative traits

For qualitative traits, which for linkage studies is almost synonymous of binary traits (i.e. disease in the medical field), non-parametric testing for linkage is usually done by comparing the average observed IBD sharing with its expected value under the assumption of no linkage. In designs where only one type of independent relative pairs is collected (e.g. affected sib-pair designs, ASP), this test based on deviation of

IBD sharing uses 1 degree of freedom (df), while a totally model-free ASP analysis necessitates a 2-df test [Risch, 1990]. Although the recognition of constraints for the parameters reduces the space of alternatives [Holmans, 1993], the higher level of significance required for the 2-df test often annihilates the gain in non-centrality parameter and the 1-df test appears to be a good testing strategy for a wide range of genetic models. Different types of independent relative pairs (e.g. affected sib pairs, discordant sib pairs, affected cousins) can be combined by using a weighted average of the excess IBD sharing of each kind; whatever the weights, provided markers segregate in a Mendelian fashion, the test will have adequate type I error, however its optimality will depend on how close the chosen relative weights are from the true relative excesses in IBD sharing at the causative locus [Teng and Siegmund, 1997].

Although less attractive than when disease inheritance is clearly Mendelian, larger families are sometimes sampled in linkage studies for complex traits. In that case, IBD-based tests can be generalized by the use of sensible scoring functions of the different IBD configurations in a pedigree [Whittemore and Halpern, 1994; Kong and Cox, 1997]. Alternatively, locally optimal tests based on the likelihood of the IBD configuration in each pedigree may be derived. The tests are pedigree-specific and only optimal if the true relative weights of the different parameters are known but sensible guesses provide decent efficiency across a wide range of genetic models [Teng and Siegmund, 1997]. As in the case of families consisting only of pairs of relatives, combining families of different types is a matter of assigning relative weights to the family-specific tests.

The incorporation of covariate information into disease linkage studies has been an active area of research in the past few years [Schaid et al., 2003]. The usual approach amounts to regressing the IBD sharing on the covariates of interest in a linear or non-linear fashion [Olson, 1999]. At least for categorical covariates, the approach can be made non-parametric at the cost of an increase in the number of parameters, however parsimonious models are needed in order to carry out efficient inference. Age is a crucial covariate to take into account in order to include unaffected individuals in a linkage study. Another way to approach the problem is to use the disease age of onset as the possibly censored endpoint.

Significance level

Since the position of the true locus is often completely ignored, the whole genome is scanned using a linkage statistic on a grid of chromosomal positions, this multiplicity of tests increases the false positive rate. The tests at neighboring positions are highly correlated so a Bonferroni correction of the α level of each test is too conservative. Asymptotic arguments based on the theory of Gaussian processes leads to approximate thresholds for the non-parametric methods statistics [Lander and Green, 1987; Feingold et al., 1993]. These thresholds rely on the Haldane's mapping function, they depend on the type of families studied (which determines the correlation structure of the process) and the degrees of freedom for the test; although they are derived under the idealized assumption of a dense map of completely informative markers, the thresholds seem to be only slightly conservative when applied to discrete evenly distributed maps of partially informative markers [Teng and Siegmund, 1998]. Due to a tradition dating back to the early days of parametric linkage [Morton, 1955], statistical significance of linkage tests is usually presented as a LOD score (originally a $\log_{10}$ of the odds that a locus is linked versus unlinked) which is obtained by dividing a $\chi^2_{[1]}$-distributed statistics by $2 \times \ln(10)$. In current practical situations of human sib-pair linkage studies, a LOD score of 3 or higher gives a rule of thumb for declaring that a 1-df statistics based on average IBD sharing is significant.

In practice, various types of families are often combined, marker information varies across the genome and the assumptions underlying the linkage model (eg. normality in variance components model) might not be fulfilled. Nowadays, researchers tend to base their assessment of significance on simulations. Given the 'experimental conditions' of a study (marker map characteristics, pedigree structures and patterns of genotype missingness), marker genotypes can be simulated under the null hypothesis of no linkage i.e. by simply obeying the rules of Mendelian segregation. In that way, provided the linkage statistic can be quickly computed, the null distribution of the statistic may be obtained at any point on the genome. This method, sometimes called gene-dropping, therefore yields point-wise empirical p-values. The number of times the statistic exceeds a certain threshold on a given chromosome can be counted (note that this entails the choice of a minimal distance for considering two consecu-

tive peaks as separate). By combining the corresponding independent p-values on all chromosomes, one can obtain a genomewide assessment of significance.

1.3 Issues in linkage mapping

Linkage analysis has been successful in the gene mapping of hundreds of mendelian diseases, however application of the same methodology in the search for genes responsible for complex traits has proved extremely disappointing. Most studies often provide only suggestive evidence for linkage, and when clearly significant, replication of the findings appears to be the exception rather than the rule.

Failure of the linkage approach to gene-mapping of complex traits is often attributed to locus heterogeneity i.e. the fact that the loci influencing a trait differ across families or groups of families [2]. This is indeed a problem likely to be more acute in linkage studies of complex traits where data from numerous small families are gathered as opposed to a small number of large families. A direct corollary of locus heterogeneity is that linkage studies are under-powered. In fact, due to the polygenic nature of complex traits, most studies probably lack the sample size to detect the inherent small gene effects.

One obvious way to tackle the problem of heterogeneity is to refine the definition of a phenotype by defining more homogeneous clinical subgroups, so instead of sampling breast cancer patients, geneticists successfully selected families with early-onset breast cancer. Researchers also try to select phenotypes that are likely to be more closely related to a biological mechanism than a broadly defined disease itself. For instance different plasma lipid levels can be measured in the search for genes involved in obesity.

One strategy for improving power is to resort to selective genotyping [Risch and Zhang, 1995] i.e. to only genotype families whose extreme phenotypic values promise to deliver high linkage information. Another natural route for solving the issue of power is by a sufficient increase of the sample size. Collaborative efforts such as the GenomEUtwin project (`http://www.genomeutwin.org/`) are being set up in order to gather sufficient data from different centers. This obviously calls for meta-

[2] Another type of heterogeneity called allelic heterogeneity refers to a situation where different allelic mutations at the same locus contribute to a phenotype, however, linkage analysis is immune to this type of heterogeneity

analytic methodologies routinely used in the field of clinical trials.

It is also felt that the models underlying the linkage methods are too simplistic, for instance, important covariates or interactions are often ignored. Although biologically plausible, incorporation of gene-gene interactions in models for linkage analysis is unlikely to yield substantial benefit [Tang and Siegmund, 2002; Purcell and Sham, 2004]. Using covariate information appears to be a more promising path towards a refinement of the methods [Peng et al., 2005].

1.4 This thesis

This thesis presents some attempts to improve the current design and analysis of linkage studies for complex traits. The statistical methodology adopted is driven by the fact that genes involved in complex traits have small effects, it therefore seems legitimate to use score tests [Cox and Hinkley, 1974] because of their local optimality properties. In addition, score tests often give rise to tractable expressions, in the context of linkage these can be meaningfully interpreted in terms of regressions and quickly computed which is a crucial feature in genetics.

Chapter 2 deals mainly with the analysis of quantitative traits in families that have been selected based on their trait values. We derive a general score test for linkage in arbitrary pedigrees which is based on the likelihood conditional on the phenotypic values. Although the derivation of the test relies on the normally distributed variance components model, its size is robust to deviations from normality. Under local alternatives and assuming the variance components model correctly specifies the distribution of the phenotype, the test has some optimality properties. In addition, the value of the test's Fisher information provides an indication of the informativeness of each family and can be used as a criterion for genotyping selection. The test is adapted to the case of binary data via a liability threshold model.

Chapter 3 advocates the use of selected families in the mapping of complex traits using twins. The methodology relies on the informativeness criterion derived in chapter 2, but we quantify the potential gains obtained using a series of examples of quantitative and qualitative phenotypes that are relevant to the GenomEUtwin project.

Chapter 4 addresses the issue of genotyping error in linkage analysis. We first

study analytically the impact of genotyping error on linkage and provide formula for the bias incurred. These results provide insights into some empirical findings, in particular, we are able to explain the differences in impact of genotyping error in random and selected designs. Finally, we suggest a robust modification of the usual linkage test based on a genomic control of the excess IBD sharing, it provides robustness against genotyping error as well as against other processes whose effect is to distort the expected value of the IBD sharing.

Chapter 5 is concerned with the (in)validity of a range of standard methods when marker information is incomplete, in particular circumstances where the generalized estimating equations method for gene localization [Liang et al., 2001] fails are identified.

Chapter 6 transfers standard meta-analytic techniques to the field of QTL mapping. The field has some specificities that can be accommodated, in particular, the problem of genetic locus heterogeneity is looked at carefully. In absence of covariate observations at the individual level and under a homogeneous model, the meta-analytic approach is asymptotically equivalent to an analysis of a pooled data set but it is logistically much easier to carry out.

Finally, in chapter 7, we develop an approximate score test for linkage in the rich class of generalized linear models. It is based on a pseudo-likelihood of the data and although unlikely to be optimal in all situations, the test has the advantage of being tractable and to have a robust type I error. It provides a simple way to incorporate known covariate effects into linkage analysis and is applicable to arbitrary pedigrees.

The last chapter is a conclusion where I draw a perspective of the role of linkage in gene mapping.

Chapter 2

Score Test for Detecting Linkage to Complex Traits in Selected Samples

Abstract

We present a unified approach to selection and linkage analysis of selected samples, for both quantitative and dichotomous complex traits. It is based on the score test for the variance attributable to the trait locus and applies to general pedigrees. The method is equivalent to regressing excess IBD sharing on a function of the traits. It is shown that, when population parameters for the trait are known, such inversion does not entail any loss of information. For dichotomous traits, pairs of pedigree members of different phenotypic nature (e.g. affected sib pairs and discordant sib pairs) can easily be combined as well as populations with different trait prevalences.

This chapter has been published as: J. Lebrec, H. Putter and J.C. van Houwelingen (2004). Score Test for Detecting Linkage to Complex Traits in Selected Samples. *Genetic Epidemiology* **6** (2), 97–108.

2.1 Introduction

In complex traits where the effect of each contributing locus is very small, the sample sizes needed to carry out linkage analysis usually result in costs far beyond research budgets, even when using new high throughput genotyping technologies [Risch, 2000]. Geneticists have been aware of this fact for a while and many designs and selection strategies have been proposed [Risch and Zhang, 1995; Dolan and Boomsma, 1998a; Purcell et al., 2001]. In the search for genes, prior to any linkage study, researchers usually gather evidence of heritability for the trait of interest. This is often done in twin studies including both monozygotic and dizygotic twins from the general population. In addition to heritability of the trait, these studies provide precise population marginal means, variability and twin-twin correlation estimates for the trait of interest.

Complex traits have small locus effect and this is probably why the search for the corresponding susceptibility loci has proved so disappointing. However this is also the reason why a score test constitutes a promising testing strategy in this context since it has local optimality properties [Cox and Hinkley, 1974]. In this article, using the variance components framework we give a general formulation for a score test to detect linkage to a putative quantitative trait locus under selective sampling based on the trait values of the pedigree members. We give simple formulae for the test in a number of commonly used designs (sibships and nuclear families of arbitrary size). Using a liability threshold model, we extend our results to dichotomous traits. In particular, they apply to sib pair designs where different types of pairs (e.g. affected and discordant sib pairs) can be combined in an optimal way, and subpopulations with different disease prevalences can be incorporated in a straightforward manner. Our approach provides a unified framework in which both optimal selection and subsequent analysis are combined in a natural way, both for quantitative and dichotomous traits.

2.2 Score test for quantitative traits in selected samples

Model

Our starting point is the variance components model, where we assume that $\mathbf{x} = (x_1, \ldots, x_m)'$, the vector of phenotypes of the pedigree members, has been standardized so that it has mean vector 0 and variances equal to 1. The $m \times m$ matrix $\boldsymbol{\pi}$ contains the identity-by-descent (IBD) information at a marker, more precisely $[\boldsymbol{\pi}]_{jk} = \pi_{jk}$ is the proportion of alleles shared IBD by pedigree members j and k. For now, we assume that the marker map is fully informative, the consequences of relaxing this assumption will be examined in Section 2.6. The variance components model specifies that the conditional distribution of the standardized $\mathbf{x}$ given IBD information $\boldsymbol{\pi}$ follows a normal distribution with zero mean and variance-covariance matrix $\boldsymbol{\Sigma}$ given by

$$[\boldsymbol{\Sigma}]_{jk} = \begin{cases} a^2 + c^2 + e^2 \;=\; 1 \;, & \text{if } j = k \;, \\ (\pi_{jk} - \mathbf{E}\pi_{jk})q^2 + (\mathbf{E}\pi_{jk})a^2 + c^2 \;, & \text{if } j \neq k \;. \end{cases}$$

where a^2 denotes the total additive genetic variance, c^2, the common-environment variance and e^2, the residual variance. This parameterization of the problem was initially introduced by Tang and Siegmund [2001] and is crucial to the obtention of simple results. For the time being we will assume absence of any dominance component of variance. We show an extension incorporating dominance variance in section 2.4. Since the trait values are standardized to unit variance, these variance components can also be interpreted as proportions of variance explained by the appropriate components. The total additive genetic variance a^2 includes both additive polygenic variance and the (additive) variance q^2 attributable to the putative quantitative trait locus (QTL). The factor $\mathbf{E}\pi_{jk}$ denotes the expected proportion of alleles shared identical by descent between pedigree members j and k; it is determined solely by the family relationship between j and k and equals twice the kinship coefficient between j and k.

The key parameter in this model is the variance component q^2 determining the presence of linkage (no linkage is equivalent to $q^2 = 0$). It is the only unknown parameter in the model and we shall denote it by γ in the sequel. Two important

properties of the variance components model are: that $\mathbf{x}$ and $\boldsymbol{\pi}$ are independent under the hypothesis of no linkage ($\gamma = 0$) and that the marginal distribution of $\boldsymbol{\pi}$ does not depend on γ.

Score test for quantitative traits

A score test for detecting linkage to quantitative traits in random samples for general pedigrees was given by Putter et al. [2002] and by Wang [2002]. Here we extend those results to a sampling scheme where data are selected based on phenotypic values. We generalize results obtained by Tang and Siegmund [2001] for sibships to arbitrary pedigrees and use the continuous case as a building block to the dichotomous case as exposed in Section 2.5.

The following expression for the score function $\ell_\gamma^{\mathbf{x}}$ in the variance components model is obtained in the appendix:

$$\ell_\gamma^{\mathbf{x}} = \frac{1}{2} \operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\pi} - \mathbf{E}\boldsymbol{\pi})(\boldsymbol{\Sigma}^{-1}\mathbf{x}\mathbf{x}' - \mathbf{I})\right) \ .$$

Here $\operatorname{tr}(A)$ stands for the trace (sum of the diagonal elements) of matrix A. Using elementary matrix theory, in particular $\operatorname{tr}(AB) = \operatorname{tr}(BA)$ and $\operatorname{tr}(AB) = \operatorname{vec}(A')'\operatorname{vec}(B)$ (here $\operatorname{vec}(A)$ places the n columns of the $m \times n$ matrix A into a vector of dimension $mn \times 1$), this score function can be rewritten as

$$\ell_\gamma^{\mathbf{x}} = \frac{1}{2} \operatorname{vec}(\mathbf{C})'\operatorname{vec}(\boldsymbol{\pi} - \mathbf{E}\boldsymbol{\pi}) \tag{2.1}$$

with $\mathbf{C} = \boldsymbol{\Sigma}^{-1}\mathbf{x}\left(\boldsymbol{\Sigma}^{-1}\mathbf{x}\right)' - \boldsymbol{\Sigma}^{-1}$. Note that the $\boldsymbol{\pi} - \mathbf{E}\boldsymbol{\pi}$ matrix has all diagonal elements equal to 0.

For selected samples, the conditional distribution of IBD sharing $\boldsymbol{\pi}$ given the trait values $\mathbf{x}$ gives a natural framework for testing linkage [Sham et al., 2000; Dudoit and Speed, 2000] and we shall refer to this setting as the *selection model*. It turns out that the score function for this *selection model*, and for the *joint model* of $\mathbf{x}$ and $\boldsymbol{\pi}$ remains the same. As we show below, this is true for any joint model of $\mathbf{x}$ and $\boldsymbol{\pi}$ under the following general conditions, which are satisfied for the variance components model:

1. $\mathbf{x}$ and $\boldsymbol{\pi}$ are independent at $\gamma = 0$ and
2. the marginal distribution of $\boldsymbol{\pi}$ does not depend on γ.

We now turn to the proof of our previous statement regarding the equality of the scores for the selection model and the joint model. We denote the conditional distribution of $\mathbf{x} \,|\, \boldsymbol{\pi}$ and $\boldsymbol{\pi} \,|\, \mathbf{x}$ by $f_\gamma(\mathbf{x} \,|\, \boldsymbol{\pi})$ and $f_\gamma(\boldsymbol{\pi} \,|\, \mathbf{x})$ respectively, and the joint distribution of $\mathbf{x}$ and $\boldsymbol{\pi}$ by $f_\gamma(\mathbf{x}, \boldsymbol{\pi})$. The subscript γ expresses the dependence of those distributions on γ. The marginal distributions of $\mathbf{x}$ and $\boldsymbol{\pi}$ are denoted by $f_\gamma(\mathbf{x})$ and $f(\boldsymbol{\pi})$ respectively. With this notation, the score function for γ in the $\mathbf{x} \,|\, \boldsymbol{\pi}$ model is denoted by $\ell_\gamma^{\mathbf{x}}$, so $\ell_\gamma^{\mathbf{x}} = \frac{\partial}{\partial \gamma} \log f_\gamma(\mathbf{x} \,|\, \boldsymbol{\pi})$; and in the selection model by $\ell_\gamma^{\boldsymbol{\pi}}$, so $\ell_\gamma^{\boldsymbol{\pi}} = \frac{\partial}{\partial \gamma} \log f_\gamma(\boldsymbol{\pi} \,|\, \mathbf{x})$. By Bayes' rule, we have

$$f_\gamma(\boldsymbol{\pi} \,|\, \mathbf{x}) = \frac{f_\gamma(\mathbf{x}, \boldsymbol{\pi})}{f_\gamma(\mathbf{x})} = \frac{f_\gamma(\mathbf{x} \,|\, \boldsymbol{\pi})\ f(\boldsymbol{\pi})}{\int f_\gamma(\mathbf{x} \,|\, \boldsymbol{\pi})\ f(\boldsymbol{\pi})\ d\boldsymbol{\pi}}\ . \tag{2.2}$$

As a result,

$$\begin{aligned} \ell_\gamma^{\boldsymbol{\pi}} = &\ \frac{\partial}{\partial \gamma} \log f_\gamma(\mathbf{x} \,|\, \boldsymbol{\pi}) - \frac{\partial}{\partial \gamma} \log \left(\int f_\gamma(\mathbf{x} \,|\, \boldsymbol{\pi}) f(\boldsymbol{\pi})\ d\boldsymbol{\pi} \right) \\ = &\ \ell_\gamma^{\mathbf{x}} - \frac{\partial}{\partial \gamma} \log \left(\int f_\gamma(\mathbf{x} \,|\, \boldsymbol{\pi}) f(\boldsymbol{\pi})\ d\boldsymbol{\pi} \right)\ . \end{aligned} \tag{2.3}$$

For the score test for linkage in selected samples, we need this score function evaluated at $\gamma = 0$. Since score functions have mean 0, the second term $\frac{\partial}{\partial \gamma} \log \left(\int f_\gamma(\mathbf{x} \,|\, \boldsymbol{\pi}) f(\boldsymbol{\pi})\ d\boldsymbol{\pi} \right)$ equals the expectation of $\ell_\gamma^{\mathbf{x}}$ under $\boldsymbol{\pi} \,|\, \mathbf{x}$ evaluated at $\gamma = 0$. Since $\mathbf{x}$ and $\boldsymbol{\pi}$ are independent at $\gamma = 0$, this is just the distribution $\boldsymbol{\pi}$ (independent of γ). As a result we obtain,

$$\ell_\gamma^{\boldsymbol{\pi}} = \ell_\gamma^{\mathbf{x}} - \mathbf{E}_{\boldsymbol{\pi}} \ell_\gamma^{\mathbf{x}}\ .$$

Hence, in our case $\ell_\gamma^{\boldsymbol{\pi}} = \ell_\gamma^{\mathbf{x}}$, since $\ell_\gamma^{\mathbf{x}}$ is already, due to the parameterization used, centered with respect to the distribution of $\boldsymbol{\pi}$. The score $\ell_\gamma^{\mathbf{x}}$ is also centered with respect to the distribution of $\mathbf{x}$. Looking back at equation (2.2), we see that the score function for γ in the joint model of $\mathbf{x}$ and $\boldsymbol{\pi}$ also equals $\ell_\gamma^{\mathbf{x}} = \ell_\gamma^{\boldsymbol{\pi}}$. This has the important consequence that there is no loss of information by basing inference only on the conditional distribution of $\mathbf{x} \,|\, \boldsymbol{\pi}$ for random samples, or only on the distribution of $\boldsymbol{\pi} \,|\, \mathbf{x}$, the selection model for selected samples.

Fisher's information $\mathcal{I}_\gamma^{\boldsymbol{\pi}} = \mathbf{E}\left(-\frac{\partial^2}{\partial \gamma^2} \log f_\gamma(\boldsymbol{\pi} \,|\, \mathbf{x}) \right)$ for γ in the selection model is also the variance of the score function $\mathrm{var}_{\boldsymbol{\pi}}(\ell_\gamma^{\boldsymbol{\pi}})$ and is thus given by

$$\mathcal{I}_\gamma^{\boldsymbol{\pi}} = \frac{1}{4}\ \mathrm{vec}(\mathbf{C})'\ \mathrm{var}_{\boldsymbol{\pi}}\left(\mathrm{vec}(\boldsymbol{\pi})\right)\ \mathrm{vec}(\mathbf{C})\ . \tag{2.4}$$

The exact calculation of $\mathrm{var}_{\boldsymbol{\pi}}(\mathrm{vec}(\boldsymbol{\pi}))$ involves enumeration of all joint probabilities $\mathbf{P}(\pi_{ij}, \pi_{kl})$ for each possible inheritance vector in a pedigree. In practice, this is efficiently achieved through the use of the `--ibd` and `--matrices` options in the `MERLIN` software [Abecasis et al., 2002] with a pedigree file describing the appropriate pedigree structure and one marker with all values as missing. Note that under the assumption of complete IBD information, Fisher's information as given in Formula (2.4) can be directly used as a criterion for selection of the most informative individuals based on trait values.

The score test statistic z is formed by adding the scores from independent pedigrees and dividing by the square root of its variance under the null hypothesis:

$$z = \frac{\sum_i \ell^{\boldsymbol{\pi}}_{\gamma,i}}{\sqrt{\sum_i \mathcal{I}^{\boldsymbol{\pi}}_{\gamma,i}}} \,. \tag{2.5}$$

Under the null hypothesis of no linkage, z has asymptotically a standard normal distribution. The test is one-sided, only positive values of z being regarded as evidence for linkage. In other words, z_+^2 defined as being equal to 0 if $z \leq 0$ and to z^2 if $z > 0$ is asymptotically distributed as $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$.

Formulae (2.1) and (2.4) provide an interpretation of this score test in terms of regression. Similar to Sham et al. [2002], the numerator of the score test statistic z can be interpreted as an estimate of the slope of the regression through the origin of excess IBD sharing on a function of the trait values. The dependent variables are the observed excess IBD sharing between all $\frac{m(m-1)}{2}$ pairs of members in pedigree of size m while corresponding observations of the explanatory variable are quadratic functions of the original trait values as defined above. Those results are applicable to general pedigrees but take a very simple and appealing form in sib pairs and some other specialized cases as shown below. The slope estimate of the score test statistic is standardized by the square root of Fisher's information, but this standardization can also be interpreted as the standard error of the slope estimate of the numerator under the null hypothesis.

2.3 Special designs

In this section we give explicit formulae for the score test in general sibships and nuclear families. The interpretation of the test in terms of regression for sib pairs provides interesting insight into the relation of our method with the so called Haseman-Elston regressions and helps us understand why these optimal methods for random samples turn out to be sub-optimal when data are subject to selection unless modified as in Sham and Purcell [2001]. We refer the reader to Skatkiewicz et al. [2003]; Cuenco et al. [2003] for a comprehensive review and numerical comparison of methods for selected sib pairs.

Sibships

In a **sibship of size** m consisting of m siblings, $\mathbf{\Sigma}$ is given by

$$[\mathbf{\Sigma}]_{jk} = \begin{cases} 1 & \text{if } j = k \\ (\pi_{jk} - \frac{1}{2})\gamma + \frac{1}{2}a^2 + c^2 & \text{if } j \neq k \ . \end{cases} \tag{2.6}$$

Hence, for $\gamma = 0$, with $\rho = \frac{1}{2}a^2 + c^2$,

$$\mathbf{\Sigma} = (1-\rho)\mathbf{I} + \rho\mathbf{J} \ \text{ so } \ \mathbf{\Sigma}^{-1} = \frac{1}{1-\rho}\left(\mathbf{I} - \omega_m \mathbf{J}\right) \ , \tag{2.7}$$

with $\omega_m = \frac{\rho}{1+(m-1)\rho}$ where $\mathbf{I}$ is the $m \times m$ identity matrix and $\mathbf{J}$ is the $m \times m$ matrix whose elements are all equal to 1. It can be shown mathematically that the elements of the matrix $\mathbf{C} = \mathbf{\Sigma}^{-1}\mathbf{x}\left(\mathbf{\Sigma}^{-1}\mathbf{x}\right)' - \mathbf{\Sigma}^{-1}$ are given by

$$C_{ij} = \frac{1}{(1-\rho)^2}\left(x_i x_j - m\omega_m \bar{\mathbf{x}}(x_i + x_j) + (m\omega_m \bar{\mathbf{x}})^2\right) + \frac{1}{1-\rho}\omega_m \ . \tag{2.8}$$

Under the assumption of perfect marker information, the IBD distributions are uncorrelated for sib pairs within a sibship and have mean $\frac{1}{2}$, the score function is thus given by

$$\ell_\gamma^{\boldsymbol{\pi}} = \sum_{1 \le i < j \le m} C_{ij} \ \left(\pi_{ij} - \frac{1}{2}\right)$$

and Fisher's information by

$$\mathcal{I}_\gamma^{\boldsymbol{\pi}} = \frac{1}{8} \sum_{1 \le i < j \le m} C_{ij}^2 \ .$$

In **sib pair** designs, the two by two covariance matrix $\mathbf{\Sigma}$ is given by

$$\begin{pmatrix} 1 & \gamma(\pi - \frac{1}{2}) + \rho \\ \gamma(\pi - \frac{1}{2}) + \rho & 1 \end{pmatrix} .$$

The score function and information in $\gamma = 0$ are

$$\begin{aligned} \ell^{\pi}_{\gamma}(x_1, x_2; \rho) &= (\pi - \frac{1}{2})\ C(x_1, x_2; \rho) \\ \mathcal{I}^{\pi}_{\gamma}(x_1, x_2; \rho) &= \frac{1}{8}\ C^2(x_1, x_2; \rho) \end{aligned}$$

where

$$C(x_1, x_2; \rho) = \frac{(1+\rho^2)x_1x_2 - \rho(x_1^2 + x_2^2) + \rho(1-\rho^2)}{(1-\rho^2)^2} .$$

The score test in a sample of n independent sib pairs with phenotypes $(x_{i1}, x_{i2})_{i=1,\ldots,n}$ is given by

$$\frac{\sum_{i=1}^{n} \left(\pi_i - \frac{1}{2}\right)\ C(x_{i1}, x_{i2}; \rho)}{\sqrt{\frac{1}{8}\sum_{i=1}^{n} C^2(x_{i1}, x_{i2}; \rho)}}$$

and its robust version by

$$\frac{\sum_{i=1}^{n} (\pi_i - \frac{1}{2})\ C(x_{i1}, x_{i2}; \rho)}{\sqrt{\sum_{i=1}^{n} \left(\pi_i - \frac{1}{2}\right)^2\ C^2(x_{i1}, x_{i2}; \rho)}} .$$

The score test in that instance simply is the regression of the excess IBD sharing $\pi - \frac{1}{2}$ on a function of the trait values $C(\mathbf{x}; \rho)$ through the origin. This method was already proposed by Tang and Siegmund [2001] and Sham and Purcell [2001]. In a recent numerical comparison of methods for selected samples, Skatkiewicz et al. [2003] and Cuenco et al. [2003] showed that it has good properties in finite samples for extreme proband ascertained sib pairs and discordant sib pairs designs. The same test was also motivated heuristically using an approximation for excess IBD sharing in Putter et al. [2003].

In selected samples, one crucial feature of this regression as far as power is concerned, is that it is constrained through the origin. Indeed, the variance of the slope estimate in an unconstrained regression, which is inversely proportional to $\sum_i (C_i - \bar{C})^2 = \sum_i C_i^2 - n\bar{C}^2$, will always be greater than its constrained version, whose variance is inversely proportional to $\sum_i C_i^2$. The contour plot of C is displayed in Figure 2.1 for $\rho = 0.2$ and $\rho = 0.5$, with the corresponding trait values density indicated in gray scale (the density plots were generated using the scatterplots function

of Eilers and Goeman [2004]). It clearly shows that extreme concordant sib pairs have moderately large positive C values whereas extremely discordant sib pairs have large negative C values. As long as sib pairs are selected so that $\bar{C}$ is close to 0, whether the regression is constrained through the origin or not is irrelevant. However, should one consider only extremely discordant pairs, then $\bar{C}$ is negative and the power can increase dramatically, when using methods for selected samples.

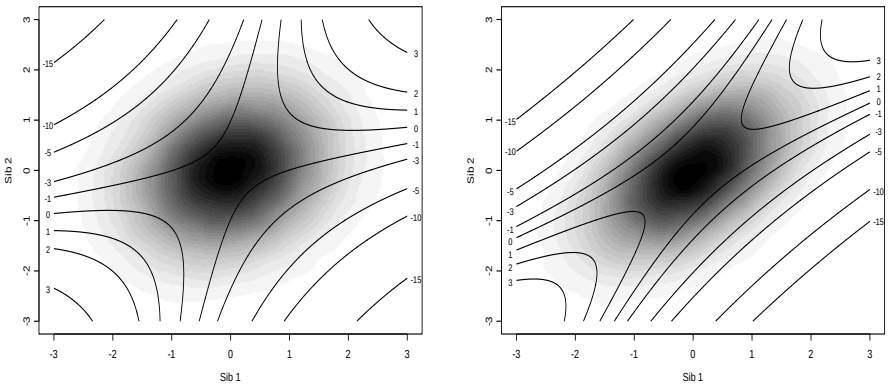

Figure 2.1: Joint distribution of sib trait values $\mathbf{x}$ (gray scale) and contour plot of $C(\mathbf{x}, \rho)$ ($\rho = 0.2$, left panel and $\rho = 0.5$, right panel)

Nuclear families

We now consider a general **nuclear family** with m sibs with trait value vector $\mathbf{x}_s$ and two parents with trait value vector $\mathbf{x}_p$, then the variance-covariance matrix $\mathbf{\Sigma}$ can be partitioned as

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{ss} & \mathbf{\Sigma}_{sp} \\ \mathbf{\Sigma}_{ps} & \mathbf{\Sigma}_{pp} \end{pmatrix} .$$

The sib-sib submatrix $\mathbf{\Sigma}_{ss}$ is the only submatrix to contain the linkage parameter γ. At $\gamma = 0$, $\mathbf{\Sigma}_{ss}$ is the same as (2.6) and (2.7) with ρ replaced by $\rho_{ss} = \frac{1}{2}a^2 + c^2$. The other submatrices are given by $\mathbf{\Sigma}_{sp} = \mathbf{\Sigma}'_{ps} = \rho_{sp}\mathbf{J}_{m2}$ and $\mathbf{\Sigma}_{pp} = (1 - \rho_{pp})\mathbf{I}_2 + \rho_{pp}\mathbf{J}_{22}$. Here, $\mathbf{I}_m$ is the identity matrix of dimension m and $\mathbf{J}_{ml}$ is the matrix of dimension $m \times l$ with all elements equal to 1. The parameter ρ_{sp} denotes the parent-sib trait

correlation and ρ_{pp} the father-mother trait correlation, both of which are assumed to be known. The correlations ρ_{ss}, ρ_{sp} and ρ_{pp} are given by 0.5, 0.5 and 0 times the additive genetic variance respectively, plus a scalar times the common environment variance. For ρ_{ss}, this multiplication factor will be 1 but we allow for smaller and mutually different factors for ρ_{sp} and ρ_{pp}. Matrices $\mathbf{\Sigma}_{sp}$ and $\mathbf{\Sigma}_{pp}$ do not involve the linkage parameter γ because there is no variation in IBD sharing between sibs and parents, nor between the two parents assuming they do not share alleles identical by descent. In practice however, parents are often genotyped because they are helpful in determining the IBD sharing of the siblings. With those conventions and using a similar reasoning as in (2.2) and (2.3), one can show that the score function for γ in the $\boldsymbol{\pi} \,|\, \mathbf{x}_p, \mathbf{x}_s$ model equals the score function for γ in the $\mathbf{x}_s \,|\, \boldsymbol{\pi}, \mathbf{x}_p$ model; in other words, the parents' phenotypes can simply be considered as 'covariates' in the analysis. Now, using standard results on conditional normal distributions, it turns out that

$$\mathbf{x}_s \,|\, \boldsymbol{\pi}, \mathbf{x}_p \;\sim\; N\left(\beta \bar{\mathbf{x}}_p,\; \mathbf{\Sigma}_{ss} - \rho_{sp}\beta \mathbf{J}_{mm}\right) \text{ with } \beta = \frac{2\rho_{sp}}{1+\rho_{pp}} \;,$$

thus

$$\left(\mathbf{x}_s - \beta \bar{\mathbf{x}}_p\right) \;/\; \left(1 - \rho_{sp}\beta\right)^{1/2} \,|\, \boldsymbol{\pi}, \mathbf{x}_p \;\sim\; N\left(0, \mathbf{\Sigma}_C\right) \;,$$

where $\mathbf{\Sigma}_C$ has diagonal elements equal to 1 and off-diagonal elements equal to

$$\left((\pi_{jk} - \frac{1}{2})\gamma + \rho_{ss} - \rho_{sp}\beta \right) \;/\; \left(1 - \rho_{sp}\beta\right) \;.$$

Finally, the score obtains as

$$\ell_\gamma^{\boldsymbol{\pi}} = \left(1 - \rho_{sp}\beta\right)^{-1} \sum_{1 \le i < j \le m} C_{ij} \left(\pi_{ij} - \frac{1}{2} \right)$$

and the information as

$$\mathcal{I}_\gamma^{\boldsymbol{\pi}} = \left(1 - \rho_{sp}\beta\right)^{-2} \frac{1}{8} \sum_{1 \le i < j \le m} C_{ij}^2 \;,$$

with C_{ij} given by formula (2.8) with $\mathbf{x} = \left(\mathbf{x}_s - \beta \bar{\mathbf{x}}_p\right) \;/\; \left(1 - \rho_{sp}\beta\right)^{1/2}$ and $\rho = \left(\rho_{ss} - \rho_{sp}\beta\right) \;/\; \left(1 - \rho_{sp}\beta\right)$. In most realistic situations ρ will be smaller than ρ_{ss}. The effect of including the parents on values of C is shown graphically in Figure 2.2. When the parent-sib trait correlation ρ_{sp} is small, whether parents are included or not

affects C mainly through the distortion of ρ. However when ρ_{sp} is substantial (e.g. high heritability or high household effect) and the parents' average trait values is high (or low), the effect is to shift the contour of C towards the north east quadrant (or south west quadrant) i.e. concordant siblings with non extreme values become valuable, whereas concordant siblings with extreme values become less attractive. For discordant pairs, the contour lines of C for average and extreme parents trait values cross, indicating that the inclusion of the extreme parents can affect C either way.

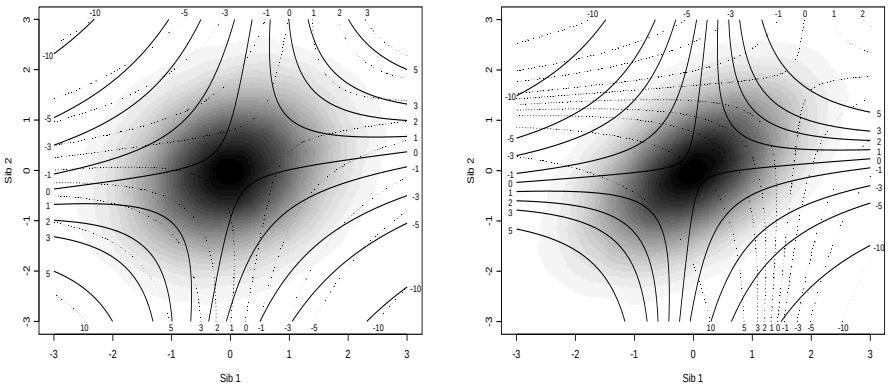

Figure 2.2: Joint distribution of sib trait values $\mathbf{x}$ (gray scale) and contour plot of $C(\mathbf{x}, \rho)$ (left panel: $\rho_{ss} = \rho_{sp} = 0.2$ and $\rho_{pp} = 0.1$, and right panel: $\rho_{ss} = \rho_{sp} = 0.5$ and $\rho_{pp} = 0.1$) for $\bar{\mathbf{x}}_p = 0$ (continuous lines, C values along vertical axis) and $\bar{\mathbf{x}}_p = 2$ (dotted lines, C values along horizontal axis)

Sibships and nuclear families of different sizes can easily be combined by weighting each family score according to its associated variance as suggested in Section 2.2.

2.4 Dominance

So far in our discussion we have neglected the effect of dominance. We show below what changes it involves in the score test compared to a fully additive model. We only consider here the most common design which allows evaluation of dominance variance component in non-inbred pedigrees: sibships consisting only of dizygotic twins or full

siblings. In presence of dominance, the conditional covariance $\mathbf{\Sigma}$ given the IBD status $\boldsymbol{\pi}$ becomes

$$[\mathbf{\Sigma}]_{jk} = \begin{cases} a^2 + d^2 + c^2 + e^2 \ = \ 1 \ , & \text{if } j = k \ , \\ (\pi_{jk} - \frac{1}{2})q^2 + (\mathbf{1}_{\{\pi_{jk}=1.0\}} - \frac{1}{4})t^2 & \text{if } j \neq k \ . \\ \quad + \frac{1}{2}a^2 + \frac{1}{4}d^2 + c^2 \ , & \end{cases}$$

where d^2 denotes total dominance variance and t^2 represents the proportion of total variance attributable to the dominance component at the locus of interest.

We re-parameterize the model as in Tang and Siegmund [2001] so as to make the terms involving π_{jk} uncorrelated, with mean 0 and same variance: let $\gamma = q^2 + t^2$ and $\delta = \frac{t^2}{\sqrt{2}}$. The covariance matrix $\mathbf{\Sigma}$ then writes

$$[\mathbf{\Sigma}]_{jk} = \begin{cases} 1 \ , & \text{if } j = k \ , \\ (\pi_{jk} - \frac{1}{2})\gamma - \frac{1}{\sqrt{2}}(\mathbf{1}_{\{\pi_{jk}=0.5\}} - \frac{1}{2})\delta & \text{if } j \neq k \ . \\ \quad + \frac{1}{2}a^2 + \frac{1}{4}d^2 + c^2 \ , & \end{cases}$$

The score for γ is as in formula (2.1) (however γ is now the sum of the additive and the dominant QTL variances) and the score with respect to δ is given by

$$\ell_\delta^{\boldsymbol{\pi}} = -\frac{1}{2\sqrt{2}} \ \mathrm{vec}(\mathbf{C})' \mathrm{vec}(\mathbf{1}_{\{\boldsymbol{\pi}=0.5\}} - \frac{1}{2}) \ .$$

Due to the new parameterization, $\ell_\gamma^{\boldsymbol{\pi}}$ and $\ell_\delta^{\boldsymbol{\pi}}$ are orthogonal under complete information (this is because π_{jk} and $\mathbf{1}_{\{\pi_{jk}=0.5\}}$ are uncorrelated in sib pairs [Amos et al., 1989]), and Fisher's information in $(\gamma, \delta) = (0, 0)$ is given by

$$\mathcal{I}_{\gamma,\delta}^{\boldsymbol{\pi}} = \begin{pmatrix} \mathcal{I}_\gamma^{\boldsymbol{\pi}} & 0 \\ 0 & \mathcal{I}_\delta^{\boldsymbol{\pi}} \end{pmatrix}$$

where $\mathcal{I}_\delta^{\boldsymbol{\pi}} = \frac{1}{8}\mathrm{vec}(\mathbf{C})' \ \mathrm{var}_{\boldsymbol{\pi}} \left(\mathrm{vec}(\mathbf{1}_{\{\boldsymbol{\pi}=0.5\}})\right) \mathrm{vec}(\mathbf{C})$ and $\mathcal{I}_\gamma^{\boldsymbol{\pi}}$ is given by formula (2.4). Under the assumption of a fully informative marker map $\mathcal{I}_\gamma^{\boldsymbol{\pi}} = \mathcal{I}_\delta^{\boldsymbol{\pi}} = \frac{1}{8}\sum_{1 \leq i < j \leq m} C_{ij}^2$, $\ell_\gamma^{\boldsymbol{\pi}} = \sum_{1 \leq i < j \leq m} C_{ij} \ \left(\pi_{ij} - \frac{1}{2}\right)$ and
$\ell_\gamma^{\boldsymbol{\pi}} = -\frac{1}{\sqrt{2}}\sum_{1 \leq i < j \leq m} C_{ij} \ \left(\mathbf{1}_{\{\pi_{ij}=0.5\}} - \frac{1}{2}\right)$ with C_{ij} as in formula (2.8), and the one-sided score test of the joint null hypothesis $(\gamma, \delta) = (0, 0)$ under the constraint $0 \leq$

$\sqrt{2}\,\delta \leq \gamma$ is then given by

$$z_+^2 = \begin{cases} \frac{\ell_\gamma^{\boldsymbol{\pi}\,2}}{\mathcal{I}_\gamma^{\boldsymbol{\pi}}} + \frac{\ell_\delta^{\boldsymbol{\pi}\,2}}{\mathcal{I}_\delta^{\boldsymbol{\pi}}} \;, & \text{if } 0 \leq \sqrt{2}\,\ell_\delta^{\boldsymbol{\pi}} \leq \ell_\gamma^{\boldsymbol{\pi}} \;, \\ \frac{\ell_\gamma^{\boldsymbol{\pi}\,2}}{\mathcal{I}_\gamma^{\boldsymbol{\pi}}} \;, & \text{if } 0 < \ell_\gamma^{\boldsymbol{\pi}} \text{ and } 0 < \ell_\delta^{\boldsymbol{\pi}} \;, \\ \frac{1}{3}\,\left(\sqrt{2}\,\ell_\gamma^{\boldsymbol{\pi}} + \ell_\delta^{\boldsymbol{\pi}}\right)^2 \;, & \text{if } -\frac{1}{\sqrt{2}}\,\ell_\delta^{\boldsymbol{\pi}} < \ell_\gamma^{\boldsymbol{\pi}} < \sqrt{2}\,\ell_\delta^{\boldsymbol{\pi}} \text{ and } \ell_\delta^{\boldsymbol{\pi}} > 0 \;, \\ 0 \;, & \text{otherwise .} \end{cases}$$

The local optimality properties of the univariate score test are preserved by this statistic since it is asymptotically equivalent to the likelihood ratio test [Verbeke and Molenberghs, 2003]. Under the null hypothesis of no locus effect, z_+^2 is distributed as $(1-\kappa)\chi_0^2 + \frac{1}{2}\chi_1^2 + \kappa\chi_2^2$ with $\kappa = 0.098$ [Shapiro, 1988]. Note that this test is the same as the one proposed by Wang and Huang [2002b] (see Section 2.6 for a closer comparison).

2.5 Dichotomous traits

Zeegers et al. [2003] have developed a modified Haseman-Elston regression for binary traits and have shown that it is approximately equivalent in power to the liability-threshold variance components model. In order to apply similar ideas to those developed in previous sections to dichotomous traits we use this so-called liability threshold model. Under such setting, a continuous variable arbitrarily scaled to have mean 0 and variance 1 underlies the trait of interest. In pedigrees involving only one type of family members relationship like sibships, the model is fully characterized by two parameters: the overall prevalence of the trait K (or equivalently the liability threshold t where $K = 1 - \Phi(t)$, Φ denotes here the cumulative density function of a standard normal) and the correlation ρ between the scaled liabilities of two sibs, also known as the tetrachoric correlation for the trait of interest. Different types of family members relationship may correspond to different tetrachoric correlations. Provided population data are available, the maximum likelihood method can be used to obtain estimates of the tetrachoric correlation between different relative pairs. Approximate formulae due to Pearson [1901] appear in Sham [1998, Section 5.5.5].

The probability $p_\gamma(\mathbf{y} \,|\, \boldsymbol{\pi})$ of the affection states of the pedigree members being $\mathbf{y}$, given $\boldsymbol{\pi}$, where $\mathbf{y}$ is one of the possible phenotypes, is obtained by integration of the density $f_\gamma(\mathbf{x} \,|\, \boldsymbol{\pi})$ for the underlying liability as expressed in the variance components

setting of Section 2.2 over $R_{\mathbf{y}}$, the region corresponding to phenotype $\mathbf{y}$ on the liability scale

$$p_\gamma(\mathbf{y} \,|\, \boldsymbol{\pi}) = \int_{\mathbf{x} \in R_{\mathbf{y}}} f_\gamma(\mathbf{x} \,|\, \pi) d\mathbf{x} \ .$$

The score $\ell_\gamma^{\mathbf{y}}$ for $p_\gamma(\mathbf{y} \,|\, \boldsymbol{\pi})$ at $\gamma = 0$ equals

$$\ell_\gamma^{\mathbf{y}} = \frac{\partial}{\partial \gamma} \log p_\gamma(\mathbf{y} \,|\, \boldsymbol{\pi}) = \frac{\int_{R_{\mathbf{y}}} \frac{\partial}{\partial \gamma} f_\gamma(\mathbf{x} \,|\, \pi) d\mathbf{x}}{\int_{R_{\mathbf{y}}} f_\gamma(\mathbf{x} \,|\, \pi) d\mathbf{x}} = \frac{\int_{R_{\mathbf{y}}} \ell_\gamma^{\mathbf{x}} f_\gamma(\mathbf{x} \,|\, \pi) d\mathbf{x}}{\int_{R_{\mathbf{y}}} f_\gamma(\mathbf{x} \,|\, \pi) d\mathbf{x}} = \mathbf{E}_{\mathbf{x}} \left(\ell_\gamma^{\mathbf{x}} \,|\, \mathbf{x} \in R_{\mathbf{y}} \right) \ .$$

As for the continuous case, the score $\ell_\gamma^{\boldsymbol{\pi}}$ for γ of the selection model $\boldsymbol{\pi} \,|\, \mathbf{y}$ is equal to the score $\ell_\gamma^{\mathbf{y}}$ for the $\mathbf{y} \,|\, \boldsymbol{\pi}$ model. Using formula (2.1) and by linearity of the expectation $\mathbf{E}$,

$$\ell_\gamma^{\boldsymbol{\pi}} = \ell_\gamma^{\mathbf{y}} = \frac{1}{2} \ \mathrm{vec}(\mathbf{C_y})' \mathrm{vec}(\boldsymbol{\pi} - \mathbf{E}\boldsymbol{\pi}) \ ,$$

and

$$\mathcal{I}_\gamma^{\boldsymbol{\pi}} = \frac{1}{4} \ \mathrm{vec}(\mathbf{C_y})' \ \mathrm{var}_{\boldsymbol{\pi}} \left(\mathrm{vec}(\boldsymbol{\pi}) \right) \ \mathrm{vec}(\mathbf{C_y})$$

with $C_{\mathbf{y}} = \mathbf{E}_{\mathbf{x}}(\mathbf{C}(\mathbf{x}, \rho) \,|\, \mathbf{x} \in R_{\mathbf{y}})$.

In the case of **sib pair** designs, there are only three possible unordered phenotypes: Affected/Affected (AA), Affected/Unaffected (AU) and Unaffected/Unaffected (UU). This implies that there are only three possible values of $\mathbf{C_y}$: C_{AA}, C_{AU}, C_{UU}, each corresponding to the conditional expectation of $C(\mathbf{x}, \rho)$, given $\mathbf{x}$ in the appropriate region on the liability scale. For a data set consisting of n_{AA} affected sib pairs, n_{AU} discordant sib pairs and n_{UU} unaffected sib pairs, the score test then equals

$$z = \frac{C_{AA} \sum_{i \in AA} \left(\pi_i - \frac{1}{2} \right) + C_{AU} \sum_{i \in AU} \left(\pi_i - \frac{1}{2} \right) + C_{UU} \sum_{i \in UU} \left(\pi_i - \frac{1}{2} \right)}{\sqrt{\frac{1}{8} \left(n_{AA} C_{AA}^2 + n_{AU} C_{AU}^2 + n_{UU} C_{UU}^2 \right)}} \ ,$$

and a robust score test is given by

$$z^* = \frac{C_{AA} \sum_{i \in AA} \left(\pi_i - \frac{1}{2} \right) + C_{AU} \sum_{i \in AU} \left(\pi_i - \frac{1}{2} \right) + C_{UU} \sum_{i \in UU} \left(\pi_i - \frac{1}{2} \right)}{\sqrt{C_{AA}^2 \sum_{i \in AA} \left(\pi_i - \frac{1}{2} \right)^2 + C_{AU}^2 \sum_{i \in AU} \left(\pi_i - \frac{1}{2} \right)^2 + C_{UU}^2 \sum_{i \in UU} \left(\pi_i - \frac{1}{2} \right)^2}} \ .$$

Nowadays, the $\mathbf{C_y}$ quantities can be approximated to a sufficient degree of precision using Monte Carlo simulation techniques.

Values of C_{AA}, C_{AU} and C_{UU} are provided in Table 2.1 for typical values of the tetrachoric correlation ρ and trait prevalence K. Under this liability threshold model, the main characteristics of the sib pair designs are that UU sib pairs provide

		AA		AU		UU				AA		AU		UU	
K	ρ	Prob.	$\bar{C}$	Prob.	$\bar{C}$	Prob.	$\bar{C}$	K	ρ	Prob.	$\bar{C}$	Prob.	$\bar{C}$	Prob.	$\bar{C}$
0.001	0.1	0.0000	9.63	0.0020	-0.04	0.9980	0.00	0.05	0.1	0.0037	3.68	0.0926	-0.29	0.9037	0.02
	0.2	0.0000	8.26	0.0020	-0.06	0.9980	0.00		0.2	0.0053	3.25	0.0895	-0.38	0.9053	0.02
	0.3	0.0000	7.24	0.0020	-0.10	0.9980	0.00		0.3	0.0071	2.92	0.0857	-0.49	0.9071	0.02
	0.4	0.0000	6.43	0.0019	-0.20	0.9980	0.00		0.4	0.0094	2.67	0.0812	-0.62	0.9094	0.03
	0.5	0.0001	5.82	0.0019	-0.34	0.9981	0.00		0.5	0.0122	2.48	0.0756	-0.80	0.9122	0.03
	0.6	0.0001	5.37	0.0018	-0.55	0.9981	0.00		0.6	0.0155	2.36	0.0690	-1.06	0.9155	0.04
0.01	0.1	0.0002	6.06	0.0196	-0.12	0.9802	0.00	0.1	0.1	0.0133	2.69	0.1733	-0.41	0.8133	0.04
	0.2	0.0003	5.27	0.0193	-0.18	0.9803	0.00		0.2	0.0172	2.40	0.1656	-0.50	0.8172	0.05
	0.3	0.0006	4.66	0.0189	-0.27	0.9806	0.00		0.3	0.0216	2.18	0.1567	-0.60	0.8216	0.06
	0.4	0.0009	4.20	0.0183	-0.40	0.9809	0.00		0.4	0.0267	2.02	0.1468	-0.73	0.8266	0.06
	0.5	0.0013	3.85	0.0174	-0.57	0.9813	0.01		0.5	0.0324	1.90	0.1352	-0.91	0.8324	0.07
	0.6	0.0019	3.60	0.0163	-0.83	0.9819	0.01		0.6	0.0390	1.83	0.1220	-1.17	0.8390	0.08
0.02	0.1	0.0007	5.02	0.0386	-0.18	0.9607	0.00	0.2	0.1	0.0481	1.75	0.3038	-0.55	0.6481	0.13
	0.2	0.0011	4.39	0.0378	-0.26	0.9611	0.01		0.2	0.0568	1.58	0.2864	-0.63	0.6568	0.14
	0.3	0.0017	3.91	0.0366	-0.35	0.9617	0.01		0.3	0.0662	1.46	0.2676	-0.72	0.6661	0.15
	0.4	0.0024	3.54	0.0352	-0.49	0.9624	0.01		0.4	0.0762	1.37	0.2476	-0.85	0.6762	0.15
	0.5	0.0034	3.26	0.0332	-0.67	0.9634	0.01		0.5	0.0871	1.31	0.2256	-1.02	0.6872	0.17
	0.6	0.0046	3.07	0.0307	-0.93	0.9646	0.01		0.6	0.0992	1.29	0.2015	-1.27	0.6992	0.18

Table 2.1: Probabilities of affection states and average C values for sib pairs

very little information whereas AA sib pairs provide the most information especially as the trait becomes rare. However, it must be stressed that as the prevalence of the trait increases, AU sib pairs become more informative. If only one type of phenotype is used (say only affected sib pairs) the score test is equivalent to $z = \frac{(\bar{\pi}-\frac{1}{2})}{\sqrt{1/(8n)}}$ and the robust score test equal $z^* = \frac{(\bar{\pi}-\frac{1}{2})}{\widehat{\text{se}}(\bar{\pi})}$ which are two standardized versions of the mean IBD sharing test. These tests are well established [Blackwelder and Elston, 1985] and have been in popular use for decades. As for the continuous case the test can be seen as a regression through the origin of the excess IBD sharing on a function C of the trait, however the function C only takes a limited number of distinct values. To illustrate this regression, we generated the affection states for 10000 sib pairs using the liability threshold model with $K = 0.05$, $\rho = 0.4$ and $\gamma = 0.15$. The 150 most informative pairs were selected using the corresponding $\bar{C}^2$ obtained from table 2.1; this resulted in all 97 affected pairs and 53 random discordant pairs being selected. Figure 2.3 illustrates the regression for this simulated data set.

One attractive feature of our approach is that it naturally allows combination of sib pairs of different nature (more generally, pedigree pairs of different nature and familial relationships). Each type of pairs contributes to the deviation from average IBD sharing with a weight proportional to the average value of the C function in the corresponding region. Note that in practice, table I can also be used with pedigrees consisting of other types of relative pairs. For example, if n^c_{AA} pedigrees consisting of affected cousins also are available then their contribution to the numerator of the previous z will simply be $C_{AA} \sum_{i=1}^{n^c_{AA}} (\pi^c_i - \frac{1}{8})$ where C_{AA} is drawn from table I with K as the population prevalence of the trait and ρ equal to the trait tetrachoric correlation between cousins. Our approach also offers an elegant solution to the problem of prevalence heterogeneity in the population: if a data set consists of groups with different disease prevalence, the contribution of each group to the overall test is weighted accordingly (see Table I).

2.6 Discussion

In the context of the variance components model, we have given an expression of the score test for linkage under sample selection based on phenotype values. It is

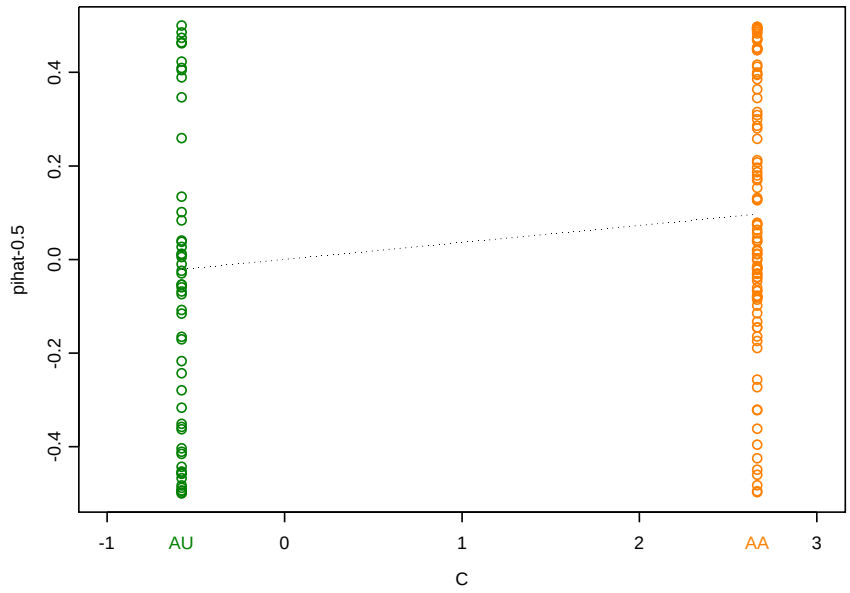

Figure 2.3: Regression of $\pi - \frac{1}{2}$ on $C(\mathbf{x}, \rho)$ for 150 selected sib pairs ($K = 0.05$, $\rho = 0.4$ and $\gamma = 0.15$)

a general expression for arbitrary pedigrees which takes a very simple form in some widely used designs. Commenges [1994] first introduced score tests in the context of linkage, however his approach is not conditional on trait values and therefore leads to reduced power in selected samples. In a recent article, Tritchler et al. [2003] give a general score test in nuclear families conditional on the trait values under the assumption that the trait distribution depends on different genetic models through the exponential family. Our results give a very similar expression to theirs. In their software implementation, they allow the population mean to be specified by the user but not the population sib-sib correlation and our understanding is that the authors attempt to estimate this correlation from the selected data, which potentially results in power loss (unless the ascertainment mechanism is known). Our approach is to fully acknowledge the fact that selected samples do not provide unbiased estimates of the population trait distribution characteristics and to assume that unbiased estimates

of the first and second moments of the population trait are available a priori. In the context of the GenomEUtwin project, where twin registries provide us with precise population mean and twin-twin correlation, this seems a realistic assumption.

The score test that we derive also has a simple interpretation in terms of regression of IBD sharing on a function of the phenotypes. Sham et al. [2002] have recently proposed a general method of analysis for quantitative linkage data which explicitly regresses IBD sharing on all possible squared sums and differences of trait values within a family. As shown in Section 2.2, the score test essentially is a regression of the excess IBD sharing on a quadratic function of the trait values whose shape depends on the normality assumption. When the data truly are normal, it seems reasonable to expect that the score test results in similar regressor as in the method of Sham et al. [2002]. We have compared the information content provided by the two methods in sibships and nuclear families of different sizes and they happen to exactly coincide. In fact, as demonstrated in a recently published paper [Chen et al., 2004], the two methods are the same for quantitative traits under an additive model (with trait correlations assumed to be the same over all pairs of relatives). The IBD covariance matrix is determined solely by family relations; no marker information is needed to compute it, which is a prerequisite to make it useful for selection prior to genotyping. Note that calculation of the information index in [Sham et al., 2002] does not require marker information either.

One possible criticism of the variance components model is that departure from the normality assumption might invalidate its results. However, the analogy of the test with regression methods, very much as the score test in unselected data coincides with the optimally weighted Haseman-Elston regression [Putter et al., 2002], pleads in favor of its robustness. In fact, as the regression interpretation of the score reveals, the test depends on the distribution of the trait values only through its second order moments. So as long as the shape of the distribution does not show any great departure from normality for those moments (e.g. heavy tail) then the test should remain valid. When the model clearly is wrong, the robust version of the test should preclude over-optimistic inference.

We showed in Section 2.2 that in the current variance components setting under which population marginal characteristics are known and the only unknown parameter

is the linkage parameter γ, there is no loss of information when conditioning on trait values. This is a direct consequence of the fact that scores for the selection model $\boldsymbol{\pi} \,|\, \mathbf{x}$, the $\mathbf{x} \,|\, \boldsymbol{\pi}$ model and the joint $(\mathbf{x}, \boldsymbol{\pi})$ model are identical. The situation becomes more complicated when population parameters are unknown and need to be conjunctly estimated.

As announced in Section 2.2, we now turn to the case of imperfect IBD information. In practice, $\boldsymbol{\pi}$ is not known with certainty. In fact, the only available data are marker information which we denote M and the phenotypes $\mathbf{x}$. Strictly speaking, the likelihood to be considered should be expressed in terms of those data, i.e. we should write $f_\gamma(M, \mathbf{x})$ for the joint distribution of M and $\mathbf{x}$ and $f_\gamma(M \,|\, \mathbf{x})$ for the conditional distribution of $M \,|\, \mathbf{x}$. It turns out that the score ℓ_γ^M for the $M \,|\, \mathbf{x}$ distribution simply becomes the weighted average of the score $\ell_\gamma^{\boldsymbol{\pi}}$ for the idealized fully informative model $\ell_\gamma^M = \sum_{\boldsymbol{\pi}} P(\boldsymbol{\pi} \,|\, M)\ \ell_\gamma^{\boldsymbol{\pi}}$ and thus, with $\hat{\boldsymbol{\pi}} = \mathbf{E}(\boldsymbol{\pi} \,|\, M)$,

$$\ell_\gamma^M = \frac{1}{2} \mathrm{vec}(\mathbf{C})' \mathrm{vec}(\hat{\boldsymbol{\pi}} - \mathbf{E}\hat{\boldsymbol{\pi}}) \ .$$

Since $\mathbf{E}\hat{\boldsymbol{\pi}} = \mathbf{E}\pi$, this result means that Formula (2.1) still holds true with imperfect data but $\boldsymbol{\pi}$ values have to be replaced by estimates given marker data available $\hat{\boldsymbol{\pi}}$. Values of $P(\boldsymbol{\pi} \,|\, M)$ and $\hat{\boldsymbol{\pi}}$ are calculated using for example the Lander-Green or Elston-Stewart algorithms [Lander and Botstein, 1989] as implemented in publicly available softwares like `GENEHUNTER` [Kruglyak et al., 1996] or `MERLIN` [Abecasis et al., 2002]. Note that this result theoretically justifies (as mentioned by Commenges [1994] and Tang and Siegmund [2001]) the use of the so-called $\hat{\pi}$ approach in variance components linkage modelling for arbitrary pedigrees. The corresponding Fisher's information is given by

$$\mathcal{I}_\gamma^M = \frac{1}{4}\ \mathrm{vec}(\mathbf{C})'\ \mathrm{var}_M\left(\mathrm{vec}(\hat{\boldsymbol{\pi}})\right)\ \mathrm{vec}(\mathbf{C}) \ .$$

Given a marker map and a certain pedigree structure, Monte Carlo simulations can be used to approximate $\mathrm{var}_M\left(\mathrm{vec}(\hat{\boldsymbol{\pi}})\right)$. A conservative alternative is to use $\mathcal{I}_\gamma^{\boldsymbol{\pi}}$ as given by Formula (2.4) instead of $\mathcal{I}_\gamma^M$ in the standardization of ℓ_γ^M. One consequence of imperfect information in the case of sibships for example is that negative terms appear on the off-diagonal components of the $\mathrm{var}_M\left(\mathrm{vec}(\hat{\boldsymbol{\pi}})\right)$ matrix. When considering both additive and dominance variance components, the scores $\ell_\gamma^{\boldsymbol{\pi}}$ and $\ell_\delta^{\boldsymbol{\pi}}$ derived

in Section 2.4 are no longer orthogonal and the use of the test as defined in that section is not optimal. It is possible to obtain the expression of a multivariate score test that is asymptotically optimal [Verbeke and Molenberghs, 2003] and whose null distribution ($(1-\kappa)\chi^2_0 + \frac{1}{2}\chi^2_1 + \kappa\chi^2_2$, where κ depends on informativeness) can be obtained using results in Shapiro [1988]. The details are beyond the scope of this article, however the results appear in Wang and Huang [2002b] who consider only random samples and therefore suggest to estimate the sib-sib correlation as well as $\mathbf{P}(\boldsymbol{\pi} = 0.5 \,|\, M)$, $\mathbf{E}(\hat{\boldsymbol{\pi}})$ and $\mathrm{var}(\hat{\boldsymbol{\pi}})$ from the data. Interestingly, our derivation shows that their approach is perfectly valid in selected samples too, provided the population sib-sib correlation is known and unbiased values for $\mathbf{P}(\boldsymbol{\pi} = 0.5 \,|\, M)$, $\mathbf{E}(\hat{\boldsymbol{\pi}})$ and $\mathrm{var}(\hat{\boldsymbol{\pi}})$ are calculated (e.g. using Monte Carlo simulation technique described above). Note that in selected samples, the use of population estimates for those 'nuisance' parameters amounts to constraining the regression through the origin and is critical in order to maintain maximum power. In practice, the asymptotic results might fail to hold in finite samples and it seems wise to use re-sampling methods (bootstrap) in order to obtain a robust assessment of significance.

By use of the liability threshold model, the continuous case extends to the case of dichotomous traits. Because of the well-known optimality properties of the score test (which is asymptotically equivalent to the likelihood-ratio test), it provides an efficient means to test for linkage in affected sib pairs and in discordant sib pairs as well as a way to combine the two types of data when needs arise. More complicated pedigrees can also be handled in a very flexible manner. In this selection framework where IBD sharing $\boldsymbol{\pi}$ is considered conditional on the trait values $\mathbf{x}$, the extension to multiple traits, in analogy with multiple regression, should be fairly straightforward.

This score test approach has been implemented into a C program calling upon the publicly available software `MERLIN` [Abecasis et al., 2002] and is available at `http://www.msbi.nl/Genetics` .

2.7 Appendix

Score test

The score function for γ in the $\mathbf{x} \,|\, \boldsymbol{\pi}$ model is denoted by $\ell_\gamma^{\mathbf{x}}$ and by definition equals $\ell_\gamma^{\mathbf{x}} = \frac{\partial}{\partial \gamma} \log f_\gamma(\mathbf{x} \,|\, \boldsymbol{\pi})$ with

$$\log f_\gamma(\mathbf{x} \,|\, \boldsymbol{\pi}) = -\frac{m}{2} \log(2\pi) - \frac{1}{2} \log(|\boldsymbol{\Sigma}|) - \frac{1}{2} \mathbf{x}' \boldsymbol{\Sigma}^{-1} \mathbf{x}$$

Standard results on matrix algebra (see, e.g. [Searle et al., 1992, Appendix M.7]) show that

$$\ell_\gamma^{\mathbf{x}} = \frac{1}{2} \left\{ \mathbf{x}' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\pi} - \mathbf{E}\boldsymbol{\pi}) \boldsymbol{\Sigma}^{-1} \mathbf{x} - \operatorname{tr}(\boldsymbol{\Sigma}^{-1} (\boldsymbol{\pi} - \mathbf{E}\boldsymbol{\pi})) \right\}$$

Because of the relation $a'b = \operatorname{tr}(ba')$, the previous equation can be rewritten

$$\ell_\gamma^{\mathbf{x}} = \frac{1}{2} \operatorname{tr} \left(\boldsymbol{\Sigma}^{-1} (\boldsymbol{\pi} - \mathbf{E}\boldsymbol{\pi}) (\boldsymbol{\Sigma}^{-1} \mathbf{x}\mathbf{x}' - \mathbf{I}) \right) \ .$$

Chapter 3

Selection Strategies for Linkage Studies using Twins

Abstract

Genetic linkage analysis for complex diseases offer a major challenge to geneticists. In these complex diseases multiple genetic loci are responsible for the disease and they may vary in the size of their contribution; the effect of any single one of them is likely to be small. In many situations, like in extensive twin registries, trait values have been recorded for a large number of individuals, and preliminary studies have revealed summary measures for those traits, like mean, variance and components of variance, including heritability.

Given the small effect size, a random sample of twins will require a prohibitively large sample size. It is well known that selective sampling is far more efficient in terms of genotyping effort.

In this paper we derive easy expressions for the information contributed by sib pairs for the detection of linkage to a quantitative trait locus (QTL). We consider random samples as well as samples of sib pairs selected on the basis of their trait values. These expressions can be rapidly computed and do not involve simulation. We extend our results for quantitative traits to dichotomous traits using the concept of a liability threshold model.

We present tables with required sample sizes for height, insulin levels and migraine, three of the traits studied in the GenomEUtwin project.

This chapter has been published as: H. Putter, J. Lebrec and J.C. van Houwelingen (2003). Selection Strategies for Linkage Studies using Twins. *Twin Research* **6** (5), 377–382.

3.1 Introduction

Genetic linkage analysis (gene mapping) has proved to be a powerful tool for the identification of genes responsible for monogenic inherited diseases such as Huntington disease and cystic fibrosis. The diseases for which the genetic basis has not yet been unravelled do not display a one-to-one correspondence between a single gene and disease status. In these complex diseases, multiple genetic loci are responsible for the disease and these genetic loci may vary in the size of their contribution, they may interact with each other and with external, environmental factors. The effect of any single one of these genes is likely to be small [Risch, 2000].

The GenomEUtwin project comprises a very large source of twins, through the union of a number of large twin registries in different countries in Europe. For the majority of these twins, data on a number of traits of interest have already been recorded. Examples include quantitative traits like height, BMI, risk factors for cardiovascular disease and qualitative traits like migraine, diabetes. Some of these traits are recorded repeatedly over time and require methods for longitudinal data, others can be thought of as having an age of onset and can be treated like survival data.

The first step in unravelling the genetic basis of a disease is to undertake a heritability study. Twin studies are ideally equipped for this purpose, because of the inherent matching for age and other environmental factors, and because of the differential degree of shared genetic variance between monozygotic (MZ) and dizygotic (DZ) twins [Boomsma et al., 2002]. For many quantitative traits of interest, twin studies (or similar studies) have given information on the distribution of the trait in the target population, in particular their mean and variance, and on the heritability.

In the planning phase of a linkage study, one of the important issues is the choice of sib pairs to be included in a scan. The good news is that for large twin registries, the number of phenotypes is in principle adequate even to detect very small genetic effects. Unfortunately, given the anticipated small genetic effect at any one disease locus, a random sample to achieve 80% power is most probably prohibitively large in terms of genotyping effort, even with the current high throughput genotyping technologies. Eaves and Meyer [1994] and Risch and Zhang [1995] showed that similar power to large random samples can be obtained by selecting only a small subset of

extreme discordant pairs. Many studies have later refined these recommendations, giving, under an assumed model, optimal selection strategies for linkage studies. The drawback of these studies is that they typically require simulation and fail to give quick, easy and insightful assessments of the amount of information that a given sib pair is expected to contribute.

In this paper, it is our aim to outline easily computable information content numbers for twins in the context of linkage twin studies for complex diseases. We start in Section 3.2 by considering quantitative traits, with given heritability, mean and variance, assuming that the effect of the quantitative trait locus is small. We replace much of the simulation employed in the above papers by explicit calculation, resulting in particularly easy expressions for the information content for DZ sib pairs. The result is an easy expression closely related to optimal Haseman-Elston regression [Sham and Purcell, 2001] and the score function for the QTL variance in a variance components model [Putter et al., 2002]. We then show in Section 3.3 how the concept of a latent underlying quantitative trait can be used to extend these results to dichotomous traits. Section 3.4 discusses issues like extended pedigrees and dominance variance.

3.2 Selection strategies for quantitative traits

Random sampling

Starting point of our selection procedure for quantitative traits is the variance components model [Schork, 1993; Amos, 1994]. We assume that the traits have been standardised so as to have zero mean and unit variance. For a DZ twin sharing i alleles identical by descent (IBD) at a particular marker locus, the distribution of their phenotypes $\mathbf{x} = (x_1, x_2)$ is assumed to follow a bivariate normal distribution with mean vector 0 and covariance matrix

$$\mathbf{\Sigma}_i = \begin{pmatrix} 1 & \rho + \frac{i-1}{2}\gamma \\ \rho + \frac{i-1}{2}\gamma & 1 \end{pmatrix} .$$

Here ρ and γ represent the proportion of this variance that can be attributed to shared components and the quantitative trait locus respectively. The parameter ρ is half of the heritability (h^2) plus the proportion of common environment variance, c^2. In what follows we consider DZ twins, since MZ twins are not informative for linkage.

We shall refer to DZ twins as sib pairs in the sequel; for our purposes there is no distinction between sib pairs and DZ twins.

The amount of information I at $\gamma = 0$ contributed by one sib pair is given by

$$I = \frac{1}{8}\frac{1+\rho^2}{(1-\rho^2)^2} \ . \tag{3.1}$$

This formula has been derived by Williams and Blangero [1999] and is a special case of our equation (3.5). The factor $1/8$ represents the variance of $\hat{\pi}$ for sib pairs for a fully informative marker [Rijsdijk et al., 2001]. This implies that an estimate of γ based on a random sample of n sib pairs will have a standard error of $\mathrm{se}(\hat{\gamma}) = \frac{1}{\sqrt{nI}}$, in the absence of nuisance parameters. This fact can be used to determine the number of sib pairs required to achieve power $1-\beta$ to detect linkage with a QTL effect size γ, using a significance level α,

$$n = \frac{(z_\alpha + z_\beta)^2}{I\gamma^2} \ . \tag{3.2}$$

Here z_α denotes the $1-\alpha$ percentile of the standard normal distribution. For a power of 80% and a significance level of 0.0001, corresponding to a lod-score of 3, this leads to $n = \frac{20.8}{I\gamma^2}$. Graphs for different values of ρ are shown in Figure 3.1.

For a quantitative trait like height, with an estimated heritability of 0.80 and an estimated common environment variance $c^2 = 0.1$, and hence a value or $\rho = 0.5$, we need to genotype approximately 7500 sib pairs or 15000 individuals to detect linkage with a moderate QTL effect of $\gamma = 0.1$. Clearly, this is not feasible, even with the current high-throughput genotyping technology.

Selective sampling

Risch and Zhang [1995] suggested selecting sib pairs for genotyping on the basis of their trait values and showed that considerably higher efficiency can be obtained by selecting extreme discordant sib pairs. Later, these recommendations have been refined, most of the papers employing simulation to calculate the information content of a sib pair [Dolan and Boomsma, 1998b; Cherny et al., 1999]. A noteworthy exception is the paper by Purcell et al. [2001], where the information content is obtained through an exact calculation that considers all possible genotypes at the quantitative trait locus. We show below a simple approach that can also be used to obtain explicit

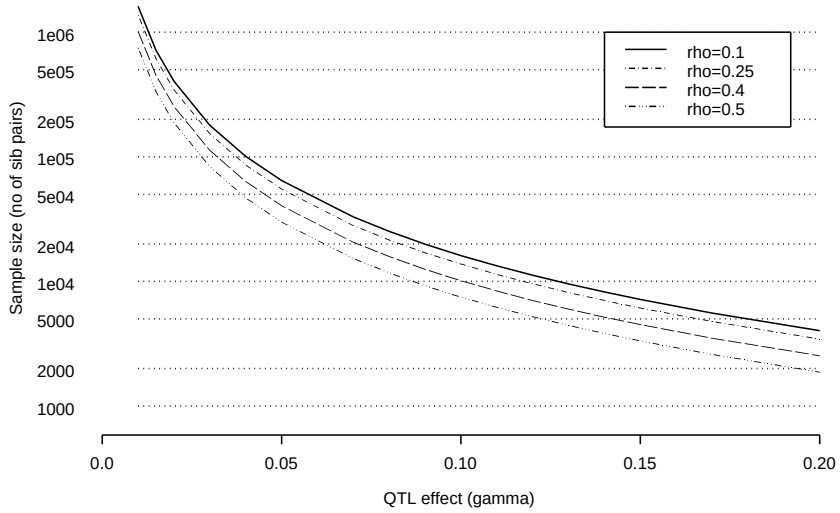

Figure 3.1: Number of sib pairs needed in a random sample to detect linkage to a quantitative trait for different values of ρ and γ. Power is 80%; significance level = 0.0001, corresponding to a lod-score of 3. For 50%, 60% and 70% power respectively, required sample sizes decrease by a factor of 1.50, 1.32 and 1.16 respectively.

expressions for the information content for a number of common designs without the need to do simulations.

The variance components model specifies the conditional distribution of the phenotypes, given the genotypes (IBD-sharing). When dealing with selected samples, it is more natural to invert the reasoning and to think of the phenotypes as given [Sham et al., 2000]. This approach is common for the analysis of dichotomous traits. Let z denote the number of alleles shared IBD by the twins at the marker locus, and $\hat{\pi}$ the proportion of alleles shared IBD. Since it is anticipated that the effect of any single gene is small, we use a linear expansion in γ along with Bayes' theorem to obtain, neglecting terms of smaller order than γ,

$$\begin{aligned} P(z=0|\mathbf{x},\gamma,\rho) &= \frac{1}{4} - \frac{\gamma}{8}C(\mathbf{x},\rho) \ , \\ P(z=1|\mathbf{x},\gamma,\rho) &= \frac{1}{2} \ , \\ P(z=2|\mathbf{x},\gamma,\rho) &= \frac{1}{4} + \frac{\gamma}{8}C(\mathbf{x},\rho) \ , \\ E(\hat{\pi}|\mathbf{x},\gamma,\rho) &= \frac{1}{2} + \frac{\gamma}{8}C(\mathbf{x},\rho) \ . \end{aligned} \tag{3.3}$$

Here,

$$C(\mathbf{x},\rho) = \frac{1}{(1-\rho^2)^2}\left((1+\rho^2)x_1x_2 - \rho(x_1^2+x_2^2) + \rho(1-\rho^2)\right)$$

is the "optimal Haseman-Elston " function [Sham and Purcell, 2001], which was shown to be the score function for the parameter γ in the variance components model [Putter et al., 2002]. Values of $C(\mathbf{x},\rho)$ range from negative to positive. Details of the derivation and extension to general pedigrees can be found in Lebrec et al. [2004].

This observation suggests using a regression method like the Haseman-Elston regression method, as already proposed by Sham et al. [2002], for the analysis of selected samples. The regression for sib pairs amounts to the inverse of the optimal Haseman-Elston regression, namely regressing $\hat{\pi}$ on $C(\mathbf{x},\rho)$. A test for linkage in this setting is a one-sided test for a positive slope in this regression. Indeed, for the case of sib pairs, our results coincide with those found in Sham et al. [2002].

In the context of regression, simple rules are available for selecting samples on the basis of the explanatory variables: since the square of the standard error of the slope of a regression of y on x is inversely proportional to $\sum(x_i-\bar{x})^2$, values of x should be

chosen as widely spaced as possible. This means that sib pairs with extreme values of $C(\mathbf{x}, \rho)$ should be selected for genotyping.

More formally, the optimal Haseman-Elston function $C(\mathbf{x}, \rho)$ determines the information of a sib pair with trait values x_1 and x_2. It is given by

$$I(\mathbf{x}, \rho) = \frac{1}{8} C^2(\mathbf{x}, \rho) \ , \tag{3.4}$$

and was obtained by Sham and Purcell [2001].

This information number is exact (at $\gamma = 0$), in contrast to the approximations used in the conditional distribution of IBD-sharing above. Figure 3.2 shows the distribution of information in a hypothetical population of standardised bivariate normal trait values with $\rho = 0.5$. Pairs are classified according to whether their information content is ranked in the top 5%, between 5% and 10% or in the remainder (i.e., not belonging to the 10% most informative). It clearly shows that both the extreme discordant and the extreme concordant pairs are most informative. The majority of the most informative pairs is discordant; in the top 5%, only about 15% is concordant, in the 5% to 10% category, about 35% is concordant.

For sib pairs chosen such that their trait values lie within a sampling region R, the average information can be computed by integrating over that region, weighted by the probability of the trait values:

$$I(R, \rho) = \int_R I(\mathbf{x}, \rho) \varphi_0(\mathbf{x}, \rho) d\mathbf{x} / \int_R \varphi_0(\mathbf{x}, \rho) d\mathbf{x} \ . \tag{3.5}$$

Here $\varphi_0(\mathbf{x}, \rho)$ denotes the bivariate normal density with mean 0, variance 1 and covariance ρ. Random sampling is a special case of this formula, since it is straightforward to show that when R is the full two-dimensional space, $I(R, \rho) = \frac{1}{8} \frac{1+\rho^2}{(1-\rho^2)^2}$. In order to select e.g. the 5% most informative sib pairs, R is the region of (x_1, x_2)-pairs with $C(x_1, x_2, \rho) \geq C_0$, where C_0 is chosen in such a way that this probability equals 5% under the null hypothesis.

Sampling over a region of sib pair trait values R, the number of sib pairs required to achieve power $1 - \beta$ to detect linkage with a QTL effect size γ, using a significance level α, then equals

$$n = \left(\frac{z_\alpha + z_\beta}{\gamma} \right)^2 / I(R, \rho) \ . \tag{3.6}$$

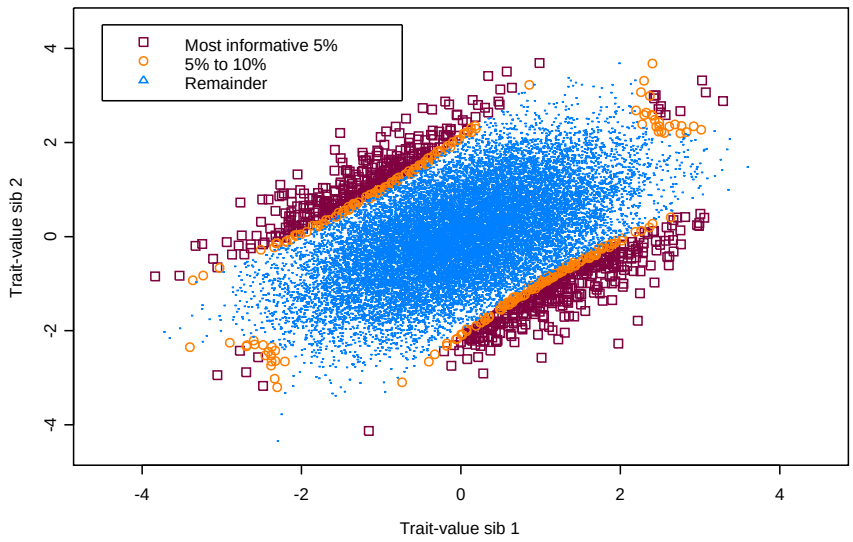

Figure 3.2: Scatterplot of trait values. Pairs are classified according to whether their information content is ranked in the top 5%, between 5% and 10% or in the remainder (not belonging to the 10% most informative).

	Height ($\rho = 0.5$) $h^2 = 0.80, c^2 = 0.10$					Insulin levels ($\rho = 0.35$) $h^2 = 0.40, c^2 = 0.15$				
QTL variance		Selection %					Selection %			
proportion (γ)	Random	10	5	2.5	1	Random	10	5	2.5	1
0.01	748180	105903	66537	43899	27648	1141429	165448	105502	71831	45494
0.02	187045	26476	16634	10975	6912	285357	41362	26375	17958	11373
0.05	29927	4236	2661	1756	1106	45657	6618	4220	2873	1820
0.10	7482	1059	665	439	276	11414	1654	1055	718	455

Table 3.1: The number of sib pairs needed to achieve 80% power to detect linkage to a quantitative trait with a significance level $\alpha = 0.0001$, for different values of γ (proportion of the variance explained by the quantitative trait locus). Height and insulin levels, two traits studied in the GenomEUtwin project are considered.

Table 3.1 shows the impact of these results on the number of sib pairs required for height and insulin levels, two quantitative traits studied in the GenomEUtwin project. For instance, for height, with a QTL variance proportion $\gamma = 0.10$, with a selection percentage of 1%, only 276 sib pairs need to be genotyped, but the trait values of 27,600 sib pairs need to be available, more than 3.5 times the amount needed for random selection. This is one reason not to go for a too restrictive selection percentage. Another, more compelling reason, is that with extreme selection percentages, the normality of the population trait values will become a crucial issue.

3.3 Selection strategies for dichotomous traits

For dichotomous traits it is convenient to think of the disease as being determined by an underlying latent quantitative trait (liability). When the value of this quantitative trait exceeds a threshold t, the individual is affected, otherwise unaffected. The threshold t is determined by the prevalence of disease K in the population of interest, through $t = \Phi^{-1}(1 - K)$, where Φ is the the distribution function of a standard normal variable. In a heritability study using twins, the heritability is estimated from the affection states of the the twins using the tetrachoric correlation of an underlying bivariate normal variable with zero mean and unit variance. The normal liability model is primarily a statistical convenience; if in reality there is no underlying normal liability in risk for an ordinal or dichotomous trait, then the model will be wrong.

The tools of Section 3.2 can be used to determine the information contributed by a twin with two affected (AA), one affected, one unaffected (AU), and two un-

	Trait I			Trait II		
latent QTL variance	$K = 5\%, \rho = 0.5$			$K = 20\%, \rho = 0.5$		
proportion (γ)	AA	AU	UU	AA	AU	UU
0.01	270122	***	***	962936	***	***
0.02	67531	649982	***	240734	403089	***
0.05	10805	103997	***	38517	64494	277326
0.10	2701	25999	***	9629	16124	69331

Table 3.2: The number of sib pairs needed to achieve 80% power to detect linkage to a dichotomous trait with a significance level $\alpha = 0.0001$, for different values of γ (proportion of the variance explained by the latent quantitative trait locus). The prevalence K and heritability approximately match that of migraine in men and women respectively. AA, AU and UU denote sib pairs with two affected, one affected and one unaffected, and two unaffected sibs respectively. *** denotes more than one million sib pairs needed.

affected (UU), given prevalence K, and tetrachoric correlation ρ (determined by the heritability). This information is

$$\frac{1}{8}\left\{\int_R C(\mathbf{x},\rho)\varphi_0(\mathbf{x},\rho)d\mathbf{x} \Big/ \int_R \varphi_0(\mathbf{x},\rho)d\mathbf{x}\right\}^2 , \tag{3.7}$$

where R is the region of (x_1, x_2)-pairs with $x_1 \geq t, x_2 \geq t$ (AA), $x_1 \geq t, x_2 < t$ (AU) or $x_1 < t, x_2 < t$ (UU). From equation (3.3) it can be seen that the expected value of $\hat{\pi}$, conditionally given that $\mathbf{x} \in R$ equals $\frac{1}{2} + \frac{\gamma}{8}\mathbf{E}(C(\mathbf{x},\rho) \,|\, \mathbf{x} \in R)$; the expression in brackets in the above expression is precisely this conditional expectation of $C(\mathbf{x},\rho)$ given $\mathbf{x} \in R$. Power calculations for dichotomous traits are very similar to (but not entirely the same as) quantitative traits using the liability threshold approach; the sampling region is now determined by affection status rather than observed trait values and does not have optimal form as in Figure 3.2. Table 3.2 shows that for dichotomous traits with low prevalence, AA sib pairs are most powerful, for traits with moderate to high prevalence, AU sib pairs however may also be quite informative.

3.4 Discussion

In this paper we have shown a simple approach to obtain explicit expressions for the information that a twin is expected to contribute towards detecting linkage to a quantitative trait. This information is based on the trait values and known values for the variance components of the trait. To achieve a given power to detect linkage to a quantitative trait with a given significance level and an anticipated proportion of the variance explained by the quantitative trait locus, the required number of sib pairs is straightforward to calculate. The expression extends to dichotomous traits through the concept of a liability, a latent underlying quantitative trait.

Earlier work uses simulation to calculate the information content of a sib pair and the number of sib pairs needed to achieve a given power [Dolan and Boomsma, 1998b; Cherny et al., 1999; Purcell et al., 2001]. For sib pairs, simulation can be replaced by calculation, as outlined below. These calculations are well known for random samples [Williams and Blangero, 1999; Rijsdijk and Sham, 2000; Rijsdijk et al., 2001] and have been pioneered for selected samples for the case of sib pairs [Sham and Purcell, 2001] and more implicitly for general pedigrees in Sham et al. [2002]. They have been implemented in `MERLIN` [Abecasis et al., 2002] through the command `MERLIN-regress`. The way they have been derived, by considering the conditional distribution of the IBD-sharing, given the phenotypes [Sham et al., 2000, 2002], also suggests methods for analysing selected samples. This is the subject of ongoing research in our group.

All expressions in Sections 3.2 and 3.3 are valid for DZ twins (sib pairs) only. It is well known however that for random samples sibships of larger sizes can achieve considerably more power than sib pairs [Dolan et al., 1999]. In a sense, a larger sibship constitutes a collection of sib pairs, and indeed the amount of information is roughly proportional to the number of sib pairs [Dolan et al., 1999; Williams and Blangero, 1999] in the sibship. Also for selective sampling, sib pairs could still be collected, even though they belong to a larger sibship. The direction taken in Section 3.2 does not readily extend to larger sibships or general pedigrees. However, the resulting expressions can be generalised more formally using efficient score functions. This approach is followed in Lebrec et al. [2004].

The score approach will also yield information content numbers for general pedi-

grees. These information content numbers can be computed in principle, but in practice the size of the pedigree may limit the calculations. Including parental information may result in a modest increase in power [Williams and Blangero, 1999]; arguably more important is the use of parental genotypes in other stages; it will increase precision of IBD-information, it can be used in quality control, and it may increase power in association studies.

The presence of dominance variance in the variance components model adds a parameter δ specifying the proportion of variance due to dominance variance of the QTL. The standardised traits of a sib pair sharing i alleles IBD will have covariance matrix

$$\mathbf{\Sigma}_i = \begin{pmatrix} 1 & \rho + \frac{i-1}{2}\gamma + (\mathbf{1}_{\{i=2\}} - \frac{1}{4})\delta \\ \rho + \frac{i-1}{2}\gamma + (\mathbf{1}_{\{i=2\}} - \frac{1}{4})\delta & 1 \end{pmatrix} .$$

For complex diseases, both γ and δ will be small, and similar calculations as in Sections 3.2 and 3.3 can be made in this case as well. The number of sib pairs needed to achieve a given power to detect linkage to a quantitative trait with a given significance level α now depends on both γ and δ through the functions $C(\mathbf{x}, \rho)$. In the case of a rare recessive allele, selection based on $C(\mathbf{x}, \rho)$ may no longer be fully informative Purcell et al. [2001]. Otherwise, dominance variance will not have a strong influence on selection, but it can influence the power.

The approach to power calculations that we took in this paper (calculating the Fisher information in an inverted variance components model, where the distribution of IBD sharing given the trait values is considered) is intimately tied to the method of analysis to be used later. As mentioned earlier, this is the subject of ongoing research in our group, but restricting the discussion to sib pairs, we note the following. It is assumed that trait values are normally distributed and have been standardised to have zero mean and unit variance. This standardisation entails subtracting the mean and dividing by the standard deviation, in the absence of covariates. Covariates can also be incorporated into both the power calculations and the analysis. Then in the standardisation the covariate values and the estimated regression coefficients (in the population!) are used instead of a common mean. Covariates can also be incorporated into the analysis of dichotomous traits; in this case not all affected sib pairs for instance will have the same C_{AA} value, but this value will now depend on the

covariate values of the sib pair. When data are not initially normally distributed, a transformation can be used in the population data to obtain approximate normality. Even in populations where the trait values are reasonably normally distributed, we think it is wise to robustify the analysis anyway, by giving sib pairs with extremely high $C(\mathbf{x}, \rho)$ values a lower weight in the inverse regression.

Chapter 4

Genomic Control for Genotyping Error in Linkage Mapping for Complex Traits

Abstract

It has been suggested that genotyping error could dramatically affect the evidence for linkage, particularly in selective designs. Using the regression-based approach to linkage, we quantify the effect of simple genotyping error models under specific selection schemes for sib pairs. We show for example, that in extremely concordant designs, genotyping error leads to over-pessimistic inference whereas it leads to increased type I error in extremely discordant designs. Perhaps surprisingly, the effect of genotyping error on inference is most severe in designs where selection is least extreme. We suggest a modification of the linkage testing procedure that accounts for genotyping errors based on a genomic estimate of the error rate.

This chapter has been submitted as: J. Lebrec, H. Putter, J.J. Houwing-Duistermaat and H.C. van Houwelingen. Genomic Control for Genotyping Error in Linkage Mapping for Complex Traits.

4.1 Introduction

In the search for genetic determinants of complex traits, the use of selective designs appears to be the only way to gain sufficient power to detect typically small gene effects in linkage studies. A few authors have shown by simulation that the impact of genotyping error on evidence for linkage could be particularly severe in affected sib-pair (ASP) designs [Douglas et al., 2000; Abecasis et al., 2001], virtually masking most of the evidence for linkage. The impact of error on quantitative traits appears to be less dramatic in random samples, however it is unclear whether the same dramatic power losses hold in selected samples.

A method of choice is now emerging for the analysis of quantitative traits arising from selected sib pairs. It boils down to a regression through the origin of excess identical by descent (IBD) sharing on a function of the trait value, whose slope is an estimate of the linkage parameter. It was first proposed by Sham and Purcell [2001] and turns out to be equivalent to a score test [Tang and Siegmund, 2001]. By use of simple genotyping error models (*population frequency error model* and *false homozygosity model*), we show analytically what effects such error generating processes (occurring at rate ϵ per sib pair) induce for an idealized fully informative marker. It is shown that it results in a reduction of the slope estimate (i.e. of the estimated linkage parameter) by a factor $1 - \frac{\epsilon}{2}$ regardless of whether sib pairs are selected or not. Since the genotyping error rate ϵ is typically small, the previous effect on the linkage test is minimal. In addition to this slope effect, the regression's intercept is modified and this may have a much more consequent effect on the test for linkage depending on the sampling scheme used to select sib pairs. Surprisingly, this simple result allows us to predict that in extremely concordant (EC) sib pairs designs and in ASP designs, the effect of genotyping error will be milder as the selection becomes more extreme. In extreme discordant (ED) designs, the effect can in theory be either over-optimistic or pessimistic depending on the definition of discordance, the genotyping error rate and the true linkage effect; in practice however, for small QTL effect, the result will be over-optimistic inference. It is argued that the basic error generating mechanisms assumed provide reasonable approximations of real-life situations. Furthermore, results obtained under the assumption of complete IBD information can be qualitatively

extended to settings where information is incomplete.

Finally, we suggest a simple genomic control for genotyping error which can easily be incorporated into the usual linkage testing procedure. This article is organized as follows: in Section 4.2, we introduce some notations and briefly sketch the inverse regression approach to linkage, in Section 4.3, we describe some common error-generating processes, in Section 4.4, we show analytically what the effect of genotyping error can be on the IBD sharing distribution and its consequence for linkage testing. Section 4.4 is devoted to studying the impact of genotyping error in common selective designs. In Section 4.5, we argue that under certain assumptions regarding the error model, one can easily implement a linkage test that incorporates a genomic control for genotyping error.

4.2 Test for linkage in selected sib pairs

We assume that the sib pair phenotypic data $\mathbf{x} = (x_1, x_2)'$ have been adjusted for any relevant covariates (e.g. sex, age, country, ...) and have been standardized so that the (known) population mean, variance and sib-sib correlation are 0, 1 and ρ respectively. In addition, let's denote by π the proportion of alleles shared identical by descent (IBD) at a certain locus by the two sibs and by $\hat{\pi}$ its estimated value given the marker information available [Kruglyak et al., 1996; Abecasis et al., 2002]. The additive variance components model assumes that $\mathbf{x}$ given IBD information π follows a normal distribution with zero mean and variance-covariance matrix given by

$$\begin{pmatrix} 1 & \gamma(\pi - \frac{1}{2}) + \rho \\ \gamma(\pi - \frac{1}{2}) + \rho & 1 \end{pmatrix} ,$$

where γ denotes the proportion of total variance explained by the putative locus. Sham and Purcell [2001] first proposed the following approach for testing linkage: regression of the estimated excess IBD sharing $\hat{\pi} - \frac{1}{2}$ through the origin of a function of the squared difference and squared sum of sib-pair phenotype values C where

$$C(x_1, x_2, \rho) = \frac{(1+\rho^2)x_1x_2 - \rho(x_1^2 + x_2^2) + \rho(1-\rho^2)}{(1-\rho^2)^2} . \tag{4.1}$$

In a sample of n independent sib pairs with phenotypes $(x_{i1}, x_{i2})_{i=1,\ldots,n}$, the test is based upon the following z statistic

$$z = \frac{\sum_i (\hat{\pi}_i - \frac{1}{2})\ C(x_{i1}, x_{i2}, \rho)}{\sqrt{\sum_i \mathrm{var}_0(\hat{\pi}_i)\ C^2(x_{i1}, x_{i2}, \rho)}}\ ,$$

it is one-sided, only positive values of z being regarded as evidence for linkage. In other words, z_+^2 defined as being equal to 0 if $z \leq 0$ and to z^2 if $z > 0$ is asymptotically distributed as $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$. For normal data, this is nothing but a score test [Tang and Siegmund, 2001] and therefore constitutes an asymptotically optimal test for linkage with small locus effect γ (see Lebrec et al. [2004] for a generalization of this score test in arbitrary pedigrees). This test is sometimes referred to as the optimal Haseman-Elston regression. In a numerical comparison of methods for selected samples, Skatkiewicz et al. [2003] and Cuenco et al. [2003] showed that this method had good properties in finite samples for extreme proband ascertained sib-pair and discordant sib-pair designs. One important feature of this regression when applied to selected samples (as far as power is concerned) is that it is constrained through the origin and this plays an important role in how genotyping error affects linkage. A different motivation for this regression through the origin was given in Putter et al. [2003] using a first order Taylor's approximation for the three IBD probabilities $\mathbf{P}(\boldsymbol{\pi} \,|\, \mathbf{x}, \gamma, \rho)$:

(4.2)

$$\begin{array}{rcl} \mathbf{P}(\boldsymbol{\pi} \,|\, \mathbf{x}, \gamma, \rho) & = & \left(\quad \mathbf{P}(\pi = 0 \,|\, \mathbf{x}, \gamma, \rho) \quad , \quad \mathbf{P}(\pi = \frac{1}{2} \,|\, \mathbf{x}, \gamma, \rho) \quad , \quad \mathbf{P}(\pi = 1 \,|\, \mathbf{x}, \gamma, \rho) \quad \right) \\ & \simeq & \left(\quad \frac{1}{4} - \frac{\gamma}{8} C(\mathbf{x}, \rho) \quad , \quad \frac{1}{2} \quad , \quad \frac{1}{4} + \frac{\gamma}{8} C(\mathbf{x}, \rho) \quad \right) \end{array} ,$$

with $C(\mathbf{x}, \rho)$ given by Formula (4.1) which implies $\mathbf{E}(\pi - \frac{1}{2} \,|\, \mathbf{x}, \gamma, \rho) = \frac{\gamma}{8}\ C(\mathbf{x}, \rho)$ when IBD information is known with certainty. This approximation is valid for small quantitative trait locus (QTL) effect γ and will be used in Section 4.4.

4.3 Genotyping error models

We consider two mechanisms for the generation of errors in marker data, namely the *population frequency error model* and the *false homozygosity model* . In those two models, we consider a single marker with m alleles and further assume that a maximum of one allelic error per sib pair can be made and that this happens

with probability ϵ. This restriction to one error per sib pair is just a first order approximation, for small ϵ, of a process where all four alleles would be allowed to be independently erroneous and does not restrict the generalizability of our results.

The *population frequency error model* re-assigns the erroneous allele (chosen at random among the four forming the sib-pair genotype) to one of the possible m alleles with probability equal to population allele frequency. One mathematical advantage of this model is that the marginal distribution of alleles and genotypes is unaltered. The *false homozygosity model* keeps homozygotes unchanged but re-assigns heterozygotes to homozygotes with alleles equal to one of the two original alleles chosen according to probabilities proportional to population allele frequencies.

To our knowledge, *false homozygosity* is a common type of error: fairly rare alleles go un-reported in samples. The *population frequency error model* provides an approximation to a process whereby alleles are misread. Errors at the two alleles of a marker's genotype might be correlated, we do not consider this type of process in details here although the effect on linkage will be qualitatively the same as in the two other models. We refer the reader to Sobel et al. [2002] for a detailed exposé on genotyping error mechanisms. Note that the two models we have chosen have been used successfully in the past in order to identify potential genotyping errors [Douglas et al., 2000; Sobel et al., 2002].

4.4 Impact of genotyping error on linkage

Effect on IBD sharing

Tests for linkage are based on the IBD sharing distribution and although errors as described in Section 4.3 are made at the genotype level (G is read as G^ϵ), the effect of errors on linkage will be entirely mediated via the distortion of the IBD distribution (the true IBD status π of two siblings may be incorrectly inferred as π^ϵ). We are therefore interested in deriving the probability distribution $\mathbf{P}(\pi^\epsilon \,|\, \pi)$, this is done by conditioning on both the true and observed genotypes as follows:

$$\mathbf{P}(\pi^\epsilon \,|\, \pi) = \sum_{G^\epsilon} \mathbf{P}(\pi^\epsilon \,|\, G^\epsilon) \sum_{G} \mathbf{P}(G^\epsilon \,|\, G)\ \mathbf{P}(G \,|\, \pi)\ .$$

Let us consider the case of complete information. This can be conceptualized

by means of an idealized marker whose number of alleles is infinite, in particular identity by state (IBS) status is equivalent to identity by descent (IBD) status. The unordered genotypes of a sib pair can be partitioned into seven exclusive classes denoted ii/ii, ii/ij, ii/jj, ii/jk, ij/ij, ij/ik and ij/kl depending on the number of homozygous sibs in the pair and the number of distinct alleles in the sib-pair genotype. Sharing 0 alleles IBD corresponds to a sib-pair genotype of the ij/kl class, should an error occur according to the *population frequency error model* then one of the four alleles would be transformed into yet another type (since the number of alleles is infinite, the probability that the new allele is read as one of i, j, k or l tends to 0), therefore the sib pair genotype will remain in the ij/kl class and the observed IBD status π^ϵ will still be 0. For the same starting genotype, an error according to the *false homozygosity model* produces an ii/jk class and π^ϵ also equals 0 therefore $\mathbf{P}(\pi^\epsilon = 0 \,|\, \pi = 0) = 1$ whatever the genotyping error mechanism considered in Section 4.3. The same line of reasoning leads to $\mathbf{P}(\pi^\epsilon = 0.5 \,|\, \pi = 0.5) = 1 - \frac{\epsilon}{2}$, $\mathbf{P}(\pi^\epsilon = 0 \,|\, \pi = 0.5) = \frac{\epsilon}{2}$, $\mathbf{P}(\pi^\epsilon = 1.0 \,|\, \pi = 1.0) = 1 - \epsilon$, $\mathbf{P}(\pi^\epsilon = 0.5 \,|\, \pi = 1.0) = \epsilon$. Those results can be summarized by the transition matrix below, where the (i, j) element is equal to $\mathbf{P}(\pi^\epsilon = (j-1)/2 \,|\, \pi = (i-1)/2)$

$$\mathbf{P}(\boldsymbol{\pi}^\epsilon \,|\, \boldsymbol{\pi}) = \begin{pmatrix} 1 & 0 & 0 \\ \frac{\epsilon}{2} & 1 - \frac{\epsilon}{2} & 0 \\ 0 & \epsilon & 1 - \epsilon \end{pmatrix} .$$

The overall effect of genotyping error is thus to reduce the observed IBD sharing. In selected samples of extremely concordant sib pairs (EC) where linkage is evidenced by excess IBD sharing, it therefore seems logical to expect a decrease in power. Conversely, in selected samples of extremely discordant sib pairs (ED) where linkage is evidenced by reduction in IBD sharing, the test might lead to increased type I error. In Section 4.4, we quantify this bias in selective samples schemes for quantitative traits under the usual assumption of a normal variance components model.

Effect on linkage

In this section, we concentrate on the case where IBD information is complete. As exposed in Section 4.2, the test for linkage corresponds to a regression through the

origin of excess IBD sharing $\hat{\pi} - \frac{1}{2}$ on a function of phenotype values $\mathbf{C} = C(\mathbf{x}, \rho)$ with C as defined by Formula (4.1) i.e. it is based on the approximate relation

$$\mathbf{E}(\pi - \frac{1}{2} \,|\, \mathbf{x}, \gamma, \epsilon) = \frac{\gamma}{8}\, C(\mathbf{x}, \rho)\ . \tag{4.3}$$

We show in the appendix that, in presence of genotyping error at rate ϵ, this relation is changed into

$$\mathbf{E}(\pi^{\epsilon} - \frac{1}{2} \,|\, \mathbf{x}, \gamma, \epsilon) = -\frac{\epsilon}{4} + (1 - \frac{\epsilon}{2})\ \frac{\gamma}{8}\ C(\mathbf{x}, \rho)\ . \tag{4.4}$$

If we were to know ϵ, we could correct for it in the regression and the loss in efficiency would only be due to the $1 - \frac{\epsilon}{2}$ term preceding γ and would therefore be minimal.

We may ignore genotyping error altogether. In the appendix, we derive a general expression (Equation (4.9)) for the probability of rejecting the null hypothesis of no linkage under this scenario. For small values of the error rate ϵ, the following first order approximation obtains

$$\Phi\left(\Phi^{-1}(\alpha) + \gamma \mathcal{I}^{1/2}\right) - \epsilon\ \mathcal{I}^{1/2} \left(\frac{\gamma}{2} + 2\frac{\overline{C}}{\overline{C^2}}\right) \times \phi\left(\Phi^{-1}(\alpha) + \gamma \mathcal{I}^{1/2}\right)\ , \tag{4.5}$$

where α is the nominal type I error rate for the linkage test with a true quantitative trait locus effect γ, $\overline{C}$ is the average of the $C(x_{i1}, x_{i2}, \rho)$ values (given by Equation (4.1)) among a sample of n sib pairs, $\mathcal{I} = \frac{n}{8}\ \overline{C^2}$ is the sample's Fisher's information for the linkage parameter γ, Φ is the cumulative density function of the standard normal distribution and ϕ is the corresponding density function. The first term $\Phi\left(\Phi^{-1}(\alpha) + \gamma \mathcal{I}^{1/2}\right)$ in this expression gives the value of this probability in absence of genotyping error while the second term is the deviation from this reference value; in particular, when $\gamma = 0$, it expresses the actual type I error as a deviation from the nominal type I error rate: $\alpha - 2\epsilon\ \frac{\overline{C}}{\overline{C^2}} \mathcal{I}^{1/2} \times \phi\left(\Phi^{-1}(\alpha)\right)$.

In extremely concordant (EC) designs, $\overline{C}$ is positive while it is negative in extremely discordant (ED) designs, inference will therefore be too conservative in EC designs and too liberal in ED designs. In random samples and under the variance components model, C is a score function hence $\mathbf{E}(C) = 0$ therefore its sample estimate $\overline{C}$ will be small and the effect of genotyping error will be minimal. The same finding would hold for any ascertainment scheme where $\overline{C} = 0$.

We now quantify the effect of genotyping error on power and type I error under specific designs. The distortion of the linkage test in presence of genotyping error

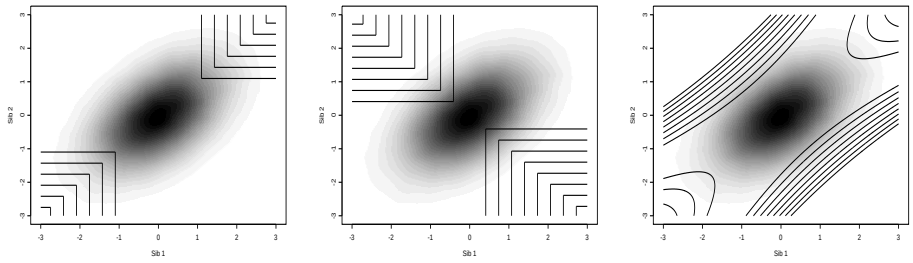

Figure 4.1: Three selective schemes: extremely concordant(EC), extremely discordant(ED) and most informative ($\mathcal{I}$) all for 10%. Joint distribution of sib trait values in gray scale for $\rho = 0.5$ (generated using the scatterplots function of Eilers and Goeman [2004])

depends heavily on the design-specific quantity $\overline{C}/\overline{C^2}$; given an ascertainment scheme corresponding to a certain region of the possible trait values, it is simple to use Monte Carlo methods to determine the expected $\overline{C}/\overline{C^2}$ value in that region. In table 4.1, we considered three different ascertainment schemes: extremely concordant (EC), extremely discordant (ED) and most informative ($\mathcal{I}$) as shown in Figure 4.1. For example, in the $\text{EC}_{10\%}$ scheme with sib-sib trait correlation $\rho = 0.5$, only sib pairs whose trait values (x_1, x_2) fulfill $x_1 > t$ and $x_2 > t$ or $x_1 \leq -t$ and $x_2 \leq -t$ where $t = t_{EC}(10\%, \rho = 0.5) = 0.136$ are retained (the value of t is such that on average 10% of the overall population is sampled). Analogously for ED, sib pairs whose trait values belong to regions defined by $x_1 > t$ and $x_2 \leq -t$ or $x_1 \leq -t$ and $x_2 > t$ are selected. The $\mathcal{I}$ scheme selects the most informative sib pairs determined using the quantiles of Fisher's information ($\mathcal{I} \propto C^2(x_1, x_2, \rho)$) distribution for the linkage parameter γ [Lebrec et al., 2004]. For example, if the percentage selected equals 10% and $\rho = 0.5$ then sib pairs whose trait values fulfill $C^2(x_1, x_2, \rho = 0.5) > 4.36$ would be selected. This sampling scheme combines both EC and ED sib pairs and constitutes a refinement of the so-called EDAC designs [Gu et al., 1996].

Table 4.1 allows us to draw three main conclusions relating to the main bias caused by the intercept mis-specification in the usual linkage testing procedure:

1. It is negative in EC designs and positive in ED designs, positive but without substantial influence for $\mathcal{I}$ designs,

ρ	sel.	EC	ED	$\mathcal{I}$	sel.	EC	ED	$\mathcal{I}$	sel.	EC	ED	$\mathcal{I}$
0.1	1%	0.27	-0.23	-0.07	10%	0.47	-0.40	-0.06	30%	0.65	-0.53	-0.04
0.2		0.29	-0.21	-0.13		0.50	-0.36	-0.11		0.69	-0.46	-0.07
0.3		0.30	-0.19	-0.15		0.52	-0.32	-0.14		0.71	-0.39	-0.09
0.4		0.31	-0.17	-0.14		0.53	-0.28	-0.16		0.69	-0.32	-0.11
0.5		0.32	-0.14	-0.12		0.52	-0.24	-0.17		0.62	-0.25	-0.11
0.6		0.31	-0.12	-0.10		0.47	-0.19	-0.15		0.50	-0.19	-0.10

Table 4.1: Average values for the $\overline{C}/\overline{C^2}$ term determining bias

2. It is more pronounced as the designs becomes less extreme for both EC and ED,

3. It is fairly independent of sib-sib trait correlation ρ for EC designs while it decreases with ρ for ED designs.

Overall, for small QTL effects γ, genotyping error will lead to conservative inference in EC designs and to liberal inference in ED designs. In Figure 4.2, we show the theoretical type I error rate and probability of rejecting the null hypothesis (obtained via Formula (4.9)) for different sampling schemes under perfect IBD information. We have used a QTL explaining 10% of the total trait variance, a trait sib-sib correlation equal to 0.3 and error rates equal to 0.01, 0.02 and 0.05. Although the power is not too badly affected at least for small error rates, genotyping error substantially affects the type I error rate, this may lead to far too liberal inference in ED designs, this deterioration of the size of the test becomes more acute as sample size increases.

Incomplete IBD information

We saw in Section 4.4 that genotyping error not only deteriorated the slope of the linkage signal but also introduced an intercept in the regression of excess IBD sharing on the optimal Haseman-Elston trait function $C(\mathbf{x}, \rho)$. In the case of complete information and at least for the *population frequency error model* and *false homozygosity model* , the perturbation caused by the error processes only depended on the error rate ϵ through the functions given in Equation (4.3). In real-life situations, IBD information is incomplete, but under the usual variance components additive model and

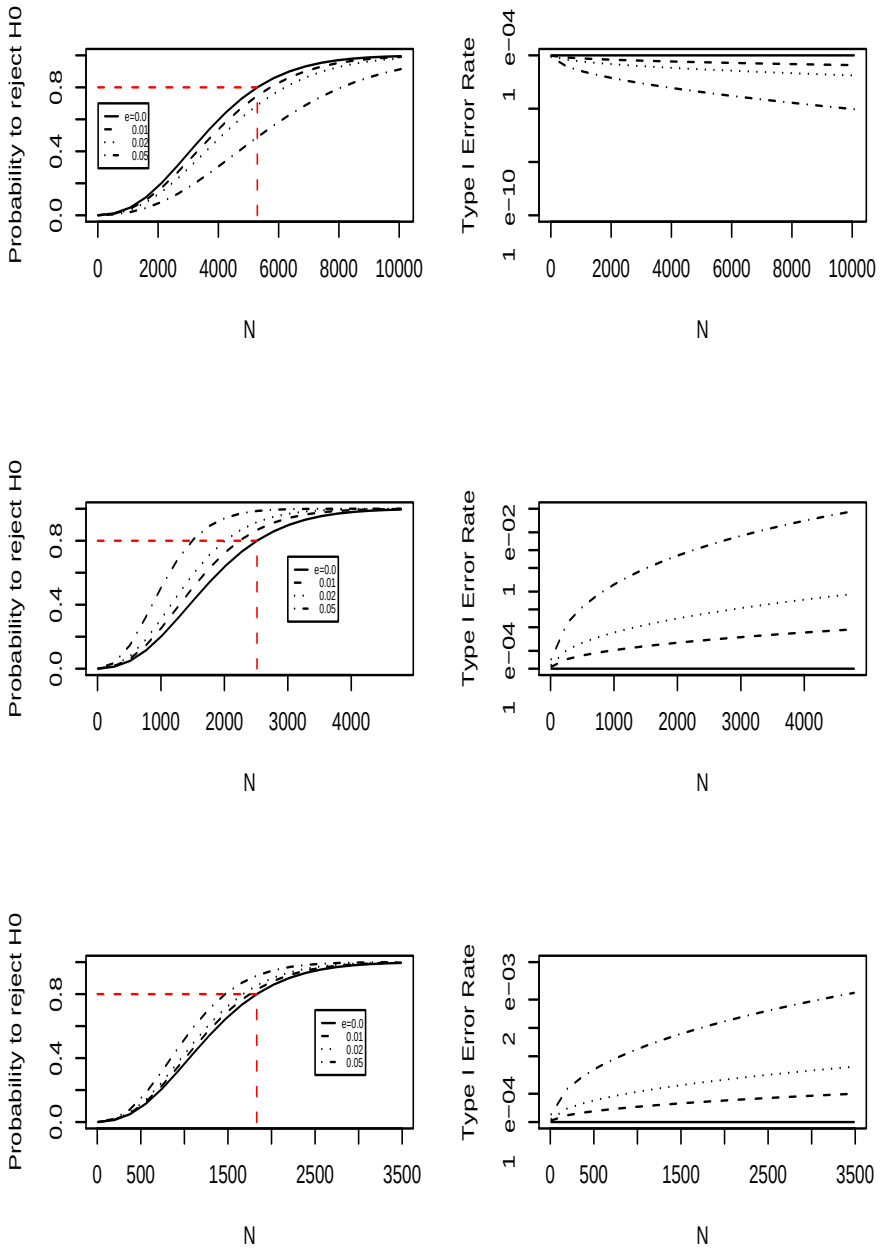

Figure 4.2: Effect of genotyping error on test for linkage in EC (top), ED (middle) and $\mathcal{I}$ (bottom) designs

in absence of genotyping errors, the excess IBD sharing is approximately related to the QTL effect γ and the optimal Haseman-Elston trait function $C(\mathbf{x}, \rho)$ through the regression (this is shown for an approximate additive model as given by Formula (4.2) in the appendix of Lebrec et al. [2006])

$$\mathbf{E}(\hat{\pi} - \frac{1}{2} \,|\, \mathbf{x}, \gamma, \epsilon) \simeq \mathrm{var}_0(\hat{\pi})\gamma \; C(\mathbf{x}, \rho) \;,$$

and the effect of genotyping error is to modify this regression into

$$\mathbf{E}(\hat{\pi}^{\epsilon} - \frac{1}{2} \,|\, \mathbf{x}, \gamma, \epsilon) \simeq a(\epsilon) + b(\epsilon) \; \mathrm{var}_0(\hat{\pi})\gamma \; C(\mathbf{x}, \rho) \;. \tag{4.6}$$

For simple cases, e.g. a single equi-frequent allele marker, explicit formulae can be derived for a and b; in general though, those functions will depend in a complex manner on the genotyping error mechanism but also on the markers' map and no explicit forms will be available. When multi-point marker data are used to infer IBD sharing, errors tend to propagate around markers and one can expect a more severe effect of genotyping error compared to single-point algorithms. As mentioned earlier, for small QTL effects, most of the impact on linkage in selected samples will be due to the intercept mis-specification in the linkage regression, we therefore focus on this issue.

In random samples or under the null hypothesis of no linkage, the sample mean excess IBD $\overline{\hat{\pi}^{\epsilon}} - \frac{1}{2}$ (averaged across families) provides an estimate of the intercept $a(\epsilon)$. We simulated three different marker map configurations in 10000 sib pairs without parents and quantified by how much IBD sharing was reduced on average under the *population frequency error model* and the *false homozygosity model* (error rates=0.01 and 0.05). MapH and MapL had eleven equi-frequent allele markers located 10cM apart, markers had 10 alleles in MapH and 2 alleles in MapL. MapM only had six markers 20cM apart with 5,2,5,2,2 and 5 alleles on the six markers (from left to right). The results are displayed in Figure 4.3 along with the corresponding map information content as defined in Kruglyak and Lander [1995] (wiggly curves in bottom part of each figure, scale on the right y-axis), for clarity and because results were very similar, we have omitted the curves corresponding to the *false homozygosity model* . One clear trend is that IBD is most affected by genotyping error in areas where marker information is high. Furthermore, even for small error rates, the decrease in

IBD sharing is substantial.

4.5 Genomic control for genotyping error

As we have seen in previous sections, the main effect of genotyping error is to modify the intercept in the regression used to test for linkage. In order to obtain more robust inference, it therefore seems natural to try and constrain the regression through its correct origin a. In this section, we propose a completely data-driven strategy for doing this.

At any position, the sample mean IBD sharing has variance $\mathrm{var}_0(\hat{\pi})/n$ where n is the number of sib pairs available. If we knew that the position is unlinked or if the sample of sib pairs was random then the deviation of this mean from $\frac{1}{2}$ would provide an estimate of the intercept a in the linkage regression. Unfortunately, detection of a position-specific intercept corresponding to typical error rates would require a sample size of order 10^4, a number that is almost never reached in linkage studies. In order to obtain an intercept estimate $\hat{a}$ with sufficient precision, it is therefore essential to combine information across positions. The value of IBD sharing at positions outside of the neighborhood of influencing loci (those positions are subsequently referred to as unlinked) across the genome may serve as control in the test for linkage, this concept of genomic control has been used to robustify the analysis of association studies by Devlin and Roeder [1999].

Ad-hoc method

Let's assume that the proportions of alleles shared IBD $\hat{\pi}$ is inferred at a series of approximately regular positions indexed by t across the whole genome. Let y_t be the sample mean (among families) excess IBD at position t i.e. $y_t \equiv \overline{\hat{\pi}_t^{\varepsilon}} - \frac{1}{2}$. Under the variance components model and for small QTL effect γ, equation (4.6) implies that

$$E(y_t) \simeq \begin{cases} a \ , & \text{if position } t \text{ is unlinked} \ , \\ a + \frac{b}{8}\gamma\overline{C} \ , & \text{if position } t \text{ is linked} \ . \end{cases}$$

In random samples or in any sample where $\overline{C} \simeq 0$, taking the average of y_t across positions provides and estimate of a. In selected samples, we can use a trimmed version of the mean of y, for example a 20%-trimmed mean of the $(y_t)_t$ series (i.e.

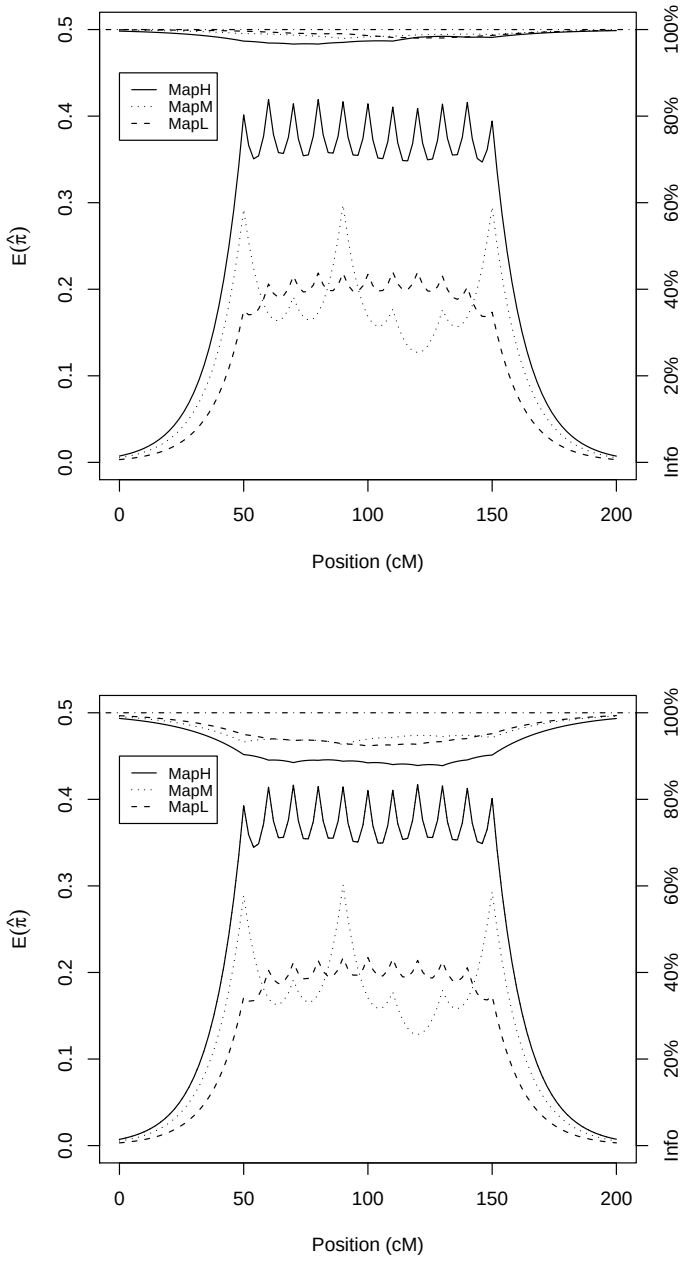

Figure 4.3: Effect of genotyping error on IBD sharing and corresponding map information content in simulated data - Error rates $\epsilon = 0.01$ (top) and $\epsilon = 0.05$ (bottom)

the mean of the y_t values after removing the 20% lowest and and 20% highest values) will provide a robust genomic estimate $\hat{a}$ of a. Because $a \leq 0$ and $\overline{C}$ is positive and negative in EC designs and ED designs respectively, $\hat{a}$ could be refined by trimming off only the 20% highest and lowest y_t values respectively before taking the mean. Of course, how much we trim is arbitrary but 20% can safely be taken as a conservative value for oligogenic traits.

An ad-hoc implementation of the concept of genomic control is then to plug in the estimate of the intercept $\hat{a}$ into the linkage regression (4.6). Since most of the bias in the inference is due to the intercept mis-specification, the precise estimate obtained by pooling across the genome will eliminate it. The implicit assumption that we make in this genomic control approach is that the regression intercept is the same at all positions.

Empirical Bayes

The method in the previous section can be formalized using an empirical Bayes inferential procedure in order to compute the posterior probability that a position is linked. Having set a minimum level of evidence for deciding whether a position is linked, the values of y_t at unlinked positions could be pooled and the estimate thus obtained plugged into the linkage regression as in the previous section. The approach is borrowed from the microarrays literature [Efron and Tibshirani, 2002] and our problem is analogous to the estimation of the proportion of true null hypotheses in false discovery rates testing rules.

We assume that the prior density f of the average excess IBD sharing $y = (y_t)_t$ is given by a mixture distribution

$$f(y) = \alpha_0 f_0(y) + (1 - \alpha_0) f_1(y) \ .$$

Here, α_0 denotes the prior probability that a position is unlinked (a conservative value would be $\alpha_0 = 1$) and $f_0(y)$ is the corresponding prior probability distribution of y, while $f_1(y)$ denotes the prior probability distribution of y at a linked position. Using Bayes' theorem, the following posterior distribution obtains

$$\mathbf{P}(\text{position } t \text{ linked} \,|\, y_t) = 1 - \frac{\alpha_0 f_0(y_t)}{f(y_t)} \ .$$

Non-parametric density estimation techniques such as kernel density estimation may be used to estimate $f(y)$ from the data without having to specify $f_1(y)$. Unless the positions where IBD is inferred are chosen far apart, the observations will not be independent but this does not invalidate the method. It suffers one inherent limitation though: the effective sample size is small in a human genome (choosing positions every 50cM produces only approximately 70 almost independent observations) and this limits our ability to estimate $f(y)$ precisely. Since $\text{var}(y_t) = (8n)^{-1}$, the prior $f_0(y)$ could be chosen as an $N(a_0, (8n)^{-1} + \tau^2)$ where a_0 would reflect our prior knowledge about the intercept a and τ^2 the associated uncertainty.

Instead of applying this empirical Bayesian framework to the average excess IBD sharing $(y_t)_t$, we can apply it directly to linkage statistics such as the QTL effect estimates $\hat{\gamma}_t = \frac{\sum_i (\pi_i^{\varepsilon} - \frac{1}{2})\mathbf{C}_i}{\frac{1}{8}\sum_i C_i^2}$ whose expectation is calculated in the Appendix. Since $\text{var}(\hat{\gamma}_t) = (\frac{1}{8}\sum_i C_i^2)^{-1}$, priors $f_0(y)$ of the form $N(a_0, (\frac{1}{8}\sum_i C_i^2)^{-1} + \tau^2)$ are possible although asymmetric versions that favor negative values might be more appropriate. Preliminary simulations give sensible results when the true number of linked positions is not too low ($\geq 5\%$) and the study is adequately powered, however the limited number of independent dimensions in a linkage scan is a serious limitation of this approach.

Alternatives

Alternatives to this genomic-control strategy are possible and they also boil down to constraining the linkage regression through a new origin as in the ad-hoc method, the estimation procedure can be adapted to suit particular circumstances.

Firstly, in random samples, the assumption regarding exchangeability of positions might be relaxed. Indeed, the y_t's may be used as estimates of the position-specific intercepts since a study sufficiently powered to detect linkage in random samples should provide sufficient precision. It must be noted though that the advantage of using a genomic control in random samples is limited because the impact of genotyping error is small in such designs. Secondly, one could use previous lab data to estimate by how much IBD sharing deviates from its expected value, this could also been done at each position separately provided sufficient data are available. In practice, such data might not be available or they might not trustfully reflect current error mechanisms.

4.6 Discussion

Under two basic error models, we were able to predict quantitatively the consequences of genotyping error on inference in linkage analysis. In the idealized situation of complete IBD information, both error models have the same impact on linkage analysis. As we have seen, the effect is due to a decrease in IBD sharing. A contrario, an error process which would increase IBD sharing would produce opposite results. The true error processes involved in practice are complicated mixtures of the models alluded to here. In our experience however, it seems that processes which lower IBD sharing are predominant. Because genotyping error tends to decrease the estimated number of alleles shared IBD, the effect on evidence for linkage is opposite in EC (over-pessimistic) and ED (over-optimistic) designs, it can be dramatic in typical designs and paradoxically less severe for more extreme ascertainment schemes. By analogy, for a dichotomous trait, this means that the effect of genotyping error is less severe in ASP designs for rare diseases than for common diseases. Remarkably, in designs combining both ED and EC pairs like the $\mathcal{I}$ (or EDAC designs), the competing effects of genotyping error tend to cancel each other out. We have considered here only three types of basic selection schemes however the approach can straightforwardly be applied to any arbitrary selection scheme, under a variance components model, the important quantity being $\overline{C}/\overline{C^2}$.

The genomic-control strategy that we have proposed offers a robust method for carrying out linkage analysis but obviously relies on a convenient approximation of a very complex situation. It is probably reasonable to assume that genotyping of markers with a similar degree of polymorphism (number of alleles and frequencies) within the same lab is subject to the same error process. On top of the true underlying error mechanism, in a multi-point setting, not only the number of markers but also the inter-marker distances could have an impact. Ideally, markers should have similar numbers of alleles and respective frequencies and be rather evenly distributed across the genome. Based on results from simulations presented in Section 4.4, it seems appropriate to pool estimates of regression's intercept a which correspond to areas of the genome where marker information is roughly the same. The advent of SNP chip therefore makes us confident of the applicability of our method, indeed this

new technology for linkage data holds the promise of providing marker maps with less variable information content than in classical microsatellites maps [Evans and Cardon, 2004; Schaid et al., 2004].

Elston et al. [2005] have recently pointed out that the implicit assumption made in ASP designs, that randomly sampled sib pairs share half of their alleles IBD, might not hold in practice and have argued for including discordant pairs in such studies. The approach presented here offers an alternative solution to this issue. Finally we note that, although we have only considered designs involving sib pairs, the approach naturally extends to other types of relative pairs.

Acknowledgements

We are grateful to Dr. Bas Heijmans from the section Molecular Epidemiology, Dept. of Medical Statistics and Bioinformatics, Leiden University Medical Center for discussions on genotyping error mechanisms.

4.7 Appendix

Effect of genotyping error on linkage

We show how regression (4.3) is modified in presence of genotyping error. We concentrate on the case where IBD information is complete.

By definition $\mathbf{E}(\pi^\epsilon - \frac{1}{2} \,|\, \mathbf{x}, \gamma, \epsilon) = \frac{1}{2}\ \mathbf{P}(\pi^\epsilon = \frac{1}{2} \,|\, \mathbf{x}, \gamma, \epsilon) + \mathbf{P}(\pi^\epsilon = 1 \,|\, \mathbf{x}, \gamma, \epsilon) - \frac{1}{2}$. We can then condition on the true IBD status π and use approximation (4.2) in order to evaluate the probabilities involved in the previous expression: $\mathbf{P}(\pi^\epsilon \,|\, \mathbf{x}, \gamma, \epsilon) = \sum_\pi \mathbf{P}(\pi^\epsilon \,|\, \pi)\ \mathbf{P}(\pi \,|\, \mathbf{x}, \gamma)\ \mathbf{P}(\pi^\epsilon \,|\, \pi)$. In the present case of complete information, this yields

$$\mathbf{E}(\pi^\epsilon - \frac{1}{2} \,|\, \mathbf{x}, \gamma, \epsilon) = -\frac{\epsilon}{4} + (1 - \frac{\epsilon}{2})\ \frac{\gamma}{8}\ C(\mathbf{x}, \rho)\ . \tag{4.7}$$

Probability to reject H_0

We derive an approximate formula for the probability of rejecting the null hypothesis of no linkage if we ignore genotyping error.

As we have seen earlier, testing for linkage boils down to regression (4.3). Let's denote by $\hat{\gamma}$, the estimate of the slope in the regression through the origin of a sam-

ple $\left(\pi_i - \frac{1}{2}\right)_{i=1,\ldots,n}$ on the corresponding $C_i = (C(x_{i1}, x_{i2}, \rho))_{i=1,\ldots,n}$ and by $\hat{\gamma}^\epsilon$, the estimate of the slope in the same regression but where the response is replaced by $\left(\pi_i^\epsilon - \frac{1}{2}\right)_{i=1,\ldots,n}$.

$$\hat{\gamma} = \frac{\sum_i (\pi_i - \frac{1}{2})\ C_i}{\frac{1}{8}\sum_i C_i^2} \quad \text{and} \quad \mathbf{E}(\hat{\gamma}\,|\,\mathbf{x}, \gamma) \simeq \gamma$$

i.e. $\hat{\gamma}$ is an approximately unbiased estimate of γ. However it appears that $\hat{\gamma}^\epsilon = \frac{\sum_i (\hat{\pi}_i^\epsilon - \frac{1}{2})\ C_i}{\frac{1}{8}\ \sum_i C_i^2}$ is biased since

$$\begin{aligned} \mathbf{E}(\hat{\gamma}^\epsilon\,|\,\mathbf{x}, \gamma, \epsilon) = &\quad \frac{\sum_i \mathbf{E}(\pi_i^\epsilon - \frac{1}{2}\,|\,\mathbf{x}, \gamma)\ C_i}{\frac{1}{8}\sum_i C_i^2} \\ \simeq &\quad (1 - \frac{\epsilon}{2})\ \gamma\ - \frac{\epsilon}{4}\ \frac{\overline{C}}{\overline{C^2}}\ . \end{aligned} \tag{4.8}$$

The bias in $\hat{\gamma}^\epsilon$ depends on two factors: the genotyping error rate ϵ and the selection procedure of sib pairs (which determines $\overline{C} = \frac{1}{n}\sum_i C_i$ and $\overline{C^2} = \frac{1}{n}\sum_i C_i^2$). Whatever the ascertainment scheme used (in particular in random samples), the estimate of γ is systematically biased downwards by a factor $1 - \frac{\epsilon}{2}$; then, depending on the sign and value of $\overline{C}/\overline{C^2}$, $\hat{\gamma}^\epsilon$ can be further decreased or increased. For complex traits and thus small QTL effects γ, the intercept mis-specification will have a greater impact than the bias in the slope. The test for linkage is based on the standardized slope estimate $\frac{\hat{\gamma}^\epsilon}{\sqrt{\mathrm{var}_0(\hat{\gamma}^\epsilon)}} = \frac{\hat{\gamma}^\epsilon}{\sqrt{\mathrm{var}_0(\pi^\epsilon)\overline{C^2}}}$, since $\mathrm{var}_0(\pi) = \frac{1}{8}$ is practically unchanged by genotyping error ($\mathrm{var}_0(\pi^\epsilon) = \frac{1}{8} - \frac{\epsilon^2}{16}$), the probability of rejecting the null hypothesis is given by

$$\Phi\left(\Phi^{-1}(\alpha) + (1 - \frac{\epsilon}{2})\gamma\mathcal{I}^{1/2} -\ 8\ \frac{\epsilon}{4}\frac{\overline{C}}{\overline{C^2}}\mathcal{I}^{1/2}\right)\ , \tag{4.9}$$

where $\mathcal{I} = \mathrm{var}_0(\hat{\gamma})^{-1} = \frac{n}{8}\ \overline{C^2}$ is the sample's Fisher's information for the linkage parameter γ, α is the nominal type I error rate for the linkage test with a true quantitative trait locus effect γ and Φ is the cumulative density function of the standard normal distribution. A first order Taylor approximation of (4.9) yields Formula (4.5).

Chapter 5

Potential Bias in Generalized Estimating Equations Linkage Methods under Incomplete Information

Abstract

The mean identity-by-descent (IBD) specification used in the Generalized Estimating Equations (GEE) methodology for linkage is only valid, strictly speaking, under the assumption of fully polymorphic markers. In practice, markers often provide only partial IBD information which can potentially result in inconsistency of the locus location and gene effect estimates obtained by the GEE method. Using both simulations and theory, we identify some realistic conditions about marker information under which the validity of the GEE linkage methods may be arguable. Namely, researchers should not trust the GEE parameters' estimates and their associated confidence intervals in areas of the genome where IBD information is sparse or when this information changes abruptly. We show that properly standardized statistics based on IBD sharing provide a valid alternative.

This chapter has been published as: J. Lebrec, H. Putter and J.C. van Houwelingen (2006). Potential Bias in Generalized Estimating Equations Linkage Methods under Incomplete Information. *Genetic Epidemiology* **30** (1), 94–100.

5.1 Introduction

Since Liang et al. [2001] introduced the use of Generalized Estimating Equations (GEE) with the purpose of estimating the position of a locus linked to a trait, there has been increasing interest in this methodology. The approach has attractive features, in particular, it allows researchers to set a confidence interval around the estimate of the locus position. In the meantime, some refinements and extensions of the approach are being developed: covariates can be introduced [Glidden et al., 2003; Chiou et al., 2005], the methodology can be extended to two linked loci in the region [Biernacka et al., 2005] and to general pedigrees [Schaid et al., 2005], and it bears potential for a wider use in the future. Strictly speaking, the GEE linkage method is only valid when markers are fully polymorphic, in other words, when identity-by-descent (IBD) status at markers is known with certainty. As far as we are aware, little has been done to assess how robust the method is under more realistic conditions of marker information. Indeed, among the aforementioned articles, those that included simulations almost always generated complete IBD data at markers. The only exception is Biernacka et al. [2005] who recognized that the use of non-fully informative marker maps produced biased estimates of the genetic effects but hardly any bias in the estimate of locus position, however they only looked at evenly distributed marker maps. In this article, we identify some realistic conditions about marker information under which the validity of the GEE linkage methods may be arguable, properly standardized statistics based on IBD sharing provide a valid alternative. In the 'Methods' section, we review the principles of the GEE method and show why it may lead to biased and inconsistent estimation and we prove that some more classical approaches do not suffer the same drawback under certain conditions. The 'Results - Monte Carlo simulations' section is devoted to simulations that illustrate the findings of the previous section in a range of realistic scenarios. Finally, in the 'Discussion' section, we discuss our findings and their possible practical impact on linkage analysis.

5.2 Methods

The GEE methodology

We start by recalling the principle of the GEE methodology as applied to linkage mapping. For affected sib pairs (ASP) the method is based on the mean specification of the excess IBD sharing at markers as

$$\mathbf{E}(\pi_t - \frac{1}{2} \mid \text{ASP}) = \frac{1}{8}(1 - 2\theta_{t,\tau})^2 \; C \; = \mu_t(\tau, C), \tag{5.1}$$

where π_t denotes the true proportion of alleles shared IBD at marker or position t, τ the position of the true and only locus in the region, $\theta_{t,\tau}$ the recombination fraction between locations t and τ, while C reflects the genetic model (note here that C in the previous equation is 4 times the C parameter used in Liang et al. [2001]). We stress that the derivation of this result assumes that markers are fully polymorphic. In practice, IBD is uncertain and is estimated using multipoint marker data, it is well known that the consequence of incomplete information is to shrink the estimated IBD towards its null value $\frac{1}{2}$, as a result the previous mean model might be erroneous. We distinguish the true (often unobserved) proportion of alleles shared IBD π from its estimated counterpart by the use of the notation $\hat{\pi}$.

We assume that we have data from $i = 1, \ldots, N$ ASPs available at marker positions $t_1, \ldots, t_M$ with corresponding IBD sharing estimates $\hat{\boldsymbol{\pi}}_i = (\hat{\pi}_{i,t_1}, \ldots, \hat{\pi}_{i,t_M})'$, where $'$ denotes the transpose of a matrix (bold letters indicate a matrix or a vector as opposed to a scalar). We denote by V the $M \times M$ working variance-covariance matrix for $\hat{\boldsymbol{\pi}}_i$ while $\boldsymbol{\mu} = \boldsymbol{\mu}(\tau, C) = (\mu_{t_1}, \ldots, \mu_{t_M})'$ then estimation of the parameters τ and C is carried out by solving the following GEE

$$\sum_{i=1}^{N} \left(\frac{\partial \boldsymbol{\mu}}{\partial(\tau, C)} \right)' \; \mathrm{V}^{-1} \; (\hat{\boldsymbol{\pi}}_i - \boldsymbol{\mu}(\tau, C)) = 0 \; .$$

The theory developed by Liang and Zeger [1986] ensures that as long as the mean of the observations is correctly specified (i.e. $\mathbf{E}(\hat{\boldsymbol{\pi}}_i) = \boldsymbol{\mu}(\tau, C)$), the GEE estimators of τ and C converge towards the true locus position and genetic effects as the sample size N increases. A specification of V as the true variance-covariance matrix of the observations $\hat{\boldsymbol{\pi}}_i$ in terms of the unknown parameter τ and C was given

in Liang et al. [2001] (again, under complete information) but is not essential to the consistency of the procedure, it only affects its efficiency. In addition, an asymptotically robust variance-covariance matrix for the estimates $(\hat{\tau}, \hat{C})'$ can be computed as $\hat{\Sigma} = \hat{\Sigma}_1^{-1}\hat{\Sigma}_2\hat{\Sigma}_1^{-1}$ with

$$\begin{aligned}\hat{\Sigma}_1 &= N\left(\frac{\partial \boldsymbol{\mu}}{\partial(\tau, C)}\right)' \mathrm{V}^{-1}\left(\frac{\partial \boldsymbol{\mu}}{\partial(\tau, C)}\right) \\ \hat{\Sigma}_2 &= \sum_{i=1}^{N}\left(\frac{\partial \boldsymbol{\mu}}{\partial(\tau, C)}\right)' \mathrm{V}^{-1}\left(\hat{\boldsymbol{\pi}}_i - \boldsymbol{\mu}(\hat{\tau}, \hat{C})\right)\left(\hat{\boldsymbol{\pi}}_i - \boldsymbol{\mu}(\hat{\tau}, \hat{C})\right)' \mathrm{V}^{-1}\left(\frac{\partial \boldsymbol{\mu}}{\partial(\tau, C)}\right) ,\end{aligned}$$

where $\frac{\partial \boldsymbol{\mu}}{\partial(\tau, C)}$ and possibly V are evaluated in $(\hat{\tau}, \hat{C})$.

An accurate IBD specification under incomplete information

The relation $\mathbf{E}(\hat{\pi}) = \mu(\tau, C)$ between the mean of the estimated IBD sharing and the locus position τ and gene effect C, exactly true when IBD is perfectly known, is only approximate under incomplete information. In fact, Teng and Siegmund [1998] have shown that a theoretical mean IBD specification can also be derived under incomplete information, namely for a one-locus (located at τ) additive model on the IBD scale (which is approximately true for a wide range of disease models; exactly true if $\lambda_S = \lambda_O$ [Risch, 1990]) such that

$$\left\{\begin{array}{lcl}\mathbf{P}(\pi_\tau = 0 \,|\, \mathrm{ASP}) &=& \frac{1}{4} - \frac{1}{8}C \\ \mathbf{P}(\pi_\tau = \frac{1}{2} \,|\, \mathrm{ASP}) &=& \frac{1}{2} \\ \mathbf{P}(\pi_\tau = 1 \,|\, \mathrm{ASP}) &=& \frac{1}{4} + \frac{1}{8}C ,\end{array}\right. \tag{5.2}$$

the expected observed excess IBD sharing at any arbitrary position t is given by

$$\mathbf{E}(\hat{\pi}_t - \frac{1}{2} \,|\, \mathrm{ASP}) = \mathrm{cov}_0(\hat{\pi}_t, \hat{\pi}_\tau)\; C \;, \tag{5.3}$$

where the covariance $\mathrm{cov}_0(\hat{\pi}_t, \hat{\pi}_\tau)$ is taken under the null hypothesis (It therefore only depends on marker map characteristics, pedigree structure and possibly missing genotype patterns). For the sake of completeness, we show a proof of this crucial result in the appendix. The correct specification of the mean IBD sharing as a function of the locus position τ and genetic effect C is essential in order to obtain valid estimates by the GEE method. Comparison of Equations (5.3) and (5.1) allows one to evaluate the discrepancy between the correct IBD specification and the one used

in the GEE linkage methods. For illustration purposes, we have displayed two typical extreme examples in Figure 5.1 assuming the true locus is at $\tau = 25$cM. Under incomplete information, the variances $\text{var}_0(\hat{\pi}_t)$ and $\text{var}_0(\hat{\pi}_\tau)$ are reduced from their fully polymorphic value $\frac{1}{8}$ while the correlation $\text{cor}_0(\hat{\pi}_t, \hat{\pi}_\tau)$ is increased compared to its complete information value $(1-2\theta_{t,\tau})^2$; the net effect is a decrease of $\text{cov}_0(\hat{\pi}_t, \hat{\pi}_\tau)$. The exact relationship between $\text{cov}_0(\hat{\pi}_t, \hat{\pi}_\tau)$ and τ is complex in general, however the covariance is taken under the null hypothesis and can therefore easily and accurately be calculated by Monte Carlo simulations (or gene dropping simulations) as advocated in Lebrec et al. [2004]: we used the `--simulate` option in `MERLIN` to generate marker data for a few thousand sib pairs and calculated the sample covariance between $\hat{\pi}_t$ and $\hat{\pi}_\tau$ after obtaining multipoint estimates of IBD sharing by use of the `--kin` option in `MERLIN` (in general, one such simulation has to be done for each type of pedigree and missing genotype pattern). Note that $\text{var}_0(\hat{\pi}_t)$ can be computed at any arbitrary position t in a similar manner. We have displayed three possible IBD mean specifications in Figure 5.1: the correct one, $\text{cov}_0(\hat{\pi}_t, \hat{\pi}_\tau)C$, labelled 'T&S', the one under complete information, $\frac{1}{8}(1-2\theta_{t,\tau})^2 C$, labelled 'GEE' and a third one, $(1-2\theta_{t,\tau})^2\sqrt{\text{var}_0(\hat{\pi}_t)\text{var}_0(\hat{\pi}_\tau)}C$, labelled 'Var Corrected' that corrects for the incomplete marker information by using the correct variances $\text{var}_0(\hat{\pi}_t)$ and $\text{var}_0(\hat{\pi}_\tau)$ but keeping the correlation as in the ideal situation of complete information (i.e. too low).

In the symmetric information case (Left panel: two markers with 10 equi-frequent alleles at 20cM and 40cM), the location estimate will in practice incur little harm (but the estimate of C will). In presence of asymmetric information (Right panel: two markers with 2 and 10 equi-frequent alleles at 20cM and 40cM respectively), the true expected excess IBD is lower at marker A than at marker B although τ is closer to A, however the true expected excess IBD sharing as per 'GEE' is grossly misspecified since expected IBD is supposed to be much higher at A than at B, the location estimate will be biased towards the more informative marker B, the 'Var Corrected' specification does a better job at approaching the true IBD mean specification but is not accurate.

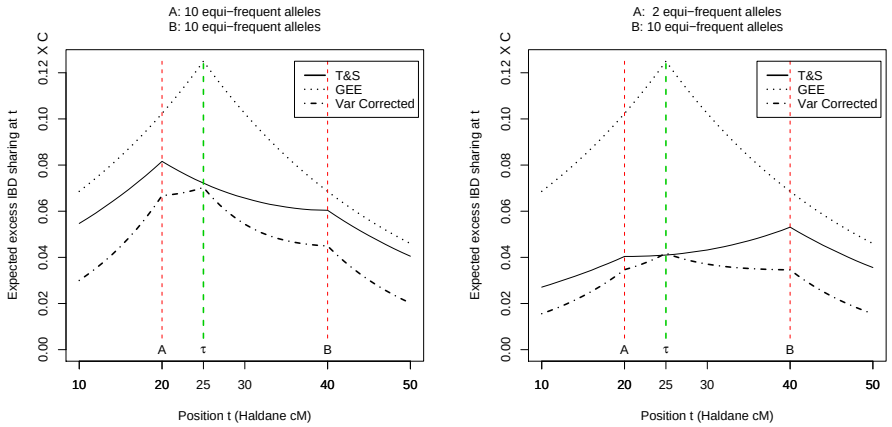

Figure 5.1: Comparison of different mean specifications for excess IBD sharing at position t ($\mathbf{E}(\hat{\pi}_t - \frac{1}{2} \,|\, \text{ASP})$) - 'T&S' (the correct one): $\text{cov}_0(\hat{\pi}_t, \hat{\pi}_\tau)C$, 'GEE' (assumes complete information): $\frac{1}{8}(1 - 2\theta_{t,\tau})^2 C$ and 'Var Corrected': $(1 - 2\theta_{t,\tau})^2 \sqrt{\text{var}_0(\hat{\pi}_t)\text{var}_0(\hat{\pi}_\tau)}C$.

A consistent score test

Feingold et al. [1993] have shown that under a complete high-resolution map, the global test for linkage based on excess IBD sharing given by the supremum of $Z_t = \frac{\sum_{i=1}^{N} \pi_{t,i} - \frac{1}{2}}{\sqrt{N\frac{1}{8}}}$ over the putative chromosomal positions t is the log-likelihood ratio test of a Gaussian process for testing the null hypothesis of no linkage and therefore provides a consistent estimate of the true disease locus location τ. When information is incomplete, a similar test was proposed by Teng and Siegmund [1998] as the maximum of $\hat{Z}_t$ across marker positions with

$$\hat{Z}_t = \frac{\sum_{i=1}^{N} \hat{\pi}_{t,i} - \frac{1}{2}}{\sqrt{\sum_{i=1}^{N} \text{var}_0(\hat{\pi}_{t,i})}} \,,$$

where $\text{var}_0(\hat{\pi}_{t,i})$ may be computed as in subsection 'An accurate IBD specification under incomplete information'. Although their test was based on evaluation of $\hat{Z}_t$ across marker positions only, there is no practical reason for such a restriction when IBD is calculated using multipoint methods and one can in theory calculate $\hat{Z}_t$ on an arbitrarily fine grid of putative locations. Assuming the locus is at τ, the statistic $\hat{Z}_\tau$ turns out to be the *score test* [Cox and Hinkley, 1974] for the C parameter in the

additive model (5.2) [1] and we refer to this test as such in the sequel. One obvious estimator of the locus position is the location $t = \hat{\tau}$ where $\hat{Z}_t$ is maximized in the chromosomal region of interest. We are unaware of a formal proof that as in the case of a high-resolution map, $\hat{\tau}$ provides a consistent estimate of the true locus position, although this is probably known from experience. It turns out to be a corollary of relation (5.3) as we show in an appendix. In addition, one can obtain bootstrap confidence intervals (CI) by resampling with replacement among the N sib pairs and recalculating $\hat{\tau}$ such that $Z_{\hat{\tau}} = \sup_t \hat{Z}_t$ in each new sample. In fact, this score test is also the score test corresponding to the exponential model used by Kong and Cox [1997] although they prefer to use the corresponding likelihood ratio test. It is perhaps worth stressing that the standardization used in $\hat{Z}_t$ is crucial to the consistency of the method, older non-parametric linkage (NPL) methods for ASPs were based on excess IBD sharing only (i.e. the numerator of $\hat{Z}_t$) and the corresponding maximum LOD score thus gave inconsistent estimates of the position under uneven incomplete information even when IBD estimation was done in a multipoint fashion.

5.3 Results - Monte Carlo simulations

In order to assess the impact of incomplete information in practice, we carried out a number of simulations: we generated data from a simple one-locus bi-allelic (disease allele D frequency=0.1) additive model (penetrances=0.0, 0.5 and 1.0 in dd, Dd and DD genotypes resp.; $\lambda_S = \lambda_O = 3.25$). A set of 11 equally-spaced markers spanned a $0 - 100$cM region and the locus was positioned between the 5^{th} and 6^{th} marker at either 42.5cM, 45cM or 47.5cM. We looked at three distinct marker maps (mapH, mapM and mapL) reflecting an increasing degree of systematic differences in marker information; the last six markers always had 10 equi-frequent alleles whereas the first five markers had 8 equi-frequent alleles in mapH, 4 equifrequent alleles in mapM and 2 equi-frequent alleles in mapL. Finally, for each scenario, we considered three sample sizes $N = 100$, 200 and 500 ASPs without parents. In all methods of analysis described below, multipoint IBD estimation was carried out using `MERLIN` [Abecasis et al., 2002]. The locus position and genetic effect were estimated according to the

[1] More precisely, in the model $\mathbf{P}(g \,|\, \text{ASP}) = \sum_{l=0,\frac{1}{2},1} \mathbf{P}_0(g \,|\, \pi_\tau = l)\ \mathbf{P}(\pi_\tau = l \,|\, \text{ASP})$ where g is the multipoint marker information available and $\mathbf{P}(\pi_\tau = l \,|\, \text{ASP})$ is given by model (5.2).

GEE method using `GeneFinder` [Liang et al., 2001], both asymptotic and bootstrap 95% confidence intervals (CI) were calculated. We also carried out two classical analyses for ASP: on a fine grid of chromosomal positions (every cM), we calculated the Kong and Cox [1997] test and the score test $\hat{Z}_t$ defined in subsection 'A consistent score test', the positions where the respective maximum of these two statistics were attained provided position estimates for the locus. In addition, for the score test, we calculated 95% ordinary bootstrap CIs by resampling among the N ASPs. All results are presented in table 5.1.

The GEE estimates of the location are subject to bias which increases as the asymmetry in marker map becomes stronger and which does not decrease with increasing sample size. Although this bias might be considered small, it leads to lower than nominal coverage probability even for the bootstrap CIs, this coverage probability can potentially decrease further as the sample size goes up. Note that a bootstrap algorithm adjusting for bias [Wehrens et al., 2000] could be used here. In contrast, the location estimates obtained by the score test have low bias (probably due to the discrete nature in the search for the supremum of $\hat{Z}_t$ and inaccuracy in calculating $\mathrm{var}_0(\hat{\pi}_t)$) independent of the marker map, the corresponding bootstrap CIs have close to nominal coverage probability.

5.4 Discussion

The GEE methodology offers an attractive and flexible framework for fine mapping of disease loci and its use will likely continue to spread in the coming years. Researchers should therefore all the more be aware of its limitations. Estimates of disease locus position (as well as genetic effect) and associated confidence intervals obtained by existing GEE methods should not be trusted in areas of the genome where IBD information is sparse in particular when this information changes abruptly. In these instances, properly standardized classical methods based on excess IBD sharing, when applied on a fine grid of locations, do provide consistent estimates of the location. Associated confidence intervals with correct coverage probability can also be obtained by re-sampling techniques such as the bootstrap.

The reason for underrating the issue of incomplete information has probably to

			GEE			Score		Kong & Cox
True location	Map (Information Content[a])	N	Average Estimate (cM)	95% Asymptotic CI coverage (%)	95% Bootstrap CI coverage (%)	Average Estimate (cM)	95% Bootstrap CI coverage (%)	Average Estimate (cM)
42.5cM	MapL	100	46.4	71.7	78.9	42.4	94.9	42.4
	(34-84%)	200	46.4	58.2	63.8	42.3	94.2	42.2
		500	46.3	27.8	32.9	42.2	95.4	42.2
	MapM	100	43.9	84.9	89.8	41.9	95.7	41.9
	(55-84%)	200	44.1	83.3	86.1	42.1	94.4	42.2
		500	44.2	76.5	78.5	42.2	94.8	42.3
	MapH	100	43.1	85.9	92.1	42.3	95.4	42.0
	(66-84%)	200	43.0	86.3	92.0	42.1	95.7	42.0
		500	43.1	88.3	90.4	42.3	94.6	42.3
45cM	MapL	100	48.2	78.3	84.7	45.7	96.5	45.4
	(34-84%)	200	48.1	75.8	77.3	45.4	96.1	45.3
		500	47.8	51.6	53.3	45.1	98.0	45.1
	MapM	100	46.4	80.9	90.2	45.0	95.1	45.2
	(55-84%)	200	46.1	90.9	91.9	45.1	97.5	45.0
		500	46.0	91.3	90.0	44.9	96.8	44.9
	MapH	100	45.2	85.1	92.6	45.1	97.6	45.0
	(66-84%)	200	45.0	94.7	95.3	45.0	96.6	44.9
		500	45.1	96.4	95.5	45.0	97.3	45.0
47.5cM	MapL	100	49.6	79.2	89.0	47.9	94.9	47.9
	(34-84%)	200	49.5	76.5	86.2	47.8	94.9	47.7
		500	49.3	78.2	80.7	48.0	94.2	47.8
	MapM	100	48.2	84.5	91.8	47.8	94.7	47.9
	(55-84%)	200	48.1	84.6	91.2	47.8	95.8	47.8
		500	47.9	90.1	92.6	47.7	94.9	47.7
	MapH	100	47.3	86.3	92.8	48.0	95.4	47.7
	(66-84%)	200	47.4	87.4	93.5	48.0	95.5	47.9
		500	47.3	91.6	94.4	47.9	95.3	47.7

Table 5.1: Results of simulations. [a] Information content is expressed as the range of average information content as defined in Kruglyak and Lander [1995] over the 0-100cM region.

do with the nature of the linkage mapping process which usually involves two stages: following a first low-density scan, higher-density genotyping is carried out in one or several promising regions. In this case, IBD information can be fairly accurately determined and the GEE methodology is directly applicable. The advent of SNP chip data for linkage has the potential to provide marker maps with not only higher but also less variable information content [Evans and Cardon, 2004; Schaid et al., 2004] than in classical microsatellites maps, this could potentially increase the reliability of the GEE method in the future. Of course, SNP chip data can only hold such a promise if the data are used in a multipoint fashion for IBD estimation which requires the careful elemination of markers in linkage disequilibrium. However, there are specific situations where similar scenarios to those chosen in our simulations will occur. For example, researchers sometimes embark on collaborative projects (or meta-analysis) whereby several already existing genomewide scans are pooled together in the hope to gain sufficient power (e.g. GenomEUtwin project). In the search for complex traits (with inherent small genetic effects), this second strategy is likely to become more popular. Those distinct scans are often carried out using different marker maps and their pooling will inevitably give rise to regions with heterogeneous IBD information at least in part of the large pooled data set. For those reasons, we believe that the scenarios envisaged in our simulations (and perhaps even more extreme ones as we have personally experienced) are realistic and that our findings have practical implications.

5.5 Appendix

Expected IBD sharing in ASP

We show a proof of the result concerning the expected excess IBD sharing in ASPs under incomplete information. This result is actually due to Teng and Siegmund [1998]. Recall first that $\hat{\pi} = \hat{\pi}(g) = \mathbf{E}_0(\pi \,|\, g) = \frac{1}{2}\ \mathbf{P}_0(\pi = \frac{1}{2} \,|\, g) + \mathbf{P}_0(\pi = 1 \,|\, g)$ where g is the multipoint marker genotype information available (the subscript 0 indicates

a probability $\mathbf{P}_0$ or expectation $\mathbf{E}_0$ independent of the disease locus), then:

$$\begin{aligned}
\mathbf{E}(\hat{\pi}_t - \frac{1}{2} \,|\, \mathrm{ASP}) &= \sum_g (\hat{\pi}_t(g) - \frac{1}{2})\ \mathbf{P}(g \,|\, \mathrm{ASP}) \\
&\quad \text{where } g \text{ spans all possible multipoint genotype configurations,} \\
&= \sum_g (\hat{\pi}_t(g) - \frac{1}{2}) \sum_{l=0,\frac{1}{2},1} \mathbf{P}(g, \pi_\tau = l \,|\, \mathrm{ASP}) \\
&= \sum_g (\hat{\pi}_t(g) - \frac{1}{2}) \sum_{l=0,\frac{1}{2},1} \mathbf{P}(g \,|\, \pi_\tau = l, \mathrm{ASP})\ \mathbf{P}(\pi_\tau = l \,|\, \mathrm{ASP}) \\
&= \sum_g (\hat{\pi}_t(g) - \frac{1}{2}) \sum_{l=0,\frac{1}{2},1} \mathbf{P}_0(g \,|\, \pi_\tau = l)\ \mathbf{P}(\pi_\tau = l \,|\, \mathrm{ASP}) \\
&\quad \text{since markers are in full linkage equilibrium with true locus,} \\
&= \sum_g (\hat{\pi}_t(g) - \frac{1}{2}) \sum_{l=0,\frac{1}{2},1} \frac{\mathbf{P}_0(\pi_\tau = l \,|\, g)}{\mathbf{P}_0(\pi_\tau = l)}\ \mathbf{P}_0(g)\ \mathbf{P}(\pi_\tau = l \,|\, \mathrm{ASP})\ .
\end{aligned}$$

Now replacing the probabilities for unobserved IBD sharing $\mathbf{P}(\pi_\tau = l \,|\, \mathrm{ASP})$ by their values under the additive model introduced above and bearing in mind that $\hat{\pi}_\tau - \frac{1}{2} = \frac{1}{2}[\mathbf{P}_0(\pi_\tau = 1 \,|\, g) - \mathbf{P}_0(\pi_\tau = 0 \,|\, g)]$, it is straightforward to show that

$$\begin{aligned}
\mathbf{E}(\hat{\pi}_t - \frac{1}{2} \,|\, \mathrm{ASP}) &= \sum_g (\hat{\pi}_t - \frac{1}{2})\ \mathbf{P}_0(g) + C \sum_g (\hat{\pi}_t - \frac{1}{2})(\hat{\pi}_\tau - \frac{1}{2}) \mathbf{P}_0(g) \\
&= 0 + \mathrm{cov}_0(\hat{\pi}_t, \hat{\pi}_\tau)\ C\ .
\end{aligned}$$

Consistency of score test

We prove here the consistency of the score test in the estimation of the locus position under an additive model. Let us consider $Y_t = \mathrm{var}_0(\hat{\pi}_t)^{-1/2}\ \left(\hat{\pi}_t - \frac{1}{2}\right)$ then

$$\begin{aligned}
\mathbf{E}(Y_t) &= \mathrm{var}_0(\hat{\pi}_t)^{-1/2}\ \mathbf{E}(\hat{\pi}_t - \frac{1}{2}) \\
&= \mathrm{var}_0(\hat{\pi}_t)^{-1/2}\ \mathrm{cov}_0(\hat{\pi}_t, \hat{\pi}_\tau)\ C \\
&= \mathrm{cor}_0(\hat{\pi}_t, \hat{\pi}_\tau)\ \mathrm{var}_0(\hat{\pi}_\tau)^{1/2}\ C \\
&= \mathrm{cor}_0(\hat{\pi}_t, \hat{\pi}_\tau)\ \mathrm{var}_0(\hat{\pi}_\tau)^{-1/2}\ \mathbf{E}(\hat{\pi}_\tau - \frac{1}{2}) \\
&< \mathbf{E}(Y_\tau) \text{ for } t \neq \tau
\end{aligned}$$

Since $\mathrm{cor}_0(\hat{\pi}_t, \hat{\pi}_\tau)$ is strictly monotonic in t, $Y_\tau - Y_t$ has a strictly positive mean μ and finite variance σ^2. By the Central Limit Theorem, we then have that the sequence $(Z_\tau - Z_t)(N) = N^{-1/2}(Y_\tau - Y_t)(N)$ converges in distribution to $\mathbf{N}(N^{1/2}\sigma\mu, \sigma^2)$ thus

$\mathbf{P}(Z_t(N) < Z_\tau(N)) \to 1$ as $N \to +\infty$ for all $t \neq \tau$. This proves the consistency of the estimate of locus position $t(N)$ taken such that $Z_{t(N)} = \sup_t Z_t(N)$.

Chapter 6

Classical Meta-Analysis Applied to Quantitative Trait Locus Mapping

Abstract

We describe how classical methods for meta-analysis of clinical trials can be adapted to the problem of pooling evidence from different linkage studies. Provided individual QTL estimates and associated standard errors are available on a common chromosomal grid, estimates can be pooled under the assumption of size homogeneity or heterogeneity of the QTL effects while homogeneity can itself be tested. We show also how a simple two-point mixture distribution can be employed as a novel way to allow for between-study locus heterogeneity. The methods may be applied to studies having different marker maps, family structures or different sampling schemes. Finally, we illustrate the methodology using seven data sets for height originating from the GenomEUtwin project and representing 3212 informative families from Australia, Denmark, Finland, The Netherlands, Sweden and the United Kingdom.

This chapter will be submitted as: J.J.P. Lebrec, D.I. Boomsma, K. Christensen, N.G. Martin, N.L. Pedersen, M. Perola, T.D. Spector, H. Putter and H.C. van Houwelingen. Classical Meta-Analysis Applied to QTL mapping - Genomewide Linkage Scan for Height in the GenomEUtwin Project.

6.1 Introduction

Individual loci influencing a complex quantitative trait are most likely to explain only a small proportion of its total variance. Most linkage studies published to date only consist of a few hundred pedigrees with a limited number of individuals and consequently little power to detect linkage of any but the largest QTLs. In order to enhance power, it is now common practice to retrospectively pool evidence for linkage from several different studies. The populations used in each of the studies often have different genetic backgrounds and a locus affecting the trait of interest in one population might have no effect in another one; we will refer to this type of heterogeneity as *locus heterogeneity*. In other instances, the same locus may influence the trait in all populations, but there are many reasons to believe that the size of the effect will vary. For instance, the frequency of the causal allele may be much smaller in some populations or it may interact with other loci, or with environments and risk factors. We will refer to this type of heterogeneity as *size heterogeneity*. Besides those biological sources of heterogeneity, some common logistic sources of variation often arise: typically, genotyping will have been carried out on different marker maps (and even when identical markers are used, their allele frequencies may vary across populations) and families may have been sampled according to different schemes. More simply, the phenotypes measured may vary in their method of collection from study to study. When the raw data are available, one obvious way to gather evidence from several studies is to pool the data into a meta-file and proceed with an overall analysis. In the case of linkage studies with different marker maps, the data manipulations involved are very tedious. Besides, running standard methods of analysis on such large data files usually requires uncommon computing capacities. Of course another simple reason for favoring meta-analysis is that researchers usually simply cannot access the raw data for each study and have to be content with individual test statistics along with (at best) parameter estimates.

We refer the reader to Dempfle and Loesguen [2003] and Rao and Province [2001] for recent overviews of meta-analytic methods for linkage studies. Although widely applicable, rank-based methods such as the GSMA [Wise et al., 1999] are sub-optimal compared to approaches based on the pooling of estimates of a common linkage pa-

rameter. The idea of pooling different estimates of a common linkage effect across studies is not new although it has only been described for sib pair designs to date. Gu et al. [1998] use the excess IBD sharing as a common effect, but their approach appears to be limited to studies with the same marker maps. Li and Rao [1996] and Etzel and Guerra [2002] both use the slope in a classical Haseman-Elston regression as a common effect, the former suffering the same restriction as Gu et al. [1998] regarding location of markers. Interestingly in the latter, the authors explicitly adjust for the (study-specific) marker to locus distance and allow for heterogeneity across studies by means of a random effect. Unfortunately, they do not seem to correctly take into account the within-study dependence structure between markers. We therefore advocate an alternative approach.

In the case of quantitative traits, a natural estimate of common linkage effect is the proportion of total variance explained by a putative location. Classical methods of meta-analysis originally introduced in the field of clinical trials [DerSimonian and Laird, 1986] can be adapted to linkage studies. The sufficient statistics used to perform such approaches are the QTL estimates and their associated standard errors on a common grid of putative locations. It is a well known fact in the biostatistical literature that in absence of individual covariates and under the assumption of homogeneity, pooled data and meta-analytic approaches are equivalent [Olkin and Sampson, 1998], we show in an appendix that a similar result holds for linkage studies.

Assuming that QTL effect estimates and standard errors are available for all studies on a *common grid* of locations, we start in Section 6.2 'Homogeneity' by describing the traditional meta-analytic approach in the context of linkage, while in 'A two-point mixture for locus heterogeneity' we introduce a simple finite mixture model to account for potential locus heterogeneity. In 'Individual analyses', we review the methods which should be used for the analysis of individual studies in order to yield the relevant statistics required for meta-analysis. The methodology is then applied to a genomewide linkage scan for height in seven data sets from six different countries in Section 6.3. Finally, in Section 6.4, we discuss a few practical and methodological issues and briefly compare our findings for height to previous scans.

6.2 Methods

The classical meta-analytic method

Meta-analytic methods are described in full detail in Normand [1999] and van Houwelingen et al. [2002], for example. In this section, we recall briefly how meta-analysis is classically carried out and introduce some refinement specific to linkage studies. We assume that at a given common putative position, each study (indexed by $i = 1, \ldots, K$) provides a consistent estimate $\hat{\gamma}_i$ of the true QTL effect γ_i and an associated standard error s_i.

Homogeneity

Under homogeneity, the effects γ_i's are assumed to be equal to a common value γ so that $\hat{\gamma}_i \sim N(\gamma, s_i^2)$. The corresponding maximum likelihood estimator of γ is therefore given by the weighted average

$$\hat{\gamma}_{\text{hom}} = \frac{\sum_i \hat{\gamma}_i / s_i^2}{\sum_i 1/s_i^2} \text{ with standard error } SE_{\text{hom}} = 1/\sqrt{\sum_i 1/s_i^2}\,. \tag{6.1}$$

The corresponding one-sided statistic

$$\left(z^+_{\text{hom}}\right)^2 = \begin{cases} \left(\hat{\gamma}_{\text{hom}}/SE_{\text{hom}}\right)^2, & \text{if} \hat{\gamma}_{\text{hom}} > 0 \\ 0 & \text{if} \hat{\gamma}_{\text{hom}} \leq 0 \end{cases}$$

follows the mixture distribution $\frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1$ under the null hypothesis, where χ^2_0 denotes the degenerate density with all mass in 0. Of course, one can calculate the corresponding LOD_{hom} score as $\left(z^+_{\text{hom}}\right)^2 / (2 \times \log 10)$.

Test for heterogeneity

Even when the same locus is affecting a trait in different populations, it seems difficult to believe, for reasons given in Section 6.1, that the QTL effects are all equal. In the setting introduced earlier, this situation of *size heterogeneity* can be tested:

$$\begin{aligned} \text{H}_0 &: \gamma_{\text{hom}} = \gamma_1 = \gamma_2 = \cdots = \gamma_K \\ \text{H}_1 &: \text{at least one } \gamma_i \text{ is different ,} \end{aligned}$$

the hypothesis of homogeneity H_0 can be tested using the following statistic

$$X^2 = \sum_{i=1}^{K} \frac{(\hat{\gamma}_i - \hat{\gamma}_{\text{hom}})^2}{s_i^2}$$

whose approximate null distribution is χ^2_{K-1}. In practice, any test for heterogeneity is likely to have little power because individual studies tend to have low precision. Nonetheless, the test can formally suggest heterogeneity in some instances, as will be seen in Section 6.3. Note that the X^2 statistic has an appealing interpretation (at least for researchers with experience in parametric linkage), indeed it can be re-written as

$$\begin{aligned} X^2 &= \sum_{i=1}^{K} \frac{\hat{\gamma}_i^2}{s_i^2} - \frac{\hat{\gamma}_{\text{hom}}^2}{(\sum_i 1/s_i^2)^{-1}} \\ &= 2 \times \log 10 \times \Big(\sum_{i=1,\ldots,K} \text{LOD}_i - \text{LOD}_{\text{hom}} \Big) \\ &= 2 \times \log 10 \times \Big(\sum_{i=1,\ldots,K} \text{LOD}_i - \text{LOD}_{\text{pool}} \Big) \end{aligned}$$

where LOD_{pool} corresponds to the analysis of the pooled meta-file (the fact that $\text{LOD}_{\text{pool}} = \text{LOD}_{\text{hom}}$ is shown in the appendix). In other words, the individual LODs add up only when the effect is perfectly homogeneous.

Size heterogeneity in locus effect

The classical way to allow for heterogeneity between studies is to introduce an additional layer in the earlier homogeneous model by assuming that the study specific effects γ_i's themselves arise from a normal distribution with common mean γ and a between study variance σ^2. This is referred to as a normal mixture model (or random effect model) and results in marginal distributions for the observations given by $\hat{\gamma}_i \sim N(\gamma, s_i^2 + \sigma^2)$. If the between study variance σ^2 were known, the estimate of γ would be

$$\hat{\gamma}_{\text{het}}(\sigma^2) = \frac{\sum_i w_i \hat{\gamma}_i}{\sum_i w_i} \text{ with } w_i = \frac{1}{\sigma^2 + s_i^2} \text{ and with standard error } SE_{\text{het}} = 1/\sqrt{\sum_i w_i}\,,$$

so one way to carry out estimation is by maximization of the profile log-likelihood $\max_{\sigma^2} l\left(\hat{\gamma}_{\text{het}}(\sigma^2), \sigma^2\right)$. In the context of linkage where the actual effects γ_i's are stan-

dardized variance components themselves, all γ_i's should be equal to 0 with probability 1 (i.e. $\sigma^2 = 0$) under the null hypothesis (and not just arising from a $N(0, \sigma^2)$). The test for linkage is then given by the corresponding log-likelihood difference

$$2 \times \left[\max_{\sigma^2} l\left(\hat{\gamma}_{\text{het}}(\sigma^2), \sigma^2\right) - l\left(\gamma = 0, \sigma^2 = 0\right) \right]$$

so that evidence for heterogeneity potentially contributes to the rejection of the null hypothesis of no linkage. The use of the usual mixture $\frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1$ for the null distribution of this non-standard likelihood is probably anti-conservative, the correct asymptotic distribution is given by a mixture $(\frac{1}{2} - p)\chi^2_0 + \frac{1}{2}\chi^2_1 + p\chi^2_0$ [Self and Liang, 1987]. However, asymptotic results are unlikely to be useful since we typically have very few observations (i.e. studies) to pool together. In practice, we use the anti-conservative limits dictated by the $\frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1$ mixture as a screening tool and resort to parametric bootstrapping for refinement of the level of significance once interesting positions have been identified.

A two-point mixture for locus heterogeneity

In some cases, the previous model will not be adequate to model differences between studies because heterogeneity is qualitative rather than quantitative, in other words the locus influences the trait in some studies/populations and not at all in others. In analogy to what is done routinely at the family level in parametric linkage (e.g. Ott [1999], see also Holliday et al. [2005] for a recent application) and can be done in the variance components setting [Ekstrøm and Dalgaard, 2003], one can fit a two-point mixture model at the study level as follows: $\hat{\gamma}_i \,|\, \gamma_i \sim N(\gamma_i, s_i^2)$ with

$$\gamma_i = \begin{cases} \gamma, & \text{with probability } \alpha; \\ 0, & \text{with probability } 1 - \alpha \end{cases}$$

so that

$$\hat{\gamma}_i \sim \alpha N(\gamma, s_i^2) + (1 - \alpha) N(0, s_i^2) \ .$$

The basic idea is that only a proportion α of the studies show linkage to the putative locus and γ is the QTL effect among those studies only. For estimation purposes, this mixture of normal distributions naturally lends itself to the EM algorithm [Dempster and Laird, 1977]. Denoting by $\phi(x; \mu, \sigma^2)$ the normal density function with mean μ

and variance σ^2, the E (estimation) step at stage $k+1$ of the iterative procedure consists in calculating the posterior probabilities $\tau_i^{(k+1)}$'s that the $\hat{\gamma}_i$'s have arisen from a normal distribution with mean $\gamma^{(k)}$ given the prior mixing proportion $\alpha_{(k)}$ i.e.

$$\tau_i^{(k+1)} = \frac{\alpha^{(k)}\phi(\hat{\gamma}_i, \gamma^{(k)}, s_i^2)}{\alpha^{(k)}\phi(\hat{\gamma}_i, \gamma^{(k)}, s_i^2) + (1-\alpha^{(k)})\phi(\hat{\gamma}_i, 0, s_i^2)} \ ,$$

whereas the M (maximization) step gives the updated parameters $\alpha^{(k+1)}$ and $\gamma^{(k+1)}$ as

$$\begin{aligned} \alpha^{(k+1)} &= \sum_{i=1}^{K} \tau_i^{(k+1)}/K \\ \gamma^{(k+1)} &= \frac{\sum_{i=1}^{K} \hat{\gamma}_i \tau_i^{(k+1)}/s_i^2}{\sum_{i=1}^{K} \tau_i^{(k+1)}/s_i^2} \ . \end{aligned}$$

Note that the value of $\tau_i^{(k+1)}$ at convergence gives the posterior probability that study i is linked. The model parameters α and γ are constrained in $[0,1]$ and $[0,+\infty[$ respectively and although the EM estimation procedure described above ensures that $\alpha \in [0,1]$, the estimate of γ will sometimes be negative in which case we set it to 0. Under usual regularity conditions, the corresponding likelihood-ratio test would be asymptotically distributed as a $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ under the null hypothesis. However, here the situation is further complicated by the fact that the model parameters are not identifiable under the null hypothesis (indeed if $\gamma = 0$, any choice of α will give the same likelihood). One way to tackle this problem is to slightly modify the likelihood as done by Chen et al. [2001] and derive corresponding simple asymptotics, but for the same reason alluded to earlier, we prefer to resort to parametric bootstrapping techniques in order to assess significance of the likelihood-ratio test.

Individual analyses

The basic ingredients of a classical meta-analysis are study specific quantitative trait locus effects' estimates $\hat{\gamma}_i$'s in the $i = 1, \ldots, K$ studies available and their associated standard errors s_i's on a *common* fine *grid* of genome locations. In this section, we explain how to do this in practice and make the adjustment for varying information across studies more explicit.

General approach

For random samples of the trait values, the variance components method [Almasy and Blangero, 1998; Amos, 1994] is the standard way of testing for linkage to a quantitative trait. Unfortunately, the emphasis of most computer programs implementing the variance components method has been placed on testing rather than estimating and they rarely provide both quantitative trait locus effect estimates and associated standard errors. In the context of linkage, two exceptions that we know of are the `MENDEL` [Lange et al., 2001] and `Mx` [Neale et al., 1999] softwares. However, in principle, this is not so much of a problem because asymptotic standard errors s can be obtained provided the quantitative trait locus effect estimate $\hat{\gamma}$ is present (and differs from 0) in addition to its statistical significance [1]. At positions where the quantitative trait locus estimate is 0, one could interpolate values of s at neighboring positions where $\hat{\gamma} \neq 0$. One problem with the variance components method, as far as meta-analysis is concerned, is that $\hat{\gamma}$ is constrained to remain positive and pooling of several imprecise estimates $\hat{\gamma}_i$'s could result in a positively biased estimate of the true quantitative trait locus effect γ. Whenever possible, we would personally favor adequate regression or score test approaches [Lebrec et al., 2004] to linkage whose slope is equal to $\hat{\gamma}$ and is allowed to be negative. As shown by Putter et al. [2002], such approaches are equivalent to the variance components method.

When data are selected based on phenotype values (selected sample), the variance components method is no longer valid and appropriate methods that take into account the sampling scheme need to be employed. These so-called inverse regression methods first introduced by Sham and Purcell [2001] have been implemented in `MERLIN-regress` [Sham et al., 2002] and apply to both random and selected samples in arbitrary pedigrees. A typical output from the software will provide a signed estimate of the quantitative trait locus effect $\hat{\gamma}$ and associated standard error s at an arbitrary grid of positions. One outstanding problem with `MERLIN-regress` is the use of an imputed covariance for IBD sharing which can lead to bias in estimation especially in genome areas where markers information is very low. In practice, one

[1] the standard error s of the quantitative trait locus effect estimate $\hat{\gamma}$ is obtained using the approximate relation $(\hat{\gamma}/s)^2 \simeq \chi^2$ with $\chi^2 = \text{LOD} \times 2 \log 10$

clear indication that the imputed covariance is not a good approximation is when the software either gives out QTL estimates larger than 1 with huge associated LOD scores (e.g. tails of chromosomes 8 and 19 in the Finnish data sets - Figure 6.6) or no estimates at all (NA). In our experience, marker maps and densities often vary quite widely and we inevitably end up with areas of the genome with scarce information. Ideally we would therefore recommend using an implementation of the inverse regression approach where a precise approximation of the variance of the IBD estimates is obtained by Monte Carlo simulations. We have been in contact with the authors of `MERLIN-regress` and we hope that the Monte Carlo calculation of the IBD covariance will be implemented as an option in the software in the near future.

Special case: sib pair designs

In order to show how we adjust for differing marker maps, we now outline the inverse regression approach in the simplest and most widespread case of sib pair studies. The trait values $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)'$ are assumed to have been standardized and to follow the usual additive variance components model i.e. the vector $\mathbf{x}$ is assumed to follow a bivariate normal distribution with mean 0 and covariance matrix Σ

$$\Sigma = \begin{bmatrix} 1 & \gamma(\pi - \frac{1}{2}) + \rho \\ \gamma(\pi - \frac{1}{2}) + \rho & 1 \end{bmatrix} .$$

Here π is the proportion of alleles shared identical by descent measured exactly at the quantitative trait locus position and γ therefore represents the proportion of total variance explained by the quantitative trait locus, ρ is the marginal sib-sib correlation for the trait of interest. We show in the appendix an extension of a relation first shown in Putter et al. [2003] under complete information, it gives an approximate regression (valid for small values of γ) between excess IBD sharing and a function of the phenotype trait values which is the basis of the inverse regression approach:

$$\mathbf{E}(\hat{\pi} - \frac{1}{2} \,|\, \mathbf{x}, \gamma) \simeq \gamma \ \mathrm{var}_M(\hat{\pi}) \ C(\mathbf{x}, \rho)$$

where

$$\hat{\pi} = 0.5 \times \mathbf{P}(\pi = 0.5 \,|\, M) + 1 \times \mathbf{P}(\pi = 1 \,|\, M)$$

is the usual estimate of IBD sharing given marker data M available while

$$C(\mathbf{x},\rho) = \left[(1+\rho^2)x_1x_2 - \rho(x_1^2+x_2^2) + \rho(1-\rho^2)\right]/(1-\rho^2)^2$$

and is sometimes referred to as the optimal Haseman-Elston function. For a sample of $j = 1,\ldots,N$ sib pairs, the method of least squares provides an approximately consistent estimate of γ given by

$$\hat{\gamma} = \frac{\sum_{j=1}^{N}(\hat{\pi}_j - \frac{1}{2})C(\mathbf{x}_j,\rho)}{\mathrm{var}_M(\hat{\pi}) \times \sum_{j=1}^{N} C^2(\mathbf{x}_j,\rho)}\ , \tag{6.2}$$

$$\text{with standard error } s = \left(\mathrm{var}_M(\hat{\pi}) \times \sum_{j=1}^{N} C^2(\mathbf{x}_j,\rho)\right)^{-1/2}\ . \tag{6.3}$$

Here $\mathrm{var}_M(\hat{\pi})$ represents the variance of $\hat{\pi}$ with respect to the probability of marker alleles and would equal $\frac{1}{8}$ under complete information. It depends on the pedigree structure, the markers' characteristics (i.e. allele frequencies and inter-marker distances) and the missing pattern of genotypes, and although an exact calculation is extremely tedious it can be closely approximated by simple Monte Carlo simulations. We show in Figure 6.1 how widely the measure of information may vary within and between studies. It is therefore crucial to appropriately account for this variation when estimating γ, failure to do so may introduce bias in the QTL estimates.

Retrospective analysis of an individual study

Often, the only data at hand are QTL estimates ($\hat{\gamma}$'s) and their standard errors (s's) on an original grid of locations which is not the common one we wish to use in the meta-analysis; typically this original grid would be a set of say $t = 1,\ldots,M$ markers' positions. If the characteristics of the original map are available, we show how to obtain QTL estimates and associated standard errors on this new common grid of locations.

For the sake of simplicity, we stick to sib-pair designs as in the previous section. Given the $M \times 1$ vector of original $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_t)_{t=1,\ldots,M}$ and associated standard errors $(s_t)_{t=1,\ldots,M}$, the best linear approximation of the QTL effect $\hat{\gamma}_q$ at an arbitrary

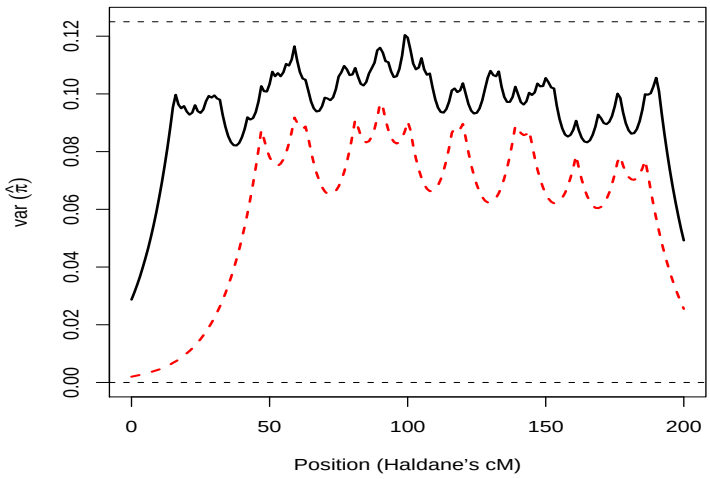

Figure 6.1: Chromosome 6 - Markers' information in the 'AUS' map (continuous line) and the 'NL1' map (dotted line)

position denoted q is given by a weighted least squares estimate

$$\begin{aligned} \hat{\gamma}_q &= \frac{\omega_q' V^{-1} \hat{\gamma}}{\omega_q' V^{-1} \omega_q} \ , \\ \text{with standard error } \ s_q &= \left(\omega_q' V^{-1} \omega_q\right)^{-1/2} \ . \end{aligned}$$

Here V denotes the variance-covariance matrix of the vector $\hat{\gamma}$ under the null hypothesis of no linkage and is given by

$$V_{kl} = \begin{cases} \text{var}_M(\hat{\pi}_k)^{-1} & \text{if } k = l \\ \text{cov}_M(\hat{\pi}_k, \hat{\pi}_l) \ (\text{var}_M(\hat{\pi}_k)\text{var}_M(\hat{\pi}_l))^{-1} & \text{if } k \neq l \end{cases} \ ,$$

and ω_q is the $M \times 1$ vector whose k^{th} element is given by

$$\omega_{q,k} = \frac{\text{cov}_M(\hat{\pi}_k, \hat{\pi}_q)}{\text{var}_M(\hat{\pi}_q)} \ .$$

All the var_M and cov_M terms can be calculated by Monte Carlo simulations provided the map characteristics and pedigree structure are known.

In the idealized case of a saturated map which would supply perfect IBD knowledge at any location on a chromosome, all var_M terms are equal to $\frac{1}{8}$ and $\text{cov}_M(\hat{\pi}_{t_1}, \hat{\pi}_{t_2}) =$

$\frac{1}{8}(1-2\theta_{t_1,t_2})^2$, where θ_{t_1,t_2} is the recombination fraction between loci at t_1 and t_2 [Risch, 1990]. Taking the off-diagonal terms in V to be equal to 0 (i.e. assuming that markers are not linked), one obtains the estimate of QTL effect advocated by Etzel and Guerra [2002] (with the between-study variance $\sigma^2 = 0$). In the context of meta-analysis, it is important to properly account for differences in marker information between studies, unless the marker maps are close to saturated in all studies. Remarkably, the elements needed to calculate $\hat{\gamma}_q$ and s_q at any arbitrary location are just the corresponding estimates at M marker locations and map characteristics, none of the subject-specific data (traits values, individual IBD estimates $\hat{\pi}_i$) are needed.

6.3 Results

We applied the methods described in Section 6.2 to seven data sets (labelled 'FIN', 'DK', 'NL1', 'NL2', 'S', 'UK' and 'AUS' for Finland, Denmark, The Netherlands(x2), Sweden, the United Kingdom and Australia respectively) gathered by members of the GenomEUtwin project with phenotypic information on height (see Silventoinen et al. [2003] for heritability study). The data available for linkage analysis consisted essentially of sibships and nuclear families of varying sizes and are summarized in Table 6.1. Genotyping had been carried out using different marker maps and densities across studies but we actually had access to the raw data sets and could therefore easily obtain QTL estimates and standard errors on a common grid of positions. This was done using the inverse regression method implemented in `MERLIN-regress` with heritability values equal to twice the country specific opposite sex sib-sib correlations observed in the large sample data published in Silventoinen et al. [2003] with an upper boundary of 0.99 (heritability values used were thus 0.98, 0.99, 0.86, 0.86, 0.99, 0.99 and 0.92 for the 'FIN', 'DK', 'NL1', 'NL2', 'S', 'UK' and 'AUS' data sets respectively). Since the data could be considered random samples of height measurements in each country, we also carried out a variance components analysis as implemented in `MERLIN`, this was done as a check of the `MERLIN-regress` analysis because of its sometimes erratic behavior in particular in the tails of the chromosomes (e.g. see chromosomes 8 and 19 in the Finnish data set). Finally, we analyzed the X-chromosome using the variance components method implemented in `MINX` (`MERLIN` in X). When

the QTL variance γ_i was estimated as 0 in the X chromosome, it was not possible to derive the asymptotic standard error s_i according to the method described in Section 6.2 'General approach'. For those positions, we either interpolated the values of s_i at other positions by simply taking the average s_i in data set i, or (when QTL variance was estimated as 0 at all positions as in the 'DK' and 'NL1' data sets) we just used the s_i values of the 'FIN' data set because those three data sets had rather comparable information on other chromosomes. We realize that those approximations might seem very crude however it is clear that the results of the subsequent pooled analysis of X are qualitatively robust.

Family type	FIN	DK	NL1	NL2	S	UK	AUS
2 sibs	346	313	25	94	51	1107	603
2 sibs + parents	0	0	77	44	0	0	185
3 sibs	14	0	13	0	0	0	84
3 sibs + parents	0	0	45	0	0	0	40
4 sibs	16	0	11	0	0	0	26
4 sibs + parents	0	0	11	0	0	0	22
5 sibs+	10	0	9	0	0	0	6
5 sibs+ + parents	0	0	4	0	0	0	7
Total number of families	386	313	195	138	51	1107	1022*

Table 6.1: Informative data available for linkage analysis - * 'AUS' also contains 49 non-nuclear families

Prior to linkage analysis, raw phenotypic data were adjusted for sex and age, within country. For that purpose, separately for each data set and for each sex within each data set, we fitted the following linear mixed model to the height measurements of relatives j and k in family i:

$$\begin{cases} \text{height}_{ij} = \mu + \beta \times \text{age}_{ij} + \epsilon_{ij} \\ \text{height}_{ik} = \mu + \beta \times \text{age}_{ik} + \epsilon_{ik} \end{cases} \quad \text{with} \begin{cases} \text{var}(\epsilon_{ij}) = a^2 + e^2 \\ \text{cov}(\epsilon_{ij}, \epsilon_{ik}) = \mathbf{E}(\pi_{jk}) a^2 \end{cases}$$

where $\mathbf{E}(\pi_{jk})$ equals the expected IBD sharing between relatives j and k i.e. twice their kinship coefficient. Estimation of the models' parameters was carried out using the $-a-$ option of the QTDT software [Abecasis et al., 2000] and the corresponding

standardized residuals obtained as $\left(\text{height}_{ij} - \hat{\mu} - \hat{\beta} \times \text{age}_{ij}\right) / \sqrt{\hat{a}^2 + \hat{e}^2}$ were then used as phenotypes in the linkage analysis.

We present graphically the results of two chromosomes which are interesting from the methodological point of view: chromosome 2 (Figure 6.2) and chromosome 7 (Figure 6.3). In the region of chromosome 2 around 200cM, QTL estimates vary quite widely across studies which is also suggested more formally by the heterogeneity test. It is also clear that we are in presence of qualitative heterogeneity since although the effect is undeniable in 'FIN' and perhaps present in 'NL2' and 'AUS' it seems to be completely absent in the four other data sets. As a result, the significant signal observed in the Finnish study has disappeared in the homogeneous model while the normal mixture and the two-point mixture somehow recover it.

Similar outputs are displayed for chromosome 7 in Figure 6.3. In the region just right of 0cM, heterogeneity of QTL effects is not as obvious as in the previous example and in fact the pooled homogeneous analysis enhances statistical significance. Note that the QTL estimates obtained by the two other methods coincide with those under the homogeneous model as well as the corresponding LOD scores although LOD scores do not follow the same null distribution.

Summary results over the whole genome are presented in Figures 4–8. Position on the chromosomes is expressed in cumulative Kosambi's cM. Data set specific QTL estimates and corresponding LOD scores are displayed in Figures 6.4, 6.5 and 6.6, 6.7 (for both `MERLIN-regress` and variance components analyses) respectively while similar outputs for the pooled analysis appear on Figures 6.9 and 6.10 (continuous blue line: homogeneity model, broken green line: random effect model and broken green line: 2-point mixture model). The test for heterogeneity is shown for the whole genome in Figure 6.8.

The highest autosomal pooled LOD score (bootstrap adjusted 2-point mixture LOD=2.11, unadjusted LOD=2.34) is obtained at 48cM on chromosome 5 with $\hat{\alpha} = 0.15 \simeq \frac{1}{7}$ indicating that only data set 'NL1' appears to be linked. The second highest score (unadjusted 2-point mixture LOD=2.06) is obtained at 208cM on chromosome 2 and pools evidence from the 'FIN' and 'NL2' data sets ($\hat{\alpha} = 0.24$). There are seven other somewhat less convincing peaks (LOD score between 1 and 2) on chromosomes

5,7,8,11 and 15. In addition, chromosome X provides undeniable proof for linkage in two locations (pooled LOD scores around 3 or beyond at 70 cM and 145cM) while there is suggestion of a third peak at 110cM, all this evidence for linkage appears to come from the Finnish data set only ($\hat{\alpha} \simeq \frac{1}{7}$).

A glance at the whole genome reveals that positions at which the three methods differ are fairly rare in the present analysis despite the fact that estimates of variance appear to vary a lot between studies. This is partly due to the relative small size of each data set which does not allow to clearly establish heterogeneity between studies. Once all data from the GenomEUtwin project are gathered, the three methods that we have described here are likely to yield quite different results. It must be noticed that the overall pooling exercise may appear fairly disappointing since there are very few locations where statistical significance is enhanced i.e. where the pooled LOD score is higher than the maximum of the individual LOD scores. Two such locations are the beginning of chromosomes 7 and 11 and correspond to fairly small QTL effects (pooled estimates between 5 and 10 % of total variance), such effect sizes would require sample sizes in the order of 30000 (unselected) sib pairs in order to have a decent chance to formally detect linkage [Putter et al., 2003].

6.4 Discussion

We have detailed how classical meta-analytic methods can be adapted to linkage provided consistent estimates of QTL effects along with standard errors are available for each study on a common grid of positions. The methods required to obtain such summary statistics are now well developed and their software implementation has been publicly available for a number of years. We realize, however, that most published studies to date will not have sufficient information in order to carry out the method advocated here. Indeed, it is still common practice nowadays in the literature, even for QTL mapping where the effect to be estimated is fairly uncontroversial, to publish statistics conveying statistical significance only (i.e. LOD scores) without any idea of the actual effect estimate. This heavily hinders powerful pooling of the many small linkage studies available in the community. Gu et al. [1998] presented guidelines on how to report linkage studies that would enable future meta-analysis using IBD

sharing as a common linkage parameter. Since the analysis tools are available (e.g. `MERLIN-regress`), it should be expected by journals that researchers publish QTL effects and associated standard errors (at least as add-on information) on a grid of locations.

We have demonstrated (see Appendix - Equivalence meta-analysis / pooled data set) that under the assumption of homogeneity and in absence of individual covariates, there is simply no advantage in analyzing a meta-file where the raw data from each separate study would be pooled. This is particularly relevant given the enormous effort required to combine data from individual sources into such a meta-file. In practice, pooling of data is a sequential process and having to re-create a meta-file each time extra data is available would become a major burden. In the purely meta-analytic approach that we advocate, addition of new data poses no problem. In fact for the homogeneous analysis, all that is needed for a re-analysis with extra data (i.e. γ_{extra} and s_{extra}) is the previous homogeneous QTL effect estimate $\hat{\gamma}_{\text{hom}}$ and its associated standard error SE_{hom}. Note that the two methods described to allow for heterogeneity between studies would still require the same summary study specific QTL effect estimates and standard errors.

Given the small individual study sizes one typically encounters, any test for heterogeneity of quantitative trait locus effects across studies is bound to suffer from a lack of power. This is reflected in the test for heterogeneity as well as in the estimate of the between study variance component σ^2 which very rarely differs from 0. Note that the classical random effects model is probably not the most appropriate in the case of linkage, indeed the fact that the quantitative trait locus effect is a variance component precludes it from being negative (which is not impossible under the normal mixture model) and suggests that the random effects γ_i's could be more appropriately modelled as arising from a Γ distribution. Another way of testing locus heterogeneity is to formally test whether $\alpha > 0$ in the two-point mixture of Section 6.2 'A two-point mixture for locus heterogeneity'.

The idea of applying the concept of finite mixture models to meta-analysis is also not new [Bohning et al., 1998] although it is new for meta-analysis of linkage studies as far as we are aware. It is based on the simple idea that only studies with a positive

effect should be pooled together to provide evidence for linkage. Instead of doing this 'by hand', we let the data decide which study exhibits positive linkage. In our data example, when locus heterogeneity appeared to be present, the resulting LOD score was always lower than the LOD score obtained in one of the studies showing strong linkage, however it need not be so in general as the next example shows. Take five studies with the following estimates of QTL effects $\hat{\gamma} = (0, 0, 0.2, 0.2, 0.2)$ and associated standard errors $s = (0, 0, 0.1, 0.1, 0.1)$, the statistical significance of the individual studies is given by χ^2 statistics equal to $(0, 0, 4, 4, 4)$. The maximum likelihood estimates of α and γ in the two-point mixture model are 1.0 and 0.12 respectively with a corresponding likelihood ratio test of 7.2. We calculated the significance of such a value by parametric bootstrapping and the corresponding value for a χ^2 distribution is 6.6 which remains higher than 4. Therefore given sufficient precision of the individual studies, allowance for heterogeneity can enhance statistical significance of individual studies.

We have implemented the three methods described in Section 6.2 along with the test for heterogeneity and the parametric bootstrapping for evaluation of significance in R. The programs are available at `http://www.msbi.nl/Genetics/`.

The two dutch data sets 'NL1' and 'NL2' that we have used were also part of the data in Willemsen et al. [2004] although they also included phenotypic information from untyped individuals in their analysis. The highest pooled peak that we found at 48cM on chromosome 5 actually corresponds to the 'NL1' data set only and was also identified by Willemsen et al. [2004], the nearest QTL identified in that region until now was at 69cM in a Swedish population [Hirschhorn et al., 2001]. The peak on chromosome 2 is a replication of findings made in the population of the Botnia region in Finland. The other suggestive peaks at the beginning of chromosome 8, on chromosome 11 and 15 appear to be replications of previous findings too. However, peaks at the end of chromosome 5, on chromosome 7 and in the middle of chromosome 8 have not been identified before as far as we are aware. We refer the reader to [Willemsen et al., 2004] for a recent overview of QTLs involved in height. The genomewide results in Figures 6.6 and 6.7 also highlight a couple of additional peaks which seem to be purely country specific, like the start of chromosome 9 in the two

Dutch data sets and chromosomes 6, 14 and 16 in the Finnish data set. Finally, the most convincing evidence of linkage comes from the X chromosome in the Finnish data sets with two substantial peaks which appear to replicate findings in Deng et al. [2002].

Overall, this pooling exercise may appear disappointing since statistical significance was enhanced in only two visible locations over the whole genome. Nevertheless, there are two lessons to be learnt from this experience. Firstly, allowance for heterogeneity has the potential to help in detecting loci with either *locus heterogeneity* or *size heterogeneity* but then sufficient sample size is required in the individual studies in order to detect heterogeneity. Secondly, when the sample size of individual studies are small, pooling will enhance statistical significance if the effects are similar across studies, the most subtle QTL effects are probably more likely to fulfill this assumption of homogeneity. We still have not reached the sample sizes required to detect such small effect sizes. When the full data potentially available in the GenomEUtwin project are gathered, we will hopefully be in a position to find QTLs involved in common complex traits.

6.5 Appendix

Expected IBD sharing under incomplete information

We derive here the expected IBD sharing for sib pairs under incomplete information and assuming that $\hat{\pi}$ is being measured exactly at the locus. Recall first that $\hat{\pi} = \mathbf{E}(\pi \,|\, g) = \frac{1}{2}\,\mathbf{P}(\pi = \frac{1}{2} \,|\, g) + \mathbf{P}(\pi = 1 \,|\, g)$ where g is the genotype information available.

$$\begin{aligned}
\mathbf{E}(\hat{\pi} - \frac{1}{2} \,|\, \mathbf{x}, \gamma) &= \sum_g (\hat{\pi} - \frac{1}{2})\ \mathbf{P}(g \,|\, \mathbf{x}) \\
&\qquad \text{where } g \text{ spans all possible multipoint genotype configurations} \\
&= \sum_g (\hat{\pi} - \frac{1}{2}) \sum_{l=0,\frac{1}{2},1} \mathbf{P}(g, \pi = l \,|\, \mathbf{x}) \\
&= \sum_g (\hat{\pi} - \frac{1}{2}) \sum_{l=0,\frac{1}{2},1} \mathbf{P}(g \,|\, \pi = l, \mathbf{x})\ \mathbf{P}(\pi = l \,|\, \mathbf{x})
\end{aligned}$$

Now since markers are in full linkage equilibrium with the true locus, we have

$$\begin{aligned}\mathbf{E}(\hat{\pi}-\frac{1}{2}\,|\,\mathbf{x},\gamma) &= \sum_g(\hat{\pi}-\frac{1}{2})\sum_{l=0,\frac{1}{2},1}\mathbf{P}(g\,|\,\pi=l)\;\mathbf{P}(\pi=l\,|\,\mathbf{x})\\ &= \sum_g(\hat{\pi}-\frac{1}{2})\sum_{l=0,\frac{1}{2},1}\frac{\mathbf{P}(\pi=l\,|\,g)}{\mathbf{P}(\pi=l)}\;\mathbf{P}(g)\;\mathbf{P}(\pi=l\,|\,\mathbf{x})\;.\end{aligned}$$

Using a first order Taylor approximation for $\mathbf{P}(\pi\,|\,\mathbf{x})$ under an additive model introduced in Putter et al. [2003]: $\mathbf{P}(\pi=0\,|\,\mathbf{x},\gamma,\rho)\simeq\frac{1}{4}-\frac{\gamma}{8}C(\mathbf{x},\rho)$, $\mathbf{P}(\pi=\frac{1}{2}\,|\,\mathbf{x},\gamma,\rho)\simeq\frac{1}{2}$ and $\mathbf{P}(\pi=1\,|\,\mathbf{x},\gamma,\rho)\simeq\frac{1}{4}+\frac{\gamma}{8}C(\mathbf{x},\rho)$, it is now easy to show that

$$\begin{aligned}\mathbf{E}(\hat{\pi}-\frac{1}{2}\,|\,\mathbf{x}) &= \sum_g(\hat{\pi}-\frac{1}{2})\;\mathbf{P}(g)+\gamma C(\mathbf{x})\;\sum_g(\hat{\pi}-\frac{1}{2})^2\;\mathbf{P}(g)\\ &= 0+\mathrm{var}(\hat{\pi})\;\gamma C(\mathbf{x})\;.\end{aligned}$$

Equivalence meta-analysis / pooled data set

This appendix presents a formal proof that, under the assumption of homogeneity of the QTL effect across studies, meta-analysis of the data as advocated in Section 6.2 'Homogeneity' is equivalent to an analysis of the individual raw data. For this purpose, we place ourselves in the case where the QTL effect γ is small and score test or inverse regression strategies are optimal [Lebrec et al., 2004]. Without loss of generality, we look at the special case of sib-pair designs and use assumptions and notations introduced in Section 6.2 'Special case: sib pair designs' with the addition that the subscript $i=1,\ldots,K$ stands for the K studies available. In this context, the test for linkage is a simple regression through the origin of excess identical by descent sharing on the optimal Haseman-Elston function of the standardized trait values. The proof is somewhat trivial: all we show is that the regression of the meta-file consisting of the K individual data sets is just the weighted average of the individual regressions given by Formula (6.1). In the meta-file, the data consist of the response variable $\left(\hat{\pi}_{ij}-\frac{1}{2}\right)_{ij}$ excess identical by descent sharing measured in the sib pair $j=1,\ldots,N_i$ in study $i=1,\ldots,K$ and the corresponding regressor equal to the product of the phenotype function trait value $C_{ij}=C(\mathbf{x}_{ij1},\mathbf{x}_{ij2},\rho_i)$ by marker information $\mathrm{var}_{M_i}(\hat{\pi}_{ij})$. The notation stresses the fact that the sib-sib correlation ρ_i and the marker information M_i are study-specific. The response variable will in general have different variance

across studies so the estimate of the QTL effect γ is given by the weighted least squares method as:

$$\hat{\gamma}_{\text{pool}} = \frac{\sum_{i=1}^{K}\sum_{j=1}^{N_i}(\hat{\pi}_{ij}-\frac{1}{2})C_{ij}}{\sum_{i=1}^{K}\sum_{j=1}^{N_i}\text{var}_{M_i}(\hat{\pi}_{ij})C^2(\mathbf{x}_{ij},\rho_i)} ,$$

and using notations introduced in Section 6.2 'Homogeneity', we have

$$\hat{\gamma}_{\text{pool}} = \frac{\sum_i \hat{\gamma}_i/s_i^2}{\sum_i 1/s_i^2} = \hat{\gamma}_{\text{hom}} .$$

Acknowledgements

We are much indebted to the different twin registries who contributed the data used in the genomewide scan for height, their collaborative effort was key to this study. We also wish to thank the researchers at KTL (Finnish National Public Health Institute) in particular Tero Hiekkalinna and Sampo Sammalisto for their enormous effort in harmonizing the data from the different contributors.

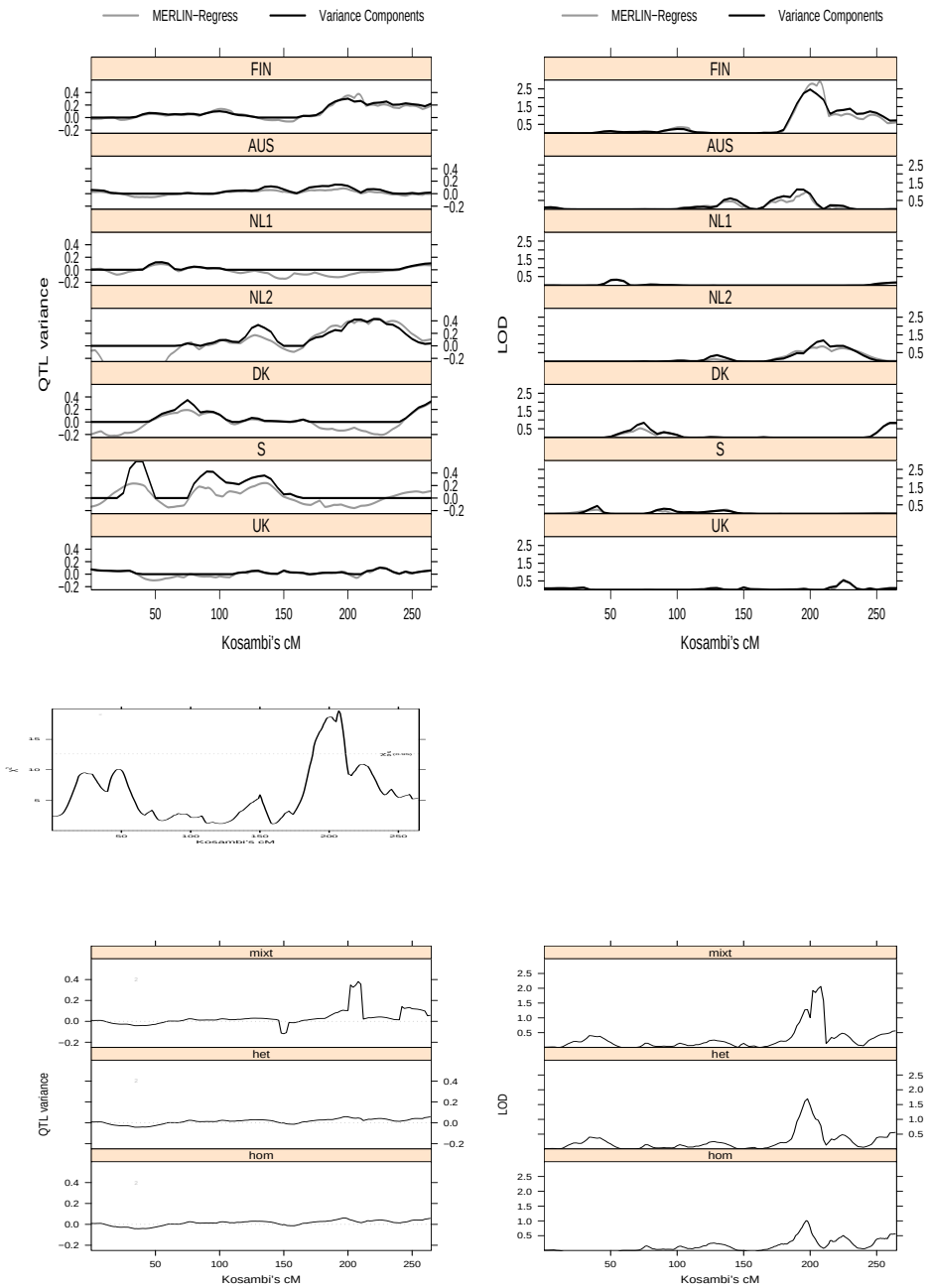

Figure 6.2: Chromosome 2 - QTL analysis for height - Individual analyses (top), test for heterogeneity (middle) and meta-analyses (bottom)

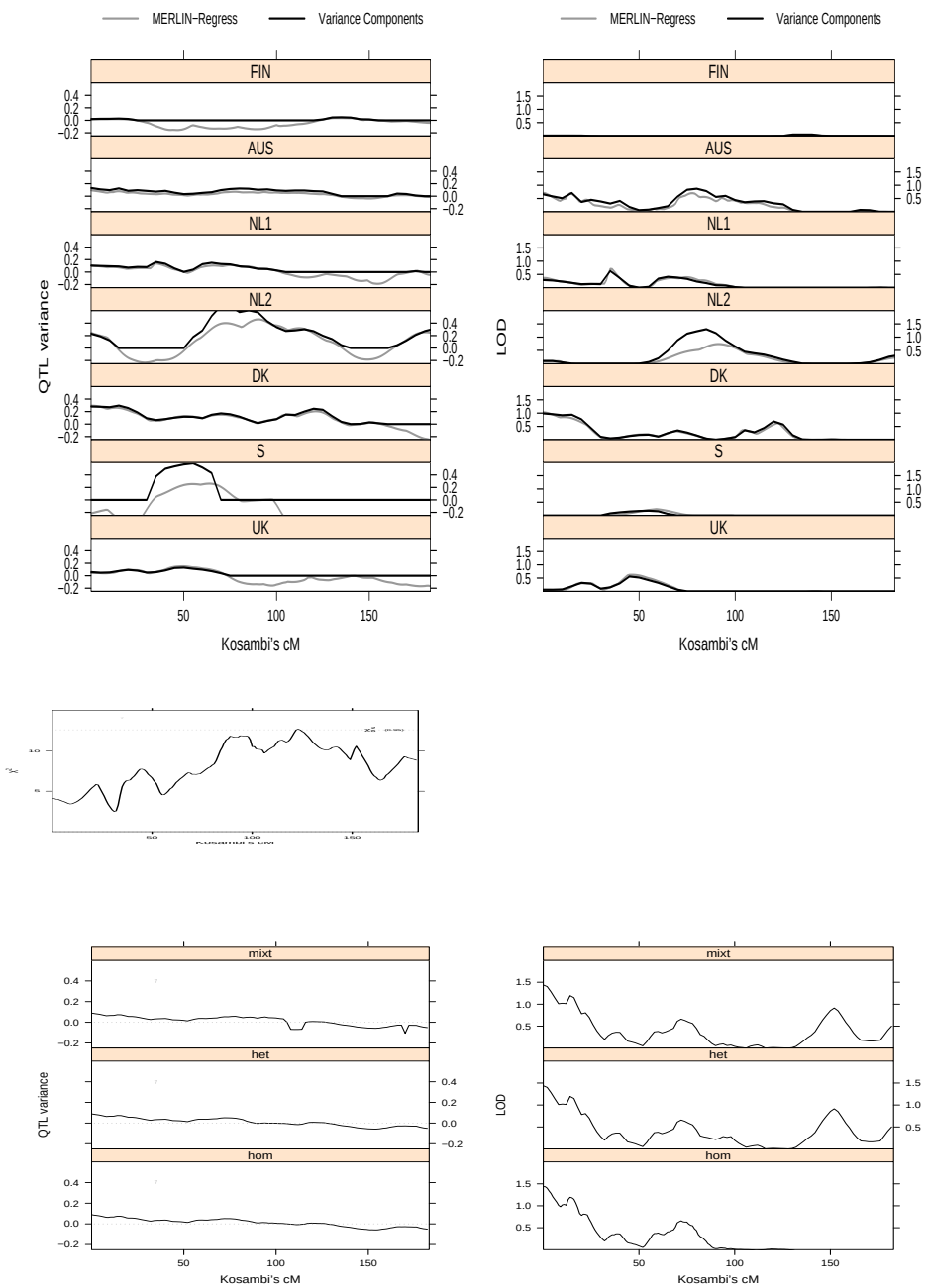

Figure 6.3: Chromosome 7 - QTL analysis for height - Individual analyses (top), test for heterogeneity (middle) and meta-analyses (bottom)

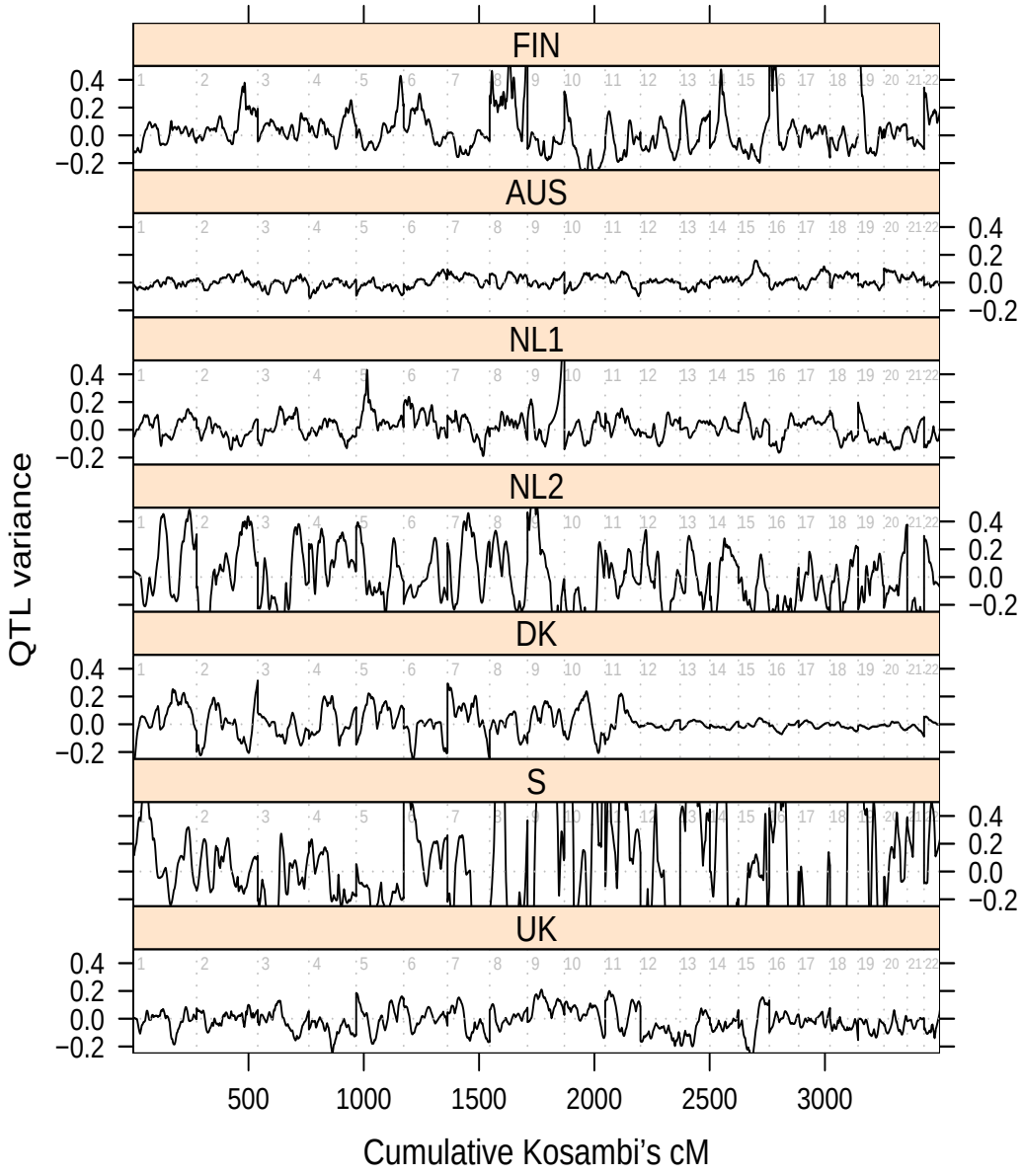

Figure 6.4: Genomewide `MERLIN-regress` QTL Variance for height - Individual studies

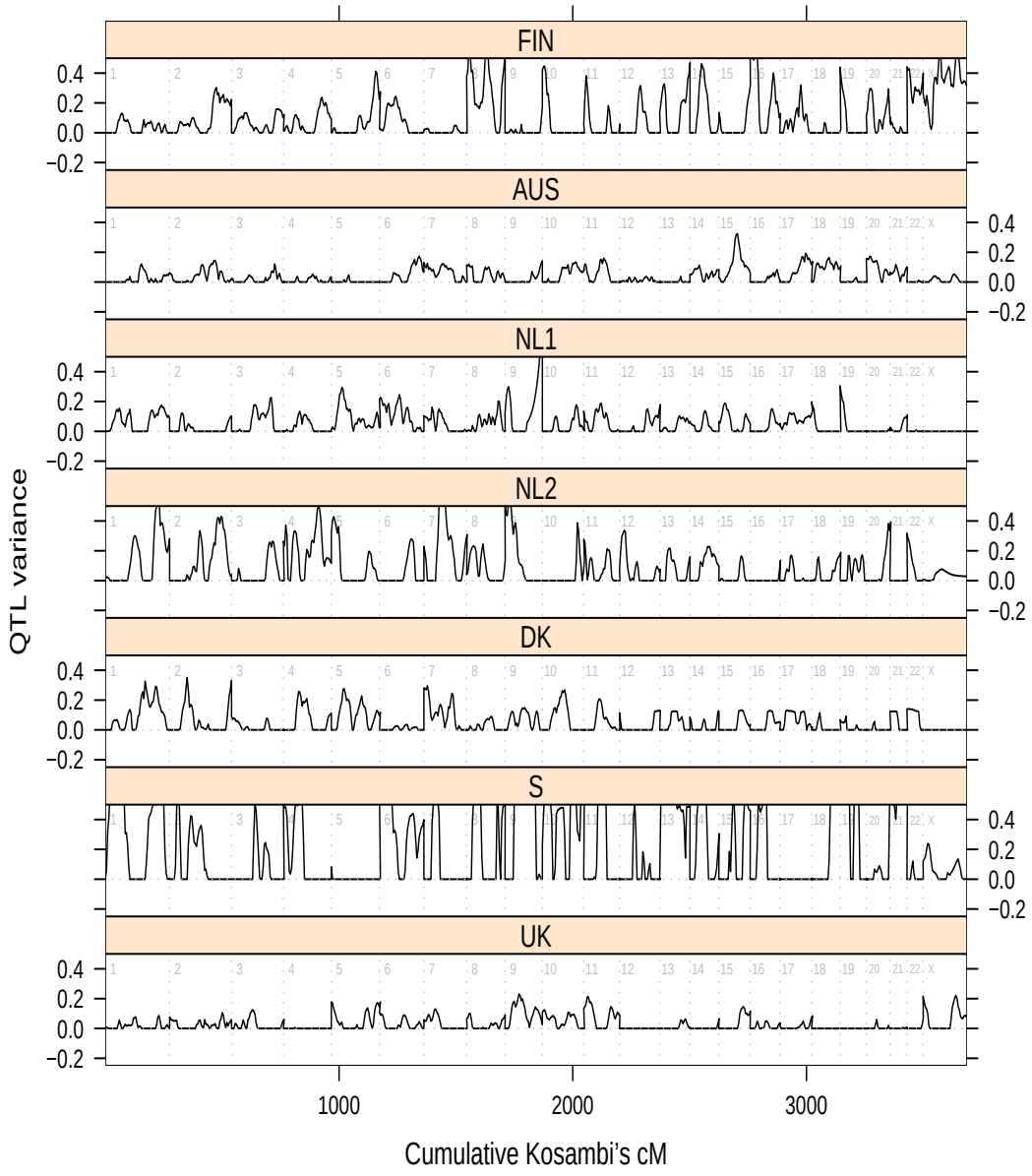

Figure 6.5: Genomewide Variance Components QTL Variance for height - Individual studies

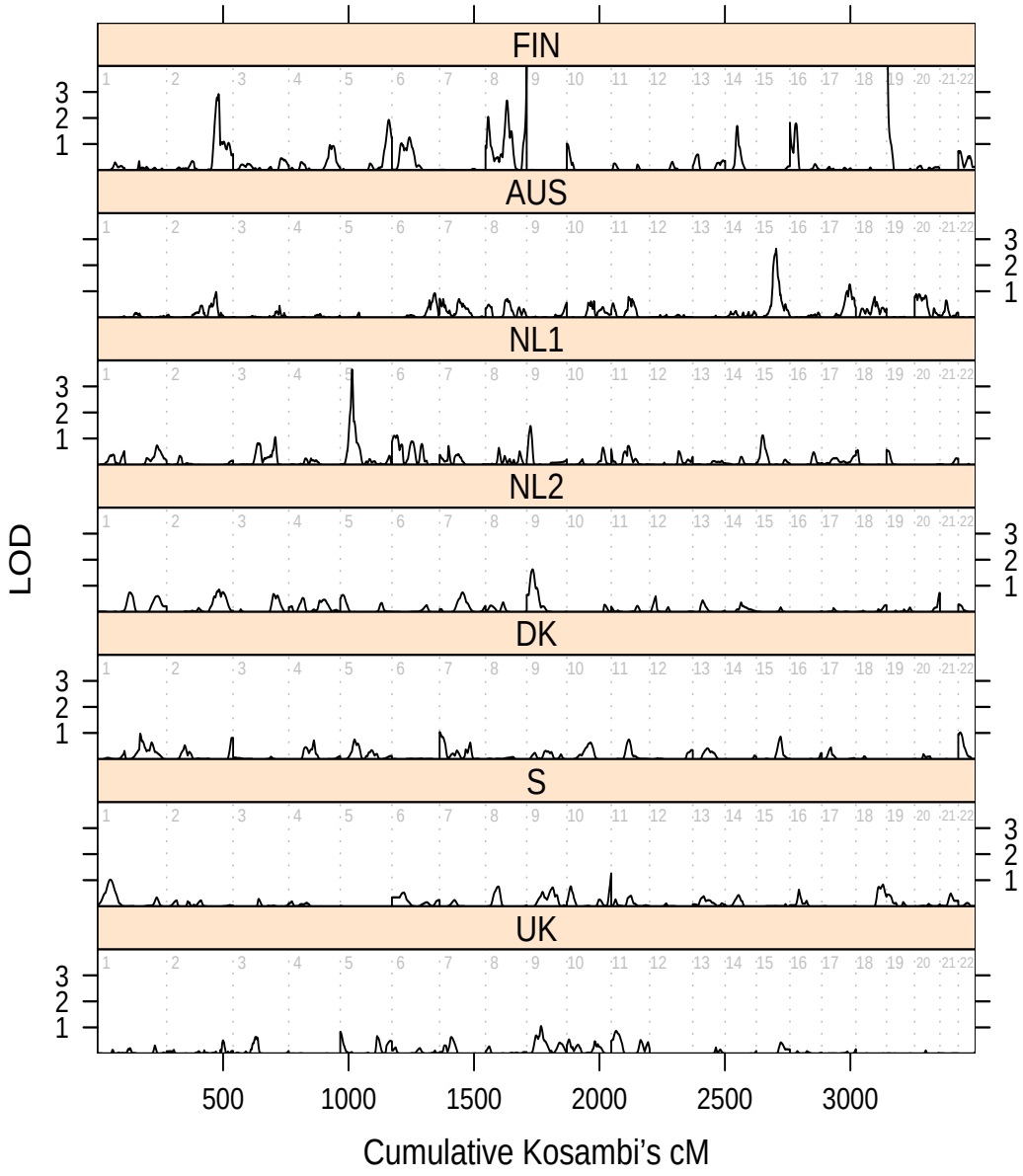

Figure 6.6: Genomewide `MERLIN-regress` LOD scores for height - Individual studies

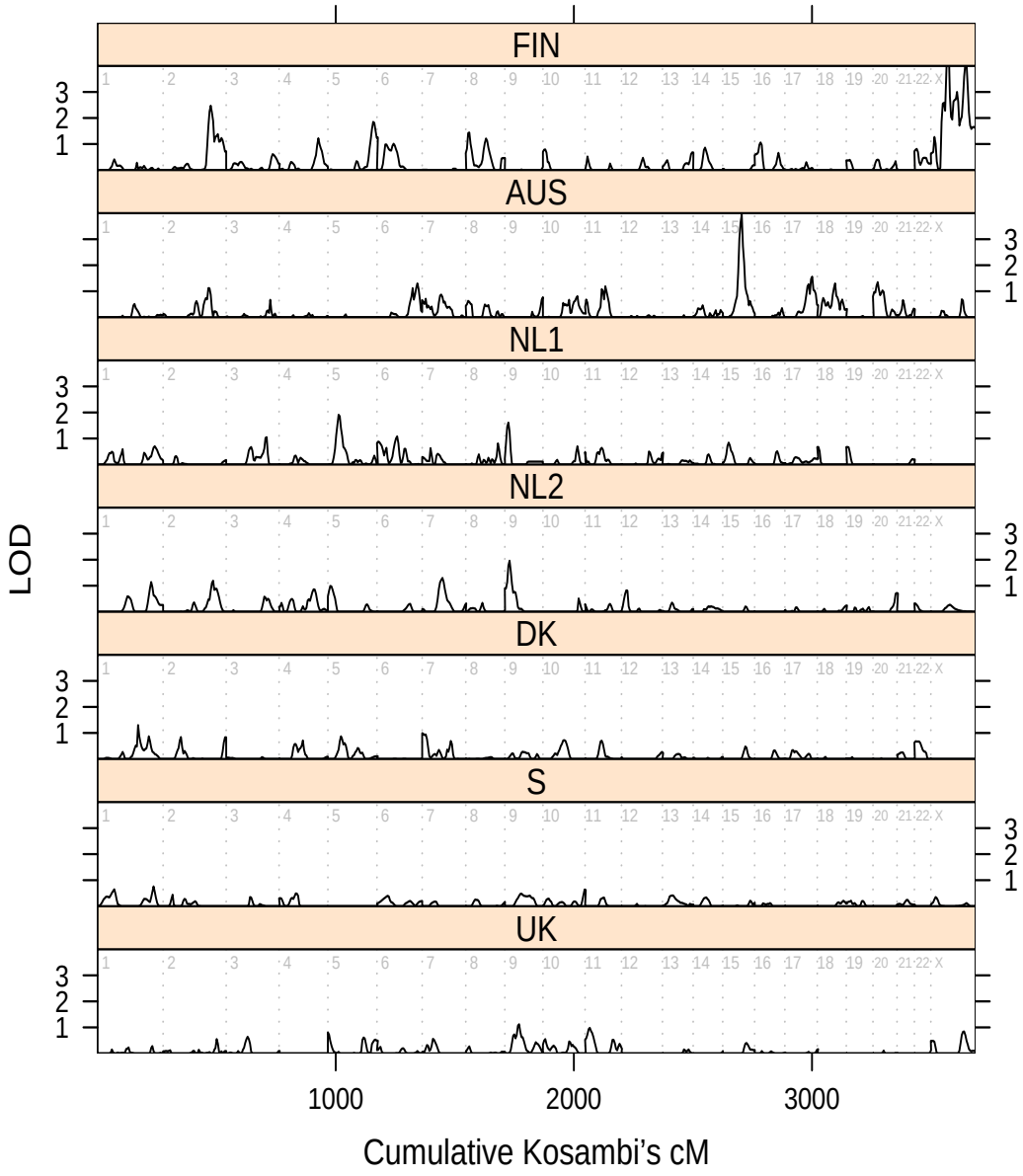

Figure 6.7: Genomewide Variance Components LOD scores for height - Individual studies

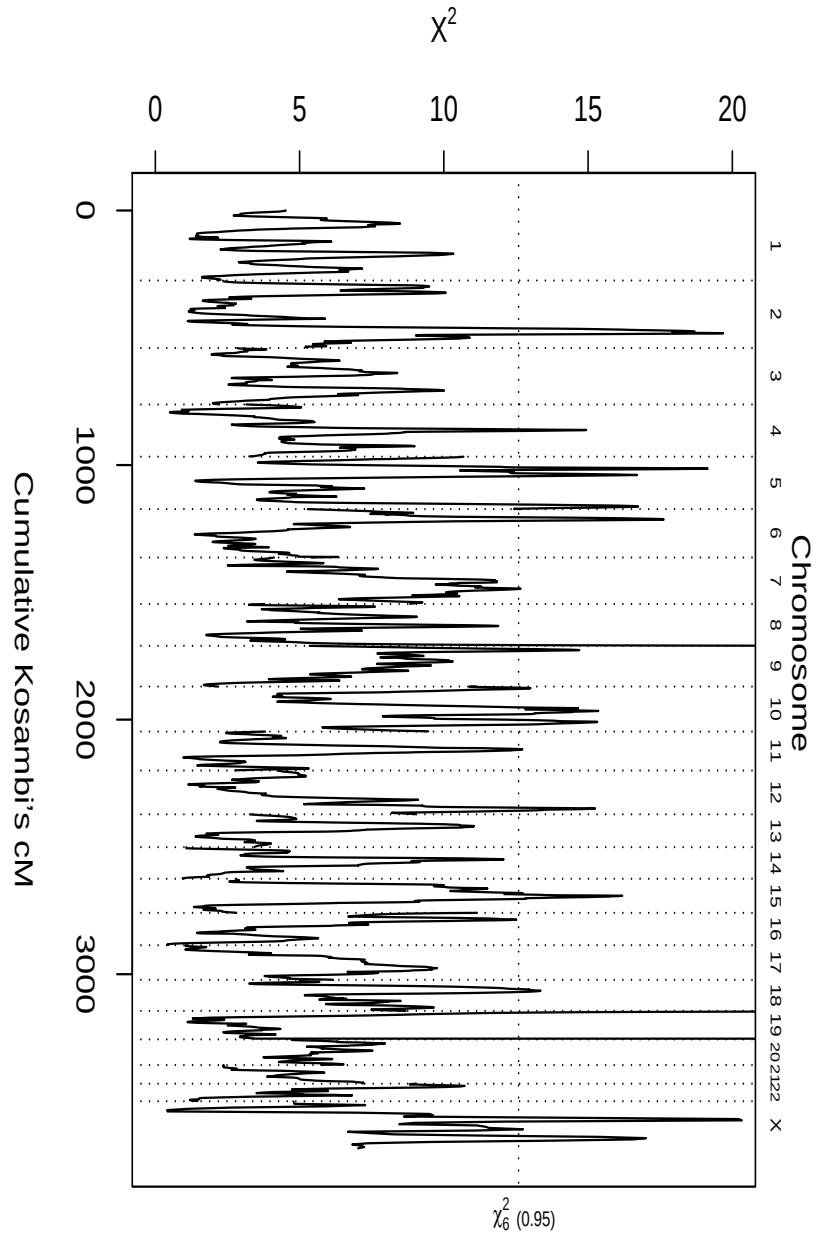

Figure 6.8: Genomewide χ^2 Test for heterogeneity in QTL analysis of height

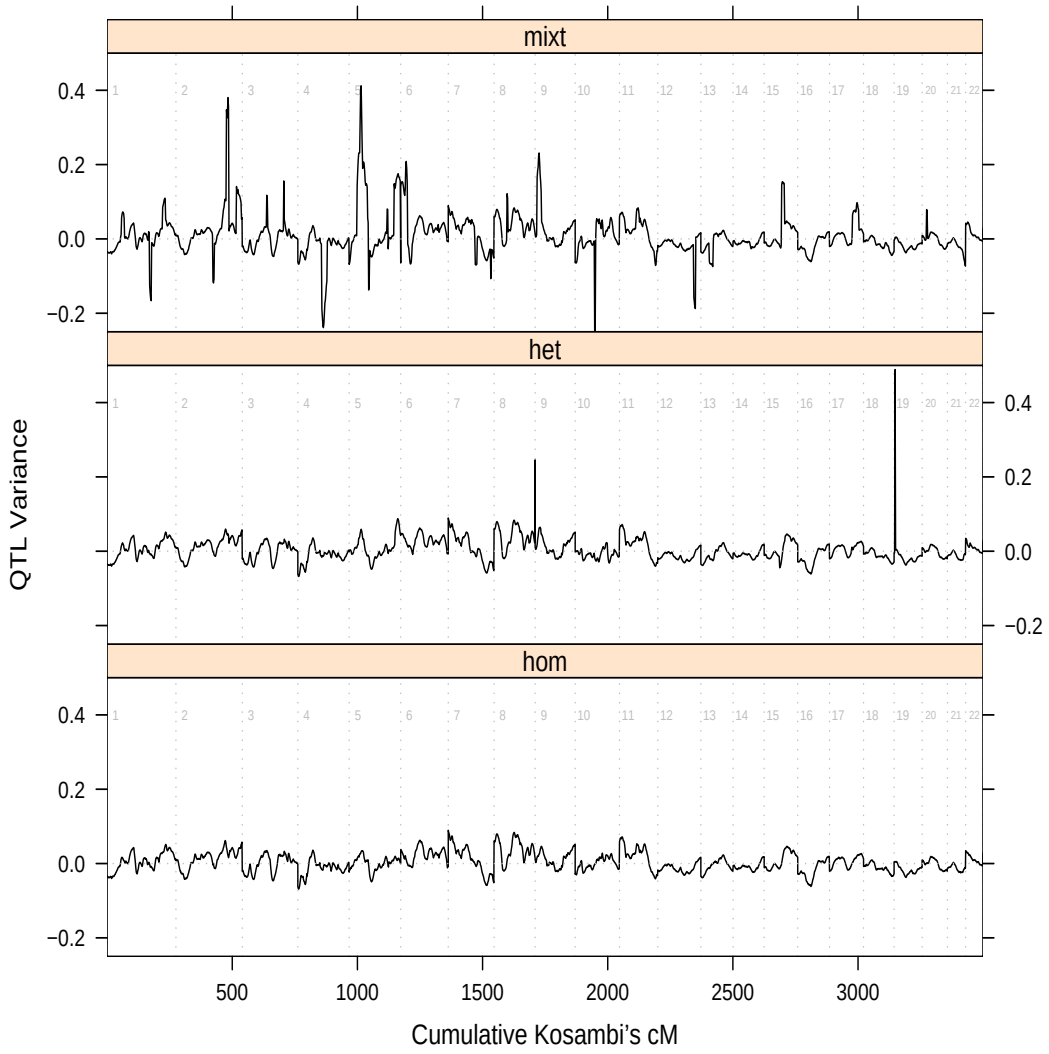

Figure 6.9: Genomewide Meta-analysis for height - QTL Variance Estimates for 2-point mixture model (top), heterogeneity model (middle) and homogeneous model (bottom)

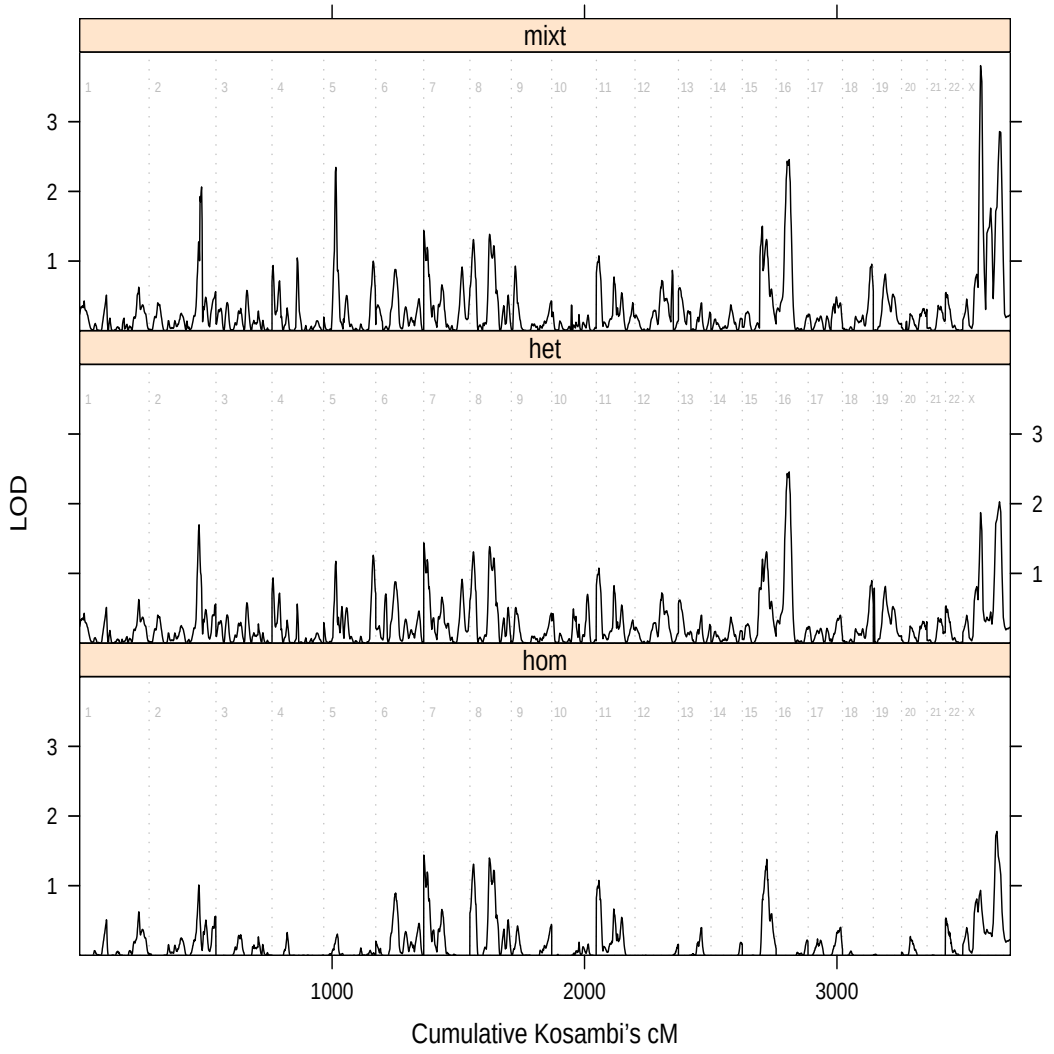

Figure 6.10: Genomewide Meta-analysis for height - LOD Scores for 2-point mixture model (top), heterogeneity model (middle) and homogeneous model (bottom)

Chapter 7

Score Test for Linkage in Generalized Linear Models

Abstract

We derive a test for linkage in a Generalized Linear Mixed Model (GLMM) framework which provides a natural adjustment for marginal covariate effects. The method boils down to the score test of a quasi-likelihood derived from the GLMM, it is computationally inexpensive and can be applied to arbitrary pedigrees. In particular, for binary traits, relative pairs of different nature (affected and discordant) and individuals with different covariate values can be naturally combined in a single test. The model introduced could explain a number of situations usually described as gene by covariate interaction phenomena, and offers substantial gains in efficiency compared to methods classically used in those instances.

7.1 Introduction

For binary traits, most linkage methods that allow for covariates focus on models where the identity-by-descent (IBD) probabilities are allowed to depend on those covariates (e.g. , Olson [1999]). This is often the most straightforward way to go because linkage studies for binary traits usually consist of families which have been selected based on their phenotypic values such as affected sib pairs (ASP) designs and effect of covariates at the population level cannot be estimated based on such data.

This chapter has been accepted for publication in *Human Heredity* as: J.J.P. Lebrec and H.C. van Houwelingen. Score Test for Linkage in Generalized Linear Models.

In many instances, however, some knowledge about the marginal effect of important covariates can often be gathered from either population-based studies or a literature review. Nevertheless, existing methods fail to integrate such external knowledge. An area where incorporation of covariates is a burning problem is late onset diseases, in fact, incorporation of population estimates of onset for the disease is not just a way to refine the analysis, it also allows inclusion of unaffected individuals. This can result in substantial gains in power, especially when traits are fairly common. In the case of continuous traits, the variance components model (and related regression methods) is widely accepted as the model of choice for testing for linkage with a putative locus. In this setting, the effect of important covariates is often modeled through a linear model while the covariance structure is left untouched. In contrast, the variance-covariance structure and the mean of binary and count data are intrinsically dependent and it is unclear how incorporation of covariates in the marginal probabilities impact linkage testing.

The Generalized Linear Mixed Models (GLMM) framework offers a natural and flexible extension of the variance components setting to categorical endpoints such as binary, count and survival data and accommodates covariate effects and arbitrary family structures. In accordance with the biometrical view of trait architecture [Fisher, 1918], small covariate effects contribute additively to the formation of a trait. Coupled with a variance components structure used to described the remaining correlation between relatives in a family, we obtain a parsimonious representation of the correlation between relatives. This unobserved latent process is linked to the actual trait values via a traditional Generalized Linear Model (see Section 7.2). In fact, this type of models have already been used for estimation of the heritability of binary traits [Burton et al., 1999; Houwing-Duistermaat et al., 2000; Noh et al., 2005] as well as for linkage of longitudinal continuous [Palmer et al., 2003] data and survival data [Scurrah et al., 2000]. Although appealing GLMMs are in general difficult to fit with family data. Besides we favor simple mathematically tractable expressions for a test, this is to reduce computational burden, but even more importantly, because we would like to get insight into the properties of this model when used in linkage studies. In stark contrast with the above cited approaches, we do not make any attempt to directly use the GLMM for inference but we resort to an approximation of the corresponding

likelihood (a quasi-likelihood). Indeed, our inference for linkage is based on a score test for the variance component corresponding to linkage in this quasi-likelihood (see Section 7.3). We assume that all segregation parameters in the GLMM have been obtained from external data and are therefore treated as nuisance parameters when testing for linkage. Estimation of such parameters in a GLMM is a notoriously difficult problem (at least for binary responses), we therefore propose an ad-hoc estimation procedure which appears to yield reasonable estimates in practice (see Section 7.4). Although the procedure does not always yield a unique set of parameters, we argue that our linkage test only weakly depends upon the parameters' choice and that its size is always preserved. The test is in fact a weighted regression of the deviation in IBD sharing on the trait values (in the same spirit as the pair-wise IBD scoring functions introduced by Whittemore and Halpern [1994] for affected relative pairs), which guarantees fast computations. Finally, in Section 7.5, we illustrate how the test could be used in linkage studies for two diseases: migraine and breast cancer. In those two examples we quantify the potential gains obtained compared to approaches that would either ignore covariates or estimate covariate effects from the linkage data only. In the discussion, we identify situations where covariate adjustment is likely to help improving the power of linkage studies.

7.2 Model

The generalized linear mixed model

Conditional on unobserved latent variables and observed covariate values, our model is specified by a generalized linear model (GLM). All information about the genetic relationship between individuals is incorporated in the latent variables just in the same way as in the variance components model for continuous traits. Formally, we consider the trait values $\mathbf{y} = (y_1, \ldots, y_m)$ of m relatives in a family whose values for k covariates are gathered in an $m \times k$ matrix $\mathbf{X}$. Conditional on a vector of random effects $\mathbf{b} = (b_1, \ldots, b_m)$ and a vector of covariate effects β, the y_i's are independently distributed according to a density function f from the canonical exponential family (to simplify notations, we have omitted the dispersion parameter), more precisely f

has the following form

$$\log f(y_i \,|\, \beta, b_i) = y_i \times (\mathbf{x}_i\beta + b_i) + a(y_i) - \psi(\mathbf{x}_i\beta + b_i)$$

where the first two derivatives of ψ determine the first and second moments of the GLM i.e. $\psi'(\mathbf{x}_i\beta + b_i) = \mathbf{E}(\mathbf{y}_i \,|\, \beta, b_i)$ and $\psi''(\mathbf{x}_i\beta + b_i) = \text{var}(\mathbf{y} \,|\, \beta, b_i)$. This type of models includes the logistic model for binary or binomial data, Poisson model for count data, continuous data (provided the dispersion parameter is known) as well as piecewise exponential hazards models for survival data [Agresti, 2002, pp.388-389]. The fixed effects β therefore model the effect of covariates while the dependence structure between relatives is entirely induced through the covariance of the random effects $\mathbf{b}$ which are assumed to follow a multivariate normal distribution with mean 0 and variance-covariance matrix $\mathbf{R}(\theta)$ where θ is the set of variance components. In the simple case of sibships the variance-covariance structure of $\mathbf{b}$ is described by a compound symmetry structure

$$\mathbf{R} = \mathbf{R}(\theta) = \sigma^2 \begin{pmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho \\ \rho & \dots & \rho & 1 \end{pmatrix} .$$

The exact marginal density $l(\beta, \theta)$ of the observations $\mathbf{y}$ is obtained by integration of the random effects $l(\beta, \theta) = \mathbf{E}_{\mathbf{b}}(\prod_{i=1,\dots,m} f(y_i \,|\, \beta, b_i))$ which entails calculation of a multivariate integral of potentially high dimension (for extended families).

GLMM for linkage

Our primary interest is on testing for linkage and we will therefore assume that all nuisance parameters i.e. the fixed covariate effects β and the marginal part of the covariance structure $\mathbf{R}(\theta)$ are known. We delay resolution of this problem to Section 7.4. We denote by γ the proportion of the random effects total variance σ^2 explained by the putative locus and focus our attention on this parameter by partitioning the set of variance components as (θ, γ). In analogy with the variance components model for continuous traits, we model linkage by specifying the conditional covariance structure $\mathbf{R} = \mathbf{R}(\theta, \gamma)$ of the random effects $\mathbf{b}$ given IBD information $\boldsymbol{\pi}$ within each family.

The $m \times m$ matrix $\boldsymbol{\pi}$ contains the identity-by-descent (IBD) information at a putative chromosomal position, more precisely $[\boldsymbol{\pi}]_{jk} = \pi_{jk}$ is the proportion of alleles shared IBD by pedigree members j and k and

$$[\mathbf{R}]_{jk} = \begin{cases} a^2 + c^2 = \sigma^2 \ , & \text{if } j = k \ , \\ (\pi_{jk} - \mathbf{E}\pi_{jk})\gamma\sigma^2 + (\mathbf{E}\pi_{jk})a^2 + c^2 \ , & \text{if } j \neq k \ . \end{cases}$$

where a^2 denotes the total additive genetic variance and c^2, the common-environment variance, on the underlying random effect scale.

7.3 Test for linkage

Quasi-likelihood for variance components

In an appendix, we show how the following quasi-likelihood for the data $\mathbf{y}$ can be obtained

$$\mathbf{y} \sim N\left(\boldsymbol{\psi\prime}(\mathbf{X}\beta) \ , \ \boldsymbol{\Psi}''(\mathbf{X}\beta) + \boldsymbol{\Psi}''(\mathbf{X}\beta).\mathbf{R}(\theta,\gamma).\boldsymbol{\Psi}''(\mathbf{X}\beta) \right) \ , \tag{7.1}$$

where $\boldsymbol{\psi}'(\mathbf{X}\beta)$ denotes the vector whose i^{th} element is given by $\psi'(\mathbf{x}_i\beta)$ and $\boldsymbol{\Psi}''(\mathbf{X}\beta)$ denotes the diagonal matrix whose i^{th} diagonal element is given by $\psi''(\mathbf{x}_i\beta)$. Note that this is not a normal approximation of the marginal likelihood, the normal shape is naturally obtained via a 2^{nd} order Taylor approximation of an exponential family likelihood in the canonical form. This quasi-likelihood can also be motivated by an approximate marginal model of the GLMM as in [Breslow and Clayton, 1993] and is the basis of the marginal quasi-likelihood (MQL) fitting algorithm. Another less crude approximation of the marginal likelihood could be based on a 1^{st} order Laplace approximation however this would render the approach mathematically intractable. Quasi-likelihood (7.1) is only accurate for small values of the random effects, hence small values of their variance σ^2; nonetheless, however accurate this approximation, the approach that we propose in Section 7.3 provides an 'unbiased' testing strategy.

Score test

For mathematical convenience, we use the quasi-likelihood for variance components introduced in Section 7.3 but expressed in terms of the first-order maximum-likelihood

estimates $\mathbf{z} = \frac{\mathbf{y}-\psi'(\mathbf{X}\beta)}{\psi''(\mathbf{X}\beta)}$ of the random effects $\mathbf{b}$. Denoting $\mathbf{\Sigma} = \mathbf{R}(\theta,\gamma) + \mathbf{\Psi}''^{-1}(\mathbf{X}\beta)$, this quasi-likelihood writes

$$\log ql(\mathbf{z},\gamma \,|\, \boldsymbol{\pi}) = -\frac{m}{2}\log(2\pi) - \frac{1}{2}\log(|\mathbf{\Sigma}|) - \frac{1}{2}\mathbf{z}'\mathbf{\Sigma}^{-1}\mathbf{z} \ .$$

We show in an appendix that the score function ℓ_γ for γ can then be written as

$$\ell_\gamma = \frac{1}{2}\ \mathrm{vec}(\mathbf{C})' \ .\ \mathrm{vec}(\boldsymbol{\pi} - \mathbf{E}\boldsymbol{\pi}) \tag{7.2}$$

with $\mathbf{C} = \mathbf{\Sigma}^{-1}\mathbf{z}\left(\mathbf{\Sigma}^{-1}\mathbf{z}\right)' - \mathbf{\Sigma}^{-1}$ and $\mathbf{\Sigma}$ taken in $\gamma = 0$. Here $\mathrm{vec}(\mathbf{C})$ places the n columns of the $m \times n$ matrix $\mathbf{C}$ into a vector of dimension $mn \times 1$, it contains weights for the pairwise IBD sharing $\mathrm{vec}(\boldsymbol{\pi} - \mathbf{E}\boldsymbol{\pi})$. Note that the $\boldsymbol{\pi} - \mathbf{E}\boldsymbol{\pi}$ matrix has all diagonal elements equal to 0. Our test for linkage is a weighted average of the different excess IBD sharing between all pairs of relatives in the pedigree. Linkage studies often include families which have been selected on the basis of their phenotypic values and it is sometimes unclear what the exact ascertainment scheme used is. A valid analysis of the data therefore requires that inference be carried out conditional on observed phenotypic values. Given the parametrization used above, accepting the quasi-likelihood $ql = ql(\mathbf{z} \,|\, \boldsymbol{\pi}, \gamma)$ as the model generating the "phenotypic data" $\mathbf{z}$ and relying on known nuisance parameters (β and θ), it turns out that the score function $\frac{\partial \log \mathbf{P}(\boldsymbol{\pi} \,|\, \mathbf{z},\gamma)}{\partial\gamma}$ evaluated at $\gamma = 0$ of the corresponding inverse likelihood of IBD sharing $\boldsymbol{\pi}$ conditional on transformed trait values $\mathbf{z}$ is simply equal to the same ℓ_γ function (see [Lebrec et al., 2004] for a proof). This justifies the use of this score statistic in selected samples. When the likelihood conditional on trait values is considered, the corresponding Fisher's information $\mathcal{I}_\gamma = \mathbf{E}\left(-\frac{\partial^2}{\partial\gamma^2}\log \mathbf{P}_\gamma(\boldsymbol{\pi} \,|\, \mathbf{z}, \gamma = 0)\right)$ for γ is also the variance of the score function $\mathrm{var}(\ell_\gamma \,|\, \mathbf{z}, \gamma = 0)$ and is thus given by

$$\mathcal{I}_\gamma = \frac{1}{4}\ \mathrm{vec}(\mathbf{C})' \ .\ \mathrm{var}\left(\mathrm{vec}(\boldsymbol{\pi}) \,|\, \gamma = 0\right) \ .\ \mathrm{vec}(\mathbf{C}) \ . \tag{7.3}$$

For a set of independent $p = 1, \ldots, P$ families with corresponding standardized trait values $\mathbf{z}_1, \ldots \mathbf{z}_P$, we therefore test for linkage using the statistic

$$T_+^2 = \begin{cases} 0 \ , & \text{if } \sum_{p=1}^P \ell_{\gamma,p} \leq 0 \\ \frac{\left(\sum_{p=1}^P \ell_{\gamma,p}\right)^2}{\sum_{p=1}^P \mathcal{I}_{\gamma,p}} \ , & \text{otherwise} \end{cases} \ ,$$

which is is asymptotically distributed as $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ under the null hypothesis (H_0) of no linkage. Indeed, the score conditional on trait values is unbiased since $\mathbf{E}(\ell_\gamma \,|\, \mathbf{z}, \gamma =$

$0) = 0$ (the term involving $\boldsymbol{\pi}$ in ℓ_γ is centered) and the standardization used (i.e. conditional on trait values $\mathbf{z}$) ensures that the test has variance 1 under H_0. Note that this would not necessarily be the case conditional on IBD sharing $\boldsymbol{\pi}$ (i.e. $\mathbf{E}(\ell_\gamma \,|\, \boldsymbol{\pi}, \gamma = 0) \neq 0$) because of model mis-specification.

Special case of relative pairs

Although the test derived previously applies to arbitrary pedigrees, the rest of the paper is devoted to relative pairs. In this instance, the variance-covariance matrix of random effects is

$$\mathbf{R} = \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} ,$$

for example, in the case of sib pairs, $\sigma^2 = a^2 + c^2$ and $\rho\sigma^2 = \frac{1}{2}a^2 + c^2$. If we denote $\psi'_i = \psi'(\mathbf{x}_i\beta)$, $\psi''_i = \psi''(\mathbf{x}_i\beta)$ and $\nu_i = (\sigma^2\psi''_i)^{-1}$, the score can be written in terms of the unstandardized centered trait values (or raw residuals) $y_i - \psi'_i$ as

$$\begin{aligned} \ell_\gamma = (\pi - \mathbf{E}\pi)\times \quad & \nu_1\nu_2 \ \left\{(1+\nu_1)(1+\nu_2) - \rho^2\right\}^{-2} \\ \times \big[\ & \left\{(1+\nu_1)(1+\nu_2) + \rho^2\right\}(y_1 - \psi'_1)(y_2 - \psi'_2) \\ & -\rho(1+\nu_2)(y_1 - \psi'_1)^2 - \rho(1+\nu_1)(y_2 - \psi'_2)^2 \\ & +\rho(\sigma^2\nu_1\nu_2)^{-1}\left\{(1+\nu_1)(1+\nu_2) - \rho^2\right\} \big] \ . \end{aligned}$$

If we let both ν_1 and ν_2 tend to $+\infty$, then the excess IBD sharing $\pi - \mathbf{E}\pi$ is simply weighted by the product of the raw residuals $(y_1 - \psi'_1)(y_2 - \psi'_2)$. This means that in the context of rare diseases and affected pairs (thus $y_1 = y_2 = 1$), the effect of covariates has to be very large for the weights to substantially differ from an unweighted strategy. Letting both ν_1 and ν_2 tend to 0, the weight then becomes $(1+\rho^2)z_1z_2 - \rho(z_1^2 + z_2^2) + \rho\sigma^2(1-\rho^2)$, where the z_i's are the first-order maximum-likelihood estimates of the random effects b_i's defined in Section 7.3. This expression is closely related to a version of the so-called Haseman-Elston regressions that is optimal with normally distributed data [Sham and Purcell, 2001], the main difference lies in the use of the variances ψ''_i in the standardization of the centered trait values $y_i - \psi'_i$ instead of the usual ${\psi''_i}^{1/2}$ as in Pearson residuals.

It is interesting to look at the special case of binary traits, where $a \equiv 0$ and $\psi(t) = \log(1 + e^t)$. In this instance, the weights associated to excess IBD sharing

$\pi - \mathbf{E}\pi$ are positive for ASP and unaffected sib pairs (USP) while they are negative for discordant sib pairs (DSP). Based on approximation (7.4) used in Section 7.4, ν_1 can be shown to be approximately related to the marginal correlation via $\nu_1 \approx \rho \, \mathrm{cor}(y_1, y_2)^{-1} \, \psi_2''^{1/2} \, \psi_1''^{-1/2}$ as long as σ^2 is not too large. This provides us with an order of magnitude for the ν_i parameters. For example, if the covariate values are the same for both individuals, ν is simply proportional to the inverse of the trait marginal correlation, which itself is an increasing function of both the prevalence and the recurrence risk ratio $\lambda_S = \mathbf{P}(\text{sib 1 is affected and sib 2 is affected})/\mathbf{P}(\text{sib 1 is affected})\,\mathbf{P}(\text{sib 2 is affected})$. For rare diseases, the ν_i parameters will likely be very large and weights given to the excess IBD sharing will be approximately equal to $(y_1 - \psi_1')(y_2 - \psi_2') \approx (y_1 - \mathbf{E}y_1)(y_2 - \mathbf{E}y_2)$ as pointed out in the previous paragraph. In this rare disease case, a direct application of the optimal Haseman-Elston regression for normally distributed data [Sham and Purcell, 2001] would lead to a weighting scheme approximately equal to the product of the Pearson residuals $(y_1 - \mathbf{E}y_1)/(\mathbf{E}y_1(1 - \mathbf{E}y_1))^{1/2} \times (y_2 - \mathbf{E}y_2)/(\mathbf{E}y_2(1 - \mathbf{E}y_2))^{1/2}$. Since the denominators $(\mathbf{E}y_i(1 - \mathbf{E}y_i))^{1/2}$ change rapidly as the trait becomes rare, the weight given to rare phenotypic values will be too extreme compared to those given to common trait values.

7.4 Estimation of segregation parameters

Estimation in GLMM has been the subject of intense research in the past decade and has proved notoriously difficult. Direct computation of the marginal likelihood can in principle be carried out by quadrature methods but are computationally burdensome, for that reason, approximate methods such as penalized quasi-likelihood (PQL) [Breslow and Clayton, 1993] have been proposed, unfortunately they are known to yield severely biased estimates, especially with binary endpoints. Another route is Bayesian fitting via Markov chain Monte Carlo algorithms. We refer the reader to `www.mlwin.com` for a list and review of possible softwares. Practical solutions appear to be problem-specific and a few authors have dealt with this problem in the case of family data [Burton et al., 1999; Houwing-Duistermaat et al., 2000; Noh et al., 2005]. Besides, in some instances (e.g. , when sib-pair data only are available), the GLMM may lack identifiability. We therefore propose the approximate

method described in Section 7.4. There is an extra difficulty in the case of binary data and we propose an ad-hoc solution which appears to yield sensible guesses of the nuisance covariance parameters θ and fixed effects β as far as the interest lies in testing for linkage: although the procedure of Section 7.4 does not give a unique choice of parameters, we argue that the actual linkage test is fairly insensitive to that specification.

General case

We first consider the case of a homogeneous population (i.e. no covariates) where three nuisance parameters need to be estimated, namely, the fixed effect β that reflects the overall level for the trait of interest, the variance σ^2 of the underlying random effect and the correlation ρ between the random effects in a pair of relatives. The marginal covariance relates to $\rho\sigma^2$ through the following approximate relation

$$\text{cov}(Y_1, Y_2) \approx \psi_1''(\beta)\psi_2''(\beta)\rho\sigma^2 \ , \tag{7.4}$$

and the marginal variance to β and σ^2 via

$$\text{var}(Y) \approx \psi''(\beta) + \psi''(\beta)^2\,\sigma^2 \ , \tag{7.5}$$

while the marginal mean can be either approximated as

$$\mathbf{E}(Y) \approx \psi'(\beta) + \frac{\sigma^2}{2}\psi'''(\beta) \ ,$$

or calculated exactly as $\mathbf{E}(\psi'(\beta + b))$ by univariate integration. Together, these three relations allow estimation of ρ, σ^2 and β.

In the case of a heterogeneous population, the simplest approach is to define relatively homogeneous strata and to apply the procedure described in the previous paragraph in each stratum separately. The series of ρ and σ^2 estimates are then averaged using the frequency of each stratum in the overall population as weight. Given those final estimates of ρ and σ^2, a second round of stratum-specific β values can then be computed.

Special case of Binary data

Relation (7.5) reflects over-dispersion in the marginal distribution i.e. the fact that the relation $\text{var}(Y) = \psi''(\beta)$ is violated, unfortunately, this does not apply to the

binary case where $\text{var}(Y) \equiv \mathbf{E}(Y)(1-\mathbf{E}(Y))$ and there can be no such thing as over-dispersion. We can still use relation (7.4) to estimate σ^2 for fixed values of ρ and the corresponding β by univariate integration of $\psi'(\beta+b)$ in each stratum. As in the general case, the values for σ^2 are averaged across strata and the stratum-specific fixed effects β are re-computed with the average σ^2 as input. This estimation procedure is therefore conditional on an arbitrarily chosen value for ρ.

For common diseases such as migraine (see Section 7.5), we can carry out a more formal procedure based on maximum likelihood. For binary traits, the data consists of stratum-specific 2×2 tables indexed by t. If we use the following notation for the cell numbers in a given 2×2 table t: n_{11}^{t} for affected-affected pairs, n_{10}^{t} for affected-unaffected , n_{01}^{t} for unaffected-affected and n_{00}^{t} for unaffected-unaffected and if $\hat{p}_{\cdot\cdot}^{\text{t}}(\sigma^2, \hat{\beta}(\sigma^2))$ denote the corresponding GLMM probabilities, then the log-likelihood of the data is given by

$$\sum_{\text{table t}} n_{11}^{\text{t}} \log \hat{p}_{11}^{\text{t}} + n_{10}^{\text{t}} \log \hat{p}_{10}^{\text{t}} + n_{01}^{\text{t}} \log \hat{p}_{01}^{\text{t}} + n_{00}^{\text{t}} \log \hat{p}_{00}^{\text{t}} \ .$$

If the trait is common, the GLMM probabilities $\hat{p}_{\cdot\cdot}^{\text{t}}(\sigma^2, \hat{\beta}(\sigma^2))$ can be calculated reasonably fast by Monte Carlo simulations and the maximization with respect to σ^2 is possible. Again, this maximization is carried out for a chosen ρ so this strategy offers a compromise between a full maximization of the marginal likelihood and the ad-hoc method of the previous paragraph.

Although the estimation approach described above is not optimal (in the sense that it is not guaranteed to yield maximum likelihood estimators), its merit is that it quickly provides sensible estimates of the nuisance parameters. The information available is often so sparse that the value of the likelihood depends very weakly (if at all) on the chosen value for ρ. In fact, as the next series of examples illustrates, the choice of ρ seems to have a limited impact on the test for linkage. In Table 1, we computed the relative weights of discordant pairs "AU" and unaffected pairs "UU" compared to affected pairs "AA" for three different values of the random effects' correlation ρ in a wide range of 2×2 tables (i.e. choices of prevalence K and recurrence risk ratios λ_S). In each scenario, we used approximation (7.4) to obtain estimates of the random effect total variance σ^2. As long as ρ is chosen not too small and that the recurrence ratio is not too large, the relative weights given to discordant

		σ^{2*}			AU			UU		
K	λ_S	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.8$
0.01	1.1	0.5	0.2	0.1	-0.01	-0.01	-0.01	0.00	0.00	0.00
0.01	1.2	1.0	0.4	0. 3	-0.01	-0.01	-0.01	0.00	0.00	0.00
0.01	1.5	2. 6	1.0	0. 6	0.00	-0.01	-0.01	0.00	0.00	0.00
0.01	2.0	5.1	2.0	1.3	0.00	-0.01	-0.01	0.00	0.00	0.00
0.01	3.0	10.2	4.1	2.6	0.00	0.00	-0.01	0.00	0.00	0.00
0.05	1.1	0. 6	0.2	0.1	-0.05	-0.05	-0.05	0.00	0.00	0.00
0.05	1.2	1. 1	0.4	0.3	-0.04	-0.05	-0.06	0.00	0.00	0.00
0.05	1.5	2. 8	1.1	0. 7	-0.03	-0.05	-0.06	0.00	0.00	0.00
0.05	2.0	5.5	2.2	1. 4	-0.02	-0.05	-0.06	0.00	0.00	0.00
0.05	3.0	11.1	4.4	2.8	-0.01	-0.0 3	-0.06	0.00	0.00	0.00
0.10	1.1	0.6	0.2	0. 2	-0.10	-0.11	-0.12	0.01	0.01	0.01
0.10	1.2	1. 2	0.5	0.3	-0.09	-0.11	-0.12	0.01	0.01	0.01
0.10	1.5	3.1	1. 2	0. 8	-0.0 6	-0.11	-0.13	0.00	0.01	0.01
0.10	2.0	6.2	2.5	1. 5	-0.0 4	-0.11	-0.14	0.00	0.01	0.01
0.10	3.0	12.3	4.9	3.1	-0.0 2	-0. 09	-0.15	0.00	0.00	0.01
0.20	1.1	0.8	0.3	0.2	-0.23	-0.26	-0.27	0.05	0.06	0.06
0.20	1.2	1. 6	0.6	0.4	-0.21	-0.26	-0.28	0.04	0.05	0.06
0.20	1.5	3. 9	1. 6	1.0	-0.1 7	-0.2 8	-0.32	0.02	0.05	0.06
0.20	2.0	7.8	3.1	2.0	-0.1 3	-0.2 9	-0.3 8	0.01	0.04	0.06
0.20	3.0	15.6	6 .2	3. 9	-0. 09	-0.2 8	-0.4 5	0.00	0.03	0.06
0.30	1.1	1.0	0.4	0.3	-0.40	-0.45	-0.47	0.14	0.17	0.18
0.30	1.2	2.0	0.8	0.5	-0.38	-0.47	-0.50	0.12	0.17	0.18
0.30	1.5	5.1	2.0	1. 3	-0.33	-0.51	-0. 60	0.09	0.16	0. 20
0.30	2.0	10.2	4. 1	2. 6	-0.27	-0.54	-0.72	0.06	0.16	0.22
0.30	3.0	20.4	8.2	5.1	-0.2 2	-0.5 6	-0.9 0	0.0 3	0.1 5	0.2 6

Table 7.1: Relative weights for Discordant (AU) and unaffected (UU) pairs (compared to affected pairs) for a range of 2×2 tables - * σ^2 obtained using approximation (7.4)

pairs and to a lesser extent, to unaffected pairs depend only weakly upon the initial choice for ρ, although the dependence becomes stronger as the prevalence of the trait increases. When comparing the relative weights of affected pairs for different prevalences/recurrence risk ratios, the dependence is even less noticeable (data not shown). Based on this study, we would advise the choice of a moderate to large value for ρ (0.5 to 0.8) since we favor the corresponding small values for σ^2 (indeed, the quasi-likelihood is based on an approximation valid for small values of σ^2 and so is relation (7.4) used for estimating σ^2).

7.5 Examples

Application to a common disease: Migraine

Migraine is known to be much more frequent in women than in men. In this section, we describe how sex could be accounted for in a linkage study for migraine and quantify the potential gains/losses incurred under different strategies including the

	U m	A m	U f	A f
U m	0.06	-0.60	0.11	-0.33
A m	.	2.71	-1.12	1.57
U f	.	.	0.25	-0.63
A f	.	.	.	1.00

Table 7.2: Relative weights C_i for all sex-sex (f:female and m:male) sib pair combinations (A: Affected and U: Unaffected)

score test presented in Section 7.3. Based on sex-specific prevalence and recurrence risk estimates derived from published data in the Dutch population [Mulder et al., 2003], we first obtain estimates of the segregation parameters ρ, σ^2 and β using the procedure described in Section 7.4. Using possible values of excess IBD sharing, we then quantify the gain obtained by accounting for sex with the score test described above. Mulder et al. [2003] fitted a liability threshold model (i.e. with sex-specific thresholds and a common tetrachoric correlation) to the data. The sex of siblings in a pair defines three possible strata or 2×2 tables, we focused on the Dutch population in the age group 36-68 years old and used the model parameters' estimates to reconstruct those three tables. For the Dutch population, the prevalence for migraine was approximately 0.34 in women and 0.17 in men and the values for λ_S were 1.31, 1.45 and 1.65 in female-female, male-female and male-male sib pairs respectively. Assuming that the three corresponding 2×2 tables were present in proportions $\frac{1}{4}$, $\frac{1}{2}$ and $\frac{1}{4}$ in the overall population, we estimated σ^2 as $\hat{\sigma}^2 = 3.3$ and $\hat{\beta} = (-2.40, -1.03)$ for $\rho = 0.5$ according to the formal maximum-likelihood based method described in Section 7.4. Based on this set of nuisance parameter estimates we calculated the weights for all possible types of sib pairs in the linkage test, these are displayed in table 7.2.

Note, first of all, that affected (and unaffected) sib pairs have positive weights while discordant sib pairs have negative weights. Male-male affected pairs are given much more weight than female-female affected pairs, while the trend is opposite for discordant pairs. One interesting feature is that male-female affected-unaffected pairs are given much more weight than female-male affected-unaffected pairs since the phe-

notypic discordance is more likely to be due to genetic factors in the former than in the latter.

We now compare four possible strategies when testing for linkage in presence of covariates. We define homogeneous groups (indexed by g) of relative pairs (i.e. families) depending on their phenotypic values (AA, AU or UU) and (categorical) covariate values. The excess or reduction in IBD sharing in each group can be parameterized as $\mathbf{E}(\pi - \mathbf{E}\pi \mid \text{ group } g) = \theta\delta_g$ where δ_g can be positive or negative while $\theta \geq 0$. A test for linkage corresponds to testing $\theta = 0$ versus $\theta > 0$. In all tests outlined below, we assume that the sign of δ_g is known (+ for AA and UU and − for AU pairs), depending on what we know or assume about the $|\delta_g|$'s, four testing strategies can be derived:

1. All $|\delta_g|$'s are taken as being equal,
2. The ratios of the $|\delta_g|$'s are known, this is an ideal situation that will serve as reference in our comparison,
3. The $|\delta_g|$'s are estimated from the data,
4. The ratios of the $|\delta_g|$'s are assumed to be given by the score test of Section 7.3.

All four tests but 3. are asymptotically distributed as $\frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1$ under the null hypothesis of no linkage. For test 3., a penalty has to be paid for estimating the weights and the corresponding null distribution is $\frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_G$ where G is the total number of homogeneous groups considered.

To keep things simple in our numerical comparison of the tests when applied to migraine data, we focused on designs with only sib pairs and two groups ($G = 2$). We compared the efficiency of tests 1., 3. and 4. relative to reference test 2. . To do so, we computed the non-centrality parameters (NCP) for the equivalent χ^2 linkage tests. If C_g denotes the assumed values for the true relative excess IBD sharing δ_g, then all tests but 3. are based upon the following statistic T

$$T = \frac{\sum_g \sum_{i \in \mathrm{g}} C_g(\pi_i - \frac{1}{2})}{(\mathrm{var}(\pi) \times \sum_g N_g C_g^2)^{1/2}} \ ,$$

where N_g denotes the number of families in group g and $N = \sum_g N_g$. For complex traits and thus small gene effect, the variance of π under the alternative hypothesis

is close to its value under the null $\mathrm{var}(\pi \,|\, \text{group g}) \simeq \mathrm{var}(\pi)$ so we have the following approximation:

$$\mathbf{E}(T^2) \simeq 1 + N \times \frac{\left(\sum_g f_g C_g (\mathbf{E}(\pi_g) - \frac{1}{2})\right)^2}{\mathrm{var}(\pi) \times \sum_g f_g C_g^2} \ , \text{ where } f_g = \frac{N_g}{N} \ ,$$

and the sample size for the corresponding 1 d.f. test is inversely proportional to the non-centrality parameter in the previous expression. Asymptotically, the estimates for the weights in test 3. should be very close to their true values, the relative loss of efficiency in test 3. relative to test 2. (where true weights are assumed to be known) is therefore only due to the additional degrees of freedom (d.f.=2 here) of the test. In the context of scan for linkage, using a conservative point-wise type I error rate of 10^{-4}, this loss amounts to about 20%. In the sequel, relative efficiency is expressed as the ratio of sample size in test 2. to sample size in the test of interest.

Using the GLMM described in Section 7.2 (with $\rho = 0.5$, $\sigma^2 = 3.3$ and $\hat{\beta} = (-2.40, -1.03)$ as previously estimated), we mimicked a situation where 10% of the total variance of the random effect is explained by the putative locus while the rest of the variance is either explained by common environment or other unlinked loci [1]. Using Monte Carlo simulations, we closely approximated the average IBD sharing for three types of sib pairs, namely AA male-male, AA female-female and discordant sib pairs AU female-male. In figure 7.1, we display the relative efficiency of the previously defined tests 1., 3. and 4. relative to 2. for two types of study designs: one mixing AA male-male and AA female-female (left-hand side, scenario 1) and one mixing AA male-male and AU female-male (right-hand side, scenario 2). In scenario 1, the 2 degrees of freedom test (test 3.) always fails in improving efficiency compared to a 1 d.f. test with no weight (test 1.) while the score test based on the quasi-likelihood of the GLMM (test 4.) almost always yields improved efficiency with gains close to an ideal strategy (test 2.). In scenario 2, the 2 degrees of freedom test does yield gains in efficiency compared to test 1. that ignores covariates (note that this test can incur efficiency loss up to almost 40% in this situation) when the mixing proportions of AmAm and AfUm are not too extreme, however our test 4. does uniformly better than any of these two tests with losses in efficiency no larger than approximately 10%.

[1] Note that for other values of the proportion of total random effect variance γ explained by the putative locus, the same relative efficiency results hold approximately as long as γ is not too large

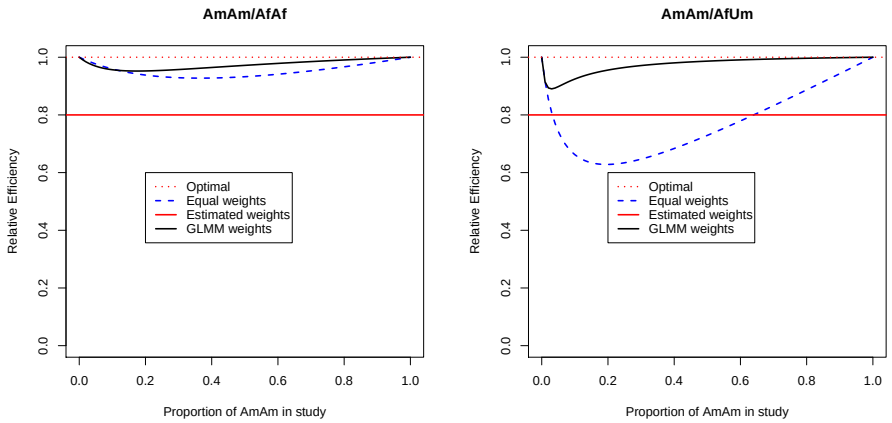

Figure 7.1: Relative efficiency in migraine example - Left: $\mathbf{E}(\pi_1 - \frac{1}{2}) = 0.0033$ in AmAm and $\mathbf{E}(\pi_2 - \frac{1}{2}) = 0.0019$ in AfAf and Right: $\mathbf{E}(\pi_1 - \frac{1}{2}) = 0.0033$ in AmAm and $\mathbf{E}(\pi_2 - \frac{1}{2}) = -0.0008$ in AfUm.

Application to breast cancer

We put ourselves in a situation where ASP's for breast cancer status have been gathered among sib pairs of all ages classified in eight classes (see Table 7.3). The disease status is positive if a woman currently has or has had breast cancer during her life time. For simplicity, we assume that both siblings belong to the same age class. The question is how to weight the excess IBD sharing in each age class effectively.

The genetics of breast cancer is often described using Claus model [Claus et al., 1991] which we will use as the basis for estimation of segregation parameters. Claus model is based on a one-locus model with a rare autosomal dominant allele (q=0.0033) leading to an increased risk of breast cancer. The cumulative probability of a woman to be affected is a function of a woman's age (see Table 2 in [Claus et al., 1991]), based on this model, we derived the prevalence and the recurrence risk ratio (λ_S) for each age class, thereby closely reproducing observed values. Following the informal approach described in Section 7.4, we estimated the variance of the random effects σ^2 in each age-specific 2×2 table based on a correlation equal to $\rho = 0.5$ and used the average value across tables $\hat{\sigma}^2 = 1.96$ (and corresponding age-specific fixed effects).

Age (Years)		Based on Claus model	Based on fitted GLMM	
	K(%)	λ_S	λ_S	Test relative weights
20-29	0.03	10.34	8.	1.70
30-39	0.36	5.97	2.32	1.38
40-49	1.62	2.64	2.26	1.21
50-59	3.09	1.93	2.04	1.11
60-69	5.38	1.44	1.83	1.05
70-79	8.55	1.34	1.70	1.01
80+	13.12	1.15	1.56	1.00

Table 7.3: Prevalence, λ_S in Claus and GLM models, stratum-specific GLMM weights

The series of λ_S's that this GLMM yields is displayed in Table 7.3, it is flatter than the observed ones because the GLMM is stretched to its maximum capacity in order to cover such a wide λ_S-range.

The relative weights for ASP of each age category are given in the last column of Table 7.3, they are fairly mild compared to the large differences observed in λ_S. An approach that would use time of onset rather than current status data is likely to be more efficient, however it is conceptually more complicated. As for migraine, we limited our quantitative comparison to ASP designs with data consisting of two groups: we chose the two most extreme age categories with a relative weight of 1.70. We closely approximated excess IBD sharing in the two age categories in the same way as for the previous example i.e. by mimicking a model where the putative locus explained 10% of the total variance of the random effect while the rest of the variance can be conceived as arising either from a common environment or other unlinked loci [2] under the fitted GLMM. Under this model, our approximate score test 4. is the one closest to the ideal test 2. ; test 3. sometimes performs better than test 1. however this advantage would disappear if data consisted (more realistically) of sib pairs in all age categories (see Fig. 7.2).

[2] but note that the same remark regarding relative efficiency holds as for the migraine example

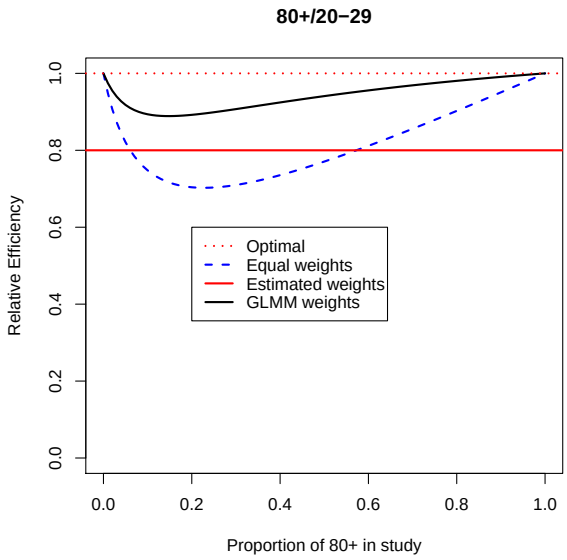

Figure 7.2: Relative efficiency in breast cancer - $\mathbf{E}(\pi_1 - \frac{1}{2}) = 0.017$ and $\mathbf{E}(\pi_2 - \frac{1}{2}) = 0.005$ in "20-29" and "80+", resp.

7.6 Discussion

Based on the GLMM, we have derived a test for linkage which makes adjustment for known marginal covariate effects. Our approach is motivated by the fact that the effect of important covariates on the marginal distribution of a trait is often known via data external to the linkage study itself, and these should be incorporated in the linkage analysis. We elude the difficult and computationally intensive problem of making exact inference based on the likelihood of the GLMM by using a quasi-likelihood, our test is then based upon a score test for the linkage parameter in this quasi-likelihood and turns out to be a tractable statistic, in fact, a simple weighted average of the excess IBD sharing between all pairs of relatives in a family. In that respect, it is reminiscent of approximate likelihoods based on pairwise joint distributions used, for example, with correlated binary data [le Cessie and van Houwelingen, 1994]. As noted by Cox and Reid [2004], the use of such pseudo-likelihoods does not only alleviate the computational burden, it also enhances the robustness of the method to model specification. It must be recognized, however, that in absence of covariates, better family-specific tests that take the full IBD distribution into account can be derived [Teng and Siegmund, 1997]. If the GLMM correctly describes the data, we can draw two general conclusions about the effect of covariate adjustment in linkage studies for binary traits. For rare traits where only affected pairs of individuals are informative, the effect of covariates needs to be huge in order for any covariate-adjustment to yield substantial power gains. Indeed, the excess IBD sharing differs only a little between covariate-specific types of affected pairs. For common traits, the gains are more easily achieved. Firstly, because discordant pairs can be more confidently included in the analysis if relevant covariates (e.g. age and sex) are taken into account, and those pairs do become informative in common traits. Secondly, because the ratios of deviations in IBD sharing between phenotypic-covariate specific strata are more likely to be large for such traits.

The test is applicable in arbitrary pedigrees, and in the case of binary traits, it allows incorporation of both affected and unaffected individuals. This way of handling the issue of covariates in binary traits, contrasts with existing methods that only use the linkage data available and model the probability of IBD sharing as a

function of covariates. The most general representative of this type of models (i.e. which in principle can handle arbitrary pedigrees and both affected and unaffected individuals) is undoubtedly the conditional logistic model [Olson, 1999; Greenwood and Bull, 1999]. It is implemented in the `LODPAL` program of the `S.A.G.E.` software but as far as we are aware (true for version 5.1), the current implementation suffers from a few important limitations: the program assumes that all pairs of relatives are independent, the covariates have to be pair-specific, when both affected and discordant pairs are analyzed together, the program cannot handle covariates. These issues do not arise in our approach. The strength of methods that let IBD sharing depend upon covariate values invariably turns into a weakness (unless differences between covariate-specific groups are very large) as the number of covariates increases because the d.f. of the corresponding test for linkage increases too. We overcome this problem by incorporating external data and by specifying a model where differences in IBD sharing naturally arise. The way we handle covariates by feeding some covariate-adjusted residuals into the linkage analysis is conceptually similar to the method advocated for sibships by Alcais [2001]. For general pedigrees however, as far as we are aware, our test actually appears to be the only available practical way to simultaneously adjust for covariates and to include both affected and unaffected individuals. In late onset diseases, the suspicion that younger unaffected individuals might become affected at a later age can explicitly be incorporated using age as a covariate. We have treated all segregation parameters required by the GLMM as known parameters and although unbiased estimates could be difficult to obtain, we propose an estimation procedure that circumvents this problem. As long as interest lies in testing for linkage and not in actually estimating segregation parameters, this procedure appears to be acceptable in that: 1) it does not affect the size of our test 2) the test itself is fairly insensitive to the non-unique choices of nuisance parameter values. By illustrating the use of our method in both common and relatively rare diseases, we have shown the order of magnitude for the gains that could be expected in some specific scenarios. We note that the GLMM model does not explicitly incorporate potential gene by covariate interaction in its structure, this is not to say that it forbids this phenomenon, indeed, the recurrence risk ratios and IBD sharing induced by the model clearly vary depending on covariate values. However, purely

for mathematical convenience, we have assumed that on the latent scale, there was no interaction between the gene at the putative location and the covariate. Actually, recent developments published by Peng et al. [2005] explicitly account for such interactions and these authors have derived the corresponding score test for linkage. The gene by covariate interaction could be explicitly incorporated into the GLMM model in a similar way (via the $\mathbf{R}$ matrix of variance-covariance of random effects) and the corresponding test would obtain analogously. We note that in practice the IBD status is not known exactly but has to be estimated from marker data, the consequence for the score test is that π has to be replaced by its estimated version $\hat{\pi}$ in equation (7.2) and that the corresponding var$(\hat{\pi})$ has to be used in the standardization of the test. This last term depends on the family structure, the marker allele frequencies, their position and the possible genotype missingness pattern, and in practice we approximate its true value using Monte Carlo simulations as implemented in an executable C program calling upon the `MERLIN` [Abecasis et al., 2002] software and available at `http://www.msbi.nl/Genetics/`. Currently, the GLMM test prescribed in this manuscript is only available as R code from the authors. Finally, we remark that although we have focused on the use of our test with binary traits, the approach can directly be applied to other traits whose distribution is in the canonical exponential family, in particular to count data with a Poisson distribution as well as survival data.

7.7 Appendix

Derivation of the quasi-likelihood

We use a 2^{nd} order Taylor approximation of the conditional log-likelihood $\log f(y \mid \beta, b)$ introduced in Section 7.2 around $\mathbf{b} = 0$ to obtain a quasi-likelihood for the data $\mathbf{y}$ in

a family:

$$\begin{aligned}
\log f(\mathbf{y} \mid \beta, \mathbf{b}) &= \sum_{i=1}^{m} \log f(y_i \mid \beta, b_i) \\
&\simeq \sum_{i=1}^{m} \log f(y_i \mid \beta, b_i = 0) + b_i(y_i - \psi\prime(\mathbf{x}_i\beta)) - \frac{1}{2} b_i^2 \psi\prime\prime(\mathbf{x}_i\beta) \\
&\simeq \sum_{i=1}^{m} \log f(y_i \mid \beta, b_i = 0) - \frac{1}{2}\psi\prime\prime(\mathbf{x}_i\beta)\left(b_i - \frac{y_i - \psi\prime(\mathbf{x}_i\beta)}{\psi\prime\prime(\mathbf{x}_i\beta)}\right)^2 \\
&\quad + \frac{1}{2}\psi\prime\prime(\mathbf{x}_i\beta)\left(\frac{y_i - \psi\prime(\mathbf{x}_i\beta)}{\psi\prime\prime(\mathbf{x}_i\beta)}\right)^2 .
\end{aligned}$$

In the previous expression, only the second term involves $\mathbf{b}$ which shows that when β is regarded as constant, $\log f(\mathbf{y} \mid \beta, \mathbf{b})$ behaves as if

$$\frac{\mathbf{y} - \boldsymbol{\psi}\prime(\mathbf{X}\beta)}{\boldsymbol{\psi}''(\mathbf{X}\beta)} \mid \mathbf{b} \sim N(\mathbf{b}, \boldsymbol{\Psi}''(\mathbf{X}\beta)^{-1})$$

where $\boldsymbol{\Psi}\prime\prime(\mathbf{X}\beta)$ denotes the diagonal matrix whose i^{th} diagonal element is given by $\psi\prime\prime(\mathbf{x}_i\beta)$. We can now easily integrate the random effects $\mathbf{b} \sim N(0, \mathbf{R}(\theta, \gamma))$ out and $\log f(\mathbf{y} \mid \beta)$ as a function of θ can be regarded as the value of the density for multivariate normal $N(0, \mathbf{R}(\theta, \gamma) + \psi\prime\prime(\mathbf{X}\beta)^{-1})$ in the data points $\frac{\mathbf{y} - \boldsymbol{\psi}\prime(\mathbf{X}\beta)}{\boldsymbol{\psi}\prime\prime(\mathbf{X}\beta)}$:

$$\frac{\mathbf{y} - \boldsymbol{\psi}\prime(\mathbf{X}\beta)}{\boldsymbol{\psi}''(\mathbf{X}\beta)} \sim N(0, \mathbf{R}(\theta, \gamma) + \boldsymbol{\Psi}''(\mathbf{X}\beta)^{-1}) .$$

Score test

In analogy with the case of normally distributed phenotypes [Lebrec et al., 2004], standard results on matrix algebra (see, e.g. [Searle et al., 1992, Appendix M.7]) lead to

$$\ell_\gamma^{\mathbf{z}} = \frac{\partial \log ql}{\partial \gamma} = \frac{1}{2}\left\{\mathbf{z}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\pi} - \mathbf{E}\boldsymbol{\pi})\boldsymbol{\Sigma}^{-1}\mathbf{z} - \mathrm{tr}(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\pi} - \mathbf{E}\boldsymbol{\pi}))\right\}$$

Because of the relation $a'b = \mathrm{tr}(ba')$, the previous equation can be rewritten

$$\frac{\partial \log ql}{\partial \gamma} = \frac{1}{2}\ \mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\pi} - \mathbf{E}\boldsymbol{\pi})(\boldsymbol{\Sigma}^{-1}\mathbf{z}\mathbf{z}' - \mathbf{I})\right) .$$

Here $\mathrm{tr}(A)$ stands for the trace (sum of the diagonal elements) of matrix A. Using elementary matrix theory, in particular $\mathrm{tr}(AB) = \mathrm{tr}(BA)$ and $\mathrm{tr}(AB) = \mathrm{vec}(A')'\mathrm{vec}(B)$ (here $\mathrm{vec}(A)$ places the n columns of the $m \times n$ matrix A into a vector of dimension

$mn \times 1$), this score function can be rewritten as

$$\ell_\gamma^{\mathbf{z}} = \frac{1}{2} \text{ vec}(\mathbf{C})' \text{ . vec}(\boldsymbol{\pi} - \mathbf{E}\boldsymbol{\pi})$$

with $\mathbf{C} = \boldsymbol{\Sigma}^{-1}\mathbf{z}\left(\boldsymbol{\Sigma}^{-1}\mathbf{z}\right)' - \boldsymbol{\Sigma}^{-1}$.

Approximation used in segregation parameters estimation

The marginal covariance can be partitioned as

$$\begin{aligned} \text{cov}(Y_1, Y_2) &= \mathbf{E}\left(\text{cov}(Y_1, Y_2 \mid \beta_1, \beta_2, b_1, b_2)\right) + \text{cov}\left(\mathbf{E}(Y_1 \mid \beta_1, b_1), \mathbf{E}(Y_2 \mid \beta_2, b_2)\right) \\ &\approx 0 + \text{cov}\left(\psi'(\beta_1) + b_1\psi''(\beta_1),\ \psi'(\beta_2) + b_2\psi''(\beta_2)\right) \ , \end{aligned}$$

using a 1^{st} order Taylor expansion of $\psi'(\beta_i + b_i)$. It follows that $\text{cov}(Y_1, Y_2) \approx \psi''(\beta_1)\ \psi''(\beta_2)\ \rho\sigma^2$. The approximation $\text{var}(Y) \approx \psi''(\beta) + \psi''(\beta)^2\sigma^2$ obtains in the same manner by setting $\rho = 1$ and taking a 1^{st} order Taylor approximation of $\text{var}(Y \mid \beta, b) = \psi''(\beta + b) \approx \psi''(\beta) + b\ \psi'''(\beta)$.

For the marginal mean, we have

$$\begin{aligned} \mathbf{E}(Y) &= \mathbf{E}\left(\mathbf{E}(Y \mid \beta, b)\right) \\ &\approx \mathbf{E}\left(\psi'(\beta) + b\psi''(\beta) + \frac{b^2}{2}\psi'''(\beta)\right) \\ &\approx \psi'(\beta) + \frac{\sigma^2}{2}\psi'''(\beta) \ . \end{aligned}$$

Chapter 8

Conclusion

Searching for genes responsible for complex traits is proving extremely challenging, and this drawback is an incredible incentive for research in statistical methodology. Even in the relatively ancient field of linkage mapping, researchers have not yet exhausted the possibilities for methodological improvements. This thesis presents some statistical methods aimed at refining the design and analysis of linkage studies.

The score test developed in chapter 2 and the associated selective genotyping procedures of chapter 3 provide a strategy for better use of resources and valid testing in such selective designs for arbitrary pedigrees. Our test is almost identical to that of Sham et al. [2002] who motivated it in terms of regression. The fact that it is a score test of the variance components model gives a sound theoretical justification for its use. It also makes interesting refinements more obvious, for example, different common environments may be accommodated for different types of paired relatives. The software implementation of the test Sham et al. [2002] in `MERLIN-regress` suffers one important drawback due to the way the covariance matrix of the test under the null hypothesis is approximated. Unfortunately, there is no fast general solution for a correct approximation of this covariance, the solution that we have implemented in a C program calling upon `MERLIN` for IBD computations is based on Monte-Carlo simulations. The program will be useful for all linkage tests based upon IBD sharing and its use is therefore not limited to continuous traits. Linkage studies involving only one type of selected families such ASP designs rely too heavily on ideal situations unlikely to be true in practice such as absence of genotyping errors or strict adherence to law of segregation. The genomic control strategy proposed in chapter 4 offers the promise of a more robust inference. The pooling of existing linkage studies is essential in order to reach a critical sample size, the meta-analytic techniques of

chapter 6 can easily be applied once the important effect of partial marker information has been understood (chapter 5). The problem of heterogeneity may be alleviated by incorporation of important covariate information into linkage studies, chapter 7 offers a simple and general way to do so.

The software implementation of the methods developed in preceding chapters are available at `http://www.msbi.nl/Genetics/` and include:

- Approximation of the covariance of IBD sharing by Monte Carlo simulations (C program),
- Score test for quantitative traits (chapter 2) in arbitrary pedigrees (C program),
- Meta-analytic models (chapter 6) and data-reading tools (R-code).

The issue of statistical significance has been mostly overlooked in this thesis. One may argue that this is not really a crucial issue in the linkage mapping of complex traits where power is much more problematic. Indeed, even in the case of a highly heritable trait such as height, the meta-analysis of chapter 6 which gathered data equivalent to more than 4300 sib pairs failed to provide any consistent evidence for linkage. In the light of the sample size calculations of chapter 3 and given the effect sizes actually observed (i.e. QTL effects between 5 and 10%), this result appears less surprising: an unselected design, under perfect model specification, requires at least 7500 sib pairs (and more realistically 30000) in order to have a decent chance to detect such effects. Until we can genotype such large numbers of individuals routinely, selective designs offer an attractive sampling scheme. Geneticists are sometimes reluctant from using such designs because they fear that the genes involved in the formation of extreme phenotypes might be different from those contributing to the phenotype in a more standard range. This is a legitimate concern but it is not always recognized that this criticism equally applies to unselected designs. Indeed, most of the linkage information in random samples comes from extreme families.

The issue of heterogeneity is ubiquitous in linkage studies with thousands of families possibly arising from different populations. The methods presented in chapter 6 where heterogeneity between different linkage studies is explicitly modelled can, in principle, be directly applied to the problem of heterogeneity between families. The

consequences of heterogeneity on power can thus be alleviated, it will nonetheless be reduced compared to an ideal homogeneous situation. The next natural step is to gain understanding in heterogeneity by including covariate information. Family-specific covariates can be readily incorporated using more advanced meta-analytic techniques such as meta-regression [van Houwelingen et al., 2002]. Individual-specific covariate marginal effects are routinely incorporated into linkage studies for continuous phenotypes and chapter 7 offers a solution for traits of other types. Further substantial gains in power will only be obtained by explicitly incorporating gene by covariate interactions into linkage analysis. The effect of a chosen covariate should be substantial and its value should vary within families in order to yield added-value.

Linkage studies has been the main tool for generating hypotheses in the positional approach to gene mapping. The advent of the SNP technology has switched the emphasis to association scans in unrelated subjects (case-control designs), however this methodology is particularly vulnerable to the confounding effect of population stratification; besides its advantage in terms of efficiency heavily rests on the presence of strong LD between genotyped SNPs and causal variants. The recognition of these facts has spurred new enthusiasm into family-based studies, although those studies are primarily aimed at detecting association, they provide new opportunities for applying and improving linkage methods. In fact, even when strong association with one or several SNPs has been established, it is often not straightforward to actually pinpoint the gene(s) involved, it becomes then tempting to use linkage in order to confirm the implication of a chromosomal region identified by association methods in family studies. Several genes under a linkage peak may influence a trait and although one gene may have already been identified, it seems natural to test this hypothesis formally. The manicheism between linkage and association scans is now becoming obsolete, it is clear that no one approach is uniformly optimal and in fact the former should be used to enhance the latter.

One crucial problem in the elucidation of the epidemiology of common diseases is the integration of knowledge from different sources and nature. Knowledge from gene-expression, proteomics and gene ontology data need to be pooled together with genetic data if we want to efficiently gather scientific evidence. Finally and notwithstanding the biological importance of identifying genes, these are bound to have small effects

at the population level, and it seems unlikely that such discoveries will revolutionize public health policies.

Bibliography

Abecasis GR, Cardon LR, and Cookson WO 2000. A general test of association for quantitative traits in nuclear families. Am J Hum Genet 66:279–292.

Abecasis GR, Cherny SS, and Cardon LR 2001. The impact of genotyping error on family-based analysis of quantitative traits. European Journal of Human Genetics 9:130–134.

Abecasis GR, Cherny SS, Cookson WO, and Cardon LR 2002. Merlin - rapid analysis of dense genetic maps using sparse gene flow trees. Nature Genet 30:97–101.

Agresti A 2002. Categorical Data Analysis. New York: Wiley Series in Probability and statistics. Second Edition.

Alcais L A Abel 2001. Incorporation of covariates in multipoint model-free linkage analysis of binary traits: how important are unaffecteds? European Journal of Human Genetics 9:613–620.

Almasy L and Blangero J 1998. Multipoint quantitative-trait linkage analysis in general pedigrees. Am J Hum Genet 62:1198–1211.

Amos CI 1994. Robust variance-components approach for assessing genetic linkage in pedigrees. Am J Hum Genet 54:535–543.

Amos CI, Elston RC, Wilson AF, and Baileywilson JE 1989. A more powerful robust sib-pair test of linkage for quantitative traits. Genet Epidemiol 6:435–449.

Biernacka JM, Sun L, and Bull SB 2005. Simultaneous localization of two linked disease susceptibility genes. Genet Epidemiol 28:33–47.

Blackwelder WC and Elston RC 1985. A comparison of sib-pair linkage tests for disease susceptibility loci. Genet Epidemiol 2:85–97.

Bohning D, Dietz E, and Schlattmann P 1998. Recent developments in computer-assisted analysis of mixtures. Biometrics 2:525–536.

Boomsma D, Busjahn A, and Peltonen L 2002. Classical twin studies and beyond. Nat Rev Genet 3:872–882.

Breslow NE and Clayton DG 1993. Approximate inference in Generalized Linear Mixed

models. J Amer Stat Assoc 88.

Burton PR, Tiller KJ, Gurrin LC, Cookson WOCM, Musk AW, and Palmer LJ 1999. Genetic variance components analysis for binary phenotypes using generalized linear mixed models (GLMMs) and gibbs sampling. Genet Epidemiol 17:118–140.

Chen H, Chen J, and Kalbfleisch J 2001. A modified likelihood ratio test for homogeneity in finite mixture models. J R Statis Soc B 63:19–29.

Chen W, Broman K, and Liang K 2004. Quantitative trait linkage analysis by generalized estimating equations: unification of variance components and Haseman-Elston regression. Genet Epidemiol 26:265–272.

Cherny SS, Purcell S, Rijsdijk F, Hewitt JK, and Sham PC 1999. Selecting maximally informative sibships for QTL linkage analysis. Behav Genet 29:352–352.

Chiou JM, Liang KY, and Chiu YF 2005. Multipoint linkage mapping using sibpairs: Nonparametric estimation of trait effects with quantitative covariates. Genet Epidemiol 28:58–69.

Claus B, Risch N, and Douglas Thompson W 1991. Genetic analysis of breast cancer in the cancer and steroid hormone study. Am J Hum Genet 48:232–242.

Commenges D 1994. Robust genetic linkage analysis based on a score test of homogeneity: the weighted pairwise correlation statistic. Genet Epidemiol 11:189–200.

Cox DR and Hinkley DV 1974. Theoretical Statistics. London: Chapman and Hall.

Cox DR and Reid N 2004. A note on pseudolikelihood constructed from marginal densities. Biometrika 91:729–737.

Cuenco KT, Skatkiewicz JP, and Feingold E 2003. Recent advances in human quantitative-trait-locus mapping: comparison of methods for selected sibling pairs. Am J Hum Genet 73:863–873.

Dempfle A and Loesguen S 2003. Meta-analysis of linkage studies for complex diseases: an overview of methods and a simulation study. Ann Hum Genet 68:69–83.

Dempster AP and Laird DB 1977. Maximum likelihood from incomplete data via the em algorithm (with discussion). Journal of the Royal Stat Soc, Series B 39:1–38.

Deng HW, Xu FH, Liu YZ, Shen H, Deng HY, Huang QY, Liu YJ, Conway T, Li J, Davies K, and Recker RR 2002. A whole-genome linkage scan suggests several genomic regions potentially containing qtls underlying the variation of stature. Am J Med Genet 113:29–39.

DerSimonian R and Laird N 1986. Meta-analysis in clinical trials. Controlled Clinical Trials 7:177–188.

Devlin B and Roeder K 1999. Genomic control for association studies. Biometrics 55:997–1004.

Dolan CV and Boomsma DI 1998a. Optimal selection of sib pairs from random samples for linkage analysis of a qtl using the edac test. Behav Genet 28:197–206.

Dolan CV and Boomsma DI 1998b. Optimal selection of sib pairs from random samples for linkage analysis of a QTL using the EDAC test. Behav Genet 28:197–206.

Dolan CV, Boomsma DI, and Neale MC 1999. A simulation study of the effects of assignment of prior identity-by-descent probabilities to unselected sib pairs, in covariance-structure modeling of a quantitative-trait locus. Am J Hum Genet 64:268–280.

Douglas JA, Boehnke M, and Lange K 2000. A multipoint method for detecting genotyping errors and mutations in sibling-pair linkage data. Am J Hum Genet 66:1287–1297.

Dudoit S and Speed TP 2000. A score test for the linkage analysis of qualitative and quantitative traits based on identity by descent data from sib-pairs. Biostatistics 1:1–26.

Eaves L and Meyer J 1994. Locating human quantitative trait loci - guidelines for the selection of sibling pairs for genotyping. Behav Genet 24:443–455.

Efron B and Tibshirani R 2002. Empirical bayes methods and false discovery rates for microarrays. Genet Epidemiol 23:70–86.

Eilers PHC and Goeman JJ 2004. Enhancing scatterplots with smoothed densities. To appear in Bioinformatics 20 .

Ekstrøm CT 2004. Multipoint linkage analysis of quantitative traits on sex-chromosomes. Genet Epidemiol 26:218–230.

Ekstrøm CT and Dalgaard P 2003. Linkage analysis of quantitative trait loci in the presence of heterogeneity. Hum Hered 55:16–26.

Elston RC, Song D, and Iyengar SK 2005. Mathematical assumptions versus biological reality: Myths in affected sib pair linkage analysis. Am J Hum Genet 76:152–156.

Etzel CJ and Guerra R 2002. Meta-analysis of genetic-linkage analysis of quantitiative-trait loci. Am J Hum Genet 71:56–65.

Evans DM and Cardon LR 2004. Guidelines for genotyping in genomewide linkage studies: Single-nucleotide-polymorphism maps versus microsatellite maps. Am J Hum Genet

75:687–692.

Feingold E, Brown PO, and Siegmund D 1993. Gaussian models for genetic analysis using complete high-resolution maps of identity by descent. Am J Hum Genet 53:234–251.

Fisher RA 1918. The correlation between relatives on the supposition of mendelian inheritance. Trans of the Royal Soc of Edinburgh 52:399–433.

Glidden DV, Liang KY, Chiu YF, and Pulver AE 2003. Multipoint affected sibpair linkage methods for localizing susceptibility genes of complex diseases. Genet Epidemiol 24:107–117.

Greenwood C and Bull S 1999. Analysis of affected sib pairs, with covariates - with and without constraints. Am J Hum Genet 64:871–885.

Gu C, , Todorov A, and Rao DC 1996. Combining extremely concordant sibpairs with extremely discordant sibpairs provides a cost effective way to linkage analysis of qtls. Genet Epi 13:513–533.

Gu C, Province M, Todorov A, and Rao DC 1998. Meta-analysis methodology for combining non-parametric sibpair linkage results: genetic homogeneity and identical markers. Genet Epi 15:609–626.

Haseman JK and Elston RC 1972. The investigation of linkage between a quantitative trait and a marker locus. Behav Genet 2:3–19.

Hirschhorn JN, M LC, Daly MJ, Kirby A, Schaffner S, Burtt N, Altshuler D, Parker A, Rioux J, Platko J, Gaudet D, Hudson T, Groop L, and Lander E 2001. analysis of stature in multiple populations reveals several regions with evidence of linkage to adult height. Am J Hum Genet 69:106116.

Holliday E, Mowry B, Chant D, and Nyholt D 2005. The importance of modelling heterogenity in complex disease: application to NIMH schizophrenia genetics initiative data. Human Heredity To appear.

Holmans P 1993. Asymptotic properties of affected-sib-pair linkage analysis. Am J Hum Genet 52:362–374.

Houwing-Duistermaat JJ, van Houwelingen HC, and de Winter JP 2000. Estimation of individual genetic effects from binary observations on relatives applied to a family history of respiratory illnesses and chronic lung disease of newborns. Biometrics 56:808–814.

Kong A and Cox NJ 1997. Allele-sharing models: Lod scores and accurate linkage tests. Am

J Hum Genet 61:1179–1188.

Kruglyak L, Daly MJ, Reeve-Daly MP, and Lander ES 1996. Parametric and nonparametric linkage analysis: a unified multipoint approach. Am J Hum Genet 58:1347–1363.

Kruglyak L and Lander ES 1995. Complete multipoint sib-pair analysis of qualitative and quantitative traits. Am J Hum Genet 57:439–454.

Lander ES and Botstein D 1989. Mapping mendelian factors underlying quantitative traits using rflp linkage maps. Genetics 121:185–199.

Lander ES and Green P 1987. Construction of multilocus genetic maps in humans. Proc Natl Acad Sci USA 84:2363–2367.

Lange K, Cantor R, Horvath S, Perola M, Sabatti C, Sinsheimer J, and Sobel E 2001. Mendel version 4.0: A complete package for the exact genetic analysis of discrete traits in pedigree and population data sets. Amer J Hum Genetics 69(supplement):504–504.

Lange K, Westale J, and Spence MA 1976. Extensions to pedigree analysis .3. variance components by scoring method. Ann Hum Genet 39:485–491.

le Cessie S and van Houwelingen JC 1994. Logistic regression for correlated binary data. Applied Statistics 43:95–108.

Lebrec J, Putter H, and van Houwelingen JC 2004. Score test for detecting linkage to complex traits in selected samples. Genet Epidemiol 27:97–108.

Lebrec J, Putter H, and van Houwelingen JC 2006. Potential bias in Generalized Estimating Equations linkage methods under incomplete information. Genet Epidemiol 30:94–100.

Li Z and Rao DC 1996. Random effects model for meta-analysis of multiple quantitative sib pair linkage studies. Genet Epidemiol 13:377–383.

Liang KY, Chiu YF, and Beaty TH 2001. A robust identity-by-descent procedure using affected sib pairs: Multipoint mapping for complex diseases. Hum Hered 51:64–78.

Liang KY and Zeger SL 1986. Longitudinal data analysis using generalized linear models. Biometrika 73:13–22.

Morton NE 1955. Sequential tests for the detection of linkage. Am J Hum Genet 7:277–318.

Mulder E, van Baal C, Gaist D, Kallela M, Kaprio J, Svensson D, Nyholt D, Martin N, MacGregor A, Cherkas L, and Boomsma D 2003. Genetic and environmental influences on migraine: a twin study across six countries. Twin Research 6:422–431.

Neale MC, Boker S, Xie G, and Maes H 1999. Mx: Statistical Modeling. Box 126 MCV,

Richmond, VA 23298: Department of Psychiatry.

Noh M, Yip B, Lee Y, and Pawitan Y 2005. Multicomponent variance estimation for binary traits in family-based studies. Genet Epidemiol 30:37–47.

Normand ST 1999. Meta-analysis: formulating, evaluating, combining and reporting. Statistics in Medicine 18:321–359.

Olkin I and Sampson A 1998. Comparison of meta-analysis versus analysis of variance of individual patient data. Biometrics 54:317–322.

Olson J 1999. A general conditional-logistic model for affected-relative pair linkage studies. Am J Hum Genet 65:1760–1769.

Ott J 1999. Analysis of human genetic linkage, ed. 3. Baltimore: Johns Hopkins University Press.

Palmer LJ, Scurrah K, Tobin M, Patel S, Celedon J, Burton P, and ST W 2003. Genome-wide linkage analysis of longitudinal phenotypes using sigma2a random effects (ssars) fitted by gibbs sampling. BMC Genet 31:Suppl 1:S12.

Pearson K 1901. On the correlations of characters not quantitatively mesurable. Phil Trans of the RSS, A 195:1–47.

Peng J, Tang H, and Siegmund D 2005. Genome scans with gene-covariate interaction. Genet Epidemiol 29:173–184.

Purcell S, Cherny SS, Hewitt JK, and Sham PC 2001. Optimal sibship selection for genotyping in quantitative trait locus linkage analysis. Human Heredity 52:1–13.

Purcell S and Sham PC 2004. Epistasis in quantitative trait locus linkage analysis: interaction or main effect? Behavior Genetics 34:143–152.

Putter H, Lebrec J, and van Houwelingen JC 2003. Selection strategies for linkage studies using twins. Twin Research 6:377–382.

Putter H, Sandkuijl LA, and van Houwelingen JC 2002. Score test for detecting linkage to quantitative traits. Genet Epidemiol 22:345–355.

Rao DC and Province MA 2001. Genetic dissection of complex traits. Academic Press.

Rijsdijk FV, Hewitt JK, and Sham PC 2001. Analytic power calculation for QTL linkage analysis of small pedigrees. Eur J Hum Genet 9:335–340.

Rijsdijk FV and Sham PC 2000. Effects of pedigree structure and genetic model on the power of QTL linkage analysis. Behav Genet 30:417–417.

Risch N 1990. Linkage strategies for genetically complex traits. ii. the power of affected relative pairs. Am J Hum Genet 46:229–241.

Risch N and Zhang H 1995. Extreme discordant sib pairs for mapping quantitative trait loci in humans. Science 268:1584–1589.

Risch NJ 2000. Searching for genetic determinants in the new millennium. Nature 405:847–856.

Schaid DJ, Guenther J, Christensen G, Hebbring S, Rosenow C, Hilker C, McDonnell S, Cunningham J, Slager S, Blute M, and Thibodeau SN 2004. Comparison of microsatellites versus single-nucleotide polymorphisms in a genome linkage screen for prostate cancer-susceptibility loci. Am J Hum Genet 75:948–965.

Schaid DJ, Olson JM, Gauderman WJ, and Elston RC 2003. Regression models for linkage: issues of traits, covariates, heterogeneity, and interaction. Hum Hered 55:86–96.

Schaid DJ, Sinnwell JP, and Thibodeau SN 2005. Robust multipoint identical-by-descent mapping for affected relative pairs. Am J Hum Genet 76:128–138.

Schork NJ 1993. Extended multipoint identity-by-descent analysis of human quantitative traits - efficiency, power, and modeling considerations. Am J Hum Genet 53:1306–1319.

Scurrah K, Palmer L, and Burton P 2000. Variance components analysis for pedigree-based censored survival data using generalized linear mixed models (glmms) and gibbs sampling in bugs. Genet Epidemiol 19:127–148.

Searle SR, Casella G, and McCulloch CE 1992. Variance Components. New York: Wiley.

Self SG and Liang KY 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. J Am Stat Assoc 82:605–610.

Sham PC 1998. Statistics in Human Genetics. Applications of Statistics. London: Arnold.

Sham PC and Purcell S 2001. Equivalence between Haseman-Elston and Variance-Components linkage analyses for sib-pairs. Am J Hum Genet 68:1527–1532.

Sham PC, Purcell S, Cherny SS, and Abecassis GR 2002. Powerful regression-based quantitative-trait linkage analysis of general pedigrees. Am J Hum Genet 71:238–253.

Sham PC, Zhao JH, Cherny SS, and Hewitt JK 2000. Variance-components qtl linkage analysis of selected and nonnormal samples: Conditioning on trait values. Genet Epidemiol 19:S22–S28.

Shapiro A 1988. Towards a unified theory of inequality constrained testing in multivariate

analysis. International Statistical Review 56:49–62.

Silventoinen K, Sammalisto S, Perola M, Boomsma DI, Cornes BK, Davis C, Dunkel L, de Lange M, Harris JR, Hjelmborg JVB, Luciano M, Martin NG, Mortensen J, Nisticò L, Pedersen N, Skytthe A, Spector TD, Stazi MA, Willemsen G, and Kapprio J 2003. Heritability of adult body height: a comparative study of twin cohorts in eight countries. Twin Research 6:399–408.

Skatkiewicz JP, Cuenco KT, and Feingold E 2003. Recent advances in human quantitative-trait-locus mapping: comparison of methods for discordant sibling pairs. Am J Hum Genet 73:874–885.

Sobel E, Papp J, and Lange K 2002. Detection and integration of genotyping errors in statistical genetics. Am J Hum Genet 70:496–508.

Tang HK and Siegmund D 2001. Mapping quantitative trai loci in oligogenic models. Biostatistics 2:147–162.

Tang HK and Siegmund D 2002. Mapping multiple genes for complex or quantitative traits. Genet Epidemiol 22:313–327.

Teng J and Siegmund D 1997. Combining information within and between pedigrees for mapping complex traits. Am J Hum Genet 60:979–992.

Teng J and Siegmund D 1998. Multipoint linkage analysis using affected relative pairs and partially informative markers. Biometrics 54:1247–1265.

Tritchler D, Liu Y, and Fallah S 2003. A test of linkage for complex discrete and continuous traits in nuclear families. Biometrics 59:382–392.

van Houwelingen JC, Arends LR, and Stijnen T 2002. Advanced methods in meta-analysis: multivariate approach and meta-regression. Statistics in Medicine 21:589–624.

Verbeke G and Molenberghs G 2003. The use of score tests for inference on variance components. Biometrics 59:254–262.

Wang K 2002. Mapping quantitative trait loci using multiple phenotypes in general pedigrees. Human Heredity 55.

Wang K and Huang J 2002a. A score-statistic approach for the mapping of quantitative-trait loci with sibships of arbitrary size. Am J Hum Genet 70:412–424.

Wang K and Huang J 2002b. Score test for mapping quantitative-trait loci with sibships of arbitrary size when the dominance effect is not negligible. Genet Epidemiol 23:498–412.

Wehrens R, Putter H, and Buydens LMC 2000. The bootstrap: a tutorial. Chemometrics and Intelligent Lab Sys 54:35–52.

Whittemore AS and Halpern J 1994. A class of tests for linkage using affected pedigree members. Biometrics 50:118–127.

Willemsen G, Boomsma DI, Beem AL, Vink JM, Slagboom PE, and Posthuma D 2004. Qtls for height: results of a full genome scan in dutch siblings. European Journal of Human Genetics pp. 1–9.

Williams JT and Blangero J 1999. Power of variance component linkage analysis to detect quantitative trait loci. Am J Hum Genet 63:545–563.

Wise LH, Lanchbury JS, and Lewis CM 1999. Meta-analyses of genome searches. Ann Hum Genet 63:263–272.

Zeegers M, Rice J, Rijsdijk F, Abecasis G, and Sham P 2003. Regression-based sib pair linkage analysis for binary traits. Human Heredity 55:125–131.

Samenvatting

In dit proefschrift worden manieren beschreven om de huidige opzet en analyse van studies naar de koppeling van genen (*linkage*) met complexe eigenschappen te verbeteren. In *linkage* onderzoek wordt gebruik gemaakt van genetische merkers (*markers*). Met deze *markers* kan men genetische overeenkomstigheden tussen verwanten meten. Door deze genetische gelijkenis te vergelijken met fenotypische overeenkomsten, kunnen regio's op het chromosoom geïdentificeerd worden waarin genen liggen die bijdragen tot de vorming van het betreffende fenotype. Hoewel, deze methode erg succesvol bleek bij het in kaart brengen van genen en eigenschappen, die volgens de wet van Mendel overerven, schiet hij bij meer complexe overervingpatronen vaak tekort. Omdat betrokken genen maar een zeer beperkte invloed hebben op complexe eigenschappen is men m.b.t *linkage* studies intrinsiek beperkt. Deze intrinsieke beperking rechtvaardigt een heel eigen statistische benadering, welke de basis vormt van de methodologie beschreven in dit proefschrift. Voor het toetsen van hypothesen kunnen score toetsen worden gebruikt [Cox and Hinkley, 1974]. De lokale optimaliteit eigenschappen van deze toetsen blijken zeer geschikt in de context van compexe eigenschappen. Bovendien hebben zij vaak een herkenbare uitdrukking en kunnen zij geïnterpreteerd worden in termen van regressieanalyse, waardoor zij in principe snel uit te rekenen zijn. Dit laatste is van groot belang in genetisch onderzoek waarbij vaak grote hoeveelheden data geanalyseerd moeten worden.

In **hoofdstuk 1** wordt een inleiding gegeven over de genetische mechanismen die ten grondslag liggen aan *linkage*. Tevens volgt een korte beschouwing over de traditionele methodologie en wordt een samenvatting gegeven van belangrijke nog onopgeloste vraagstukken.

Hoofdstuk 2 behandelt hoofdzakelijk de analyse van kwantitatieve eigenschappen die gemeten zijn in geselecteerde families. Hierbij is de selectie gebaseerd op de waarde van een betreffend kenmerk. Een score toets gebaseerd op de conditionele likelihood gegeven de fenotypische waarden, wordt afgeleid. Deze toets kan gebruikt worden

bij data uit willekeurige stambomen. Hoewel bij de afleiding van de toets wordt aangenomen dat het model normaal verdeelde variatiecomponenten bevat, is de type I fout van de toets robuust tegen afwijkingen van deze normale verdeling. Onder de aanname dat het model de verdeling van het fenotype goed beschrijft, heeft de toets optimale eigenschappen voor lokale alternatieven. Bovendien geeft de waarde van de bijbehorende Fisher informatie van de toets een indicatie in hoeverre elke familie informatief is. Deze Fisher informatie kan gebruikt worden als criterium voor het selecteren van individuen voor genotyperingen. Verderop in het hoofdstuk wordt een aangepaste versie van de toets gegeven voor binaire gegevens. Een model met een onderliggende continue latente variabelen wordt gebruikt waarbij deze variabelen in twee klassen wordt verdeeld door een drempelwaarde te creëren. Dit model wordt liability treshold model genoemd.

Hoofdstuk 3 bepleit het gebruik van geselecteerde families bij het in kaart brengen van genen voor complexe eigenschappen waarbij tweelingen worden gebruikt. Met behulp van de methodologie, welke gebaseerd op het informatiecriterium dat in hoofdstuk 2 is afgeleid, worden potentile voordelen gekwantificeerd door gebruik te maken van een serie voorbeelden van kwantitatieve en kwalitatieve fenotypen, welke relevant zijn voor het GenomEUtwin project.

Hoofdstuk 4 behandelt het probleem van genotyperingsfouten binnen *linkage* onderzoek. Het effect van genotyperingsfouten op *linkage* studies wordt beschreven door en formule te creëren, die de vertekening die optreedt door genotyperingsfouten weer kan geven. Deze formule geeft inzicht in enkele van de empirische bevindingen, in het bijzonder verklaart het de rol van genotyperingsfouten in onderzoeksopzetten met selecte versus aselecte data. Ten slotte wordt een voorstel gedaan tot een robuuste aanpassing van de gebruikelijke *linkage* toetsen gebaseerd op een genoom wijde controle van het overschot van allelen, die een kopie zijn van een zelfde voorouderlijk allel. Allelen die een kopie zijn van een zelfde voorouderlijk allel worden identical by descent genoemd. Deze aanpassing geeft niet alleen robuustheid tegen genotyperingsfouten, maar ook tegen andere processen die de verwachte waarde van deze fractie verstoren.

Hoofdstuk 5 bespreekt de (on)juistheid van aan aantal standaard methoden welke gebruikt worden als markerinformatie niet volledig is. Het probleem van gevallen waarbij de methode van gegeneraliseerde schattingsvergelijkingen (*generalized esti-*

mating equations) voor het in kaart brengen van genen faalt [Liang et al., 2001] wordt uitgelicht.

Hoofdstuk 6 vertaalt de standaard meta-analyse technieken naar het onderzoeksveld van het in kaart brengen van genen die een rol spelen bij kwantitatieve eigenschappen (*quantitative trait loci (QTLs)*). Dit onderzoeksgebied heeft een aantal specifieke kenmerken waarbij aanpassingen nodig zijn. Het probleem van heterogeniteit in genetische loci wordt nader toegelicht. Wanneer er geen co-variabelen geobserveerd zijn op individu nivo en onder een homogeen model, is de meta-analytische aanpak asymptotisch equivalent aan de analyse van samengevoegde databestanden, maar is logistiek veel eenvoudiger uit te voeren.

Ten slotte wordt in **hoofdstuk 7** een score toets voor *linkage* analyse in de grote klasse van de algemene lineaire modellen beschreven. Deze benadering is gebaseerd op een pseudo-likelihood van de gegevens. Hoewel deze test waarschijnlijk niet optimaal is in alle situaties, heeft deze test het voordeel herkenbaar te zijn en een robuuste type I fout te hebben. Het levert een eenvoudige manier om het bekende effect van co-variabelen te implementeren in *linkage* analyse en is toepasbaar voor willekeurige stambomen.

Het proefschrift wordt afgesloten met conclusie, waarin ik een perspectief schets van de methodologie die een rol speelt in *linkage* bij het in kaart brengen van genen.

Curriculum Vitae

Jérémie Lebrec was born on June 3, 1974 in Rennes (France). In 1992, he passed his Baccalaureate at Lycée J. Loth, Pontivy and subsequently undertook undergraduate studies in mathematics at the University of Rennes. During the course of his studies, he spent one year at the University of Cantábria, Santander (Spain) where he first got acquainted with the discipline statistics. After his graduation in 1997, he embarked on a Master of Science course in statistics at University College London (U.K.) in which he graduated (with Distinction) in 1998 (dissertation under supervision of Prof. Stephen Senn). He then joined the pharmaceutical industry (SmithKline Beecham Pharmaceuticals and Pfizer, R&D) where he worked on the design and analysis of clinical trials. In 2002-03, he learned the basics of clinical research and biostatistics as applied to the field of oncology at the European Organization for Research and Treatment of Cancer data center in Brussels (Belgium). Hans van Houwelingen and the late Lodewijk Sandkuijl then offered him the opportunity to start his doctoral research in the area of statistical genetics.

The work gathered in this thesis was carried out in the period 2003-2006 at the Dept. of Medical Statistics and Bioinformatics, Leiden University Medical Center as part of a European Union funded project (GenomEUtwin). During this research period, the author presented his work at several international conferences including in Odense (Denmark), Cardiff (U.K.) and Montréal (Canada).

Published and submitted chapters

Some chapters in this thesis have already been published in scientific journals, others have been submitted or are about to be submitted for publication:

Chapter 2: J. Lebrec, H. Putter and J.C. van Houwelingen (2004). Score Test for Detecting Linkage to Complex Traits in Selected Samples. *Genetic Epidemiology* **27** (2), 97–108.

Chapter 3: H. Putter, J. Lebrec and J.C. van Houwelingen (2003). Selection Strategies for Linkage Studies using Twins. *Twin Research* **6** (5), 377–382.

Chapter 4: J.J.P. Lebrec, H. Putter, J.J. Houwing-Duistermaat and H.C. van Houwelingen. Genomic Control for Genotyping Error in Linkage Mapping for Complex Traits. Submitted.

Chapter 5: J. Lebrec, H. Putter and J.C. van Houwelingen (2006). Potential Bias in Generalized Estimating Equations Linkage Methods under Incomplete Information. *Genetic Epidemiology* **30** (1), 94–100.

Chapter 6: J.J.P. Lebrec, D.I. Boomsma, K. Christensen, N.G. Martin, N.L. Pedersen, M. Perola, T.D. Spector, H. Putter and H.C. van Houwelingen. Classical Meta-Analysis Applied to QTL mapping - Genomewide Linkage Scan for Height in the GenomEUtwin Project. To be submitted.

Chapter 7: J.J.P. Lebrec and H.C. van Houwelingen. Score Test for Linkage in Generalized Linear Models. Accepted for publication in *Human Heredity*.