Statistical methods for analysing complex genetic traits

El Galta, Rachid

Statistical methods for analysing complex genetic traits

Rachid el Galta

Statistical methods for analysing complex genetic traits

PROEFSCHRIFT

ter verkrijging van de graad van Doctor
aan de Universiteit Leiden,
op gezag van de Rector Magnificus Dr. D. D. Breimer,
hoogleraar in de faculteit der Wiskunde en Natuurwetenschappen
en die der Geneeskunde,
volgens besluit van het College voor Promoties
te verdedigen op woensdag 27 september 2006
klokke 13:45 uur

door

Rachid el Galta

geboren te Elkbab, Marokko, in 1972

Promotiecommissie

Promotores:	Prof. dr. J. C. van Houwelingen Prof. dr. T. Stijnen · *Erasmus Medical Center, Rotterdam*
Co-promotor:	Dr. J. J. Houwing-Duistermaat
Referent:	Prof. dr. H. Bickeböller · *University of Göttingen, Göttingen*
Overige leden:	Prof. dr. C. M. van Duijn · *Erasmus Medical Center, Rotterdam* Prof. dr. E. P. Slagboom Prof. dr. F. R. Rosendaal

This work was financially supported by a Program Grant from the Netherlands Organization for Scientific Research (NWO 912-03-014). Publication of this thesis was supported by the department of Medical Statistics, LUMC.

Contents

CHAPTER 1

Introduction

1.1 Introduction

The focus of this thesis is on statistical methods for complex genetic traits. A genetic trait is complex when genetic and environmental factors are involved. Researchers seek to identify genetic factors causing such traits using family as well as population-based approaches. Since many genes are likely to be involved in the etiology of a complex trait, the contribution of each single locus can be small and therefore, such genes are difficult to detect. Many existing statistical methods have been devised to study simple Mendelian disorders and so they are not suitable for studying complex genetic traits. Therefore, studying such traits necessitates the development of new statistical methods. A strategy for studying complex genetic traits is to perform linkage analysis to identify regions, and then carry out association analysis to verify whether a candidate gene is involved.

In this chapter we briefly describe and discuss several statistical methods and analyses used in genetic studies. For a reference book of statistical methods in human genetics we refer to Elston et al. (2002). Recently Forabosco et al. (2005) provided a valuable review of methods for linkage analysis and association studies. Throughout this chapter we also refer to more subject-specific reviews. Finally we close this chapter with an outline of the content of the next chapters.

1.2 Familial Correlation

In genetic epidemiology, researchers aim to identify genetic factors involved in the etiology of traits. Any evidence of correlation between phenotypes and genotypes is suggestive of the presence of such genetic factors. However, before genetic markers are typed, investigators should assure that the trait clusters within families. Closely related individuals tend to have similar phenotypes compared with non related individuals or distant relatives. Although the presence of such a clustering may be due to shared environmental factors rather than shared genetic factors, the absence of familial clustering implies that it is of no benefit to continue the study. Its presence after adjusting for environmental factors implies that genetic factors may play a role in producing a predisposition to a trait.

Many methods have been proposed in the literature for the analysis of familial correlation. Liang and Beaty (2000) discussed methods of analysis of aggregation for the family case-control design. For binary traits, familial aggregation is often measured by odds ratios using logistic regression models that allow for dependence between family members (Bonney, 1986; Liang and Beaty, 1991). Generalized estimating equations (GEE)(Liang and Zeger, 1986) are often used to address the dependence between family members, when estimating the parameters. For more references about the GEE approach to deal with familial correlation for binary traits see FitzGerald and Knuiman (2000) or Tregouet and Tiret (2000). For quantitative traits, familial correlation is often studied by fitting multivariate normal models to the trait values of family members (Beaty and Liang, 1987; Rao and Wette, 1987). The covariances between relatives depend on their degrees of relationship. For randomly selected families, Houwing-Duistermaat et al. (1995) used a generalized linear mixed model (McCullagh and Nelder, 1989), and derived the score statistic to test for familial clustering within relatives based on their degrees of relationship. The score statistic does not assume any distribution of random effects. However, assessing familial aggregation in non-random pedigrees requires accounting for the sampling scheme.

Ascertainment

Often families are selected based on the phenotype of one or more family members. Family members who caused the family to be entered in a study are referred to as probands. For binary traits, a proband may be an affected subject with a disease. For quantitative traits, a proband may have a phenotype that exceeds a certain threshold. Many studies (Cardon and Fulker, 1994; Carey and Williamson, 1991; Risch and Zhang, 1995; Zhang and Risch, 1996) have shown that the power substantially increases when selected samples are used.

There are different ascertainment schemes e.g. complete ascertainment, single ascertainment and quadratic ascertainment. In complete ascertainment, all families with at least one family member with the required phenotype are equally probable to enter the study. In single ascertainment and quadratic ascertainment families enter the study with probabilities proportional to the number and the square of the number of family members having the required phenotype, respectively.

When families are selected through probands, the statistical methods used should take into account the ascertainment procedure (Elston and Sobel, 1979; Morton, 1959). For single ascertainment, many authors have proposed to condition on the observed phenotypes of the probands (Beaty and Liang, 1987; Cannings and Thompson, 1977; Hopper and Mathews, 1982). For quantitative traits, Elston and Sobel (1979) and Rao and Wette (1987) proposed to condition on the event that the phenotype of the proband exceeds a pre-specified threshold value. Rao et al. (1988) and de Andrade and Amos (2000) compared the two ways of correcting for ascertainment in the context of the variance components approach. Both studies concluded that the former adjustment for ascertainment is less efficient than the latter provided that the threshold is well known. Whereas conditioning on exceeding the threshold is less efficient if the threshold is not known.

However if single ascertainment cannot be assumed, conditioning on affected probands may not be sufficient. Ewens and Shute (1986) showed that if remaining family members are related to ascertainment, the estimates of genetic parameters may be biased. They proposed to split the family in a set of probands related to ascertainment and a set of relatives not related to ascertainment.

1.3 Linkage analysis

Once familial aggregation is established, researchers pursue the study by gathering genetic materials of family members relevant for the study. Then genome wide linkage analysis is performed to identify regions that may contain susceptibility genes. Methods for linkage analysis rely on the biological phenomenon of recombination. During meiosis, recombination occurs when homologous chromosome pairs exchange genetic material. The probability of a recombination event occurring between loci increases with the physical distance between them. Hence, alleles at close loci are more likely to be transmitted jointly to descendants than alleles at distant loci. Moreover, relatives who have similar phenotypes are expected to have inherited the same genetic materials in the vicinity of the genes that predispose to those phenotypes.

Linkage analysis has been divided into classes: parametric and non-parametric methods. Parametric linkage analysis methods are based on the analysis of the recombination between the unobserved disease locus and observed genetic markers along the human genome. They require specification of genetic parameters describing the mode of trait inheritance, such as penetrance and disease-allele frequency. When genetic parameters are correctly specified parametric linkage analysis is the most powerful. Very distant loci are expected to recombine with probability $\theta = 0.5$, while close (linked) loci recombine with probability $\theta < 0.5$. To estimate the probability of a recombination, and test the null hypothesis of no linkage, between the unknown predisposing gene locus and a marker locus the LOD score method is usually used. The LOD score is defined as follows

$$Z = \max_{0 \leq \theta \leq 0.5} log_{10}[\frac{L(\theta)}{L(\theta = 0.5)}], \tag{1.1}$$

where $L(\theta)$ is the likelihood of the observed data given assumed parameter values. A test is significant if $Z \geq 3$ (see Lander and Kruglyak (1995) for details). For a comprehensive introduction to parametric linkage analysis we refer to Ott (1999).

Although parametric linkage methods have been successful in localizing genes responsible for simple Mendelian diseases, these methods have achieved only limited success in identifying genes predisposing to complex traits (Risch, 2000). Many complex genetic traits are controlled by multiple genetic factors. For such traits the mode of inheritance is often unknown. In contrast to parametric linkage methods, non-parametric linkage methods do not require the specification of mode of inheritance. Non-parametric methods rely only on the information of sharing alleles identical by descent (IBD) between relatives at a given locus to study whether this locus is genetically linked to the unobserved disease locus. Two alleles are IBD if they are both physical copies of the same ancestral gene (Lange, 2002). Relatives may share 0, 1 or 2 allele IBD at any given locus. According to the Mendelian law of inheritance, the probabilities π_0, π_1 and π_2 of sharing 0, 1 and 2 alleles IBD under random segregation, can be calculated for any two relatives. For example, a pair of siblings shares 0, 1 and 2 alleles IBD with probabilities $\pi_0 = 0.25$, $\pi_1 = 0.5$ and $\pi_2 = 0.25$. Linkage between a disease locus and marker genotypes can be studied by comparing the observed numbers of alleles shared IBD by affected relative pairs to the expected number of alleles IBD under random segregation. An increase in the number of alleles IBD indicates the presence of a susceptibility gene in the region.

A widely used design for binary traits is the affected sibling pair design. To test for linkage between a disease locus and a marker locus, Day and Simons (1976) proposed the proportion test, which compares the observed proportion of sib pairs that share two IBDs with $\pi_2 = 0.25$, As an alternative, Green and Woodrow (1977) proposed the mean test, which compares the observed mean IBD with its null value of 0.5. Risch (1990b,c) proposed a likelihood ratio statistic that compares the likelihood of observing 0, 1 and 2 alleles IBD with the likelihood under random segregation. For a valuable review about allele sharing based test statistics for affected relative pairs and general pedigrees we refer to Shih and Whittemore (2001).

Since many marker loci should be tested, multiple testing problems arise. A classical method to adjust for multiple testing is Bonferroni correction. However, this method does not take into account the dependency between marker loci. Hence, it yields conservative p-values. Lander and Kruglyak (1995) proposed to reject the null hypothesis of no linkage for loci with p-values smaller than 0.0001.

For a large sample of affected relative pairs, Feingold et al. (1993) used a Gaussian approximation to the IBD process to test for linkage using all marker loci jointly. The authors assumed (1) the presence of at most a single disease locus on any chromomsome, (2) the Haldane's mapping function, and (3) the observed markers are dense and fully informative about the IBD status. They modelled the excess of IBD sharing at the disease locus with a parameter α. For affected sib pairs, α is equal to $\frac{\lambda_s - 1}{\lambda_s}$, with λ_s the relative risk of a sibling of an affected subject (Risch, 1990a). Since the location of the disease gene is unknown, they proposed to use the maximum of the mean IBD sharing over marker loci as the test statistic for linkage. Further, they used a Gaussian approximation based on the central limit theorem to derive its null distribution.

Recently Liang et al. (2001) introduced a GEE approach for affected sib pairs to estimate the location of a single disease gene in a candidate region previously identified by other approaches. The method also uses the IBD information on all markers simultaneously by incorporating the correlation between them. Schaid et al. (2005) extended the GEE approach to other relative pairs, and Biernacka et al. (2005) derived a similar GEE approach when two disease loci exist on the same chromosomal region of interest.

For quantitative traits, many methods have been proposed in the literature. A famous method for modeling excess IBD sharing in relative pairs is the Haseman-Elston (HE) approach (Haseman and Elston (1972), Amos and Elston (1989)). The method regresses the squared trait difference of pair members on their estimated proportion of IBD sharing. Since the square difference does not capture all the information on the linkage (Wright, 1997), Elston (2000) reconsidered the HE method by regressing the product of the trait difference and trait sum on the estimated proportion of IBD sharing. Another linkage analysis method for quantitative trait-locus is the variance components (VC) approach (Amos, 1994). VC methods fit multivariate normal models to the trait values of family members, and model the covariance between relatives conditional on their IBD sharing. For small effect size the VC approach is equivalent to the HE approach (Putter et al., 2002). A typical VC model is described as follows. Let y_k be a vector of the trait values for the kth family. The variable y_k can be written as

$$y_k = \mu + X_k\beta + q_k + g_k + e_k$$

where μ is the overall mean; β is a vector of the regression coefficients for the covariates; X_k is a design matrix of the covariate values; q_k is a vector of random genetic effects of the locus with $q_k \sim N(0, \sigma_q^2\hat{\Pi}_k)$; g_k is a vector of random effects representing the sum of residual genetic effects with $g_k \sim N(0, \sigma_g^2 R_k)$; e_k is a vector of uncorrelated variables representing residual environmental factors, with $e_k \sim N(0, \sigma_e^2 I_k)$. The covariance of y_k is

$$COV(y_k|\hat{\Pi}_k, R_k) = \sigma_q^2\hat{\Pi}_k + \sigma_g^2 R_k + \sigma_e^2 I_k$$

where R_k is the matrix of degree of relationship for the family; $\hat{\Pi}_k$ is the matrix of the estimated proportion of IBD shared by relative pairs in the family; I_k is an identity matrix. Testing for linkage is equivalent to testing the null hypothesis $H_0 : \sigma_q^2 = 0$ versus $H_1 : \sigma_q^2 > 0$. For reviews on linkage methods for quantitative traits see Feingold (2001) and Amos and de Andrade (2001).

1.4 Association studies

Genetic association studies generally aim to narrow candidate regions, which may be chosen because on the basis of their known biological function or because they were initially identified by linkage analysis. Genetic association analysis compares allele or genotype frequencies at markers in candidate regions in affected individuals with those in unaffected individuals. An allele associated with a disease should be over-represented in affected individuals. A marker may be associated with the disease because it is a disease locus, or because one of its alleles is in linkage disequilibrium (LD) with a causal variant at a disease locus. Let A and B be alleles at two distinct loci and let AB be the corresponding haplotype. A and B are said to be in LD if

$$Pr(AB) \neq Pr(A)Pr(B),$$

with $Pr(A)$, $Pr(B)$ and $Pr(AB)$ the probability of the occurrence of A, B, and AB in the population, respectively. LD suggests that two loci may be very close to one another.

Disease-marker association may also be due to population admixture. This may occur if allele frequencies differ among subpopulation groups. To eliminate false evidence for association due to population admixture, family-based designs are used. A simple and common family-based design is the case-parent trio design, which compares the frequencies of alleles transmitted to affected individuals with those non-transmitted. For a review of family-based designs and corresponding statistical methods we refer to Zhao (2000). In this thesis we do not consider further this type of design.

A classical design to perform association study is the case-control design. Unrelated cases (affected individuals) and unrelated controls (unaffected individuals) are ascertained for study. Let q_i and p_i be the frequency of the associated allele i at a marker in cases and controls respectively. Hastbacka et al. (1992) modelled the excess of the associated allele in cases by

$$q_i = p_i + \delta(1 - p_i),$$

with δ the fraction attributable at risk (Clayton, 2000). δ can serve as a measure of linkage disequilibrium. Devlin and Risch (1995) provided a detailed discussion about measures of linkage disequilibrium.

For di-allelic markers, testing for disease-marker association can be carried out using Pearson's χ^2 with one degree of freedom. For a review of linkage disequilibrium methods for di-allelic markers see Lazzeroni (2001). For multi-allelic markers, Pearson's χ^2 has degrees of freedom equal to the number of alleles minus one. When markers with many alleles are considered, sparse data may occur, making the asymptotic distribution of Pearson's χ^2 invalid. Despite the fact that Monte-Carlo simulation can be used to derive the empirical p-values, χ^2 has low power due to large degrees of freedom. As an alternative, the maximum of the chi-squared statistics of 2-by-2 tables, each of which compares one allele against the rest, can be used when at most one allele is associated (Ewens et al., 1992). Sham and Curtis (1995) proposed to use the maximum of chi-squared statistics corresponding to all possible 2-by-2 tables, comparing any combination of alleles against the rest.

An alternative to taking the maximum is to take the sum over all possibilities. When one allele is associated Terwilliger (1995) modelled the excess of the associated allele in cases by the parameter δ. Since it is unknown which marker allele is associated with the disease, the likelihood corresponding to this model is a weighted sum over all alleles i of conditional likelihoods given that allele i is over-represented in the set of cases. As weights Terwilliger (1995) used the allele frequencies in the overall population. Testing for association can be carried out by comparing the likelihood of the data under the alternative hypothesis of the presence of one associated allele with the likelihood of data under the null hypothesis of no disease-marker association.

When one allele is associated, the maximising approach and the Terwilliger's likelihood approach perform much better than Pearson's χ^2, especially for markers with many alleles. However, if more than one allele is associated with the disease, they may have low power as they are designed to test the alternative hypothesis of the presence of one associated allele (Sham et al., 1996).

When a candidate region contains multiple markers, an alternative to studying single-marker association with a disease is haplotype-based association analysis. A haplotype is a combination of alleles at multiple linked markers inherited together. When haplotypes can be observed directly they can be viewed as variants of a multiallelic maker and then methods for multiallelic markers can be used to study genetic association. However haplotypes are often not observable and they should be inferred from genotype data. Many methods for inferring haplotypes and estimating their frequencies have been developed in the literature (see Niu (2004) for a review). However, when haplotypes can not be determined with certainty, the analysis should take this uncertainty into account. Further discussions of haplotype-based methods are given by Schaid (2004).

1.5 Scope of this thesis

This thesis aims to develop new statistical methods to study genetic backgrounds of complex traits. These methods are presented in a number of subsequent chapters. The order of chapters is motivated by the commonly adopted strategy to identify genes responsible for complex genetic traits, namely starting with aggregation analysis, followed by linkage analysis and accomplished by association studies.

Chapter 2 deals with the ascertainment issue when modelling familial aggregation. A new statistical tool is proposed for testing familial clustering of binary and quantitative traits when families are selected based on the phenotypes of the probands. Familial correlation is modelled using a generalized linear mixed model. To adjust for ascertainment we condition on the phenotypes of the individuals relevant for ascertainment. Further we illustrate the methods using data on a sample of first-degree relatives of probands with type 2 diabetes mellitus.

Chapter 3 is concerned with linkage analysis. Inspired by Liang et al. (2001), we derived two global test statistics for linkage. The approach is based on allele IBD sharing and uses all markers simultaneously. Excess IBD sharing at disease locus is modelled by a parameter α (Feingold et al., 1993). The likelihood of data is the average of conditional likelihoods of data given that a marker locus is a disease locus. Either the likelihood ratio or score statistics can be used to test for linkage. Results of a simulation study of each method's performance are presented.

In Chapters 4, 5 and 6 new methods for analysis of disease association with multiallelic markers or haplotypes are described. Chapter 4 uses the semi-Bayesian likelihood approach proposed by Terwilliger (1995) to derive a score statistic for testing genetic association. The score test is simple to compute and enables us to derive empirical p-values by means of Monte-Carlo simulations. Further, we present the results of analytic as well as empirical comparisons of the performance of the score statistic and some existing statistics for multiallelic markers including Pearson χ^2. Chapter 5 presents an application of some of the methods described in chapter 4 to simulated data from Genetic Analysis Workshop 14 (Bailey-Wilson et al., 2005). The aim of this chapter is to illustrate how these methods perform when they are applied to candidate regions initially identified by means of linkage analysis. Chapter 6 describes a generalization of the method presented in chapter 4, its performance in comparison with Pearson's χ^2 and likelihood ratio proposed by Terwilliger (1995) and their application to data on thrombosis.

Chapter 7 is concerned with (1) the use of combining information from two sources (parents and teachers) when diagnosing children with attention deficit hyperactivity disorder (ADHD) in genetic studies, and (2) studying familial aggregation among three phenotypic subtypes of ADHD in an isolated population. A test statistic is described that compares the distribution of the kinship coefficient between two samples. The kinship coefficients of all pairs in each sample are obtained using the genealogical information of 22 generations.

CHAPTER 2

Score statistic to test for genetic correlation for proband-family design

R. el Galta, C.M. van Duijn, J.C. van Houwelingen and J.J. Houwing-Duistermaat

Abstract

In genetic epidemiological studies, informative families are often oversampled to increase the power of a study. For a proband-family design, where relatives of probands are sampled, we derived the score statistic to test for clustering of binary and quantitative traits within families due to genetic factors. The derived score statistic is robust to ascertainment scheme. We considered correlation due to unspecified genetic effects and/or due to sharing alleles identical by descent (IBD) at observed marker locations in a candidate region. A simulation study was carried out to study the distribution of the statistic under the null hypothesis in small data-sets. To illustrate the score statistic, data on 33 families with type 2 diabetes mellitus (DM2) were analyzed. In addition to the binary outcome DM2, we also analyzed the quantitative outcome, body mass index (BMI). For both traits, familial aggregation was highly significant. For DM2, including also IBD sharing at marker D3S3681 as a cause of correlation gave an even more significant result, which suggests the presence of a trait gene linked to this marker. We conclude that for the proband-family design the score statistic is a powerful and robust tool for detecting clustering of outcomes.

Published in *Ann Hum Genet*. 2005; **69**:373-81

2.1 Introduction

Complex genetic traits are often determined by multiple genetic and environmental factors with small effects (Lander and Schork, 1994) necessitating large-scale studies to obtain enough power to identify genetic factors. The power can also be increased by enrichment of the sample for the presence of genetic factors via selection of families with an unusual distribution of the trait (Amos and de Andrade, 2001; Carey and Williamson, 1991; Elston, 2000; Liang and Beaty, 2000; Risch and Zhang, 1995). For binary and quantitative traits, the selected families often have a high proportion of affected individuals and of subjects with phenotypes exceeding some threshold T, respectively. In this paper, we consider a proband-family design where relatives are sampled through probands. We define a set of probands to be a set of family members chosen in such a way that the remaining relatives are not related to ascertainment (Ewens and Shute, 1986). A score statistic is derived to test for clustering of a trait due to genetic factors within these families.

In family studies, a first question to be answered is whether genetic factors play a role in the observed trait. Closely related individuals should tend to have similar outcomes compared to non or distantly related individuals. Another question of interest may be whether correlation exists due to the presence of a genetic factor located in a genomic region. If a genetic factor is linked to a marker, relatives with excess sharing of marker alleles identical by descent (IBD) should tend to have similar outcomes compared to relatives with less sharing of alleles IBD.

Since the families selected through probands are non random, the statistical methods used should take into account the ascertainment procedure (Elston and Sobel, 1979; Morton, 1959). For single ascertainment, Cannings and Thompson (1977) proposed to condition on the phenotypes of the probands. However if single ascertainment cannot be assumed, conditioning on affected probands may not be sufficient. Ewens and Shute (1986) showed that if remaining family members are related to ascertainment, the estimates of genetic parameters may be biased. They proposed to split the family in a set of probands related to ascertainment and a set of relatives not related to ascertainment. For example if nuclear families with at least one affected offspring are ascertained, the set of probands should consist of all observed offspring regardless of their affection status and the only family members not related to ascertainment are the parents. Like Ewens and Shute (1986) we also assume that family members who do not belong to the set of probands are not related to ascertainment.

For randomly selected families, Houwing-Duistermaat et al. (2003, 1995) derived the score statistic to test for correlation due to genetic factors. The test can be applied before complex genetic models are fitted. The score statistic does not assume any distribution of random effects. Adjustments for covariates can be made. In this paper we derive the score statistic for clustering of binary and quantitative traits in the proband family design by using the conditional likelihood given the phenotypes of the probands. By conditioning on the phenotypes of the individuals relevant for ascertainment, the method is robust to the ascertainment scheme (Ewens and Shute, 1986). To take into account the relationship between covariates and outcome, we propose to use data from other sources, such as large-scale epidemiological studies.

Commenges et al. (1995) proposed also a score statistic for testing familial aggregation of binary traits in families ascertained via probands. However, they only considered a random intercept model, which yields equal correlation between family members and they assumed single ascertainment. Furthermore the effects of covariates on the trait are estimated under the null hypothesis using the data on the relatives. However for larger sets of probands, it can be impossible to estimate the parameters. For example for age dependent traits, the effect of age cannot be estimated from the data if the set of probands consists of all offspring and the remaining relatives are only the parents. Moreover, Rao et al. (1988) and de Andrade and Amos (2000) showed that conditioning on the trait value of the proband is not efficient for parameter estimation. Therefore we suggest to obtain parameters from an available population based study. Commenges et al. (1995) used the normal distribution for the distribution of the statistic under the null hypothesis. As alternative, we propose the scaled χ^2 distribution. By means of simulation we study the performance of the scaled χ^2 and the normal distribution in small data-sets.

As an illustration we applied the score statistics to a sample of first-degree relatives of probands with type 2 diabetes mellitus (DM2). Probands were patients with DM2 living in the GRIP population (Genetic Research in Isolated Populations), and known to be affected by the physicians participating in GRIP. To study correlation due to sharing alleles IBD at marker positions, we used genotypes at five makers located in a region earlier identified as possibly harboring a genetic factor that plays a role in the distribution of DM2 in this set of families (Aulchenko et al., 2003). We also studied body mass index (BMI) to illustrate the use of the score statistic on quantitative traits. Age and sex specific distributions were obtained from the Rotterdam Study (Hofman et al., 1991), a large population-based follow up study.

2.2 Methods

The generalized linear mixed model (GLMM)

Let $Y_k = (Y_{k,1}, ..., Y_{k,n_k})'$ be the response vector of a family with n_k relatives. We assume that each component of Y_k has a distribution f in the exponential family with a dispersion parameter ϕ. Let $\mu_{k,i}$ be the expected value of $Y_{k,i}$ which may depend on a vector of covariates $x_{k,i}$. To model genetic correlation among family members, the following generalized linear mixed model (GLMM) is proposed

$$\begin{array}{rcl} Y_{k,i} & \sim & f(\mu_{ki}, \phi) \\ \mu_{k,i} & = & E(Y_{k,i}|x_{k,i}, u_{k,i}) = h^{-1}(x_{k,i}\beta + u_{k,i}), \end{array}$$

with h a link function (McCullagh and Nelder, 1989). For a quantitative outcome h may be the identity function and for a qualitative outcome h may be the logit function. The vector β is a vector of regression coefficients and $u_k = (u_{k,1}, ..., u_{k,n_k})'$ is a random vector with mean 0 and covariance matrix $\tau^2 R_k$. If $\tau^2 = 0$ the variables $Y_{k,i}$ are independent and the model $\mu_{k,i} = h^{-1}(x_{k,i}\beta)$ is simply a generalized linear model (McCullagh and Nelder, 1989).

For u_k equal to an additive genetic effect, the correlation structure R_k has elements $R_{k,ij}$ equal to the coefficients of relationships (Sham, 1998). The coefficient of relationships can be written as follows

$$R_{k,ij} = \pi_{ij}^2 + \frac{1}{2}\pi_{ij}^1, \tag{2.1}$$

with $\pi_{ij}^l, l = 0, 1, \text{or } 2$ the probability of sharing l alleles identical by descent (IBD) between individual i and j. When genetic markers are typed, correlation due to sharing alleles IBD at marker positions may be of interest. For a certain marker, we propose to extend the correlation structure (2.1) to the following set of correlation structures:

$$R_{k,ij} = \rho(\tilde{\pi}_{ij}^2 + \frac{1}{2}\tilde{\pi}_{ij}^1) + (1-\rho)(\pi_{ij}^2 + \frac{1}{2}\pi_{ij}^1), \text{ with } 0 \leq \rho \leq 1, \tag{2.2}$$

with $\tilde{\pi}_{ij}^l, l = 0, 1, \text{or } 2$ the conditional probability of sharing l alleles IBD between relative i and j at the marker locus, given the marker data and the family structure. Now the random effect u_k represents correlation due to tight linkage between a marker and a gene involved in the etiology of the trait and due to residual genetic factors. The proportion of the additive genetic variance explained by a locus is modelled by the parameter ρ. For $\rho = 0$, R_k equals correlation structure (2.1) and no contribution of the locus to the genetic correlation is modelled. For $\rho = 1$, R_k is the mean IBD sharing matrix at a locus and hence this model assumes that the total genetic variance is explained by this locus.

The choice of ρ depends on the population and trait studied. For a genetically homogeneous population where only one gene is expected to be involved in the etiology of the studied trait, ρ may be set to 1. For complex genetic traits studied in the general population, small values of ρ may be of interest. Alternatively plots of ρ versus the p-value may be made to study if the p-value decreases by adding the correlation due to IBD sharing to the correlation structure (1).

For randomly chosen families, Houwing-Duistermaat et al. (1995) used the GLMM above and derived the score statistic to test the null hypothesis H_0 : $\tau^2 = 0$ of no correlations between relatives of randomly chosen families. The statistic is given by

$$\sum_{1 \le k \le m} Q_k = \sum_{1 \le k \le m} (Y_k - \mu_k)' R_k (Y_k - \mu_k),$$

with m the number of families.

Score statistic for proband family design

Since the score statistic adds over independent families, it suffices to give the statistic and its distribution only for a single family. Therefore we drop the family index k. Suppose a family has n members with n_p probands related to ascertainment and $n_r = n - n_p$ remaining relatives not related to ascertainment. Let $Z_i = Y_i - \mu_i$ with μ_i the known mean. To let μ_i depend on covariates a marginal model may be used (Diggle, 1994). Now let the first n_p observations belong to the set of probands then we can write $Z' = (Z^{p\prime}, Z^{r\prime})$ with $Z^p = (Z_1, ..., Z_{n_p})'$ and $Z^r = (Z_{n_p+1}, ..., Z_n)'$. Analogously we decompose the correlation matrix R into four blocks, $R = \begin{pmatrix} R^{pp} & R^{pr} \\ R^{pr\prime} & R^{rr} \end{pmatrix}$. The logarithm of the conditional likelihood $l(Z|Z^p, \tau)$ of Z given the outcomes of the probands Z^p is

$$\begin{aligned} l(Z|Z^p, \tau) &= l(Z|\tau) - l(Z^p|\tau) \\ &= \log(E_u[f(Z|u, \tau)]) - \log(E_u[f(Z^p|u, \tau)]). \end{aligned}$$

The corresponding score statistic to test $H_0 : \tau^2 = 0$ is obtained by taking the first derivative of $l(Z|Z^p, \tau)$ with respect to τ^2 at $\tau^2 = 0$

$$\frac{\partial}{\partial \tau^2} l(Z|Z^p, \tau^2)_{|\tau^2=0} = Q - E_{|\tau^2=0}(Q),$$

with Q a quadratic form equal to the stochastic part of the first derivative plus a constant. The statistic Q can be written as follows

$$Q = Z'RZ = Z^{r\prime}R^{rr}Z^r + 2Z^{p\prime}R^{pr}Z^r + Z^{p\prime}R^{pp}Z^p.$$

Note that the stochastic part of Q can be written as a sum of two statistics $L = 2Z^{p\prime}R^{pr}Z^r$ and $Q^r = Z^{r\prime}R^{rr}Z^r$. The linear term L measures correlation between probands and relatives. The quadratic form Q^r is the score statistic applied to the relatives and ignoring the probands. If the outcome of the proband is unknown, no information is available about the correlation between the proband and the family members. Furthermore from the formula for Q it is clear that information is lost when more relatives are allocated to the set of probands (see also Ewens & Shute, 1986).

Under the null hypothesis of no correlation, the conditional expectation $E(Q|Z_p)$ and variance $Var(Q|Z_p)$ of the statistic given the outcomes of the probands are

$$E(Q|Z^p) = E(Q^r) + Z^{p\prime}R^{pp}Z^p = \sum_{i=n_p+1}^{n} R_{ii}Var(Z_i) + Z^{p\prime}R^{pp}Z^p,$$

and

$$\begin{aligned} Var(Q|Z^p) &= Var(L|Z^p) + 2Cov(L; Q^r|Z^p) + Var(Q^r) \\ &= 4\sum_{j=n_p+1}^{n}(\sum_{i=1}^{n_p} Z_iR_{ij})^2 Var(Z_j) + 4\sum_{j=n_p+1}^{n}\sum_{i=1}^{n_p} Z_iR_{ij}E(Z_j^3) \\ &\quad + \sum_{i=n_p+1}^{n} R_{ii}^2(E(Z_i^4) - 3Var(Z_i)^2) + 2\sum_{i,j=n_p+1}^{n} R_{ij}^2 Var(Z_i)Var(Z_j). \end{aligned}$$

For binomially and normally distributed outcomes, formulae for the expectation and the variance of Q are given in the appendix. Asymptotically, the statistic $\frac{Q-E(Q|Z^p)}{\sqrt{Var(Q|Z^p)}}$ follows a standard normal distribution $N(0,1)$. Alternatively, the distribution of Q under $H_0 : \tau^2 = 0$ can be approximated by a scaled chi-square distribution $c\chi^2_v$ with the scale parameter c given by $c = \frac{Var(Q|Z^p)}{2E(Q|Z^p)}$ and the degrees of freedom v given by $v = \frac{2E(Q|Z^p)^2}{Var(Q|Z^p)}$ (le Cessie and van Houwelingen, 1995).

2.3 Simulation study

In order to study the performance of the $c\chi^2$ and normal distributions as approximations of the null distribution of the score statistic, we performed a simulation study. For sake of simplicity we used the data structure of our example of 33 families (see below). We generated 100,000 data sets of independently binomially distributed outcomes and 100,000 data sets of independently normally distributed outcomes. The score statistics were calculated using correlation structure (2.1) based on the coefficients of relationship. We also studied the performance of the distributions in a very small set of nine families.

In table 2.1, the actual p-values corresponding to a nominal p-value of 0.05, 0.01, 0.001 and 0.0001 are given. The results were in favour of the $c\chi^2$ distribution for both binomially and normally distributed outcomes. Even for the set of nine families, the $c\chi^2$ distribution performed very well.

TABLE 2.1: Type I error rate when using $c\chi^2$ distribution and normal distribution as approximation for the distribution of Q under the null hypothesis. The estimates are based on 100,000 simulations.

		33 families		9 families	
	nominal	$c\chi^2$	normal	$c\chi^2$	normal
Binomial (DM2)					
	0.05	0.0547	0.0606	0.0550	0.0649
	0.01	0.0143	0.0194	0.0137	0.0239
	0.001	0.0020	0.0041	0.0017	0.0070
	0.0001	0.0004	0.0011	0.0002	0.0019
Normal (BMI)					
	0.05	0.0538*	0.0615*	0.0566	0.0651
	0.01	0.0125*	0.0196*	0.0151	0.0233
	0.001	0.0016*	0.0047*	0.0027	0.0069
	0.0001	0.0002*	0.0011*	0.0004	0.0023

* based on 27 families

A data example

Description of families

To illustrate the score statistic, we used data from 79 patients with type 2 diabetes mellitus (DM2), their first-degree relatives and spouses (Aulchenko et al., 2003). These families were derived from the GRIP population (Genetic Research in Isolated Populations), an isolated village in the Southwest of the Netherlands. The GRIP population is described in detail elsewhere (Aulchenko et al., 2003; Vaessen et al., 2002; van Duijn et al., 2001). Probands are patients with DM2 treated by physicians participating in GRIP. Among the relatives are patients not related to ascertainment namely patients of other physicians and subjects who did not know that they have DM2. In a combined linkage and association study, a genome scan was carried out on these data and Aulchenko et al. (2003) found a borderline association between marker D3S3681 and DM2 (LOD score of 1.20, P=0.01).

For DM2 we analysed 33 families informative for linkage. One of these families was a combination of two nuclear families. Three families had probands with unknown disease status. In total 31 probands and 65 relatives were observed. The percentage of women was 60%. The mean age in years was 62 (range 45-94). We did not use subjects younger than 45 years, because we do not have information on the prevalence of DM2 for these age groups. In table 2.2 the number of families for combinations of number of affected relatives and number of observed relatives in the family are given. The mean size of the families was 2.9 (range 2 to 5) and the number of affected relatives per proband was 0.72 (range 0 to 4). The quantitative outcome body mass index (BMI) was known for a subset of 27 families with 46 relatives. BMI was only known for 7 probands. In this subset, the distributions of age and sex agreed with those of larger set of 33 probands. Also for this outcome we assumed that the family members are not related to ascertainment if they are not probands.

The age and gender specific distributions of DM2 and BMI were obtained using data from the Rotterdam Study (Hofman et al., 1991). The Rotterdam Study is a population based follow up study of the elderly with about 8000 subjects aged 55 and over. For both sexes, we fitted logistic and linear regression models to estimate the relationschip between age and DM2 and BMI respectively. The following marginal models were obtained

TABLE 2.2: Counts of families per combinations of number of affected relatives and number of observed relatives in the family.

		Number of relatives				
		1	2	3	4	total
	0	11	3	3	0	17
Number	1	3	4	3	0	10
of affected	2	-	1	3	1	5
relatives	3	-	-	0	0	0
	4	-	-	-	1	1
	total	14	8	9	2	33

$$logit(\mu(DM2)) = \begin{cases} -4.379 + 0.035 * age & \text{for women} \\ -3.529 + 0.025 * age & \text{for men} \end{cases} \quad (2.3)$$

$$\text{and } \mu(BMI) = \begin{cases} 25.48 + 0.018 * age & \text{for women} \\ 28.10 - 0.036 * age & \text{for men} \end{cases} \quad (2.4)$$

and $\sigma^2(BMI) = 13.62$. We used these models also for the distributions of DM2 and BMI for subjects aged between 45 and 55 years. In table 2.3, the observed and expected prevalence of DM2 and the observed and expected mean of BMI in the relatives are given. The expected values were computed using model (2.3) and (2.4) respectively. The prevalence of DM2 and the mean of BMI was higher than expected.

The conditional probabilities of sharing zero, one or two alleles IBD at marker D3S3681 and four informative proximal markers namely D3S1276, D3S3634, D3S1603, and D3S1271 were computed using the multipoint option in GENEHUNTER (Kruglyak et al., 1996) and using all available family members regardless of their age. Unfortunately no informative distal marker was available. The genetic distances between adjacent markers are 2.67, 2.67, 0.53, and 2.67 cM successively. The markers appeared to be highly informative (> 0.89), using the entropy as a measure of the informativeness (Kruglyak et al., 1996), hence also the Spearman's rank correlations between the estimated proportion of alleles shared IBD at each marker locus and the coefficient of relationship are rather small (< 0.53). The Spearman's rank correlations between the estimated proportion of alleles shared IBD at pairs of

TABLE 2.3: Observed and expected prevalence of DM2 and mean of BMI

	Observed	Expected*
Prevalence of DM2		
women (n=40)	0.35	0.10
men (n=25)	0.40	0.11
Mean of BMI (standard error)		
women (n=28)	29.27 (0.91)	26.40
men (n=18)	28.63 (0.93)	25.98

* expected values are based on the Rotterdam Study

markers varied from 0.74 (D3S3681 and D3S1271 at distance 8.54 cM) to 0.90 (D3S3634 and D3S1603 at distance 0.53 cM). Due to recombination between two physically close located markers D3S3634 and D3S1603, the estimated proportion of alleles shared IBD differed between these marker loci.

Results

We applied the score statistics to test for clustering of DM2 and BMI due to genetic factors. Correlation structure (2.1) based on familial relationship appeared to be highly significant for both traits ($P < 0.00001$). For testing correlation structure (2.2) for the five markers, plots of ρ versus minus log_{10} of p-value are given in figure 2.1 for DM2 and in figure 2.2 for BMI. All p-values were highly significant ($P < 0.00001$). Especially for DM2, adding correlation due to sharing allele IBD at marker D3S3681 to the familial correlation decreased the p-value for clustering.

2.4 Discussion

In this paper we proposed a score statistic for the proband family design to test for the presence of a prespecified correlation structure for binary and quantitative outcomes. The score statistic allows for adjusting of covariates. No assumption about the distribution of the random effects is made. Furthermore by conditioning on the trait value of all individuals related to ascertainment the method is robust to the ascertainment scheme. By means of a simulation study we showed that the $c\chi^2$ distribution performs well as an approximation of the distribution of the score statistic under the null hypothesis even in very small data-sets.

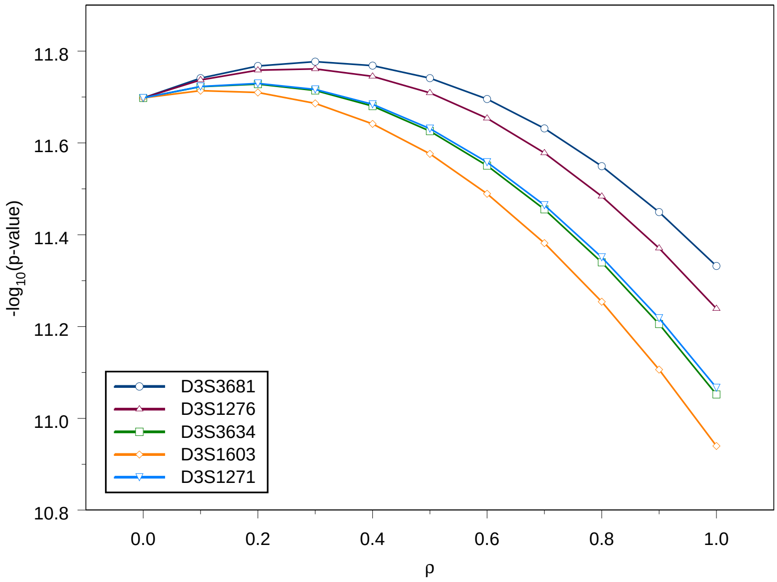

FIGURE 2.1: *For DM2 the p-values for testing correlation due to sharing alleles IBD at the five marker positions and residual genetic correlation for $\rho = 0, 0.1, ..., 1$. The parameter ρ models the proportion of genetic variances explained by IBD sharing.*

We analysed the clustering of DM2 and BMI in families of DM2 cases. Age and sex specific distributions of DM2 and BMI were obtained from the Rotterdam Study. The number of DM2 cases was higher than expected taking into account the age and sex distributions. Also the mean BMI was higher than expected in these families. This may indicate that genetic factors play a role in these families. Application of the score statistics indeed showed significant familial clustering of DM2 and BMI. Furthermore for DM2 adding IBD sharing at the five marker locations decreased the p-value. This decrease was most pronounced for marker D3S3681, which also showed some association with DM2 (Aulchenko et al., 2003). The next step in analysing these data will be estimation of the parameters. However the methodology has to be developed (see below).

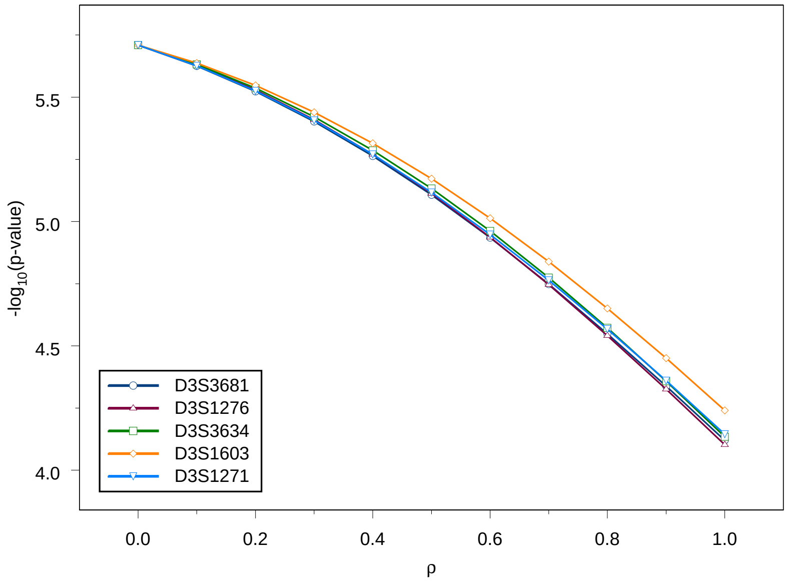

FIGURE 2.2: *For BMI the p-values for testing correlation due to sharing alleles IBD at the five marker positions and residual genetic correlation for $\rho = 0, 0.1, ..., 1$. The parameter ρ models the proportion of genetic variances explained by IBD sharing.*

In this paper we derived formulae for binomially and normally distributed outcomes. However the score statistic can be used for any distribution belonging to the exponential family. Furthermore we restricted ourselves to correlation due to additive genetic effects because for many complex traits, dominant effects are assumed to be small (Risch, 1990a). If dominant effects do exist the power will be only slightly reduced, because dominant effects only influence the correlation among pairs who can share two alleles IBD.

In addition to correlation due to any genetic component we also considered correlation partly due to excess sharing of alleles IBD in candidate regions. We made plots of ρ, the proportion of genetic variance explained by the locus versus the p-value. A decrease of the p-value at $\rho = 0$ suggests a role of a gene involved in the etiology of the trait linked to the marker locus. Note that our statistic does not test the null hypothesis of no linkage. A formal test for linkage is a score statistic for $H_0 : \rho = 0$. For quantitative traits, this score statistic corresponds to the statistic derived by Putter et al. (2002). However for binary traits, derivation of the score statistic is complex due to the non linear relation between the outcome and the random effects. Nevertheless for both quantitative and binary outcomes, the test statistic Q provides insight in the underlying correlation structure and should be used before genetic parameters are estimated.

When a trait appears to be significantly correlated, the next step is to estimate the genetic parameters modelling the covariance structure. A natural framework of estimation methods are generalized estimating equations (GEE), because they do not fully specify the distribution (see Tregouet and Tiret (2000) and Ziegler et al. (1998) for reviews on application of these methods to family studies). For quantitative traits observed in random families, Stram et al. (1993) proposed GEE for segregation analysis. For binary traits, Liang and Beaty (1991) used these methods to study the dependence within families under the assumption that the families sampled are geometrically proportional to the number of affected family members. For a case-control family design, Zhao et al. (1998) derived GEE corresponding to the likelihood proposed by Whittemore (1995). To correct for ascertainment, Whittemore (1995) proposed to use different intercepts for probands than for the remaining family members. Further research is needed to extend these methods to the more general selection scheme as considered in this paper.

Methods of estimation should allow for more than one proband per family and also for different relationships between probands and relatives. Furthermore under the alternative, estimation of the parameters modelling the mean from data on the relatives may be biased (Pfeiffer et al., 2001). Hence to adjust for covariates, estimates of parameters should be obtained from other sources. Note that biased estimates of the regression parameters will affect the correlation between residuals (Diggle and Zegger (1994, p. 63-64); Verbeke and Molenberghs (1997, p. 120-122)) and consequently, the estimates of parameters modelling the covariance.

The statistic Q measures deviation from the mean as well as clustering. Hence if the mean is invalid the type I error is inflated. Commenges et al. (1995) proposed to estimate the parameters from the data on the relatives, which is valid under the null hypothesis. However for large sets of probands estimation of the parameters modelling the mean may not be possible. Furthermore for estimation of the genetic parameters, the means should be known as pointed out above. It is natural to use estimates from other sources when our statistic is applied before models are fitted. Therefore we feel that it is important to know the effects of covariates on the trait in studied populations and to use this knowledge in analysing selected families aiming to elucidate the underlying genetic mechanisms. For our data example obtaining the age and gender distribution of DM2 and BMI from the Rotterdam Study seems to be reasonable since GRIP is a recently isolated population. We conclude that the score statistic is a good tool to study clustering of traits due to genetic factors within families selected via probands.

The analysis was performed using S-plus codes, which are available from: http://www.medstat.medfac.leidenuniv.nl/MS/

2.5 Appendix

For $Y_i \sim N(\mu_i, \sigma^2)$ the expectation of Q given Z^p is

$$E(Q|Z^p) = \sigma^2 trace(R^{rr}) = n\sigma^2,$$

and the variance of Q given Z^p is

$$Var(Q|Z^p) = 2\sigma^4 trace((R^{rr})^2) + 4\sigma^2 \sum_{j=n_p+1}^{n} (\sum_{i=1}^{n_p} Z_i R_{ij})^2.$$

For $Y_i \sim Bin(1, \mu_i)$ the expectation of Q given Z^p is

$$E(Q|Z^p) = \sum_{i=n_p+1}^{n} \mu_i(1-\mu_i),$$

and the variance of Q given Z^p is

$$\begin{aligned} Var(Q|Z^p) &= 4 \sum_{j=n_p+1}^{n} (\sum_{i=1}^{n_p} Z_i R_{ij})^2 \mu_j(1-\mu_j) \\ &+ \sum_{i=n_p+1}^{n} \mu_i(1-\mu_i)(1-6\mu_i+6\mu_i) \\ &+ 2 \sum_{i,j=n_p+1}^{n} (R_{ij})^2 \mu_i\mu_j(1-\mu_i)(1-\mu_j) \\ &+ 4 \sum_{j=n_p+1}^{n} \sum_{i=1}^{n_p} R_{ij} Z_i \mu_j(1-\mu_j)(1-2\mu_j). \end{aligned}$$

CHAPTER 3

Global tests for linkage

R. el Galta, J.C. van Houwelingen and J.J. Houwing-Duistermaat

Abstract

To test for global linkage along a genome or in a chromosomal region, the maximum over the marker locations of mean alleles shared identical by descent of affected relative pairs, Z_{max}, can be used. Feingold et al. (1993) derived a Gaussian approximation to the distribution of the Z_{max}. As an alternative we propose to sum over the observed marker locations along the chromosomal region of interest. Two test statistics can be derived. (1) The likelihood ratio statistic (LR) and (2) the corresponding score statistic. The score statistic appears to be the average mean IBD over all available marker locations. The null distribution of the LR and score tests are asymptotically a 50:50 mixture of chi-square distributions of null and one degree of freedom and a normal distribution, respectively.

We compared empirically the type I error and power of these two new test statistics and Z_{max} along a chromosome and in a candidate region. Two models were considered, namely (1) one disease locus and (2) two disease loci. The new test statistics appeared to have the right type I error. Along the chromosome, for both models we concluded that for very small effect sizes, the score test was more powerful than the other test statistics. For large effect sizes, the likelihood ratio statistic and Z_{max} were comparable and performed much better than the score test. For candidate regions of about 30 cM, all test statistics were comparable.

To be submitted for publication

3.1 Introduction

In complex genetic diseases, multiple genes are assumed to cause a predisposition to disease. Each single susceptibility gene might contribute little to disease, and therefore the statistical power to detect such a gene is low, especially if the gene is common and has low penetrance. Furthermore, some of these genes might lie on the same chromosomal region of interest.

Regions harboring genes responsible for a trait are often identified by means of genome wide linkage analysis, which studies co-segregation of an unobserved disease locus and a marker locus. Two approaches can be considered, namely parametric and nonparametric linkage analysis. Parametric methods require allele frequencies at the disease locus and the penetrances to be known. Parametric linkage analysis is the most powerful when the genetic parameters are correctly specified. For many complex traits the mode of inheritance is unknown. For such traits non-parametric methods are suitable, as they do not make any assumption about the mode of inheritance. Nonparametric methods rely only on the information of sharing alleles identical by descent (IBD) between relatives at a locus to study whether it is genetically linked to the unobserved disease locus. Linkage between a disease locus and marker genotypes can be studied by comparing the observed IBD sharing of affected relative pairs to the expected IBD sharing under random segregation. An increase in IBD sharing indicates the presence of a susceptibility gene in the region. Conventionally testing for linkage is carried out at each observed locus throughout the human genome. To adjust for multiple testing, only p-values smaller than 2.2×10^{-5} are considered to be significant (Lander and Kruglyak, 1995). In this paper we propose two global tests for linkage, which use IBD information from observed markers all together. All tests considered in this paper, assume that each of unlinked chromosomal regions carries one disease locus at most.

A global test for linkage, which tests all observed markers simultaneously, can be obtained by summing the likelihood of the data over all marker locations along the region of interest. Throughout this paper, we will refer to this approach as the averaging approach. Siegmund (2001) discussed briefly the averaging approach for complete IBD information. However, we think that this approach merits more consideration. For a candidate region, Liang et al. (2001) proposed a generalized estimating equations approach to primarily estimate the location of a disease gene. The authors used IBD information from all markers jointly. Liang's method can also be regarded as an averaging approach.

For various types of affected relative pairs, and dense and fully informative markers about IBD status, Feingold et al. (1993) proposed a global test for linkage, which appeared to be the maximum of mean IBD over all markers (Z_{max}). The authors derived a Gaussian approximation to the significance level of Z_{max} for large sample sizes. Teng and Siegmund (1998) extended this approach to relative pairs with partial information about the IBD status.

In this paper, we considered the averaging approach for both complete and partial information about the IBD status. For simplicity, we restricted to affected sibling pairs (ASP). Two test statistics were derived namely the likelihood ratio statistic and the corresponding score statistic. The score statistic appeared to be the average mean IBD over all available marker locations. The null distribution of the likelihood ratio and score statistic are asymptotically a 50:50 mixture of chi-square distributions of null and one degree of freedom and a normal distribution, respectively.

For complete IBD information, we compared empirically the type I error and power of these two test statistics and Z_{max}. To generate data two models were considered, namely (1) single-locus disease model and (2) two-locus disease model. The new test statistics appeared to have the right type I error. For both models we concluded that for a sample of 200 ASPs and a small effect sizes ($\lambda_s < 1.17$, with λ_s the siblings relative recurrence risk (Risch, 1990a)), the score test had slightly more power than the other test statistics. For large effect sizes or large sample sizes, the likelihood ratio statistic and Z_{max} were comparable and perform better than the score test. Further, we studied the effect of the information loss on the performance of the test statistics when only partial information is available.

3.2 Methods

Complete IBD information

First we describe briefly the approach proposed by Feingold et al. (1993). Let T be the length (in cM) of a chromosome of interest. Suppose that we have N affected sib pairs. For a marker locus at the position t let $X_{t,i}^k$ be the event that the members of the i^{th} sib-pair share k alleles identical by descent (IBD), for $k = 0, 1, 2$ and let $X_{t,i} = \sum_{k=0}^{2} k X_{t,i}^k$ the number of marker alleles shared IBD. Assume that all markers are fully informative about the IBD status. Hence $X_{t,i}$ has expectation $E_0[X_{t,i}] = 1$ and variance $Var(X_{t,i}) = 1/2$ under the null hypothesis of no linkage. Let $X_t = \sum_{i=1}^{N} X_{t,i}$

and $Z_t = \frac{X_t - N}{\sqrt{N}}$. Assuming Haldane's mapping function, $\{Z_t, 0 \leq t \leq T\}$ is approximately a Gaussian Markov process which has zero mean and covariance $R(t,s) = 1/2 \exp(-0.04|t-s|)$ under the null hypothesis. Under the alternative hypothesis of the presence of one susceptibility gene at the location τ on the chromosome, this process $\{Z_t, 0 \leq t \leq T\}$ is superimposed by $0.5\sqrt{N}\alpha \exp(-0.04|t-\tau|)$ with α representing the excess IBD sharing. The parameter α varies between 0 and 1. For an additive model, $\alpha = \frac{\lambda_s - 1}{\lambda_s}$ with λ_s, the sibling risk ratio (Risch, 1990a). For small α an approximation of the log likelihood of this process is

$$L(\tau, \alpha | Z_t, 0 \leq t \leq T) = L(\alpha | Z_\tau) \propto exp(\sqrt{N}\alpha Z_\tau - N\alpha^2/4). \tag{3.1}$$

It is not known which locus is the disease locus. To test the null hypothesis $H_0 : \alpha = 0$ i.e. none of the markers is linked with the disease, Feingold et al. (1993) proposed to use the maximum of the corresponding likelihood ratio over the parameters α and τ, which appeared to be

$$Z_{max} = \sqrt{\max_\tau \max_\alpha 2logL(\tau, \alpha)} = \max_\tau \sqrt{2} Z_\tau.$$

Feingold et al. (1993) derived the following approximation to calculate the significance level of Z_{max}

$$P_0(Z_{max} > b) \approx 1 - \Phi(b) + 0.04T\varphi(b)$$

with φ and Φ are the standard normal density and distribution functions, respectively.

Averaging approach

As an alternative to maximizing the conditional likelihood over τ one can take the average of conditional likelihoods given the location of the disease locus over all marker loci assuming that they are all equally likely to be in complete linkage with the disease locus. Hence the average likelihood is approximately

$$L(\alpha | Z_t, 0 \leq t \leq T) \propto \sum_t L(\tau = t, \alpha) = \sum_t exp(\sqrt{N}\alpha Z_t - N\alpha^2/4). \tag{3.2}$$

Here, we assume also the presence of a single disease locus in the region of interest. As test statistic we propose to use either the corresponding likelihood ratio test

$$\Lambda = \max_{0 \leq \alpha \leq 1} 2\log(\frac{L(\alpha)}{L(\alpha = 0)}) = \max_{0 \leq \alpha \leq 1} 2\log(\sum_t exp(\sqrt{N}\alpha Z_t - N\alpha^2/4)) \tag{3.3}$$

or the score test

$$U = \frac{\partial \log(L(\alpha = 0))/\partial\alpha}{\sqrt{-E[\partial^2 \log(L(\alpha = 0))/\partial\alpha^2]}} = \frac{\sum_t Z_t}{\sqrt{\sum_{t,s} Cov_0(Z_t, Z_s)}}, \tag{3.4}$$

with $L(\alpha)$ given in formula (3.2) and $\sum_{t,s} Cov_0(Z_t, Z_s) = \sum_{t,s} R(t,s)/2$, the variance under the null hypothesis. Since the parameter α is positive, the null distribution of Λ is approximately a 50:50 mixture of χ^2_1 distribution and a point mass at zero ($\frac{1}{2}\chi^2_1 + \frac{1}{2}\chi^2_0$). The score test U follows asymptotically the normal distribution under the null hypothesis. It is a one-sided test and rejects the null hypothesis only for positive values of U. Let $U^2_+ = U^2$ if $U > 0$ otherwise $U^2_+ = 0$. Then U^2_+ is asymptotically distributed as $\frac{1}{2}\chi^2_1 + \frac{1}{2}\chi^2_0$. Note that U^2_+ approximates Λ for α small. Hence Λ and U are locally asymptotically equivalent. The theoretical power of the score test U is given by

$$\beta = 1 - \Phi(\frac{b\sqrt{Var_0(Z)} - E_\alpha[Z]}{\sqrt{Var_\alpha(Z)}}), \tag{3.5}$$

with $Z = \sum_t Z_t$, and $E_\alpha[Z]$ and $Var_\alpha[Z]$, the expectation and the variance of Z under the alternative hypothesis. The critical value b is the normal percentile corresponding to a prespecified significance level. For instance $b = 1.64$ corresponds to a significance level of 0.05. The covariance matrix of the process $\{Z_t, 0 \leq t \leq T\}$ is given in the appendix for two models, namely (1) one disease locus on the chromosome and (2) two disease loci on the same chromosome. For the likelihood ratio test Λ, the power can be estimated by means of a simulation study. The value of α that maximize the average likelihood can be used as an estimate of the effect size of the susceptibility gene. The posterior probabilities are of interest; for instance the largest posterior probability indicates the most likely location of a disease gene.

Incomplete IBD information

Since genotyped markers are typically not fully informative about the IBD status, the IBD sharing $X_{t,i}$ can be replaced by $\hat{X}_{t,i} = E_0[X_{t,i}|G_i]$ the mean IBD under the null hypothesis of no linkage given the genotypes, G_i, of the ith sibpair for all observed markers. Given the disease locus at the location τ Teng and Siegmund (1998) used the following likelihood to derive a corresponding test statistic to Z_{max} for incomplete IBD information

$$\widetilde{L}(\tau, \alpha|G) \propto \prod_i^N (1 + \alpha(\hat{X}_{\tau,i} - 1)). \tag{3.6}$$

Similarly to the case of complete IBD information the average likelihood can be obtained by summing over all available marker positions. Hence the average likelihood is

$$\widetilde{L}(\alpha) \propto \sum_t \prod_i^N (1 + \alpha(\hat{X}_{t,i} - 1)). \tag{3.7}$$

The likelihood ratio corresponding to $\widetilde{L}(\alpha)$

$$\hat{\Lambda} = \max_{0 \leq \alpha \leq 1} 2log(\widetilde{L}(\alpha)/\widetilde{L}(\alpha = 0))$$

Let $\hat{Z}_t = \sum_{i=1}^N (\hat{X}_{t,i} - 1)/\sqrt{N}$. An estimate of the score test is

$$\hat{U} = \frac{\sum_t \hat{Z}_t}{\sqrt{\sum_{t,s} \hat{Cov}(\hat{Z}_t, \hat{Z}_s)}}$$

where the covariances are the sample covariances. The likelihood ratio $\hat{\Lambda}$ follows approximately a 50:50 mixture of χ_1^2 distribution and a point mass at zero. The score test $\hat{U}$ asymptotically follows the normal distribution.

3.3 Simulation

Complete IBD information

Single-locus disease model

Whole chromosome

For practical reasons we simulated data from a multivariate normal distribution instead of generating the process of number of IBD sharing for affected sib pairs. We generated a vector of length 100, which corresponds to IBD data on 100 equidistant markers along a chromosome of 300 cM, from a multivariate normal distribution with the mean and covariance being evaluated under the alternative hypothesis (Liang et al., 2001). Data at each position were considered as an average IBD sharing of 200 affected sib-pairs. We repeated the procedure 10,000 times. We positioned one disease gene at 75 cM on the chromosome, with varying values of α. The results are summarized in Figure 7.1. Type I error and power were calculated at significance level of 0.05. All test statistics provided reasonable type I error rates. For small effect size ($\alpha \leq 0.15$), the score statistic U appeared to have slightly more power than

Z_{max} and Λ. For larger effect size ($\alpha > 0.15$), the test statistics Λ and Z_{max} had similar power and performed better than U. From the simulation results we observed that Λ_s and Z_{max} statistics attained the power of 80 % at $\alpha \approx 0.37$, which corresponds to $\lambda_s = 1.59$. Hence, the sample size N required to achieve the power of 80 % for a given α, can be approximated by using the following formula

$$N = 200(0.37/\alpha)^2.$$

For example the sample size required to achieve a power of 80 % when $\alpha(\lambda_s) = 0.05$ (1.05), 0.1 (1.11), 0.15 (1.17) and 0.23 (1.3) are 10765, 2692, 1196 and 500 ASPs, respectively. The sample size required to attain a power of 80 % can be approximated by the using the corresponding formula to the formula (3.5). The score test U achieved a power of 80 % at $\alpha(\lambda_s) = 0.6(2.5)$. The sample size required to achieve a power of 80 % when $\alpha(\lambda_s) = 0.05$ (1.05), 0.1 (1.11), 0.15 (1.17) and 0.23 (1.3) are 28800, 7200, 3200 and 800 ASPs, respectively. Similar results were also obtained by using

$$N = 200(0.6/\alpha)^2.$$

Note that these results are only for one chromosome, and thus the sample sizes required for genome scan are much higher, see also Cordell (2001).

Candidate region

To study the performance of the test statistics in a candidate region, we generated 10000 data sets. Each data set consists of fully IBD information for 500 ASPs, on 10 equidistant markers spanning a chromosomal region of 30 cM. The disease gene was positioned at the middle of the chromosomal region, with varying values of α. The results are depicted in Figure 3.2. A nominal significance level of 0.05 is considered. All test statistics were comparable in terms of the power.

Two-locus disease model

In order to study the robustness of these test statistics we considered the presence of two disease loci on the same chromosome. Data were generated similar to the case of one disease locus from a multivariate normal distribution with the mean and covariance matrix as given in the appendix (3.5) for various values of α_1 and α_2, the parameters of increased IBD at the first and the second disease locus, respectively. Unlike for the single-locus disease model, the IBD

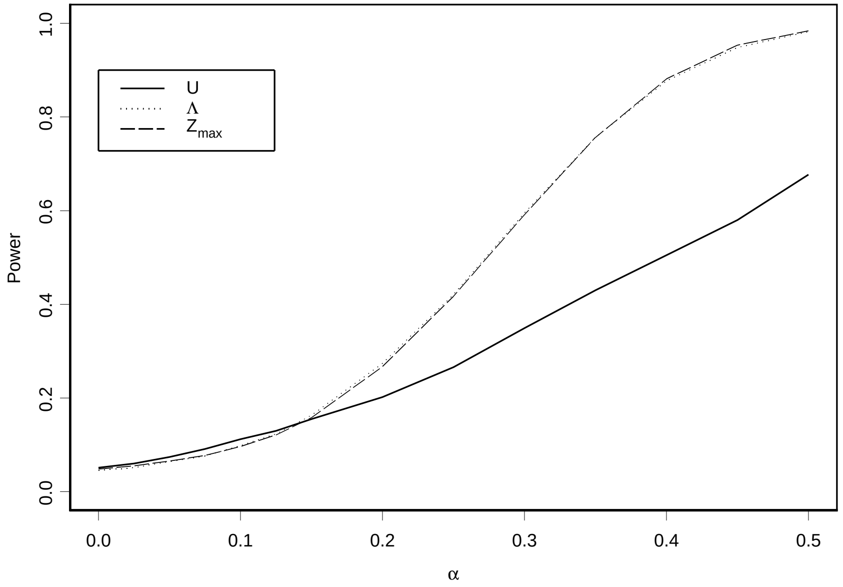

FIGURE 3.1: *Power to detect linkage due to a single disease locus at $\tau = 75$ cM when IBD data are available on 100 fully equidistant markers along a chromosome of length 300 cM for 200 ASPs.*

process is now controlled by α_1, α_2, and the covariance matrix of IBD at the disease loci through the penetrance matrix and the allele frequencies at the disease loci. The disease loci were located at $(\tau_1, \tau_2) = (90, 150)$. The results are shown in Table 3.1. A nominal significance level of 0.05 is used. Compared to the single-locus disease model, all test statistics appeared to gain power when two disease loci exist. The power of the score statistic U was especially improved. For effect sizes $\alpha_1 \leq 0.10$ and $\alpha_2 \leq 0.10$ the score statistic yielded the highest power. For larger effect sizes the corresponding likelihood ratio statistic Λ often performed the best. The Z_{max} statistic has good power relative to U.

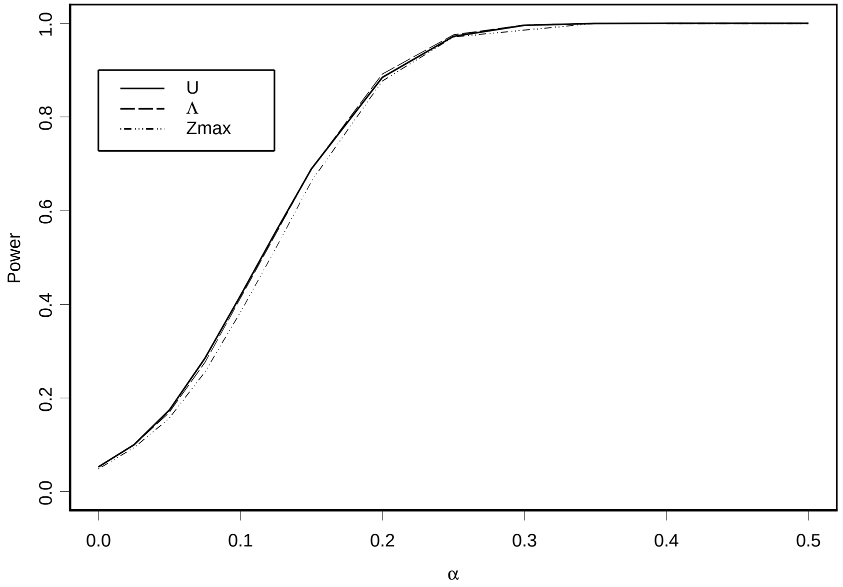

FIGURE 3.2: *Power to detect linkage due to a single disease locus at $\tau = 15$ cM when IBD data are available on 10 fully equidistant markers along a chromosome of length 30 cM for 500 ASPs.*

Partial IBD information

We generated 10000 data sets under the null model and 1000 data sets under the alternative model of one disease locus, using the ALLEGRO program (Gudbjartsson et al., 2000). Each data set consisted of 200 affected sib-pairs and their parents. We considered a chromosome with a length of 315 cM. Genotypes were simulated for 41 markers spaced about 7.8 cM on average. The disease locus was located at $\tau = 75$ cM. The two adjacent observed markers to τ were located at $t_9 = 73$ cM and $t_{10} = 81$ cM. We varied α from 0 to 0.3. Multipoint IBD's were calculated using the Merlin program (Abecasis et al., 2002). The simulation results are depicted in Figure 3.3 for affected sib-pairs, with and without information on parental genotypes at the left and right panel, respectively. A nominal significance level of 0.05 is used. Here

TABLE 3.1: Power to detect linkage due to two disease loci located on the same chromosome at $\tau_1 = 90$ and $\tau_2 = 150$, when IBD data are available on 100 fully informative markers in 200 ASPs.

p_1	p_2	α_1	α_2	λ_s	U	Λ	Z_{max}
0.033	0.033	0.05	0.05	1.051	0.10	0.07	0.07
0.05	0.033	0.10	0.05	1.082	0.14	0.12	0.12
0.05	0.05	0.10	0.10	1.113	0.18	0.16	0.15
0.063	0.031	0.15	0.05	1.113	0.19	0.21	0.21
0.064	0.05	0.15	0.10	1.146	0.23	0.24	0.22
0.065	0.065	0.15	0.15	1.183	0.28	0.30	0.28
0.08	0.066	0.20	0.15	1.226	0.35	0.44	0.42
0.081	0.081	0.20	0.20	1.267	0.44	0.54	0.51
0.092	0.068	0.25	0.15	1.267	0.43	0.59	0.59
0.096	0.084	0.25	0.20	1.317	0.48	0.65	0.63
0.1	0.1	0.25	0.25	1.370	0.55	0.70	0.67

p_1 and p_2 are the frequencies of disease alleles A and B at loci τ_1 and τ_2 respectively. The data were generated under genetic models with the following penetrance matrix

	Genotypes at τ_1		
Genotypes at τ_2	AA	Aa	aa
BB	0.95	0.95	0.95
Bb	0.95	0.09	0.09
bb	0.95	0.09	0.09

we compared the score test $\hat{U}$, the likelihood ratio $\hat{\Lambda}$ and the maximum of the score statistics $\hat{Z}_{max}$ proposed by Teng and Siegmund (1998). When parental genotypes were available, all test statistics showed the same pattern as for the perfect IBD information. The score test $\hat{U}$ had the highest power for effect sizes $\alpha < 0.15$, and for $\alpha \geq 0.15$ the test statistics $\hat{\Lambda}$ and $\hat{Z}_{max}$ performed similarly and had the highest power. When parental genotypes were not available, all test statistics appeared to be conservative and the power decreased. The type I error rates of $\hat{U}, \hat{\Lambda}$ and $\hat{Z}_{max}$ were about 0.048, 0.045 and 0.052 when parental genotypes were available respectively, and they dropped to 0.04, 0.42 and 0.039 when parental genotypes were not available. In terms of the power the $\hat{Z}_{max}$ statistic suffered the most from the loss of IBD information, whereas the score statistic was the least affected. For small $\alpha < 0.15$ the score statistic

$\hat{U}$ had the highest power, and for $\alpha \geq 0.15$ the likelihood ratio test Λ had the highest power.

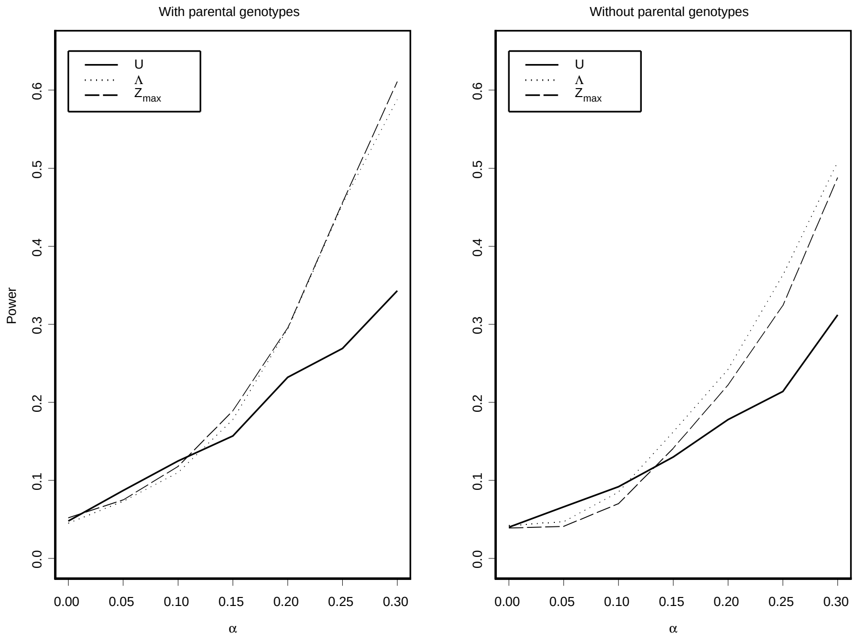

FIGURE 3.3: *Power to detect linkage due to a single disease locus when IBD information is partially available on 41 markers along a chromosome of length 315 cM in 200 ASPs.*

3.4 Discussion

In this paper, we proposed an averaging approach to test for linkage of an unobserved disease locus and a marker locus when IBD data are available on multiple markers. We first considered marker data to be fully informative about the IBD status. We assumed that all observed marker loci are equally likely to be the disease locus. As a basic model to our approach we used the model proposed by Feingold et al. (1993). The model assumes the presence of one disease locus at most on a chromosome. The likelihood of the data was computed by summing the conditional likelihood given a marker locus being the disease locus, over all observed marker loci. The corresponding likelihood ratio test or the score test can be used. Both test statistics are easy to compute and have known asymptotic distributions under the null hypothesis. Further, we adapted the method to the case of partial information on the IBD status.

We performed a simulation study to compare the performance of the averaging approach in comparison to the maximising approach (Feingold et al., 1993; Teng and Siegmund, 1998). For complete IBD information we considered a single-locus disease model as well as a two-locus disease model. The score test of the averaging approach appeared to perform slightly better when each single susceptibility gene contributed little to the disease and the sample size was small. However, this difference in power may disappear if the sample size is large. For large effect sizes ($\alpha > 0.15$), the likelihood ratio test of the averaging approach performed best when two disease loci existed on the same chromosome or when the IBD information was partial, and it had similar power to the maximising approach when IBD information was complete. In a candidate region the averaging approach has slightly higher power to detect linkage relative to the maximising approach. The score test U and the likelihood ratio Λ performed equally.

When information about IBD status is not fully known, the amount of IBD sharing is often estimated by its expectation given the available genotypes data from all observed markers assuming linkage equilibrium. However this assumption may not be valid if the marker map is dense. The less the IBD information the less accurate are the IBD estimates. Further, the variance of the estimate is unknown and it should in turn be estimated by its sample variance (Teng and Siegmund, 1998). Unlike the $\hat{Z}_{max}$ and $\hat{U}$ statistics , the likelihood ratio $\hat{\Lambda}$ does not need to estimate the variance, and therefore it may be less affected by the loss of IBD information when sample size is large. To reduce the bias one should include all available parental genotypes in analysis (Risch, 1990c). Recently, Bacanu (2005) proposed an approach to eliminate the bias due to the presence of linkage disequilibrium between adjacent markers. The author partitioned the markers into interlaced and non-overlapping subsets, and then analyzed each set separately. A final test statistic for linkage is the standardized average of subset-specific statistics.

Liang et al. (2001) proposed a GEE method to estimate the location of a single susceptibility gene in a candidate region. The authors proposed also a GEE-based test statistic of linkage. The test statistic appeared to be a weighted sum of the number of IBD sharing over the markers positions along the region of interest. This statistic assigns more weight to markers at the ends of the chromosomal region of interest. Power calculations using the formula (3.5) and the corresponding formula in Liang et al. (2001) (data not shown) show that the score test proposed in this paper has more power than the GEE-based test statistic as long as the susceptibility gene does not lie at the end of the chromosomal region of interest. Moreover the GEE-based test statistic becomes very conservative relative to the test statistic $\hat{U}$ when parental genotypes are missing (data not shown), which is in agreement with the findings of Lebrec et al. (2005).

Biernacka et al. (2005) extended the work of Liang et al. (2001) to address the presence of two disease loci on the same chromosome. The averaging approach can also be extended to test for the presence and estimate the effect sizes of two disease loci on the same chromosome. The likelihood can be obtained by taking the sum of the conditional likelihoods given two disease loci over all marker pairs. the corresponding likelihood ratio asymptotically follows a 0.25, 0.5 and 0.25 mixture of chi squares with zero, one and two degrees of freedom, respectively (Self and Liang, 1987) Moreover, the application of the approach for other relative pairs, i.e. half sibling, grandparent-grandchild, cousin pairs, etc, is straightforward.

We conclude that the averaging approach improves the power to detect linkage relative to the maximising approach when a single disease-locus of small effect size exists on a chromosome or two disease loci lie on the same chromosome.

3.5 Appendix

The expectation and the covariance of the process $\{Z_t, 0 \leq t \leq T\}$ when two disease loci exist

Suppose that two loci disease loci are located at τ_1 and τ_2. Let $\alpha_1 = E[X_{\tau_1,i} - 1]/2$ and $\alpha_2 = E[X_{\tau_2,i} - 1]/2$ be the deviation of the expectation of the number of IBD sharing at τ_1 and τ_2 from the mean under the null hypothesis. Using the fact that given the IBD at the disease loci the IBD process on the same chromosome does not involve α_1 and α_2 (Teng and Siegmund, 1998), the expectation and covariance matrix of the Gaussian process $\{Z_t, 0 \leq t \leq T\}$ can be calculated using the following formulae

$$\begin{aligned} E[Z_t] &= E[E_0[Z_t|Z_{\tau_1}, Z_{\tau_2}]] \\ Cov(Z_t, Z_s) &= E[Cov_0(Z_t, Z_s|Z_{\tau_1}, Z_{\tau_2})] + Cov(E_0[Z_t|Z_{\tau_1}, Z_{\tau_2}], E_0[Z_s|Z_{\tau_1}, Z_{\tau_2})]) \end{aligned}$$

Using the theory of the conditional multivariate normal distributions (Anderson, 1984, p. 37) and the fact that the Gaussian process is Markovian, the expectation and the covariance are

$$
\begin{aligned}
E[Z_t] &= \begin{cases} 0.5\sqrt{N}\alpha_1 e^{-0.04|\tau_1 - t|} & \text{for } t < \tau_1 \leq \tau_2 \\ 0.5\sqrt{N}\alpha_2 e^{-0.04|\tau_2 - t|} & \text{for } \tau_1 \leq \tau_2 < t \\ 0.5\sqrt{N}(\alpha_1 c_1(t) + \alpha_2 c_2(t)) & \text{for } \tau_1 \leq t \leq \tau_2 \end{cases} \\
Cov(Z_t, Z_s) &= \begin{cases} e^{-0.04|s-t|}/2 + (Var(Z_{\tau_1}) - 1/2)e^{-0.04(|s-\tau_1|+|t-\tau_1|)} & \text{for } s \leq t \leq \tau_1 \leq \tau_2 \\ e^{-0.04|s-t|}/2 + (Var(Z_{\tau_2}) - 1/2)e^{-0.04(|t-\tau_2|+|s-\tau_2|)} & \text{for } \tau_1 \leq \tau_2 \leq s \leq t \\ Cov(Z_{\tau_1}, Z_{\tau_2})e^{-0.04(|s-\tau_1|+|t-\tau_2|)} & \text{for } s \leq \tau_1 \leq \tau_2 \leq t \\ Var(Z_{\tau_1})c_3(s,t) + Cov(Z_{\tau_1}, Z_{\tau_2})c_4(s,t) & \text{for } s \leq \tau_1 \leq t \leq \tau_2 \\ Var(Z_{\tau_2})c_5(s,t) + Cov(Z_{\tau_1}, Z_{\tau_2})c_6(s,t) & \text{for } \tau_1 \leq s \leq \tau_2 \leq t \\ c(s,t)/2 + Var(Z_{\tau_1})c_1(s)c_1(t) + Var(Z_{\tau_2})c_2(s)c_2(t) & \\ +Cov(Z_{\tau_1}, Z_{\tau_2})(c_1(s)c_2(t) + c_2(s)c_1(t)) & \text{for } \tau_1 \leq s \leq t \leq \tau_2 \end{cases}
\end{aligned}
$$

with

$$
\begin{aligned}
c_1(s) &= \frac{1 - e^{-0.08|s-\tau_2|}}{1 - e^{-0.08|\tau_2-\tau_1|}} e^{-0.04|s-\tau_1|} \\
c_2(s) &= \frac{1 - e^{-0.08|s-\tau_1|}}{1 - e^{-0.08|\tau_2-\tau_1|}} e^{-0.04|s-\tau_2|} \\
c_3(s,t) &= \frac{1 - e^{-0.08|t-\tau_2|}}{1 - e^{-0.08|\tau_2-\tau_1|}} e^{-0.04|s-t|} \\
c_4(s,t) &= \frac{1 - e^{-0.08|t-\tau_1|}}{1 - e^{-0.08|\tau_2-\tau_1|}} e^{-0.04(|s-\tau_1|+|t-\tau_2|)} \\
c_5(s,t) &= \frac{1 - e^{-0.08|s-\tau_1|}}{1 - e^{-0.08|\tau_2-\tau_1|}} e^{-0.04|s-t|} \\
c_6(s,t) &= \frac{1 - e^{-0.08|s-\tau_2|}}{1 - e^{-0.08|\tau_2-\tau_1|}} e^{-0.04(|s-\tau_1|+|t-\tau_2|} \\
c(s,t) &= e^{-0.04|s-t|} - \frac{1 - e^{-0.08|s-\tau_2|}}{1 - e^{-0.08|\tau_2-\tau_1|}} e^{-0.04(|s-\tau_1|+|t-\tau_1|)} \\
&\quad - \frac{1 - e^{-0.08|s-\tau_1|}}{1 - e^{-0.08|\tau_2-\tau_1|}} e^{-0.04(|s-\tau_2|+|t-\tau_2|)}
\end{aligned}
$$

Similar formula of the expectation was also given by Biernacka et al. (2005). The variance and covariance of Z_{τ_1} and Z_{τ_2} can be directly calculated by

$$
\begin{aligned}
Cov(Z_{\tau_1}, Z_{\tau_2}) &= Cov(X_{\tau_1,1}, X_{\tau_2,1}) \\
&= 4g_{22} + 2g_{12} + 2g_{21} + g_{11},
\end{aligned}
$$

with g_{ij} the probability that a pair share i and j alleles IBD at disease locus τ_1 and τ_2 respectively:

$$\begin{aligned} g_{ij} &= P(X_{\tau_1,1}=i, X_{\tau_2,1}=j|\phi) \\ &= \frac{\sum_G f_{G_{11}\times G_{12}} f_{G_{21}\times G_{22}} P(G_{11},G_{21}|X_{\tau_1,1}=i) P(G_{12},G_{22}|X_{\tau_2,1}=j) P(X_{\tau_1,1}=i, X_{\tau_2,1}=j)}{\sum_{i,j}\sum_G f_{G_{11}\times G_{12}} f_{G_{21}\times G_{22}} P(G_{11},G_{21}|X_{\tau_1,1}=i) P(G_{12},G_{22}|X_{\tau_2,1}=j) P(X_{\tau_1,1}=i, X_{\tau_2,1}=j)} \end{aligned}$$

where the sum is taken over all possible genotypes, $G = (G_{11} \times G_{12}, G_{21} \times G_{22})$, at both disease loci τ_1 and τ_2 of the first and second relative, respectively (Biernacka, 2004). The function $f_{G_{r1}\times G_{r2}}$ is the penetrance given the genotypes G_{r1} and G_{r2} at both disease loci of the pair member r with $r = 1, 2$. The joint probabilities of sharing i and j IBD at τ_1 and τ_2 $P(X_{\tau_1,1} = i, X_{\tau_2,1} = j|\phi)$ were given by Haseman and Elston (1972). The probabilities that a pair has genotypes G_1 and G_2 given that they share j IBDs, $P(G|X_\tau = j)$, for $j = 0, 1, 2$, are summarized in Table 3.2 according to Thompson (1975).

Table 3.2: Probability $P(G_1, G_2|X_\tau = j)$, for $j = 0, 1, 2$

G_1	G_2	$X_\tau = 2$	$X_\tau = 1$	$X_\tau = 0$
A/A	A/A	p^2	p^3	p^4
A/A	A/a	0	$2p^2q$	p^3q
A/A	a/a	0	0	p^2q^2
A/a	A/A	0	$2p^2q$	p^3q
A/a	A/a	pq	$pq(p+q)$	p^2q^2
A/a	a/a	0	$2pq^2$	pq^3
a/a	A/A	0	0	p^2q^2
a/a	A/a	0	$2pq^2$	pq^3
a/a	a/a	q^2	q^3	q^4

The locus has 2 alleles A and a with frequencies p and q.

CHAPTER 4

Testing for association between a disease and a multi-allelic marker: a powerful score test

R. el Galta, T. Stijnen and J.J. Houwing-Duistermaat

Abstract

To study association between a candidate gene and a complex genetic disease, Pearson's χ^2 statistic can be applied to a m-by-2 contingency table, where the m categories correspond to m haplotypes or marker alleles. For $m > 2$, two alternative approaches for Pearson's χ^2 can be followed that are more powerful if one haplotype or marker allele is associated. For the first approach, various 2-by-2 tables are formed by combining various categories and the maximum of the corresponding chi-square statistics is considered as the final statistic. The second approach takes the average over the possible associated categories by writing down an overall likelihood. For the latter approach we propose a new score statistic, which gives more weight to haplotypes or marker alleles that are common. Since the disease allele is often not observed, the power of the various statistics depends both on the linkage disequilibrium pattern as well as the frequencies of the associated haplotype or marker allele in the cases and the controls. We heuristically compare various statistics within the two approaches and present the results of a simulation that compares the performance of all considered statistics. Finally we apply the statistics to a case control study on the association between COL2A1 gene and radiographic osteoarthritis. Our conclusion is that overall the new proposed score statistic has good power.

Submitted for publication

4.1 Introduction

As more and more single nucleotide polymorphisms (SNPs) are discovered, candidates genes will be saturated with SNPs and the focus on haplotype based analysis will increase. When the homologous chromosomes are independently transmitted to the next generation, i.e. when Hardy-Weinberg equilibrium holds, the haplotype counts can be summarized in a m-by-2 contingency table, whose columns refer to m haplotypes and whose rows refer to the disease status. When phase is unknown, the haplotype counts have to be estimated. When phase is known (see for example Uitte de Willige et al. (2005)) or when multi allelic markers are used (see for example Kizawa et al. (2005)), summarizing the data in a m-by-2 table is straightforward. A classical test statistic for the m-by-2 table is Pearson's χ^2 statistic. For a large m this statistic has low power and when the assumption can be made that one haplotype is associated with the binary trait (Terwilliger, 1995), a more specific statistic may be preferred. In this paper we consider various statistics for these m-by-2 tables.

Genetic association is a powerful approach for common associated variants (Wang et al., 2005). Usually the disease allele is not observed and the power of the study will depend on the linkage disequilibrium between the disease locus and the marker loci and on the frequencies of the associated marker allele or haplotype in the cases and the controls (Zondervan and Cardon, 2004). Note that if the disease allele is rare, it will be detectable if the unobserved disease allele has a rather large effect on the trait and is sometimes present on a common haplotype. For sake of simplicity, we describe the methods and simulations in terms of haplotypes, but they can be applied to any m-by-2 table.

To deal with the fact that the associated haplotype is unknown, two approaches may be followed. (1) For each haplotype a statistic is computed by combining the other haplotyes and the maximum of these statistics is taken as the final statistic (maximizing approach). (2) For each haplotype a conditional likelihood given that this haplotype is associated is computed. The overall likelihood is the weighted sum over all haplotypes of all these conditional likelihoods with weights equal to the prior probabilities that a haplotypes is associated. These approaches can also be followed if one allows for a few haplotypes to be associated. Then the maximum is taken over all possible 2-by-2 tables and the likelihood is computed over all possible sets of associated haplotypes.

The maximizing approach was considered by several authors. Ewens et al. (1992) proposed to use the maximum of the χ^2 statistics of 2-by-2 tables each of which compares one variant against the rest ($\hat{Z}_{max}$), when at most one variant is associated. Sham and Curtis (1995) proposed to use the maximum of χ^2 statistics corresponding to all possible 2-by-2 tables, comparing any combination of variants against the rest ($\hat{Z}_{clump}$). From a rather small simulation study, they concluded that Pearson's χ^2 and $\hat{Z}_{clump}$ should be preferred above $\hat{Z}_{max}$ for highly polymorphic markers. Intuitively, when one haplotype is associated with a disease, $\hat{Z}_{max}$ should be more powerful than Pearson's χ^2 and $\hat{Z}_{clump}$, while if more than one associated haplotype exists $\hat{Z}_{clump}$ should have more power than $\hat{Z}_{max}$. More simulations are needed to study the performance of these test statistics.

An alternative to taking the maximum is to take the sum over all possibilities. When one variant is associated Terwilliger (1995) proposed to model the excess of the associated variant in cases by the parameter λ which is the population attributable risk (Clayton, 2000). Since it is unknown which variant is associated with the disease, the likelihood corresponding to this model is a weighted sum over all variants i of conditional likelihoods given that variant i is over-represented in the set of cases. These weights represent the prior probability that a haplotype is associated to the disease. In line with the common disease common variant hypothesis (Reich and Lander, 2001) and in line with the method of Terwilliger (1995), the haplotype frequencies in controls can be used as weights. To test for association the likelihood ratio test can be used. Maximizing the log likelihood function over the haplotype frequencies and λ appears not straightforward, because the weights are equal to the haplotype frequencies, and these same haplotype frequencies are also unknown parameters in the conditional likelihood functions. In this paper we propose the corresponding score statistic and we use Monte-Carlo permutation to derive p-values (Sham and Curtis, 1995).

We first compare heuristically the power of the Pearson's χ^2, $\hat{Z}_{max}$ and $\hat{Z}_{clump}$. We then derive the new score test and describe the results of a simulation study which we performed to compare the performance of the new score test, χ^2, $\hat{Z}_{max}$ and $\hat{Z}_{clump}$. In the simulations we assumed that phase is known. In the discussion we describe how to derive p-values for the case of phase ambiguity. As an illustration we apply these test statistics to a published case-control study on association between COL2A1 gene and radiographic osteoathritis (Meulenbelt et al., 1999).

4.2 The maximising approach

Assume that Hardy-Weinberg equilibrium holds and that we have a sample of n_1 case chromosomes and of n_2 control chromosomes. Let $p = (p_1, \cdots, p_m)$ be the vector of frequencies of the m haplotypes in controls. Let $x = (x_1, \cdots, x_m)$ and $y = (y_1, \cdots, y_m)$ be the vector of haplotype counts in the cases and the controls, respectively and let n be equal to $n_1 + n_2$. Let $\hat{Z}_i$ be equal to the observed minus expected ith haplotype count $x_i - n_1\hat{p}_i$ with $\hat{p}_i = \frac{x_i+y_i}{n}$, the estimate of haplotype frequency in combined sample. Let $\hat{Z} = (\hat{Z}_1, \cdots, \hat{Z}_m)'$. Throughout the text the hat symbolˆrefers to two samples statistics emphasizing the fact that haplotype frequencies are estimated under the null hypothesis.

Testing the null hypothesis of no disease-marker association is classically performed by means of Pearson's χ^2 statistic

$$\chi^2 = \sum_{j=1}^{m} \frac{(x_j - n_1\hat{p}_j)^2}{n_1\hat{p}_j} + \sum_{j=1}^{m} \frac{(y_j - n_2\hat{p}_j)^2}{n_2\hat{p}_j} = \frac{n}{n_1 n_2} \sum_{j=1}^{m} \frac{(x_j - n_1\hat{p}_j)^2}{\hat{p}_j}.$$

An alternative test statistic is $\hat{Z}_{max}$, defined as

$$\hat{Z}_{max} = \max_{i=1\cdots m} \frac{\hat{Z}_i^2}{Var(\hat{Z}_i)}.$$

Sham and Curtis (1995) proposed the largest value of all possible χ^2 statistics of 2-by-2 tables each obtained by testing a combination of haplotype against the rest. We denote this statistics by $\hat{Z}_{clump}$ according to the program they use for the computation. In addition, they proposed to use Monte-Carlo methods to derive the empirical p-values of $\hat{Z}_{clump}$, χ^2 and $\hat{Z}_{max}$.

In order to compare these three test statistics heuristically, we rewrite them as maxima of the same expression where the maximum is taken over different sets. Pearson's χ^2 can be rewritten as:

$$\chi^2 = \max_{u \in R} \frac{(u'\hat{Z})^2}{u'Var(\hat{Z})u},$$

where R is the set of vectors with m coordinates (see appendix 4.7). Since $\sum_{i=1}^{m} \hat{Z}_i = 0$, the Pearson's χ^2 test is the Hotelling's test statistic applied to any $m-1$ coordinates of the vector $\hat{Z}$.

Now $\hat{Z}_{max}$ is the maximum value of all Pearson's χ^2 tests on 2-by-2 tables obtained by comparing any haplotype against the rest. $\hat{Z}_{max}$ can be reexpressed by

$$\hat{Z}_{max} = \max_{u \in A} \frac{(u'\hat{Z})^2}{u'Var(\hat{Z})u},$$

where A is the set of the m different permutations of the vector $(1, 0, .., 0)'$.

The $\hat{Z}_{clump}$ statistic can be given by

$$\hat{Z}_{clump} = \max_{s \subseteq (1,\dots,m)} \frac{(\sum_{i \in s} x_i - n_1 \sum_{i \in s} \hat{p}_i)^2}{n_1(1 - \sum_{i \in s} \hat{p}_i) \sum_{i \in s} \hat{p}_i} = \max_{u \in S} \frac{(u'\hat{Z})^2}{u'Var(\hat{Z})u},$$

where S is the set of vectors whose k coordinates set to 1 and $m - k$ coordinates set to 0, with $k = 1, ..., m - 1$ and s any subset of $(1, 2, ..., m)$. This implies that under the alternative hypothesis of the presence of an association, all associated haplotypes are assumed to have the same effect sizes in terms of relative risk.

Note that A is a subset of S, which in turn, is a subset of R. Hence, if only one haplotype is associated with the disease, the alternative hypothesis is properly specified by A, and then $\hat{Z}_{max}$ is likely to have more power than $\hat{Z}_{clump}$ and χ^2. However if two or more marker haplotypes are associated with the disease and have the same effect size in terms of relative risk the $\hat{Z}_{clump}$ is expected to provide more power than $\hat{Z}_{max}$ and χ^2 as the alternative hypothesis is better specified by the set S. In case of associated haplotypes with unequal effect sizes χ^2 is expected to perform the best unless the number of haplotypes is large.

4.3 The averaging approach

Assume that one of the haplotypes is over-represented in the cases. Denote this haplotype with index i. The haplotype frequencies in the cases can be modelled as $q_i = p_i + \lambda(1 - p_i)$ with $0 \leq \lambda \leq 1$ for the associated haplotype i and as $q_j = p_j - \lambda p_j$ for the remaining haplotypes with $j = 1, \cdots, m$ and $j \neq i$. Here λ is the population attributable risk. Then the conditional likelihood of data given that haplotype i is over-represented in cases is

$$L_i(x, y|\lambda, p) = (p_i + \lambda(1 - p_i))^{x_i}(1 - \lambda)^{n_1 - x_i} \prod_{j \neq i}^{m} p_j^{x_j} \prod_{j=1}^{m} p_j^{y_j}. \tag{4.1}$$

Terwilliger (1995) proposed the following likelihood, assuming that the prior probability of a marker haplotype i being associated with the disease is equal to the haplotype frequency p_i

$$L(x,y|\lambda,p) = \sum_{j=1}^{m} p_j L_j, \tag{4.2}$$

with L_j given in formula (4.1). The corresponding score statistic is an alternative to the likelihood ratio test proposed by Terwilliger (1995) for $H_0 : \lambda = 0$ versus $H_a : \lambda > 0$. Since the first derivative of the likelihood L with respect to λ at $\lambda = 0$ is equal to zero, we propose to use the second derivative of the log-likelihood with respect to λ to derive the score statistic (Dudoit and Speed, 2000; Tritchler et al., 2003). The second derivative of the log-likelihood with respect to λ evaluated for $\lambda = 0$ is

$$\frac{\partial^2}{\partial\lambda^2}\log(L(x,y|\lambda=0,p)) \quad = \quad \sum_{j=1}^{m}\frac{(x_j - n_1 p_j)^2}{p_j} - \sum_{j=1}^{m}\frac{x_j - n_1 p_j}{p_j} - n_1(m-1).$$

Derivation of the second derivative is given in the appendix (4.8). By dividing the second derivative by n_1 and then taking the stochastic part of it, the score statistic can be given by

$$S_p = X^2 - \frac{U}{n_1}, \tag{4.3}$$

with $X^2 = \sum_{j=1}^{m}\frac{(x_j - n_1 p_j)^2}{n_1 p_j}$, the one sample Pearson's χ^2 statistic (haplotype frequencies in controls are known), and $U = \sum_{j=1}^{m}\frac{x_j - n_1 p_j}{p_j}$, the score statistic obtained by replacing the weights in the likelihood (4.2) by equal weights. For equally frequent haplotypes the statistic $U = 0$, hence $S_p = X^2$. Under the null hypothesis, S_p has mean $\mathrm{E}[S_p] = m - 1$ and variance $\mathrm{Var}(S_p) = 2n_1^{-1}(n_1 - 1)(m-1)$ (see Appendix 4.8). Hence, asymptotically the statistics S_p and X^2 have the same expectation and the same variance.

When the haplotype frequencies p_i are unknown, the score statistic can be estimated by replacing the haplotype frequencies p_i by their maximum likelihood estimators under the null hypothesis $\hat{p}_i = \frac{x_i + y_i}{n}$. Now after some algebra the score statistic can be given by

$$\hat{S}_p = \chi^2 - \frac{n}{n_1 n_2}\hat{U}, \tag{4.4}$$

with χ^2, the two samples Pearson's χ^2 statistic on the m-by-2 table and $\hat{U} = \sum_{j=1}^{m}\frac{x_j - n_1\hat{p}_j}{\hat{p}_j}$. It can be shown by means of the δ-method that the score statistic $\hat{S}_p$ and Pearson's χ^2 have asymptotically the same expectation and the same variance under the null hypothesis (see appendix 4.8). For m large, the score test $\hat{S}_p$ (S_p) follows approximately a normal distribution under the null hypothesis. To ensure the validity of the asymptotic distribution, the number of cases and control chromosomes should be much larger than the number of marker haplotypes m. Nevertheless, the empirical distribution under the null hypothesis of the statistic $\hat{S}_p$ (S_p) can easily be derived by using Monte-Carlo methods (Sham and Curtis, 1995).

Under the alternative hypothesis of the presence of one positively associated haplotype i, it can be shown that the expectations of U and $\hat{U}$ are

$$\begin{aligned} \mathrm{E}[U] &= n_1\lambda(\frac{1}{p_i} - m) \\ \mathrm{E}[\hat{U}] &\approx \frac{n_1 n_2 \lambda}{n - n_1\lambda}(\frac{n}{np_i + n_1\lambda(1-p_i)} - m) \leq \frac{n_1 n_2 \lambda}{n - n_1\lambda}(\frac{1}{p_i} - m). \end{aligned}$$

This implies that the expectations of U and $\hat{U}$ are negative if the frequency of the associated haplotype p_i is larger than the inverse of the number of marker haplotypes $\frac{1}{m}$. Consequently, the score statistic $\hat{S}_p$ (S_p) becomes larger in expectation than $\chi^2(X^2)$ if the frequency of the associated haplotype is larger than $\frac{1}{m}$. Hence, for common associated haplotypes $(p_i > \frac{1}{m})$ the score statistic is expected to have higher power than Pearson's χ^2.

Terwilliger (1995) discussed the presence of more than one associated haplotype. For two positively associated haplotypes i and k he proposed the following model with two free parameters λ_1 and λ_2 with $\lambda_1 + \lambda_2 \leq 1$

$$\begin{aligned} q_i &= p_i(1-\lambda_1-\lambda_2) + \lambda_1, \\ q_k &= p_k(1-\lambda_1-\lambda_2) + \lambda_2 \text{ for } i \neq k, \\ q_j &= p_j(1-\lambda_1-\lambda_2) \text{ for } j \neq i \text{ and } j \neq k. \end{aligned}$$

The likelihood for two associated haplotypes given by Terwilliger (1995) was incorrect since the weights are prior probabilities and did not sum to 1. Therefore, we propose the following likelihood for two associated variants

$$L(x,y|\lambda_1,\lambda_2,p) = \sum_{i=1}^{m}\sum_{k=1}^{m} p_i p_k \prod_{j}^{m} q_j^{x_j} p_j^{y_j},$$

assuming $q_i = p_i(1-\lambda_1-\lambda_2) + \lambda_1 + \lambda_2$ for $i = k$. Since i and k are interchangeable with respect to λ_l for $l = 1, 2$ the following derivatives are

$$\begin{aligned}
\frac{\partial}{\partial \lambda_l} L(x, y|\lambda_1 = 0, \lambda_2 = 0, p) &= 0, \\
\frac{\partial^2}{\partial \lambda_1 \partial \lambda_2} L(x, y|\lambda_1 = 0, \lambda_2 = 0, p) &= 0, \text{and} \\
\frac{\partial^2}{\partial \lambda_l^2} L(x, y|\lambda_1 = 0, \lambda_2 = 0, p) &= n_1 X^2 - mU - n_1(m-1).
\end{aligned}$$

In contrast to the Terwilliger's likelihood ratio, $\hat{S}_p$ (equation 4.4) is the score statistic of testing no disease-marker association regardless of the potential number of associated variants.

4.4 Simulation study

The aim of the simulation study is to evaluate empirically the power of the score test $\hat{S}_p$ in comparison with the Pearson's χ^2, $\hat{Z}_{max}$ and $\hat{Z}_{clump}$ tests. We generated at least 1000 replicates from the multinomial distributions according to the models described previously. Without loss of generality we assumed that the first or first two haplotypes are associated with the disease. The remaining haplotypes were equally frequent. We varied the number of variants m from 3 to 20. The p-values of the test statistics were calculated empirically by means of 1000 Monte-Carlo permutations using a program based on the program Clump (Sham and Curtis, 1995). We used a nominal p-value of 0.05.

Type I error rate

To verify whether Monte-Carlo yields the right type I error rate of these test statistics, data sets were generated under the null model ($\lambda = 0$) each time for markers with $5, 7, 9, 11, 16$ and 20 alleles. The frequency of the first allele was set to 0.5, whereas the remaining alleles were equally frequent. The results are shown in Table 4.1. The type I error rate is approximately equal to the nominal rate for the score $\hat{S}_p$, Pearson's χ^2, and $\hat{Z}_{clump}$ tests, regardless of the number of alleles m at the marker locus, whereas the $\hat{Z}_{max}$ becomes somewhat conservative as the number of marker alleles m increases (Sham and Curtis, 1995).

Table 4.1: The type I error rates based on 10000 simulated m-by-2 tables for $\lambda = 0$ and $p_1 = 0.5$.

α	m	χ^2	$\hat{Z}_{clump}$	$\hat{Z}_{max}$	$\hat{S}_p$	m	χ^2	$\hat{Z}_{clump}$	$\hat{Z}_{max}$	$\hat{S}_p$
0.05	5	0.053	0.053	0.047	0.054	11	0.047	0.046	0.039	0.046
0.01		0.011	0.011	0.009	0.011		0.010	0.010	0.008	0.010
0.001		0.001	0.001	0.001	0.001		0.001	0.001	0.000	0.001
0.05	7	0.051	0.048	0.045	0.052	16	0.049	0.049	0.040	0.047
0.01		0.011	0.010	0.009	0.010		0.011	0.011	0.007	0.010
0.001		0.001	0.001	0.000	0.001		0.001	0.001	0.000	0.001
0.05	9	0.051	0.052	0.044	0.052	20	0.052	0.052	0.035	0.053
0.01		0.001	0.011	0.008	0.010		0.011	0.011	0.008	0.010
0.001		0.002	0.002	0.001	0.001		0.001	0.002	0.001	0.002

Single associated variant

To study the power of the statistics we first considered the model used by Terwilliger (1995) for one positively associated common haplotype. The frequency p_1 of this haplotype was 0.5 in controls. The parameter λ was fixed to 0.5, which corresponds to a haplotype frequency of 0.75 in the cases and a relative risk γ of 3. We considered 100 case chromosomes (n_1) and 100 control chromosomes (n_2). The results are shown in Table 4.2. For $m \leq 5$ all test statistics performed well; however $\hat{S}_p$ had slightly higher power than other test statistics. For $m > 5$ the score test $\hat{S}_p$ and $\hat{Z}_{max}$ tests appeared to perform better than the Pearson's χ^2 and $\hat{Z}_{clump}$ tests regardless of the number of haplotypes at the marker locus. Especially for the significant level of 0.05, $\hat{S}_p$ and $\hat{Z}_{max}$ had similar power, while for lower significant levels $\hat{S}_p$ had somewhat lower power than $\hat{Z}_{max}$. The power of Pearson's χ^2 decreased as the number of haplotypes increased.

Second, we studied the power of the test statistics for various values of the frequency of the associated haplotype (0.06 to 0.5). We chose λ so that the relative risk γ of the associated variant with respect to its absence was about 2. Because of low λ, the number of chromosomes n_1 and n_2 were now set to 200. The results are depicted in figure 4.1. Almost overall $\hat{Z}_{max}$ outperformed the other test statistics. $\hat{S}_p$ had the second best power. It performed better than Pearson's χ^2 especially when $m \geq 10$. Further it had higher power than Z_{clump} for $p_1 = 0.06$ and 0.1 while for $p_1 = 0.15$, 0.2, 0.3 and 0.4, the performances of

TABLE 4.2: The power based on 10000 simulated m-by-2 tables for $\lambda = 0.5$ and $p_1 = 0.5$.

α	m	χ^2	$\hat{Z}_{clump}$	$\hat{Z}_{max}$	$\hat{S}_p$	m	χ^2	$\hat{Z}_{clump}$	$\hat{Z}_{max}$	$\hat{S}_p$
0.05	3	0.92	0.93	0.93	0.94	9	0.76	0.85	0.91	0.88
0.01		0.79	0.80	0.80	0.83		0.55	0.68	0.79	0.73
0.001		0.54	0.55	0.55	0.58		0.30	0.42	0.57	0.48
0.05	4	0.89	0.90	0.91	0.92	11	0.72	0.84	0.93	0.88
0.01		0.73	0.76	0.77	0.79		0.51	0.67	0.82	0.72
0.001		0.47	0.5	0.53	0.54		0.27	0.40	0.60	0.48
0.05	5	0.86	0.89	0.90	0.91	16	0.64	0.82	0.95	0.88
0.01		0.69	0.74	0.77	0.77		0.42	0.63	0.85	0.72
0.001		0.43	0.49	0.54	0.53		0.20	0.36	0.61	0.46
0.05	7	0.80	0.87	0.89	0.89	20	0.60	0.82	0.95	0.88
0.01		0.61	0.70	0.76	0.74		0.38	0.62	0.85	0.72
0.001		0.35	0.45	0.55	0.50		0.20	0.35	0.62	0.46

$\hat{S}_p$ and Z_{clump} were comparable. Pearson's χ^2 performed well when $m = 5$ or when $p_1 = 0.06$.

Two associated variants

To study the performance of the test statistics when there are two haplotypes positively associated with the disease, we generated data according to the model given by (4.5). We simulated data sets for $(p_1, p_2) = (0.06, 0.06)$, (0.06,0.1), (0.1, 0.15), (0.15,0.2), (0.2,0.3), (0.3,0.4) and their corresponding excess frequencies $(\lambda_1, \lambda_2) = (0.05, 0.05)$, (0.05,0.08), (0.08,0.1), (0.1,0.15), (0.15,0.2), (0.18,0.25), respectively. The total relative risk of the two associated variants with respect to their absence was again about two (between 1.9 and 2.1). The number of chromosomes n_1 and n_2 were set to 100. The power curves are shown in figure 4.2. In contrast to the case of one associated haplotype, $\hat{S}_p$ and $\hat{Z}_{clump}$ performed now better than Z_{max}. For $m \leq 8$ and $p_1 = 0.06$, $\hat{S}_p$ appeared to have somewhat less power than Pearson's χ^2 and $\hat{Z}_{clump}$. Whereas for $m \geq 10$ $\hat{S}_p$ had somewhat more power than Pearson's χ^2 and $\hat{Z}_{clump}$. For the remaining situations ($p_1 \geq 0.1, p_2 \geq 0.15$), $\hat{S}_p$ had the best power and $\hat{Z}_{clump}$ had the second best power. The power of Pearson's χ^2 was comparable to that of $\hat{Z}_{clump}$ for $m = 5$ and it decreased with the increase of m. For

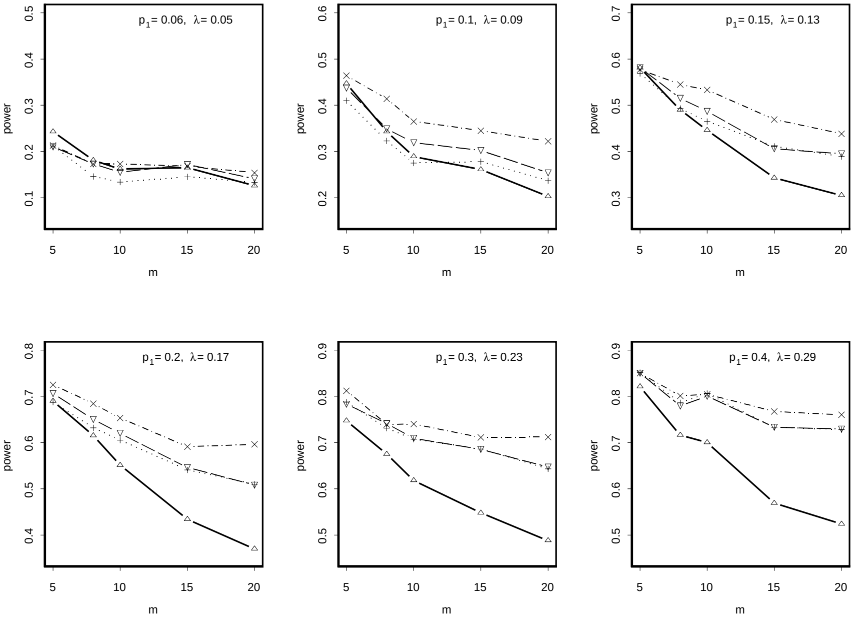

FIGURE 4.1: *Power curves of χ^2 (—△—), $\hat{Z}_{clump}$ (····+····), $\hat{Z}_{max}$ (–·–·×·–·–) and $\hat{S}_p$ (– – ▽– –) for 200 case and 200 control chromosomes and a haplotypic relative risk of about 2.*

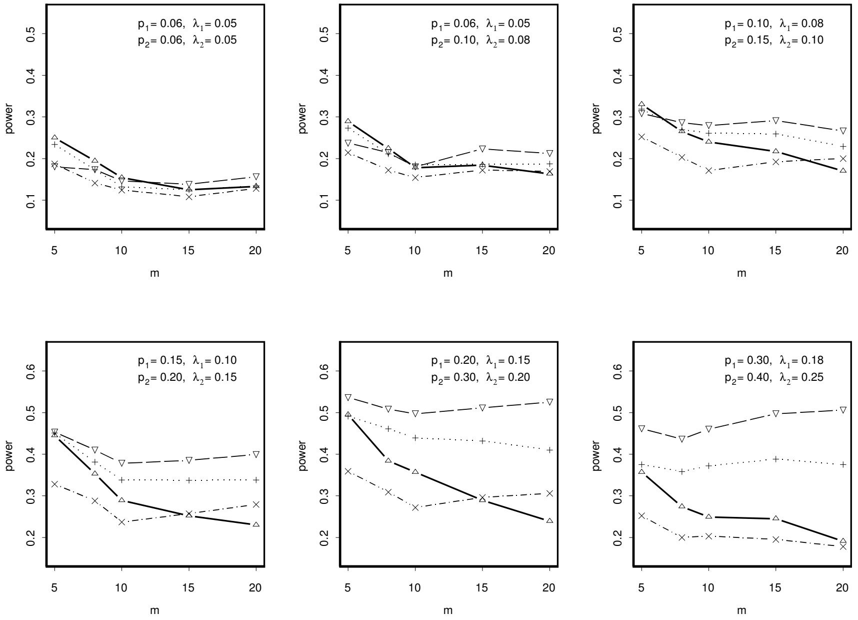

FIGURE 4.2: *Power curves at the nominal level $\alpha = 0.05$ of χ^2 (—△—), $\hat{Z}_{clump}$ (····+····), $\hat{Z}_{max}$ (–·–·×·–·–) and $\hat{S}_p$ (——▽——) for 100 case and 100 control chromosomes and a relative risk of about 2 for the combined associated haplotypes.*

most situations, it had higher power than $\hat{Z}_{max}$, which had low power.

4.5 Application to real data

To illustrate the score test with real data we used data obtained from a published study on association between the collagen type II gene(COL2A1) and radiographic osteoarthritis (ROA) (Meulenbelt et al., 1999). Osteoarthritis is a degenerative disease of the joints. The VNTR marker next to COL2A1 was typed in 820 subjects aged 50-70 years from a population-based cohort study, the Rotterdam study. Radiographs of knees, hips, hands, and spine were scored for the presence of ROA. 123 cases had ROA in at least 3 joints groups. 697 remaining subjects were used as controls. Five variants had frequencies $\geq$ 0.05. The other variants were combined into one group. Preliminary in-

TABLE 4.3: *Results of analysis of association between ROA and COL2A1 gene.*

COL2A1 haplotypes	2# cases $n_1 = 246$	2# controls $n_2 = 1394$	χ^2	$\hat{Z}_{max}$	$\hat{Z}_{clump}$	$\hat{S}_p$
13R1	114 (0.46)	581 (0.42)				
14R1	54 (0.22)	385 (0.28)				
11R1	22 (0.09)	149 (0.11)				
14R2$^+$	27 (0.11)	78 (0.06)				
13R2	12 (0.05)	85 (0.05)				
Others pooled	17 (0.08)	116 (0.08)				
P-value			0.013	0.009	0.044	0.012

spection of data suggested the existence of one positively associated variant. The frequency of the potential associated variant, 14R2, was 0.06 ($< 1/m$ with $m = 6$) in controls. Its frequency was increased in the cases with a relative risk of about 1.94 ($\lambda \approx 0.05$). Hence one might expect that Pearson's χ^2 will perform slightly better than $\hat{S}_p$ (see Figure 4.2). (Meulenbelt et al., 1999) used Terwilliger's LR. They found evidence of association between ROA and COL2A1 (P=0.03). We also applied the score $\hat{S}_p$, Pearson's χ^2, $\hat{Z}_{max}$, and $\hat{Z}_{clump}$ tests to these data. Table 4.3 shows the distribution of the variants among cases and controls and summarizes the results of applied test statistics. All test statistics indicated the presence of a significant association at the 0.05 level. The score test $\hat{S}_p$, Pearson's χ^2, and $\hat{Z}_{max}$ gave quite similar p-values of about 0.01. The $\hat{Z}_{clump}$ test yielded a quite higher p-value of about 0.04.

4.6 Discussion

In this paper, we derived a new score statistic $\hat{S}_p$, which corresponds to the likelihood ratio statistic of Terwilliger (1995). The score test is easy to compute and is asymptotically locally most powerful (Cox and Hinkley, 1974). For a single common positively associated haplotype (frequency $> 1/m$), we showed heuristically that the score test would provide more power than Pearson's χ^2 on the m-by-2 table. Further in contrast to the likelihood ratio statistic of Terwilliger, the same score test is obtained regardless of the number of associated haplotypes. For large m, the score test follows approximately a normal distribution under the null hypothesis. For small sample sizes or small m, the empirical p-values can be derived by means of Monte-Carlo methods.

By means of simulations we compared the performance of this new statistic to the existing statistics χ^2, $\hat{Z}_{max}$ and $\hat{Z}_{clump}$. The $\hat{S}_p$ gives more weight to common haplotypes, but for one associated haplotype it had similar or slightly less power than $\hat{Z}_{max}$. The power of $\hat{Z}_{max}$ was dramatically low for many considered models of two associated haplotypes. The power of Pearson's χ^2 decreased with the number of observed haplotypes, due to the increasing number of degrees of freedom.

In the simulation study we assumed that phase is known. When phase is ambiguous, the haplotype counts have to be estimated and the uncertainty in phase has to be taken into account when computing the p-value of the statistics. This can easily be incorporated when Monte Carlo methods are used by estimating the haplotype frequencies in each permutation step (see for example Becker et al. (2005)). This adjustment is less efficient than maximizing the likelihood over the haplotype frequencies and the unknown parameter λ simultaneously, but in many situations the loss of information due to phase-uncertainty is small. We recommend to consider smaller blocks or single marker methods when the loss of information due to phase uncertainty is rather large (Uh et al., 2005).

In this paper, we used the chromosome as unit of analysis and not the individual. By doing so we have to assume Hardy Weinberg equilibrium and the methods correspond to a multiplicative model for diplotypes (Sasieni, 1997). Therefore the power will decrease when the true model deviates from this multiplicative model as is the case for a recessive model. For this model we recommend to use other methods (see also Cordell and Clayton (2005)).

Like the approach of Terwilliger (1995), $\hat{S}_p$ gives more weight to common haplotypes. A positive association may be due to a common causal variant, due to a rare mutation with a rather large impact which is sometimes present on a common haplotype or due to multiple mutations on the associated haplotype. When mutations are present on more than one haplotype, the score statistic $\hat{S}_p$ can still detect this association, but identification of the causal variants will be difficult. Genetic association studies have no power to identify genes with multiple rare mutations on rare haplotypes (Zondervan and Cardon, 2004).

The advantages of the parameter λ are that it is directly related to recombination fraction and is less sensitive to haplotype frequencies than other measures (Devlin and Risch, 1995). However, when allelic association is modelled by means of λ it is not straightforward to adjust for other covariates. Houwing-Duistermaat and Elston (2001) discussed various ways to quantify allelic association and estimate the location of the gene responsible for the disease using logistic regression models. As an alternative to λ, the log relative risk as measured by the regression coefficient in the logistic model may be used to allow for adjustment of other covariates. More research is needed to build this kind of flexible models.

We conclude that overall the score statistic $\hat{S}_p$ has good power regardless of the number of observed haplotypes. When one haplotype is associated, $\hat{Z}_{max}$ performs better, but when two haplotypes are associated, $\hat{Z}_{max}$ performs dramatically bad and $\hat{S}_p$ performs well.

Software

The method described in this paper is implemented in software written in C programming language and it is based on the source of Clump program (Sham and Curtis, 1995). The C program will be available from our Web site (http://clinicalresearch.nl/personalpage/)

4.7 Appendix 1

Pearson's chi-square

Let R be the set of vectors with m coordinates and R_{-m} be the set of vectors with $m-1$ coordinates. Let $\hat{Z}_{-m}$ the vector of the first $m-1$ centered allele counts. Since $\sum_{i=1}^{m} \hat{Z}_i = 0$, it follows

$$\max_{u \in R} \frac{(u'\hat{Z})^2}{u'Var(\hat{Z})u} = \max_{v \in R_{-m}} \frac{(v'\hat{Z}_{-m})^2}{v'Var(\hat{Z}_{-m})v}.$$

The covariance matrix $Var(\hat{Z}_{-m})$ is positive definite, hence applying the extend Cauchy-Schwarz inequality (Johnson and Wichern, 1998) and noting that the maximum is attained when $v \propto Var(\hat{Z}_{-m})^{-1}\hat{Z}_{-m}$ give

$$\max_{v \in R_{-m}} \frac{(v'\hat{Z}_{-m})^2}{v'Var(\hat{Z}_{-m})v} = \hat{Z}'_{-m}Var(\hat{Z}_{-m})^{-1}\hat{Z}_{-m}.$$

After some algebra it follows

$$\max_{u \in R} \frac{(u'\hat{Z})^2}{u'Var(\hat{Z})u} = \sum_{j=1}^{m} \frac{(x_j - n_1\hat{p}_j)^2}{n_1\hat{p}_j} + \sum_{j=1}^{m} \frac{(y_j - n_2\hat{p}_j)^2}{n_2\hat{p}_j}.$$

4.8 Appendix 2

Derivation of the score statistic

The first derivative of the likelihood is

$$\begin{aligned}\frac{\partial}{\partial\lambda}L(x,y|\lambda,p) &= \sum_{i=1}^{m} p_i \frac{\partial}{\partial\lambda}L_i(x,y|\lambda,p)\\ &= \sum_{i=1}^{m}\{x_i(1-p_i)(p_i+\lambda(1-p_i))^{-1}-(n_1-x_i)(1-\lambda)^{-1}\}p_iL_i(x,y|\lambda,p).\end{aligned}$$

Hence the score statistic is

$$\begin{aligned}\frac{\partial}{\partial\lambda}\log(L(x,y|\lambda=0,p)) &= \frac{\frac{\partial}{\partial\lambda}(L(x,y|\lambda=0,p))}{L(x,y|\lambda=0,p)}\\ &= \sum_{i=1}^{m}\frac{p_i(x_i-n_1p_i)}{p_i}=0\end{aligned}$$

Therefore, the score statistic can be now obtained by evaluating the second derivative of the log-likelihood with respect to λ at $\lambda = 0$ and using the fact that the first derivative is zero

$$\frac{\partial^2}{\partial\lambda^2}\log(L(x,y|\lambda=0,p)) = \frac{\frac{\partial^2}{\partial\lambda^2}L(x,y|\lambda=0,p)}{L(x,y|\lambda=0,p)} \tag{4.5}$$

The second derivative of the likelihood is

$$\begin{aligned}\frac{\partial^2}{\partial\lambda^2}L(x,y|\lambda,p) &= \sum_{i=1}^{m}\{-x_i(1-p_i)^2(p_i+\lambda(1-p_i))^{-2}-(n_1-x_i)(1-\lambda)^{-2}\}p_iL_i(x,y|\lambda,p)\\ &+ \sum_{i=1}^{m}\{x_i(1-p_i)(p_i+\lambda(1-p_i))^{-1}-(n_1-x_i)(1-\lambda)^{-1}\}^2p_iL_i(x,y|\lambda,p).\end{aligned}$$

Therefore the second derivative at $\lambda = 0$ is

$$\frac{\partial^2}{\partial\lambda^2}L(x,y|\lambda=0,p)=\{\sum_{i=1}^{m}(x_i^2-x_i)p_i^{-1}+n_1-n_1^2\}L(x,y|\lambda=0,p). \tag{4.6}$$

Combining (4.5) and (4.6)

$$\frac{\partial^2}{\partial\lambda^2}\log(L(x,y|\lambda=0,p)) = \sum_{i=1}^{m}\frac{(x_i-n_1p_i)^2}{p_i}-\sum_{i=1}^{m}\frac{x_i-n_1p_i}{p_i}-n_1(m-1)$$

Derivation of expectation and variance of S_p

The expectation of S_p is

$$E[S_p] = E[X^2] = k - 1.$$

and the variance of $n_1 S_p$ is

$$\begin{aligned}
Var[n_1S_p] &= \sum_{i,j=1}^{m} \frac{Cov[(x_i - n_1p_i)^2, (x_j - n_1p_j)^2]}{p_ip_j} + \sum_{i,j=1}^{m} \frac{E[(x_i - n_1p_i)(x_j - n_1p_j)]}{p_ip_j} \\
&- 2\sum_{i,j=1}^{m} \frac{E[(x_i - n_1p_i)^2(x_j - n_1p_j)]}{p_ip_j} \\
&= \sum_{i,j=1}^{m} \frac{-n_1p_ip_j(1 - 2p_i - 2p_j + 6p_ip_j)}{p_ip_j} + 2\sum_{i,j=1}^{m} \frac{(n_1p_ip_j)^2}{p_ip_j} \\
&+ \sum_{i=1}^{m} \frac{n_1p_i(1 - 6p_i + 8p_i^2)}{p_i^2} + 2\sum_{i=1}^{m} \frac{n_1^2p_i^2(1 - 2p_i)}{p_i^2} + \sum_{i,j=1}^{m} \frac{-n_1p_ip_j}{p_ip_j} \\
&+ \sum_{i=1}^{m} \frac{n_1p_i}{p_i^2} - 2\sum_{i,j=1}^{m} \frac{-n_1p_ip_j(1 - 2p_i)}{p_ip_j} - 2\sum_{i=1}^{m} \frac{n_1p_i(1 - 2p_i)}{p_i^2} \\
&= 2n_1(n_1 - 1)(m - 1)
\end{aligned}$$

Note that

$$\begin{aligned}
Cov[(x_i - n_1p_i)^2, (x_j - n_1p_j)^2] &= -n_1p_ip_j\{(1 - 2p_i - 2p_j + 6p_ip_j) - 2n_1p_ip_j\} \\
&+ I_{\{j=i\}}n_1p_i\{(1 - 6p_i + 8p_i^2) + 2n_1p_i(1 - 2p_i)\} \\
E[(x_i - n_1p_i)^2(x_j - n_1p_j)] &= -n_1p_ip_j(1 - 2p_i) + I_{\{j=i\}}n_1p_i(1 - 2p_i) \\
E[(x_i - n_1p_i)(x_j - n_1p_j)] &= -n_1p_ip_j + I_{\{j=i\}}n_1p_i,
\end{aligned}$$

with $I_{\{j=i\}} = 1$ if $i = j$ otherwise 0. Hence the variance of S_p is

$$Var[S_p] = 2(m - 1)\frac{n_1 - 1}{n_1}$$

Derivation of the asymptotic expectation and variance of $\hat{S}_p$

The expectation and the variance of $\hat{S}_p$ can be given as follows

$$\begin{aligned}
E[\hat{S}_p] &= E[\chi^2] - \frac{n}{n_1n_2}E[\hat{U}] \\
Var[\hat{S}_p] &= Var[\chi^2] - 2\frac{n}{n_1n_2}Cov[\chi^2, \hat{U}] + (\frac{n}{n_1n_2})^2Var[\hat{U}]
\end{aligned}$$

By means of δ-method it can be shown that

$$\begin{aligned} E[\hat{U}] &\approx 0, \\ Var[\hat{U}] &\approx n^{-1}n_1n_2(\sum_{j=1}^{m}\hat{p}_j^{-1} - m^2), \\ Cov[\chi^2, \hat{U}] &= E[\chi^2\hat{U}] \approx 0. \end{aligned}$$

Hence for $\frac{n}{n_2} \ll \infty$ and $n_1 \to \infty$

$$\begin{aligned} E[\hat{S}_p] &= E[\chi^2] = m - 1 \\ Var[\hat{S}_p] &= Var[\chi^2] = 2(m-1) \end{aligned}$$

Chapter 5

Methods to test for association between a disease and a multi-allelic marker applied to a candidate region

R. el Galta, L. Hsu and J.J. Houwing-Duistermaat

Abstract

We report the analysis results of the GAW14 simulated micro-satellite marker data set, using replicate 50 from the Danacaa population. We applied several methods for association analysis of multi-allelic markers to case-control data to study the association between Kofendrerd Personality Disorder (KPD) and multi-allelic markers in a candidate region previously identified by the linkage analysis. Evidence for association was found for marker D03S0127 ($P < 0.01$). The analyses were done without any prior knowledge of the answers.

To appear in *BMC Genetics*

5.1 Background

Terwilliger (1995) proposed a powerful method for the association analysis between a disease and a multi-allelic marker. The model assumes that only one marker allele is associated with the disease and that any marker allele may be associated with the disease with prior probability equal to its allele frequency in the population. The excess allele in the cases is modelled by a parameter λ, the population attributable risk (Devlin and Risch, 1995). The likelihood of the data given the allele frequencies and the parameter λ is the weighted sum of the conditional likelihood functions given that an allele is associated with the disease over all marker alleles with weights equal to the allele frequencies. Hence more weight is assigned to more frequent marker alleles.

To test the null hypothesis ($\lambda = 0$) against the alternative hypothesis ($\lambda > 0$), Terwilliger (1995) proposed a likelihood ratio statistic (LR). However this statistic appeared to be conservative and computation of the maximum-likelihood estimates might be slow. Another point mentioned by Sham et al. (1996) is that this likelihood ratio test statistic might not be robust against model deviation, especially when there is more than one allele associated with the disease. With this consideration, we derived the corresponding score statistic S_p, which is a linear combination of Pearson's chi-square χ^2 and a weighted sum of observed minus expected allele counts in cases. The score test is locally most powerful and since it is evaluated under the null hypothesis, it is expected to be robust against model deviation (El Galta et al., 2004). The score statistic S_p is easy to compute, which enables one to use Monte-Carlo permutations to estimate the empirical p-value of the test statistic (Sham and Curtis, 1995). For a large number of alleles Pearson's χ^2 follows asymptotically a normal distribution (Haldane, 1939). Hence for a large sample size and for a large number of marker alleles the distribution of the score test S_p under the null hypothesis can be approximated by a normal distribution. Another alternative may be to replace the weights in the likelihood ratio statistic proposed by Terwilliger (1995) by equal weights, which might be suitable if the associated allele is less common.

For replicate 50 of the Danacaa population we applied the Pearson's χ^2, the score test and Terwilliger's LR test to the micro-satellite markers D03S0124, D03S0125, D03S0126, and D03S0127 to test their association with Kofendrerd Personality Disorder (KPD). The allelic distribution was compared between a sample of 100 cases and a sample of 50 controls. In order to ensure a high power one might either select more controls, as they are easier to ascertain than cases or to compare the allele frequencies in cases to the allele frequencies in the population if they are known. Since the allele frequencies in controls were supplied by GAW14 we considered the latter option to verify the result of markers that showed significant association with KPD.

5.2 Material and Methods

Score test

Suppose we have a multi-allelic marker. Let p_i be the frequency of ith allele in the controls. Suppose we have n_1 unrelated case chromosomes and n_2 unrelated control chromosomes. Let x_i and y_i be the ith allele counts in cases and controls respectively. The score statistic corresponding to the likelihood proposed by Terwilliger (1995) is

$$S_p = \sum \frac{(x_i - n_1 p_i)^2}{p_i} - \sum \frac{x_i - n_1 p_i}{p_i},$$

where the sum is taken over the alleles. When the allele frequencies are unknown, p_i can be estimated by the frequencies in combined sample $\frac{x_i+y_i}{n_1+n_2}$. When more than one allele is associated with the disease, the score test S_p is expected to perform better than the likelihood ratio, since it sums over the contributions of the alleles.

Data analysis

Firstly we selected four replicates from each of the Aipotu, Karangar and Danacaa populations to perform genome-wide linkage analysis, i.e. we analysed 12 replicates. Each replicate consisted of 100 nuclear families. For each replicate we applied the single-point Spairs allele-sharing scoring function (Whittemore1994) as implemented in the Merlin program (Abecasis et al., 2002) to search for regions with evidence for linkage. The parental genotypes were used to compute the probabilities of sharing 0, 1 or 2 alleles identical by descent. A region on chromosome 3 showed a significant linkage to latent disease locus for several populations at level 0.0001.

For testing the association using the proposed methods, we selected replicate 50 of the Danacaa population, as in this replicate marker D03S0127 showed highly significant linkage to the disease locus with a LOD score greater than 6 ($P < 0.0001$). Flanking markers D03S0126, D03S0125, and

D03S0124 showed borderline linkage with a LOD scores equal to 1.35, 1.45 and 2.42 respectively.

In order to obtain marker genotypes for 50 unrelated controls for the association analysis we purchased packets 149 to 153. The first affected in each family ($n = 100$) was used as a case regardless of being child or parent. We tested for the Hardy-Weinberg equilibrium to each micro-satellite marker in the controls. Then we applied the score test S_p, Pearson's chi-square χ^2, and Terwilliger's likelihood ratio (LR) to study association with KPD. For the score test S_p, Pearson's χ^2 we used Monte-Carlo permutations to estimate the empirical p-values. P-values lower than 0.05, were considered to be significant.

As an alternative to using the controls, we also used the provided allele frequencies as reference allelic frequency distribution for Pearson's χ^2, the score S_p and Terwilliger's LR. Furthermore we also used Terwilliger's LR with equal weights.

Finally additional SNP's in the vicinity of the associated marker D03S0127 were tested for association and the linkage disequilibrium between markers was studied in this region.

5.3 Results

All markers were in Hardy-Weinberg equilibrium proportions. Table 5.1 presents the p-values for the association analysis of various markers with the disease. Marker D03S0127 appeared to be highly significantly associated with the disease. The score S_p and Pearson's χ^2 gave about the same p-value ($P = 0.008$, 0.007), whereas Terwilliger's LR yielded somewhat a larger p-value ($P = 0.033$). For this marker, allele 1 and 3 were 2 and 3.3 times more often present in cases than in controls respectively, whereas allele 6 occurred approximately 2.8 times as often in controls as in cases. Marker D03S0125 showed borderline significant association with KPD. Next we repeated the analysis of association between KPD and marker D03S0127 using the provided allele frequencies of 0.070, 0.206, 0.100, 0.114, 0.048, 0.111, 0.154 and 0.197 for allele 1 to 8 respectively. Again the score statistic S_p and Pearson's χ^2 yielded similar empirical p-values ($P = 0.027$) while LR of Terwilliger and LR with equal weights gave an asymptotic p-value of 0.029 and 0.023 respectively. Compared to the given allele frequencies only allele 3 showed some excessive frequency in the cases and it occurred about 1.7 times as often in cases as in the population.

5.4 Discussion

In this paper we reported results of several methods for studying association between a disease and a multi-allelic marker. Marker D03S0127 located at chromosome 3 showed significant association with the disease. Both score S_p and Pearson's χ^2 tests gave somewhat lower p-values than the Terwilliger's LR test. Further examination shows that marker D03S0127 appeared to have two positively associated alleles. When we assumed known allele frequencies, only one allele was positively associated with the disease and all test statistics yielded similar p-values. Perhaps the fact that there are two associated alleles might be the reason why the Terwilliger's LR test yielded somewhat lager p-value in this data set. To study whether this holds in general, an extensive simulation study is needed.

In addition to Pearson's χ^2 and LR, a new test statistic was applied to the Gaw14 simulated data. The new test statistic is derived based on the score function under the null hypothesis. So it possesses the usual optimal properties as other score test statistics: locally most powerful and robust against model misspecification. In contrast to the LR test statistic the new score statistic is very easy to compute and enables to use Monte-Carlo to derive empirical p-values. Details of the derivation of this score statistic as well as a simulation study of its power will be extensively provided in another paper.

The parameter λ is a preferred measure of allelic association since it is directly related to recombination fraction and it is less sensitive to allele frequencies than other measures (Devlin and Risch, 1995). However, when allelic association is modelled by means of λ it is not straightforward to adjust for other covariates. Houwing-Duistermaat and Elston (2001) discussed various ways to quantify allelic association and estimate the location of gene responsible for disease using logistic regression models. As an alternative to λ, the log relative risk as measured by the regression coefficient in the logistic model may be used to allow for adjustment of other covariates. More research is needed to build this kind of flexible models.

Applying Pearson's chi-square with one degree of freedom to 19 SNP's revealed strong association between KPD disease and two di-allelic markers in this region: SNP B03T3056 and SNP B03T3057. Furthermore LD observed between B03T3056 and B03T3057 and B03T3056 and D03S0127 further confirms the precedent results.

Conclusions

All test statistics showed significant association between D03S0127 and KPD. Probably due to the presence of more than one positively associated allele the Pearson's χ^2 and score tests yielded lower p-values than the Terwilliger's likelihood ratio test in this data set.

TABLE 5.1: Results of association tests for multi-allelic markers

Marker	Number	Allele counts		Score S_p*	Terwillger′s	χ^{2*}
	of alleles	cases	controls		LR	
D03S0124	5	9 (0.045)	6 (0.060)	0.910	0.500	0.930
		31 (0.155)	16 (0.160)			
		28 (0.140)	11 (0.110)			
		75 (0.375)	39 (0.390)			
		57 (0.275)	28 (0.280)			
D03S0125	4	12 (0.060)	3 (0.030)	0.069	0.070	0.053
		79 (0.395)	27 (0.270)			
		55 (0.275)	31 (0.310)			
		54 (0.270)	39 (0.390)			
D03S0126	7	22 (0.110)	16 (0.160)	0.426	0.500	0.428
		8 (0.040)	5 (0.050)			
		64 (0.320)	37 (0.370)			
		16 (0.080)	10 (0.10)			
		23 (0.115)	6 (0.060)			
		51 (0.255)	18 (0.180)			
		16 (0.080)	8 (0.080)			
D03S0127	8	12 (0.060)	3 (0.030)	0.008	0.033	0.007
		33 (0.165)	20 (0.200)			
		33 (0.165)	5 (0.050)			
		23 (0.115)	13 (0.130)			
		8 (0.040)	4 (0.040)			
		14 (0.070)	20 (0.200)			
		39 (0.195)	16 (0.160)			
		38 (0.190)	19 (0.190)			

* P-values were obtained using 10000 Monte Carlo simulations

Chapter 6

Generalizing Terwilliger's likelihood approach: a new score statistic to test for genetic association

R. el Galta, S. Uitte de Willige, M.C.H. de Visser, L. Hsu, J.J. Houwing-Duistermaat

Abstract

In this report, we propose a one degree of freedom test for association between candidate genes and binary traits. We consider the situation of observed haplotypes, i.e. haplotype tagging single nucleotide polymorphisms are typed from which the haplotypes can be derived with almost 100% certainty. The method is a generalization of the approach of Terwilliger (1995) and is especially powerful for the situation of one associated haplotype. As alternative to the likelihood ratio statistic, we derive a score statistic, which is locally most powerful. By means of a simulation study, we compare the performance of the score statistic to Pearson's chi-square statistic and the likelihood ratio statistic proposed by Terwilliger (1995). We illustrate the method on three candidate genes studied in the Leiden Thrombophilia Study (Uitte de Willige et al., 2005). We conclude that the statistic follows a chi square distribution under the null hypothesis and that the score statistic has often more power than Terwilliger's likelihood ratio statistic, especially for variants with frequencies between 0.1 and 0.4 and which have a small impact on the studied disorder.

Submitted for publication

6.1 Introduction

For genetic association studies, single nucleotide polymorphisms (SNPs) are popular genetic markers, because they are stable, easier to type than the micro satellite markers and distributed with a high density over the genome (Collins et al., 1997). The pattern of variants appears to be structured with sets of physically close SNPs inherited together in blocks (Daly et al., 2001; Gabriel et al., 2002). Within a block, SNPs are highly correlated (linkage disequilibrium) and each block contains only a few common haplotypes (Gabriel et al., 2002). These haplotypes can be described by a small number of SNPs. Several methods are described to identify the minimal informative subset of SNPs, so called haplotype tagging SNPs (htSNPs) (see Sebastiani et al. (2003) and Stram (2004) for references).

In the case of zero recombination within a block, the haplotypes can be uniquely identified and n haplotypes can be described by n-1 SNPs (Bafna et al., 2003; Clark, 2004; Clayton et al., 2004). Stram (2004) introduced a measure for reliability of haplotype assignment r_h^2 (Epstein and Satten, 2003; Satten and Epstein, 2004). For r_h^2 close to one, the haplotypes are known and association between the observed haplotypes and a disease can be studied by comparing the haplotype frequencies of the gene in cases and controls. Examples in the literature of genes with known haplotypes are APOE (Fullerton et al., 2000), CRP (Carlson et al., 2005) and the fibrinogen gamma (FGG), fibrinogen alpha (FGA), and fibrinogen beta (FGB) genes (Uitte de Willige et al., 2005). Note that Carlson et al. (2005)) used the haplo.stats package (Schaid et al., 2002) which does take into account the uncertainty in phase by using the EM algorithm and then they noted that the same results could be obtained by using the htSNPs. The reason for this is that the htSNPs uniquely correspond to the haplotypes.

If a causal mutation occurred in one haplotype in the past, it would be natural to consider haplotypes rather than individual genotypes (Clark, 2004) and to assume that only one haplotype carries the causal variant (Terwilliger, 1995). The haplotype frequencies in cases can therefore be modelled by the frequencies in controls plus one additional parameter, which accounts for the excessiveness of the causal haplotype. Under this assumption, a statistic which tests the null hypothesis of this additional parameter equal to zero will have more power than the classical chi-square test, since the latter tests for any differences in frequencies between cases and controls.

Terwilliger (1995) proposed this model in the context of multi-allelic markers and used the likelihood ratio test, i.e. comparing the likelihood under the alternative to the likelihood under no association, for testing. Since it is unknown which marker allele is associated with the disease, the likelihood corresponding to this model is a weighted sum over all alleles i of conditional likelihoods given that allele i is over-represented in the set of cases. As for weights Terwilliger (1995) proposed to use the allele frequencies in the population from which the cases and controls were sampled. The excess frequency of the associated allele in cases is modelled by the parameter λ which is the fraction attributable at risk (Clayton, 2000). The corresponding log likelihood function has a number of unusual features. For example, the allele frequencies that are used as weights are unknown parameters in the conditional likelihood functions. The score function, the first derivative of the log likelihood function with respect to λ evaluated at $\lambda = 0$ is a constant zero for any observed data. Therefore, the distribution of the likelihood ratio (TLR) statistic under the null hypothesis is not straightforward and the $50 : 50$ mixture of chi square distributions of null and one degrees of freedom, which was suggested by Terwilliger (1995), appears to yield conservative p-values (Sham et al., 1996).

Under Hardy-Weinberg equilibrium and complete linkage disequilibrium, the observed haplotype frequencies in the controls agree with the population frequencies many generations ago. Hence the frequency of a haplotype in the population corresponds to the prior probability that the mutation occurred on that haplotype. If one is focussed in detecting common haplotypes with a small impact on the trait, an alternative for the haplotype frequencies is a flat prior. The advantages of using a flat prior are that the probabilities do not have to be estimated. Furthermore the first derivative of the log likelihood with respect to λ evaluated at $\lambda = 0$ is not equal to zero and the score statistic as alternative for the likelihood ratio statistic can be used. Advantages of score statistics compared to likelihood ratio statistics are that they are also locally most powerful, and because they do not need to evaluate the log likelihood under the alternative, they are often easier to compute and robust to small model deviations under the alternative (Cox and Hinkley, 1974).

In this paper we consider the score statistic as alternative to the classical chi-square and the original TLR statistic of Terwilliger (1995) in the context of haplotypes. We also include the likelihood ratio statistic corresponding to the log likelihood using equal weights. We carried out a simulation to study the performance of the four statistics under the null hypothesis and to compare the power of the four statistics under various alternatives. Finally we illustrate the proposed statistics by an association analysis of three candidate genes in the Leiden Thrombophilia Study (LETS) (Uitte de Willige et al., 2005).

6.2 Methods

Let m be the number of haplotypes describing most of the genetic variation in a gene. Assume that the haplotype frequencies are in Hardy-Weinberg equilibrium proportions. Let $p = (p_1, \cdots, p_m)$ be the vector of haplotype frequencies in controls. Assume that only one haplotype denoted with index i is over-represented in the cases, then the haplotype frequencies in the cases can be modelled as $q_i = p_i + \lambda(1 - p_i)$ and $q_j = p_j - \lambda p_j$ for $j \in (1, \cdots, i-1, i+1, \cdots, m)$. Let $x = (x_1, \cdots, x_m)$ and $y = (y_1, \cdots, y_m)$ be vectors of haplotype counts in the cases and the controls, respectively, and let n_1 and n_2 be the total number of case chromosomes and of control chromosomes, respectively, and $n = n_1 + n_2$. Then the conditional likelihood L_i given that haplotype i carries the mutation is given by

$$L_i(\lambda, p|x, y) = (p_i + \lambda(1 - p_i))^{x_i}(1 - \lambda)^{n_1 - x_i} \prod_{j \neq i}^{m} p_j^{x_j} \prod_{j=1}^{m} p_j^{y_j} \tag{6.1}$$

and the likelihood proposed by Terwilliger is equal to

$$L(\lambda, p|x, y) = \sum_{j=1}^{m} p_j L_j, \tag{6.2}$$

with L_j given in formula (6.1).

It is easy to see that likelihood function (6.2) can be generalized to the following likelihood function:

$$L(\lambda, p|x, y, w) = \sum_{j=1}^{m} w_j L_j,$$

with L_j given in formula (6.1) and $w = (w_1, \cdots, w_m)$ a vector of known positive weights restricted by $\sum_{j=1}^{m} w_j = 1$. The first derivative of the log likelihood $l(\lambda, p|x, y, w) = log(L(\lambda, p|x, y, w))$ to λ evaluated in $\lambda = 0$ is equal to

$$\begin{aligned} U_w &= \frac{\partial}{\partial \lambda} l(\lambda, p|x, y, w)_{|\lambda=0} \\ &= \sum_{j=1}^{m} \frac{w_j(x_j - n_1 p_j)}{p_j}. \end{aligned}$$

For known allele frequencies p_j, the distribution of the U_w under H_0 can be approximated by the normal distribution with zero mean and variance $\mathrm{VAR}[U_w] = n_1(\sum_{j=1}^m w_j^2 p_j^{-1} - 1)$. Note that $U_w = 0$ when for all j $w_j = p_j$.

Often the haplotype frequencies are unknown and have to be estimated from the data. Under the null hypothesis we can estimate the frequencies from the combined sample of cases and controls by $\hat{p}_j = \frac{x_j+y_j}{n}$ and an estimate of the score statistic U_w is given by

$$\hat{U}_w = \sum_{j=1}^m \frac{w_j(x_j - n_1\hat{p}_j)}{\hat{p}_j}.$$

under H_0, $\hat{U}_w$ has approximately mean equal to zero and variance

$$\mathrm{VAR}(\hat{U}_w) \approx n^{-1}n_1n_2(\sum_{j=1}^m w_j^2 p_j^{-1} - 1). \tag{6.3}$$

Note that the variance $\mathrm{VAR}(\hat{U}_w)$ is increased by n_2/n fold compared to the variance $\mathrm{VAR}(U_w)$ because the allele frequencies are estimated from the data. Now the score statistic $\hat{S}_w$ is defined by

$$\hat{S}_w = \frac{\hat{U}_w^2}{\hat{\mathrm{VAR}}(\hat{U}_w)},$$

where $\hat{\mathrm{VAR}}(\hat{U}_w)$ is obtained by replacing p_j by its estimate $\hat{p}_j$ in formula (6.3). When all haplotypes are common, a natural choice of weights is $w_j = \frac{1}{m}$.

Under the alternative hypothesis of the presence of one positively associated haplotype i, the expectation of $\hat{U}_{\frac{1}{m}}$ is

$$E_{H_A}[\hat{U}_{\frac{1}{m}}] \quad \approx \quad \frac{n_1n_2\lambda}{n - n_1\lambda}(\frac{n}{n_1q_i + n_2p_i} - m), \tag{6.4}$$

with $\frac{n_1p_i+n_2q_i}{n}$ the frequency of the associated haplotype in the combined sample and $q_i = p_i + \lambda(1-p_i)$. When $\frac{n_1p_i+n_2q_i}{n}$ is larger than $\frac{1}{m}$ this expectation becomes negative. Therefore we propose a chi-square distribution with one degree of freedom to approximated the distribution of this statistic under the null hypothesis.

6.3 Results

Simulation study

By means of a simulation study, we first evaluated the type I error rate of the score statistic $\hat{S}_{\frac{1}{m}}$, Pearson's chi square χ^2, the likelihood ratio with equal weights LR, and the Terwilliger's likelihood ratio with weights equal to p_j's TLR. For the score statistic we used the chi square distribution with one degree of freedom to approximate the distribution under the null hypothesis. For the LR and TLR statistics we used the 50:50 mixture of two chi squares with zero and one degree of freedom. We generated 10,000 samples of 200 case chromosomes and 200 control chromosomes from the multinomial distributions with probabilities $p_1 \cdots p_m$ for m equal to 4, 5, 8, 10, 15 and 20 haplotypes. Similar to the simulation described by Terwilliger, the frequency of the most common haplotype, p_1, was set to 0.5, whereas the remaining haplotypes were equally frequent $(0.5/(m-1))$. The results are shown in left columns of table 6.1.

For all m, the type I error rates of the score statistic $\hat{S}_{\frac{1}{m}}$ were maintained at the nominal error rate. For $m < 10$, the type I error rates of Pearson's chi square corresponded to the nominal level. However for larger m the type I error rates became conservative due to sparse data. For all m considered, the type I error rates for the TLR statistic were conservative (< 0.03). The type I error rates for the LR statistic were also somewhat small (≈ 0.04), but were better than the type I error rates for the TLR statistic.

To evaluate the power of the statistics, we generated 10,000 samples of n_1 case chromosomes and n_2 control chromosomes from the multinomial distributions with probabilities $p_1(1-\lambda)+\lambda$ and $p_j(1-\lambda)$ for $j = 2 \cdots m$ for cases and p_j $j = 1 \cdots m$ for controls, respectively. First, we considered the model used by Terwilliger (1995). The most common haplotype frequency p_1 in controls was again set to 0.5 and this haplotype was more frequent in cases. The parameter λ was fixed to 0.5 which corresponds to a haplotype frequency of 0.75 in the cases. The number of haplotypes m was again set to 4, 5, 8, 10, 15 and 20. The number of case and control chromosomes, n_1 and n_2 were now 100. The results are shown in the right columns of table 6.1.

For $m < 8$, the power of Pearson chi square was good. For $m \leq 8$ the power of the score statistic $\hat{S}_{\frac{1}{m}}$ was similar to the power of TLR statistic, while for $m > 8$, the TLR statistic had higher power than $\hat{S}_{\frac{1}{m}}$. For $m > 8$, the haplotype frequencies of the non associated haplotypes become too small yielding a large variance of the score statistic (see formula 6.3). The LR statistic appeared to perform worse than both $\hat{S}_{\frac{1}{m}}$ and TLR. Therefore we did not consider this statistic in the following simulations.

Second, we studied the power of the Pearson's chi square, $\hat{S}_{\frac{1}{m}}$, and TRL as function of the excess frequency λ for various values of the frequency of the associated haplotype $p_1 = 0.1, 0.2, 0.3, 0.4$ and 0.5. The remaining haplotypes were again equally frequent. We restricted ourselves to a number of observed haplotypes m of 5 and 8, because most of the genes can be described by up to 8 common variants. The parameter λ was varied between 0 and 0.5. The number of chromosomes n_1 and n_2 were 200. We used a nominal significance level of 0.05. The results are depicted in figure 6.1.

For $p_1 = 0.5$, the score statistic $\hat{S}_{\frac{1}{m}}$ and likelihood ratio TLR performed similarly, and better than Pearson's chi square. For $m = 5$ and $p_1 = 0.4$ or 0.3 and for $m = 8$ and $p_1 = 0.4$, 0.3 or 0.2, the score statistic $\hat{S}_{\frac{1}{m}}$ performed better than TLR especially for small λ. For $p_1 = 0.2$ and $m = 5$ all three statistics had similar power. For $p_1 = 0.1$ and $m = 5$ or $m = 8$, Pearson's chi-square performs similar to TLR. Both statistics performed better than $\hat{S}_{\frac{1}{m}}$ except for$\lambda \leq 0.1$ and $p_1 = 0.1$ and $m = 5$. Note that for $p_1 = 0.1$ and $m = 5$, the power of the score statistic was small around $\lambda = 0.2$. This drop in power was due to the fact that the expectation of $\hat{S}_{\frac{1}{m}}$ becomes small (see formula 6.4).

Especially for common variants with frequency p_1 of 0.3 or 0.2 and a small impact on the disease ($\lambda \leq 0.1$), the score statistic performed well. For $m = 5$ and $m = 8$ and $p_1 = 0.3$, the gain in power of the score statistic compared to TLR statistic was about 4% and 8% for λ of 0.05 and 0.1 respectively. For $m = 5$ and $p_1 = 0.2$, both statistics performed similar. For $m = 8$ and $p_1 = 0.2$ the gain in power of the score statistic was large, namely 7% and 12% for λ of 0.05 and 0.1 respectively.

Data example

We applied the three statistics to a study on association between haplotypes of fibrinogen alpha (FGA), beta (FGB) and gamma (FGG) and the risk of deep venous thrombosis in LETS (Koster et al., 1993; van der Meer et al., 1997). Fifteen haplotype tagging SNPs were typed in 474 cases and 474 controls (Uitte de Willige et al., 2005). Within the three genes, the SNPs were in high linkage disequilibrium ($r_h^2 > 0.95$). The number of common haplotypes (frequency larger than 5%) describing FGG, FGA and FGB were three, five and five respectively. Since we focus on common haplotypes, we pooled the rare haplotypes with the less frequent haplotype category with frequency larger than 5%. In this analysis we considered p-values below 0.05 to be significant. In table 6.2 the data are described and the results are given.

For all genes, haplotype H2 appeared to be more frequent in the cases than in the controls. For FGG, FGA and FGB the allelic odds ratios of presence of H2 versus the rest was 1.34, 1.29 and 1.28 respectively. Note that these odds ratios were rather similar while the p-values of the corresponding chi square statistics were different namely, 0.008, 0.051 and 0.059 respectively. The difference in p-values was caused by the difference in degrees of freedom of the chi square statistics and the frequencies of the other haplotypes. From the results of the standard chi-square statistics we concluded that only FGG was significantly associated to thrombophilia.

The p-values of the TLR were respectively 0.004, 0.021 and 0.078. These p-values were in line with the estimates of λ, namely 0.09, 0.07 and 0.05 respectively. Since FGA and FGB both had 5 variants, the frequency of the associated haplotype H2 was 0.3 and 0.2 respectively and the λ's were rather small, the score statistic should have more power than TLR for these genes.

The p-values of the score statistic $\hat{S}_{\frac{1}{m}}$ were 0.021, 0.007, 0.024 for FGG, FGA and FGB respectively. Indeed the p-values for FGA and FGB were smaller than the corresponding p-values of TLR statistic. The p-values for FGG and for FGB were larger than for FGA, because the frequencies of H2 in the combined case control sample were around $\frac{1}{m}$ (see formula 6.4). Based on the results of the score statistic, all genes were significantly associated to thrombophilia.

6.4 Discussion

In this report we have derived a new score statistic to test for association between a candidate gene and a binary trait. For candidate genes with a small impact on the disease and five to eight observed variants this new statistic appears to perform better than Terwilliger's LR statistic. Moreover the statistic is easy to compute and follows a chi square distribution under the null hypothesis. For more than eight variants, Terwilliger's LR statistic is more powerful. However by pooling less frequent haplotypes, the number of observed haplotypes is often smaller than eight.

Instead of using multi-locus haplotypes, some authors advocate to test each locus separately (Clayton et al., 2004). However, since mutations arise on haplotypes and because of the high degree of linkage disequilibrium, we prefer haplotype based tests for highly structured genes (Clark, 2004). For candidate genes that exists of several blocks we suggest to apply the test for each block separately. Alternatively the uncertainty has to be taken into account.

For multi allelic markers, Slager and Schaid (2001) and Czika and Weir (2004) proposed a multi allelic version of the trend test to test for association of genotypes. Also for genotypes at multi allelic markers, Houwing-Duistermaat and Elston (2001) considered various ways to test for association using logistic regression models. If Hardy-Weinberg equilibrium does not hold, these methods should be preferred. Further, the parametrization used has a lower bound for the parameter λ which is often larger than -1. An alternative, more symmetrical parametrization might be the log relative risk corresponding to a logistic model. Moreover, in logistic models adjustments for other covariates are easily made. More research is needed to build these kind of models and derive corresponding tests for pairs of haplotypes.

We conclude that by choosing alternative weights, in particular constant weights, in the likelihood of Terwilliger, a set of new powerful and robust statistical tests was derived. For genetic association studies aiming to identify common associated variants, we recommend to first pool rare variants and then apply both the standard Pearson's chi square statistic as well as the new score statistic. By using both statistics more insight in the data can be obtained. A program is freely available which computes the statistics and corresponding p-values.

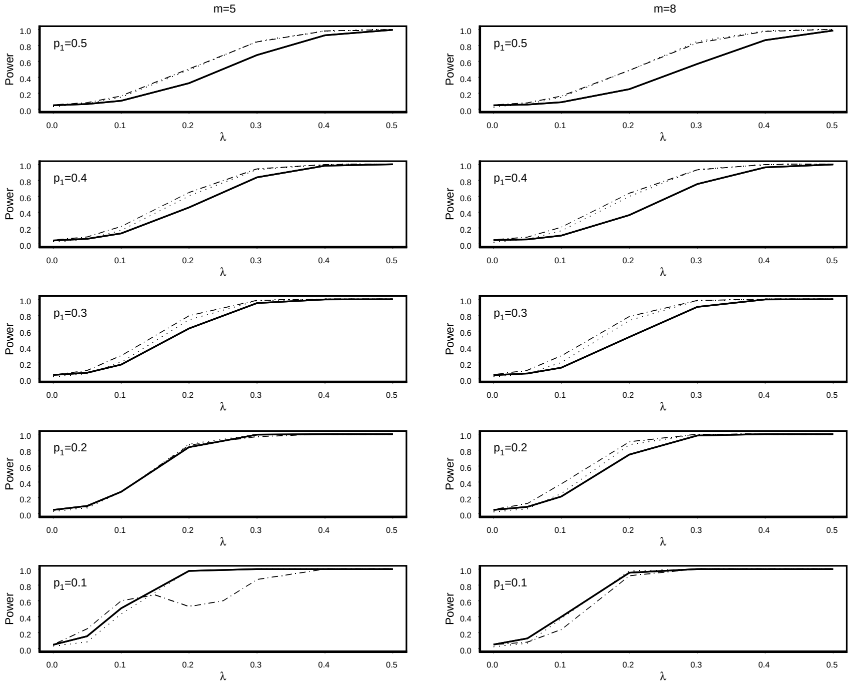

FIGURE 6.1: *Power curves of χ^2 (—), $\hat{S}_{\frac{1}{m}}$ (-·-·) and TLR (·····) as function of λ for various values of the frequency of the associated variant $p_1 = 0.1, 0.2, 0.3, 0.4, 0.5$ and $m = 5$ (left column), and $m = 8$ (right column)*

TABLE 6.1: Type I error rate and power of the statistics χ^2 , $\hat{S}_{\frac{1}{m}}$, LR, TLR.

		type I error rate				power when $\lambda = 0.5$			
m	nominal	χ^2	$\hat{S}_{\frac{1}{m}}$	LR	TLR	χ^2	$\hat{S}_{\frac{1}{m}}$	LR	TLR
4	0.05	0.053	0.052	0.042	0.032	0.91	0.94	0.91	0.95
	0.01	0.009	0.010	0.010	0.007	0.78	0.84	0.80	0.86
	0.001	0.001	0.001	0.001	0.000	0.53	0.62	0.56	0.65
5	0.05	0.053	0.048	0.040	0.032	0.88	0.94	0.89	0.95
	0.01	0.010	0.010	0.010	0.007	0.71	0.83	0.77	0.87
	0.001	0.001	0.001	0.001	0.000	0.44	0.59	0.53	0.64
8	0.05	0.048	0.049	0.038	0.035	0.77	0.92	0.86	0.95
	0.01	0.008	0.010	0.008	0.004	0.53	0.80	0.74	0.86
	0.001	0.001	0.000	0.001	0.000	0.25	0.54	0.50	0.65
10	0.05	0.045	0.051	0.034	0.023	0.70	0.90	0.84	0.95
	0.01	0.006	0.008	0.008	0.003	0.43	0.76	0.70	0.85
	0.001	0.000	0.001	0.000	0.000	0.18	0.49	0.47	0.60
15	0.05	0.043	0.048	0.041	0.020	0.58	0.86	0.80	0.95
	0.01	0.007	0.010	0.010	0.003	0.31	0.69	0.65	0.85
	0.001	0.000	0.001	0.002	0.000	0.09	0.42	0.42	0.63
20	0.05	0.043	0.052	0.045	0.021	0.48	0.84	0.77	0.95
	0.01	0.005	0.011	0.010	0.004	0.20	0.64	0.62	0.84
	0.001	0.000	0.001	0.002	0.000	0.04	0.35	0.39	0.60

TABLE 6.2: Descriptives and results of genetic association on LETS.

haplotype	case chromosomes	control chromosomes	χ^2	$\hat{S}_{\frac{1}{m}}$	TLR	$\hat{\lambda}$
FGG ($n_1 = 938, n_2 = 942$)			0.008	0.021	0.004	0.09
H1	334 (36.3)	366 (38.9)				
H2	315 (33.2)	254 (27.0)				
H3+H4	289 (30.5)	321 (34.1)				
FGA ($n_1 = 936, n_2 = 942$)			0.051	0.007	0.021	0.07
H1	270 (28.8)	266 (28.2)				
H2	320 (34.2)	270 (28.7)				
H3	95 (10.2)	117 (12.4)				
H4	100 (10.7)	121 (12.9)				
H5	151 (16.1)	168 (17.8)				
FGB ($n_1 = 936, n_2 = 932$)			0.059	0.024	0.078	0.05
H1	328 (35.0)	310 (33.3)				
H2	231 (24.7)	189 (20.3)				
H4	135 (14.4)	149 (16.0)				
H6	128 (13.7)	143 (15.3)				
H3+H5+H7	114 (13.2)	141 (15.1)				

CHAPTER 7

Phenotypic Subtypes in Attention Deficit Hyperactivity Disorder in an Isolated Population

E.A. Croes, R. el Galta, J.J. Houwing-Duistermaat, R.F. Ferdinand, S. Lopez Leon, T.A. Rademaker, M.C. Dekker, B.A. Oostra, F. Verhulst and C.M. Van Duijn

Abstract

Background: We address the use of two informants in genetic studies and whether familial aggregation is similar for the three phenotypic subtypes of ADHD. Methods: Lifetime ADHD was diagnosed in a Dutch isolated population using parents and teachers as informants, creating two subgroups (one or two informants), then further divided into three phenotypic categories (inattentive, hyperactive/impulsive, combined). Genealogy was collected for all patients. Mean kinship coefficients for the subgroups were calculated. Results: Fifteen of twenty-six children were linked to a common ancestor within ten generations. The mean kinship coefficient of patients confirmed by two informants was significantly higher than in patients only scored positive by one informant (p=0.03). All patients of the inattentive subtype were connected to a common ancestor, which was significantly higher (0.028) than expected. 81% of these patients derive of consanguineous marriages, also higher than expected. This means that recessive mutations may be involved in the inattentive subtype. These patients were more closely related than those with the other phenotypes ($p < 0.01$*). Conclusion: Our data suggests that using two informants in diagnosing ADHD helps identify a phenotype with a strong genetic component. The inattentive phenotype showed strong familial clustering and evidence of a recessive origin.*

Published in *Eur J Epidemiol*. 2005; **20**:789-94

7.1 Introduction

Attention Deficit Hyperactivity Disorder (ADHD) is one of the most common psychiatric disorders in children (Verhulst et al., 1997). It is characterised by inattention, distractibility, over-activity and poor impulse control (Barkley, 2003). Males are more frequently affected than females (Gaub and Carlson, 1997). It has been suggested that ADHD is a risk for academic problems, antisocial behaviour and substance abuse in adolescence and adulthood (Barkley, 1996; Cantwell, 1996; Hill and Schoener, 1996). There is strong familial aggregation of ADHD in families (Faraone et al., 2001). The heritability of ADHD has been estimated to be 0.50-0.98 (Faraone et al., 2000; Levy et al., 1997; Thapar et al., 1995). The mode of inheritance is thought to involve genes with dominant effects (Lopera et al., 1999), but others have argued that the inheritance is more complex because many different genes are involved. A major problem hampering genetic research of ADHD, and psychiatric genetic research in general, is the difficulty in defining the phenotype (Thapar et al., 1999). The phenotype is diverse, including patients with inattention, patients who are hyperactive/impulsive, and those with both. There may be a difference in the contribution of genes to the clinical phenotype.

Another problem to address is that for ADHD no biochemical tests or opportunities to support the diagnosis with imaging are available. Therefore, for children with ADHD, diagnostic information is based on reports of observations of behaviour in different contexts. By convention, in order to meet DSM-IV criteria for ADHD, symptoms need to be present in at least two of three settings (home, school, work) (Shaffer et al., 2000). Agreement between various informants such as parents and teachers is low, ranging between 0.30 and 0.50 (Achenbach, McConaughy, and Howell, Achenbach et al.; Ferdinand et al., 2003). Variation in the child's behavior across different situations, and differences in the way different observers judge the child's behavior, are two possible sources of cross-informant variance (van der Ende, 1999). By combining information from both parents and teachers, the validity of the diagnosis of ADHD has been found to improve (de Nijs et al., 2004; Mitsis et al., 2000; Verhulst et al., 1994). The first question we address is whether the use of two informants is helpful in genetic studies. Second we addressed the question whether familial aggregation is similar for the three phenotypic subtypes of ADHD.

7.2 Methods

Study Population

This study was conducted within the framework of the program Genetic Research in Isolated Populations (GRIP). Approximately 150 individuals founded this population in the Southwest of the Netherlands in the middle of the 18th Century. The population is characterised by minimal migration ($< 5\%$) and rapid growth (700 inhabitants in 1848 and 20,000 inhabitants at present). For this population the genealogical records are available since 1750. The GRIP population has proved to be suitable to study complex diseases such as type 1 and type 2 diabetes mellitus (Aulchenko et al., 2003; Vaessen et al., 2002). For this study, two paediatric neurologists, who obtained referrals from this genetically isolated village, asked all their patients diagnosed with ADHD to participate in this study (n=49; 22% female). Thirty-three (67.3%) patients and their parents agreed to participate.

This programme has obtained approval of the Medical Ethical Committee. All parents provided informed consent for themselves and for their children. Children over the age of eleven co-signed the informed consent.

Psychiatric Assessment

The Dutch version of the National Institute of Mental Health Diagnostic Interview Schedule for Children (NIMH DISC or DISC)-IV was used to assess DSM-IV diagnoses (Ferdinand and van der Ende, 2000; Shaffer et al., 2000). Psychologists and psychology students trained by the authors of the Dutch DISC-IV administered the DISCs. The training schedule used was similar to the schedule used by the authors of the original English version, at Columbia University, New York. To obtain information regarding a wide range of current DSM-IV Axis 1 diagnoses, parent DISCs (DISC-P) were administered during face-to-face contacts, at a community general health centre or in a children's hospital. Furthermore, lifetime ADHD symptoms were also assessed with the DISC-P. Teachers were interviewed with the ADHD section (current, not lifetime) of the teacher DISC (DISC-T) via telephone. The child version of the DISC (DISC-C) was not applied since most of the children included in our sample were too young (< 11 years of age). To assess the presence of diagnoses besides ADHD, the following diagnoses were assessed with the DISC-P: social phobia, separation anxiety disorder, specific phobia, agoraphobia, generalised anxiety disorder, panic disorder without agoraphobia, panic disorder with agoraphobia, obsessive compulsive disorder, posttraumatic stress disorder, major depressive disorder, dysthymia, bipolar disorder, oppositional disorder and schizophrenia.

Phenotypic subgroups (inattentive, hyperactive/impulsive, and combined) of ADHD were formed based on application of the DSM-IV criteria that had been assessed with the DISC. Current ADHD diagnoses were based on information from parents and teachers. Two types of ADHD diagnoses were derived: (1) 'based on one informant', and (2) 'based on two informants'. A diagnosis of ADHD based on one informant was applied when either parent or teacher scored six or more criteria positive for the inattentive, hyperactive or combined phenotype, while the other informant scored less than three criteria positive. A diagnosis of ADHD based on two informants was applied when one informant scored six or more criteria of one of the ADHD subgroups positive and the second informant scored three or more criteria positive. The threshold of '3 criteria positive' was chosen arbitrarily for the purpose of the present study. DSM-IV does not provide explicit rules for the number of criteria that need to be positive in 2 settings to obtain an ADHD diagnosis. It merely states that symptoms have to be present in at least 2 settings. If a child did not fulfil criteria for current ADHD with the DISC-P, lifetime information from the DISC-P was used to obtain a lifetime diagnosis of ADHD, based on parent information.

Genealogical information

Genealogical information comprising the name, date, and place of birth of parents, grandparents and great-grandparents was collected during a home interview. This genealogical information was extended up to 22 generations using municipal and church registers and data from a large genealogy database holding genealogical information on 60,000 individuals from this region in the Netherlands (Vaessen et al., 2002).

Statistical analysis

The relationship between two patients was expressed as the kinship coefficient. This is the probability that variation in the genome of a patient is identical by descent to a randomly drawn allele at the same locus of another patient. For example the kinship coefficient is 0.25 for sib-pairs, and 0.125 for cousins, meaning that the probability of a random allele genotyped in a sib-pair or cousin-pair to be identical by descent is 0.25 and 0.125 respectively. Kinship coefficients were calculated for all pairs of patients with PEDKIN, using all information contained in the genealogical database (Zwetselaar, 2003).

Furthermore, mean kinship coefficients as well as Inbreeding coefficients were computed for each subgroup.

The null hypothesis of no differences between kinship coefficients of two subgroups was tested using a statistic (Z) as outlined in the appendix. This statistic Z is based on the difference between the means of the logarithm of the kinship coefficients.

To assess whether the number of patients connected to a common ancestor and the number of patients derived of consanguineous marriages is larger in particular subgroups than expected based on the population structure respectively, 100 random sets of controls were sampled from the pedigree.

7.3 Results

Of the thirty-three patients who agreed to participate, two were excluded because their genealogy could not be worked up. Baseline characteristics of the remaining study population and the co-morbidity found are presented in Table 7.1. Five children did not fulfil criteria for any of the definitions of ADHD used in the present study; these were excluded from further analyses. In the remaining group of twenty-six ADHD patients, the mean age at the time of the study was 10.1 years, and 23.1 % of patients were female. Oppositional disorder (54%) and specific phobias (27%) were the most prevalent co-morbid diagnoses. Eleven patients fulfilled the DSM-IV criteria for the combined type of ADHD, twelve for the predominantly inattentive, and three for the predominantly hyperactive/impulsive type.

Based on genealogical information, fifteen out of twenty-six patients (58%) could be linked to one common ancestor within ten generations (Figure 7.1). In nine patients the inbreeding coefficient was higher than 0.001 (range 0.001 - 0.027). The parents of seven patients were related within four to seven generations (patients 1, 5, 8, 9, 10, 12, 15; Figure 7.1).

The mean kinship coefficient was highest for children with the inattentive subtype of ADHD and lowest for those with the hyperactive/impulsive subtype (Table 7.2). Children with the inattentive phenotype were significantly more closely related than those with the combined type ($p < 0.01$). All of the patients with a consistent diagnosis of the inattentive subtype were connected to a common ancestor, which is significantly higher (p=0.028) than expected based on the structure of the population. Eighty-one percent of these patients derive of consanguineous marriages which is also significantly increased (p=0.015). We further found that children with a diagnosis of ADHD confirmed by two informants (mean kinship coefficient 0.0029) were significantly more closely related than children in whom the diagnosis was only confirmed by one informant (mean kinship coefficient 0.0005; p=0.03).

TABLE 7.1: Baseline characteristics of the study sample and co-morbidities

	All	ADHD ever	No ADHD
Number of subjects	31	26	5
Mean age at examination (range)	10 (6-16)	10 (6-16)	10 (8-13)
Females (%)	22.6	23.1	20.0
Co-morbidity			
Social phobia	2	2	0
Separation anxiety disorder	4	4	0
Specific phobia	8	7	1
Agoraphobia	3	3	0
Generalised anxiety disorder	1	1	0
Panic disorder with agoraphobia	1	1	0
Obsessive compulsive disorder	2	2	0
Oppositional disorder	15	14	1
Conduct disorder	2	2	0

Co-morbidities are based on DISC-P. Selective mutism, panic disorder without agoraphobia, posttraumatic stress disorder, major depressive disorder, dysthymia, bipolar disorder, and schizophrenia were not present in any of the patients.

TABLE 7.2: Mean kinship coefficient in ADHD phenotypes*

ADHD phenotype	Number of patients	Kinship coefficient
Inattentive	12	0.0046
Hyperactive/impulsive	3	0.00002
Combined	11	0.0030

**Based on DSM-IV criteria.*

7.4 Discussion

There are two important findings in this study. First, we found that adding diagnostic information of a second informant results in a group of patients who are genetically more closely related than patients in whom the diagnosis is based on one informant. This finding indicates that a consistent diagno-

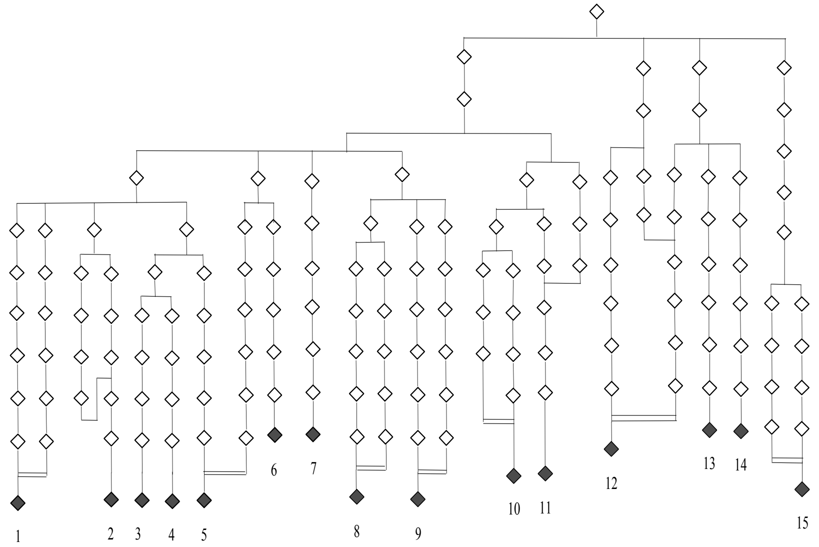

FIGURE 7.1: *Pedigree of the kindred. The symbol on top represents the common ancestor. Filled symbols indicate individuals affected with ADHD. The double line denotes a marriage between parents with a shared ancestor. A diamond symbol has been used to mask the sex of the patients, in order to protect patient confidentiality.*

sis of ADHD, confirmed by a second informant, is more suitable for future gene-finding studies. Second, we showed that children with the inattentive subtype of ADHD are in our population genetically more cluster related than those with the combined type. We confirmed previous studies which found that ADHD clusters in families (Faraone et al., 2001). We found, however, that patients with the inattentive phenotype were more closely related. Using extensive data-based genealogical information of the patients included in this study, we also found evidence of inbreeding. The presence of inbreeding strongly suggests that recessive genes are involved. In those which the diagnosis was confirmed by two informants as well as those with the combined phenotype, inbreeding was significantly increased in comparison to the control group. The chances that two similar recessive mutations are transmitted to a child are therefore much more likely when they come from the same ancestor in a pedigree with so-called "loops", than in an out-bred pedigree with non-related parents. So far, only genes with a dominant effect have been considered in the aetiology of ADHD.

A major limitation of our study is the small sample. In order to use information on genealogy we had to restrict our study to an isolated population for which we have genealogical data available. This has limited the number of patients eligible for the study. Nevertheless, the relation between the number of informants and the distance of relationship between patients was found to be statistically significant, even with this small sample size. The advantage of working with an isolated community is that we have detailed information on genealogy, which is not available in the general population. The loops identified in seven of fifteen patients who were linked to a common ancestor further suggest the involvement of a gene with a recessive effect. We have previously shown that, in this population with inbreeding, homozygosity mapping is a powerful approach in detecting genes with recessive effects (Bonifati et al., 2003). Of note is the fact that in our study the mean kinship coefficients for ADHD are relatively high. In the same genetically isolated population we studied Alzheimer's disease, which is known for its strong genetic clustering. The mean kinship coefficient for the Alzheimer's disease patients (0.0003) was found to be smaller than in any of the ADHD-subgroups (Roks et al., 2001). This suggests that the ADHD children in this population are closely related, making it suitable for future genetic studies.

Several studies addressed the question whether the number of informants confirming the ADHD diagnosis would improve the validity of the diagnosis (Achenbach, McConaughy, and Howell, Achenbach et al.; Ferdinand et al., 2003). Their findings indicate that scores of informants from different settings (e.g., home, school) may differ, either due to a different behavior of the child in these surroundings, or to differences in the interpretation of the child's behavior by the informants. Combining this unique contribution of each informant may yield a more consistent diagnosis, which also may better discriminate ADHD from other psychiatric disorders, such as conduct disorder (Crystal et al., 2001; Mitsis et al., 2000). Various studies assessed associations between type of informant (parent or teacher) and heritability of ADHD (Martin et al., 2002; Thapar et al., 2000; Todd et al., 2001). Thapar et al. found that a common genetic factor underlies both the parent-rated and teacher-rated ADHD symptoms (Thapar et al., 2000). However, they also found that additional specific genetic factors might contribute to the ADHD symptoms as rated by the teacher. Also Martin et al. concluded that ADHD diagnosed by using parent and teacher information showed a high degree of heritability (Martin et al., 2002). They suggested, however, that different genes might underlie the symptoms reported by parent versus teacher.

Another complicating factor in the search for genes involved in ADHD may be that phenotypic subtypes show differences in heritability (Neuman et al., 2001; Todd et al., 2001), as seen by our finding of the closer genetic relationships in children with inattentive and combined subtypes compared to those with the hyperactive/impulsive subtype. Also our study shows that the use of these subtypes, instead of viewing all subtypes as one single disorder, may provide the best opportunity to find genes involved in ADHD (Neuman et al., 2001; Todd et al., 2001).

In conclusion, our data showed that patients with the inattentive phenotype of ADHD were more closely related. By adding phenotypic information of a second informant a genetically more homogeneous group may result suitable for gene finding studies.

7.5 Appendix

Statistic Z to test the null hypothesis of no difference in mean of kinship coefficients between two groups.

Let d_{ij} be the natural logarithm of the kinship of pair(i, j) in group of size n_k . Let μ_k, σ_k^2 and γ_k be the mean, the variance and the covariance of two pairs with one subject in common, respectively. To test the null hypothesis $\mu_1 = \mu_2$

the following statistic Z is proposed

$$Z = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{var(\hat{\mu}_1) + var(\hat{\mu}_2)}},$$

with $\hat{\mu}_k$ the mean of $\frac{n_k(n_k-1)}{2}$ kinship coefficients in group k. The variance of $\hat{\mu}_k$ depends also on the covariance γ_k and is given by

$$var(\hat{\mu}_k) = \frac{2\sigma_k^2 + 4(n_k - 2)\gamma_k}{n_k(n_k - 1)}$$

for $k = 1, 2$. The variance $var(\hat{\mu}_k)$ is estimated by replacing σ_k^2 and γ_k by their estimators. The distribution of Z under H_0 is approximated by a standard normal distribution.

SUMMARY

Chapter 8

Summary

Complex traits are caused by multiple genetic and environmental factors, and are therefore difficult to study compared with simple Mendelian diseases. The modes of inheritance of Mendelian diseases are often known. Methods to dissect such diseases are well described in literature. For complex genetic traits, the inheritance pattern is not clear and difficult to understand, and genetic variants contributing to such traits probably have small effect sizes. Hence, searching for genes responsible for complex traits requires different strategies as well as new methods. A common strategy for mapping complex traits is as follows: (1) perform a genome-wide linkage analysis using dense genetic markers, and identify regions showing evidence of linkage, then (2) perform association analysis to refine these regions. Along these lines we propose several methods for detecting disease genes in this thesis.

In **chapter 1** we provide a general introduction to statistical tools for mapping genes responsible for complex genetic traits. Familial clustering, ascertainment issues, linkage analysis, and association analysis are considered. Furthermore, the aim and the outline of this thesis are described.

A first step in studying the genetic background of any trait is to verify whether it clusters within families. To this end, families ascertained through one or more members are often collected, especially when the trait is rare. Therefore, statistical methods that assume random families are not appropriate to this kind of data. In **chapter 2** we address this issue and develop a score statistic for testing familial clustering that takes into account the ascertainment scheme. The method does not assume any particular scheme, however, the set relevant for ascertainment (probands) should be known (Ewens and Shute, 1986). Familial correlation is modeled via a random effect parameter in the context of a generalized linear mixed model (Houwing-Duistermaat et al., 1995; le Cessie and van Houwelingen, 1995). The score statistics has the following features. It measures the proband-relative correlation as well as the relative-relative correlation. It allows for adjusting of covariates. No assumption about the distribution of the random effects is made. Furthermore, by conditioning on the trait value of all individuals related to ascertainment the method is robust to the ascertainment scheme. The score test can also be used to test the presence of correlation that is partly due to excess sharing of alleles identical by descent (IBD) in a candidate region. The method is applied to a candidate region in families of probands with type 2 diabetes mellitus.

In **chapter 3** we develop an allele sharing method to test for linkage along the genome or in a candidate region. The method considers all genetic markers jointly and hence multiple testing problems are avoided. At the disease locus, the increase of the mean of the number of alleles shared identically by descent (IBD) is modeled as a function of the relative risk ratio (Risch, 1990a). At each marker locus, the average of mean IBDs is calculated over the number of affected sibling pairs and the resulting variables behave approximately as a Gaussian Markov process along each chromosome (Feingold et al., 1993). Furthermore, the method sums the conditional likelihood of data given that a certain marker is a disease locus, over the marker positions. Either the likelihood ratio or the score statistics can be used to test for global linkage in the region of interest. Both statistics have known asymptotic distributions. For a genome-wide scan, likelihood ratio should be used, while for candidate region the score test is appropriate since it is comparable to the likelihood ratio when the assumed model is true, and it may perform better when the gene effect size is very small or the model is incorrect. However, for genome-wide scans the sample size required for detecting positive linkage is very high, especially for small effect sizes.

In **chapter 4**, **5**, and **6** we propose new statistical tools for association analysis. For testing disease association with a single nucleotide polymorphism (SNP), standard methods such as Pearson's χ^2 can be used. However, standard methods may not be suitable when testing disease association with a multi-marker. As an alternative, Terwilliger (1995) proposed a likelihood ratio test (TLR). The likelihood of the data is the weighted sum of the conditional likelihoods given that an allele is associated with the disease over all marker alleles with weights equal to the allele frequencies. In these chapters we consider testing for genetic association between a disease and either multi-allelic markers or haplotypes constructed from several flanking SNPs. However, haplotypes must be constructed with certainty.

In **chapter 4** we propose the score test corresponding to the Terwilliger's likelihood ratio. Under the null hypothesis of no genetic association the expectation and variance of the score statistic and Pearson's χ^2 are approximately equal. Under the alternative hypothesis of the presence of one associated allele, the score statistic has higher expectation than Pearson's χ^2 when the associated allele is common, which may imply higher power in favour of the score test. Furthermore we provide heuristic as well as empirical comparisons between the score test and other test statistics. We conclude that the score test has the highest power on average. Furthermore, we illustrated the methods based on a real data example.

In **chapter 5** we apply, the score test proposed in chapter 4, TLR and Pearson's χ^2 to candidate regions that initially showed evidence of linkage with Kofendrerd Personality Disorder (KPD), a fictional disorder. Data are from the GAW14 simulated micro-satellite data problem (Bailey-Wilson et al., 2005; Greenberg et al., 2005). All test statistics suggest the presence of association between KPD disease and the multi-allelic marker, D03S0127, in this region. Testing a dense map of SNPs reveal strong evidence of genetic association with SNPs B03T3056 and B03T3057, which are in linkage disequilibrium with a disease locus located at the end of chromosome 3 (Greenberg et al., 2005). The analyses are done without prior knowledge of the answers.

In **chapter 6**, we generalise the method proposed by Terwilliger (1995). Instead of allele frequencies we propose weights, which do not depend on actual data. In this light, we derive a new score statistic, which is easy to compute. The new score statistic has a power comparable to TLR statistic and sometimes even better, especially when the excess of associated allele is small and the allele frequency is relatively common. However, the score statistic may not have a reasonable power if the frequency of the associated allele is equal to the inverse of the number of marker alleles, or if there is one positively and one negatively associated allele with the same amount of allele excess. The score statistic is successfully applied to three candidate genes studied in the Leiden Thrombophilia Study (Uitte de Willige et al., 2005).

Chapter 7 describes a study of the relevance of adding information of a second informant when children are diagnosed with Attention Deficit Hyperactivity Disorder (ADHD). Especially, two points are addressed: (1) The usefulness of the use of two informants in genetic studies, and (2) whether phenotypic subtypes of ADHD cluster similarly in families. The study is conducted in a genetically isolated population in the Southwest of the Netherlands. Genealogical information is available for 22 generations. The study finds that patients that are diagnosed based on two informants are more closely related than those diagnosed based on one informant, which may be relevant for further study of the genetic background of ADHD. Moreover, the study confirms the familial clustering of ADHD, which was previously suggested by other studies. Furthermore, patients with the inattentive subtype of ADHD are found to be more closely related. The study suggests that recessive genes may be involved in causing ADHD. In order to study the closeness of familial relationship between two patient groups we propose a test statistic, which compares the mean kinship coefficients of the two groups. For large sample size, the test statistic follows the normal distribution. However, in this study all patient groups are small and hence the asymptotic distribution of the test statistic cannot be used. Nevertheless the significance level could be obtained by means of Monte Carlo simulation.

CHAPTER 9

Samevatting

Veel aandoeningen zijn het resultaat van meerdere genetische en/of omgevingsfactoren. In de genetische epidemiologie noemt men deze aandoeningen gecompliceerde erfelijke aandoeningen. Genen die leiden tot het verkrijgen van dergelijke gecompliceerde erfelijke aandoening zijn vaak moeilijk op te sporen. Zij hebben vaak lage *penetranties*. Dat wil zeggen dat sommige patiënten de genvariant wel dragen maar de ziekte niet ontwikkeld hebben. Patiënten met dezelfde ziekte kunnen ook verschillende combinaties van ziekteveroorzakende genen dragen. Men noemt dit fenomeen een heterogeniteit effect. Samenvattend is het onderliggende genetische mechanisme van een gecompliceerde erfelijke aandoening meestal onbekend, en daarom is het moeilijk te ontdekken. In tegenstelling tot gecompliceerde erfelijke ziektes, zijn de erfelijke eigenschappen van veel Mendeliaanse ziektes gemakkelijk in de kaart te brengen. Mendeliaanse ziektes worden veroorzaakt door één enkel gen en hebben daarom een duidelijk overervingpatroon. Methoden om de onderliggende genetische achtergrond van zulke ziektes te bestuderen zijn wel in de literatuur beschreven. Echter, het opsporen van gecompliceerde ziekteveroorzakende genen vergt een andere benadering en vraagt om nieuwe statistische methoden. Een veel gebruikte strategie om gecompliceerde erfelijke aandoeningen in kaart te brengen is te beginnen met een scan van het gehele genoom waarbij genetische familie verbanden geanalyseerd worden (linkage analyses.) Dit kan helpen bij het identificeren van chromosoom-regio's die mogelijk ziekteveroorzakende genen bevatten. Vervolgens voert men een associatie studie uit om dergelijke kandidaat regio's te verfijnen. Binnen deze methodologie stelt dit proefschrift zich tot doel om nieuwe statistische methoden te ontwikkelen en te evalueren, die geschikt kunnen zijn voor het analyseren van genetische gegevens van patiënten met een gecompliceerde erfelijke aandoening in zowel families als populaties.

Voordat men een genetische studie kan opzetten, moet men eerst nagaan of er wel genetische factoren betrokken zijn bij de etiologie van de bestudeerde aandoening. Dit kan gedaan worden met behulp van een analyse van de familiale aggregatie van de aandoening. Hiervoor kan men aselecte steekproeven van families verzamelen. Echter, meestal worden families van patiënten met de aandoening (probands) opzettelijk geselecteerd. Dit laatste wordt vooral gedaan om de statistische power te verhogen.

In **hoofdstuk 2** hebben wij een nieuwe statistische methode ontwikkeld om de aanwezigheid van clustering van een aandoening binnen families van probands te toetsen. Deze methode houdt rekening met de selectieprocedure. De familiale correlatie wordt gemodelleerd aan de hand van een stochastisch effect met verwachting nul, onbekende variantie en een correlatie die van tevoren aangegeven moet worden. Toetsen op de afwezigheid van een correlatie structuur is equivalent aan toetsen of de variantie van het stochastische effect gelijk aan nul is. De methode doet geen specifieke aanname over de verdeling van het stochastisch effect. Om familiale clustering te toetsen kan men de correlatie bepalen als een functie van de graad van de verwantschap tussen familieleden (kinship coefficient). De methode kan ook gebruikt worden om de genetische correlatie te toetsen die gedeeltelijk te wijten aan een bepaalde locus. Verder hebben we de methode geïllustreerd aan de hand van genetische gegevens van families van probands met ouderdomssuikerziekte.

In **hoofdstuk 3** hebben wij een niet-parametrische methode ontwikkeld om de genetische linkage globaal te toetsen over het hele genoom of op een gedeelte van een chromosoom. Deze methode is gebaseerd op een vergelijking van het aantal IBD allelen van een marker dat twee familie leden met de aandoening delen met het aantal dat ze zouden delen per toeval. IBD is de Engelse afkorting voor *identical by descent*. Het aantal IBD allelen dient als een maat van genetische gelijkenis tussen twee personen. De methode toetst alle markers tegelijkertijd zodat het meervoudige toetsingsprobleem vermeden wordt. Dit wordt bereikt door het optellen van de voorwaardelijke likelihood functies gegeven dat een marker de ziekteveroorzakende gen bevat. Wij hebben twee toets statistieken afgeleid, namelijk de likelihood ratio toets en de score toets. Verder hebben wij de werking van deze twee toetsen bestudeerd aan de hand van een simulatiestudie. Twee modellen werden beschouwd: (1) *één-locus* model en (2) *twee-locus* model. Het blijkt dat de likelihood ratio toets het beste presteert wat betreft van onderscheidend vermogen (power), terwijl de scoretoets alleen geschikt is voor het vinden van kandidaat-regio's of voor situaties waarin het effect van het ziekteveroorzakende gen heel klein is.

Hoofdstukken 4, 5 and 6 beschrijven statistische methoden voor genetische associatie analyses in de populatie. Hierbij vergelijkt men de allelfrequenties tussen een groep patiënten en een willekeurige groep uit de populatie. De nadruk in deze hoofdstukken ligt op markers met meerdere allelen of haplotypen uit meerdere *enkelvoudige nucleotide polymorphismes* (SNPs). De methodes gebruiken een *semi-Bayesiaanse* aanpak waarbij de totale likelihood de som is van de gewogen voorwaardelijke likelihood functies gegeven dat een allel (haplotype) geassocieerd is met de desbetreffende ziekte. Voor deze aanpak heeft Terwilliger (1995) de gewichten gelijk gesteld aan de allelfrequenties en de corresponderende likelihood ratio (TLR) gebruikt als toets statistiek.

In **hoofdstuk 4** hebben wij dezelfde likelihood as Terwilliger (1995) gebruikt en daaruit een scoretoets afgeleid. De scoretoets is eenvoudig en, in tegenstelling tot de likelihood ratio toets, doet hij geen aanname over het aantal geassocieerde allelen. De nulverdeling van de toets kan gevonden worden met behulp van Monte Carlo simulaties. Verder hebben wij deze toets analytisch vergeleken met de bekende Chi-kwadraat toets voor cruistabellen. Het blijkt dat de scoretoets het beter doet dan Chi-kwadraat toets wanneer het geassocieerde allel frequent voorkomt. Aan de hand van een simulatie hebben wij ook de power van de scoretoets vergeleken met andere bestaande toetsen, inclusief de Chi-kwadraat toets. Wij hebben zowel markers met één geassocieerd allel als met twee geassocieerde allelen beschouwd. Het blijkt dat de scoretoets gemiddeld de beste power heeft. Ter illustratie hebben wij de scoretoets toegepast op echte data.

Hoofdstuk 5 beschrijft de resultaten van toepassingen van de scoretoets uit hoofdstuk 4, TLR and Chi-kwadraat op kandidaat- regio's. De kandidaat-regio's zijn geïdentificeerd met behulp van linkage analyses op een aantal chromosomen. De data waren afkomstig uit het *Genetic Association Workshop* 14 gesimuleerde microsatellieten probleem (Bailey-Wilson et al., 2005; Greenberg et al., 2005). De bestudeerde ziekte was "Kofendrerd Personality Disorder" (KPD), een fictieve aandoening. Alle toegepaste toetsen duiden aan dat er mogelijk een associatie bestaat tussen KPD en de microsatelliet marker D03S0127. Verdere associatie-analyses tussen KBD en 20 SNPs rond marker D03S0127 laten zien dat er een sterke associatie is tussen KBD en de SNP B03T3057 wat eigelijk ligt naast de echte ziekte locus in het end van chromosoom 3. (Greenberg et al., 2005). De analyses werden uitgevoerd zonder enige voorkennis van de antwoorden.

Hoofdstuk 6 presenteert een generalisatie van de methode beschreven in hoofdstuk 4. Wij hebben de gewichten als allel- (haplotype-) frequenties in de totale likelihood vervangen door constante gewichten. In deze nieuwe aanpak hoeven de gewichten dus niet geschat te worden uit de data. Voor deze likelihood hebben wij een nieuwe scoretoets afgeleid. De scoretoets is eenvoudig en heeft een normale verdeling. Het specificeren van de gewichten kan gedaan worden op basis van de voorkennis dat onderzoekers bijvoorbeeld hebben opgedaan uit eerder studies. Echter, men kan ook niet-informatieve (gelijke) gewichten gebruiken wanneer er geen op voorhand informatie beschikbaar is. Dat wil zeggen dat alle allelen (haplotypes) a-priori een even grote kans hebben om geassocieerd te zijn met de ziekte. Voor gelijke gewichten hebben wij verder een uitgebreide simulatie studie uitgevoerd om de werking van deze nieuwe toets in vergelijking met Chi-kwadraat en TLR toets te bestuderen zowel onder het nulmodel als onder het alternatieve model. Wij hebben alleen markers (haplotypes) met één geassocieerde variant beschouwd. In het algemeen presteert de scoretoets het beste. De werking van scoretoets is met succes geïllustreerd in de context van haplotype analyses aan de hand van data van drie kandidaatgenen uit *Leiden Thrombophilia Studie* . De analyses zijn uitgevoerd onder de veronderstelling van perfecte informatie omtrent *haploype phase* .

Bibliography

Abecasis, G. R., S. S. Cherny, W. O. Cookson, and L. R. Cardon (2002). Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics 30*, 97–101.

Achenbach, T., S. McConaughy, and C. Howell. Child/adolescent behavioral and emotional problems: implications of cross-informant correlations for situational specificity. *Psychol Bull 101*, 213–132.

Amos, C. I. (1994). Robust variance-components approach for assessing genetic-linkage in pedigrees. *American Journal of Human Genetics 54*, 535–543.

Amos, C. I. and M. de Andrade (2001). Genetic linkage methods for quantitative traits. *Statistical Methods in Medical Research 10*, 3–25.

Amos, C. I. and R. C. Elston (1989). Robust methods for the detection of genetic-linkage for quantitative data from pedigrees. *Genetic Epidemiology 6*, 349–360.

Anderson, T. (1984). *An introduction to multivariate statistical analysis*. Second edition, Jhon Wiley & Sons, Inc.

Aulchenko, Y. S., N. Vaessen, P. Heutink, J. Pullen, P. J. L. M. Snijders, A. Hofman, L. A. Sandkuijl, J. J. Honwing-Duistermaat, M. Edwards, S. Bennett, B. A. Oostra, and C. M. van Duijn (2003). A genome-wide search for genes involved in type 2 diabetes in a recently genetically isolated population from the netherlands. *Diabetes 52*, 3001–3004.

Bacanu, S. (2005). Multipoint linkage analysis for a very dense set of markers. *Genetic Epidemiology 29*, 195–203.

Bafna, V., D. Gusfield, G. Lancia, and S. Yooseph (2003). Haplotyping as perfect phylogeny: A direct approach. *Journal of Computational Biology 10*, 323–340.

Bailey-Wilson, J. E. Almasy, L., M. de Andrade, J. Bailey, H. Bickeböller, H. J. Cordell, E. W. Daw, L. Goldin, E. L. Goode, C. Gray-McGuire, W. Hening, G. Jarvik, B. S. Maher, N. Mendell, A. D. Paterson, J. Rice, G. Satten, B. Suarez, V. Vieland, M. Wilcox, H. Zhang, A. Ziegler, and J. W. MacCluer (2005). Genetic analysis workshop 14: Microsatellite and snp marker loci for genome-wide scans. *BMC Genetics 6*, S1.

Barkley, R. (1996). *Attention deficit/hyperactivity disorder. In:Child Psychopathology*. New York: Guildford Press.

Barkley, R. A. (2003). Issues in the diagnosis of attention-deficit/hyperactivity disorder in children. *Brain and Development 25*(2), 77–83.

Beaty, T. H. and K. Y. Liang (1987). Robust inference for variance-components models in families ascertained through probands .1. conditioning on probands phenotype. *Genetic Epidemiology 4*, 203–210.

Becker, T., S. Cichon, and M. Jonson, E. Knapp (2005). Multiple testing in the context of haplotype analysis revisited: application to case-control data. *Annals of Human Genetics 69*, 747–56.

Biernacka, J. (2004). *Statistical methods for studying two linked disease genes*. Ph. D. thesis, University of Toronto.

Biernacka, J. M., L. Sun, and S. B. Bull (2005). Simultaneous localization of two linked disease susceptibility genes. *Genetic Epidemiology 28*, 33–47.

Bonifati, V., P. Rizzu, F. Squitieri, E. Krieger, N. Vanacore, J. C. van Swieten, A. Brice, C. M. van Duijn, B. Oostra, G. Meco, and P. Heutink (2003). Dj-1 (park7), a novel gene for autosomal recessive, early onset parkinsonism. *Neurological Sciences 24*(3), 159–160.

Bonney, G. E. (1986). Regressive logistic-models for familial disease and other binary traits. *Biometrics 42*, 611–625.

Cannings, C. and E. A. Thompson (1977). Ascertainment in sequential sampling of pedigrees. *Clinical Genetics 12*, 208–212.

Cantwell, D. P. (1996). Attention deficit disorder: A review of the past 10 years. *Journal of the American Academy of Child and Adolescent Psychiatry 35*(8), 978–987.

Cardon, L. R. and D. W. Fulker (1994). The power of interval mapping of quantitative trait loci, using selected sib pairs. *American Journal of Human Genetics 55*, 825–833.

Carey, G. and J. Williamson (1991). Linkage analysis of quantitative traits increased power by using selected samples. *American Journal of Human Genetics 49*, 786–796.

Carlson, C. S., S. F. Aldred, P. K. Lee, R. P. Tracy, S. M. Schwartz, M. Rieder, K. A. Liu, O. D. Williams, C. Iribarren, E. C. Lewis, M. Fornage, E. Boerwinkle, M. Gross, C. Jaquish, D. A. Nickerson, R. M. Myers, D. S. Siscovick, and A. P. Reiner (2005). Polymorphisms within the c-reactive protein (crp) promoter region are associated with plasma crp levels. *American Journal of Human Genetics 77*, 64–77.

Clark, A. G. (2004). The role of haplotypes in candidate gene studies. *Genetic Epidemiology 27*, 321–333.

Clayton, D. (2000). Linkage disequilibrium mapping of disease susceptibility genes in human populations. *International Statistical Review 68*, 23–43.

Clayton, D., J. Chapman, and J. Cooper (2004). Use of unphased multilocus genotype data in indirect association studies. *Genetic Epidemiology 27*, 415–428.

Collins, F. S., M. S. Guyer, and A. Chakravarti (1997). Variations on a theme: Cataloging human dna sequence variation. *Science 278*, 1580–1581.

Commenges, D., H. Jacqmin, L. Letenneur, and C. M. van Duijn (1995). Score test for familial aggregation in probands studies: application to alzheimers-disease. *Biometrics 51*, 542–551.

Cordell, H. J. (2001). Sample size requirements to control for stochastic variation in magnitude and location of allele-sharing linkage statistics in affected sibling pairs. *Annals of Human Genetics 65*, 491–502.

Cordell, H. J. and D. G. Clayton (2005). Genetic associaton studies. *Lancet 366*, 1121–31.

Cox, D. and D. V. Hinkley (1974). *Theoretical statistics*. Florida, USA: Chapman & Hall.

Crystal, D. S., R. Ostrander, R. S. Chen, and G. J. August (2001). Multimethod assessment of psychopathology among dsm-iv subtypes of children with attention-deficit/hyperactivity disorder: Self-, parent, and teacher reports. *Journal of Abnormal Child Psychology 29*(3), 189–205.

Czika, W. and B. Weir (2004). Properties of the multiallelic trend test. *Biometrics 52*, 69–74.

Daly, M. J., J. D. Rioux, S. E. Schaffner, T. J. Hudson, and E. S. Lander (2001). High-resolution haplotype structure in the human genome. *Nature Genetics 29*, 229–232.

Day, N. E. and M. J. Simons (1976). Disease susceptibility genestheir identification by multiple case family studies. *Tissue Antigens 8*, 109–119.

de Andrade, M. and C. I. Amos (2000). Ascertainment issues in variance components models. *Genetic Epidemiology 19*, 333–344.

de Nijs, P. F. A., R. E. Ferdinand, E. I. de Bruin, M. C. J. Dekker, C. M. van Duijn, and F. C. Verhulst (2004). Attention-deficit/hyperactivity disorder (adhd): parents' judgment about school, teachers' judgment about home. *European Child and Adolescent Psychiatry 13*(5), 315–320.

Devlin, B. and N. A. Risch (1995). A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics. 29*, 311–22.

Diggle, P. J., L. K. Y. and S. L. Zegger (1994). *Analysis of longitudinal data*. Oxford: Clarendon Press, Oxford Science Publication.

Dudoit, S. and T. P. Speed (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.

El Galta, R., T. Stijnen, and J. Houwing-Duistermaat (2004). Score statistic for analysis of association between disease and a multi-allelic marker [abstract]. *Genetic Epidemiology 27*, 268.

Elston, R., J. Olson, and L. Palmer (Eds.) (2002). *Biostatistical Genetics and Genetic epidemiology*. John Wiley & Sons, Ltd.

Elston, R. C. (2000). Introduction and overview. *Statistical Methods in Medical Research 9*, 527–541.

Elston, R. C. and E. Sobel (1979). Sampling considerations in the gathering and analysis of pedigree data. *American Journal of Human Genetics 31*, 62–69.

Epstein, M. P. and G. A. Satten (2003). Inference on haplotype effects in case-control studies using unphased genotype data. *American Journal of Human Genetics 73*, 1316–1329.

Ewens, W., R. Griffiths, S. Ethier, S. Wilcox, and J. Graves (1992). Statistical analysis of in situ hybridization data: derivation and use of the zmax test. *Genomics. 12*, 675–82.

Ewens, W. J. and N. C. E. Shute (1986). A resolution of the ascertainment sampling problem .i. theory. *Theoretical Population Biology 30*, 388–412.

Faraone, S. V., J. Biederman, E. Mick, A. E. Doyle, T. Wilens, T. Spencer, E. Frazier, and K. Mullen (2001). A family study of psychiatric comorbidity in girls and boys with attention-deficit/hyperactivity disorder. *Biological Psychiatry 50*(8), 586–592.

Faraone, S. V., J. Biederman, and M. C. Monuteaux (2000). Toward guidelines for pedigree selection in genetic studies of attention deficit hyperactivity disorder. *Genetic Epidemiology 18*(1), 1–16.

Feingold, E. (2001). Methods for linkage analysis of quantitative trait loci in humans. *Theoretical Population Biology 60*, 167–180.

Feingold, E., P. O. Brown, and D. Siegmund (1993). Gaussian models for genetic-linkage analysis using complete high-resolution maps of identity by descent. *American Journal of Human Genetics 53*, 234–251.

Ferdinand, R. and J. van der Ende (2000). Nimh disc-iv: Diagnostic interview schedule for children. *Authorized Dutch translation.: Erasmus MC-Sophia: Rotterdam, the Netherlands*.

Ferdinand, R. F., K. N. Hoogerheide, J. van der Ende, J. H. Visser, H. M. Koot, M. C. Kasius, and F. C. Verhulst (2003). The role of the clinician: three-year predictive value of parents', teachers', and clinicians' judgment of childhood psychopathology. *Journal Of Child Psychology And Psychiatry And Allied Disciplines 44*(6), 867–876.

FitzGerald, P. E. B. and M. W. Knuiman (2000). Use of conditional and marginal odds-ratios for analysing familial aggregation of binary data. *Genetic Epidemiology 18*, 193–202.

Forabosco, P., M. Falchi, and M. Devoto (2005). Statistical tools for linkage analysis and genetic association studies. *Expert Review of Molecular Diagnostics 5*, 781–796.

Fullerton, S. M., A. G. Clark, K. M. Weiss, D. A. Nickerson, S. L. Taylor, J. H. Stengard, V. Salomaa, E. Vartiainen, M. Perola, E. Boerwinkle, and C. F. Sing (2000). Apolipoprotein e variation at the sequence haplotype level: Implications for the origin and maintenance of a major human polymorphism. *American Journal of Human Genetics 67*, 881–900.

Gabriel, S. B., S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, S. N. Liu-Cordero, C. Rotimi, A. Adeyemo, R. Cooper, R. Ward, E. S. Lander, M. J. Daly, and D. Altshuler (2002). The structure of haplotype blocks in the human genome. *Science 296*, 2225–2229.

Gaub, M. and C. L. Carlson (1997). Gender differences in adhd: A meta-analysis and critical review. *Journal of the American Academy of Child and Adolescent Psychiatry 36*(8), 1036–1045.

Green, J. R. and J. C. Woodrow (1977). Sibling method for detecting hla-linked genes in disease. *Tissue Antigens 9*, 31–35.

Greenberg, D. A., J. Zhang, D. Shmulewitz, L. J. Strug, R. Zimmerman, V. Singh, and S. Marathe (2005). Construction of the model for the genetic analysis workshop 14 simulated data: genotype-phenotype relationships, gene interaction, linkage, association, disequilibrium, and ascertainment effects for a complex phenotype. *BMC Genetics 6 (Suppl 1)*, S1, in press.

Gudbjartsson, D. F., K. Jonasson, M. L. Frigge, and A. Kong (2000). Allegro, a new computer program for multipoint linkage analysis. *Nature Genetics 25*, 12–13.

Haldane, J. (1939). The mean and variance of ?2 when used as a test of homogeneity, when expectations are small. *Biometrika 31*, 346–355.

Haseman, J. K. and R. C. Elston (1972). Investigation of linkage between a quantitative trait and a marker locus. *Behavior Genetics 2*, 3–19.

Hastbacka, J., A. Delachapelle, I. Kaitila, P. Sistonen, A. Weaver, and E. Lander (1992). Linkage disequilibrium mapping in isolated founder populations - diastrophic dysplasia in finland. *Nature Genetics 2*, 204–211.

Hill, J. C. and E. P. Schoener (1996). Age-dependent decline of attention deficit hyperactivity disorder. *American Journal of Psychiatry 153*(9), 1143–1146.

Hofman, A., D. E. Grobbee, P. T. V. M. De Jong, and F. A. van den Ouweland (1991). Determinants of disease and disability in the elderlythe rotterdam elderly study. *European Journal of Epidemiology 7*, 403–422.

Hopper, J. L. and J. D. Mathews (1982). Extensions to multivariate normal-models for pedigree analysis. *Annals of Human Genetics 46*, 373–383.

Houwing-Duistermaat, J. J., C. Bijkerk, L. Hsu, T. Stijnen, E. P. Slagboom, and C. M. van Duijn (2003). A unified approach to modelling linkage to quantitative and qualitative traits. *Annals of Human Genetics 67*, 457–463.

Houwing-Duistermaat, J. J., B. H. F. Derkx, F. R. Rosendaal, and H. C. van Houwelingen (1995). Testing familial aggregation. *Biometrics 51*, 1292–1301.

Houwing-Duistermaat, J. J. and R. C. Elston (2001). Linkage disequilibrium mapping of complex genetic diseases using multiallelic markers. *Genetic Epidemiology 21S*, 527–81.

Johnson, R. and D. Wichern (1998). *Applied multivariate statistical analysis*. New jersey, USA: Prentice-Hall.

Kizawa, H., I. Kou, A. Iida, A. Sudo, Y. Miyamoto, A. Fukuda, A. Kotani, A. kawakami, S. Yamamoto, K. Uchida, A. Nakamura, K. Notoya, Y. Nakamura, and S. Ikegawa (2005). An aspartic acid repeat polymorphism in asporin inhibits chondrogenesis and increases susceptibility to osteoarthritis. *Nature Genetics 37*, 138–44.

Koster, T., F. R. Rosendaal, H. Deronde, E. Briet, J. P. van den broucke, and R. M. Bertina (1993). Venous thrombosis due to poor anticoagulant response to activated protein-cleiden thrombophilia study. *Lancet 342*, 1503–1506.

Kruglyak, L., M. J. Daly, M. P. ReeveDaly, and E. S. Lander (1996). Parametric and nonparametric linkage analysis: A unified multipoint approach. *American Journal of Human Genetics 58*, 1347–1363.

Lander, E. and L. Kruglyak (1995). Genetic dissection of complex traitsguidelines for interpreting and reporting linkage results. *Nature Genetics 11*, 241–247.

Lander, E. S. and N. J. Schork (1994). Genetic dissection of complex traits. *Science 265*, 2037–2048.

Lange, K. (2002). *Mathematical and statistical methods for genetic analysis*. Second edition, Springer-Verlag.

Lazzeroni, L. C. (2001). A chronology of fine-scale gene mapping by linkage disequilibrium. *Statistical Methods in Medical Research 10*, 57–76.

le Cessie, S. and H. C. van Houwelingen (1995). Testing the fit of a regression-model via score tests in random effects models. *Biometrics 51*, 600–614.

Lebrec, J., H. Putter, and J. C. van Houwelingen (2005). Potential bias in generalized estimating equations linkage methods under incomplete information. *To appear in Genetic Epidemiology*.

Levy, F.and Hay, D., M. McStephen, C. Wood, and I. Waldman (1997). Attention-deficit hyperactivity disorder: a category or a continuum? genetic analysis of a large-scale twin study. *J Am Acad Child Adolesc Psychiatry 36*, 737–44.

Liang, K. Y. and T. H. Beaty (1991). Measuring familial aggregation by using odds-ratio regression-models. *Genetic Epidemiology 8*, 361–370.

Liang, K. Y. and T. H. Beaty (2000). Statistical designs for familial aggregation. *Statistical Methods in Medical Research 9*, 543–562.

Liang, K. Y., Y. F. Chiu, and T. H. Beaty (2001). A robust identity-by-descent procedure using affected sib pairs: Multipoint mapping for complex diseases. *Human*

Heredity 51, 64–78.

Liang, K. Y. and S. L. Zeger (1986). Longitudinal data-analysis using generalized linear-models. *Biometrika 73*, 13–22.

Lopera, F., L. Palacio, I. Jimenez, P. Villegas, I. Puerta, D. Pineda, M. Jimenez, and M. Arcos-Burgos (1999). [discrimination between genetic factors in attention deficit] discriminacion de factores geneticos en el deficit de atencion. *Rev Neurol 28*, 660–664.

Martin, N., J. Scourfield, and P. McGuffin (2002). Observer effects and heritability of childhood attention-deficit hyperactivity disorder symptoms. *British Journal of Psychiatry 180*, 260–265.

McCullagh, P. and J. Nelder (1989). *Generalized linear models*. Second edition, London: Chapman and Hall.

Meulenbelt, I., C. Bijkerk, S. C. De Wildt, H. S. Miedema, F. C. Breedveld, H. A. Pols, A. Hofman, C. M. Van Duijn, and P. E. Slagboom (1999). Haplotype analysis of three polymorphisms of the col2a1 gene and associations with generalised radiological osteoarthritis. *Annals of Human Genetics 63*, 393–400.

Mitsis, E. M., K. E. McKay, K. P. Schulz, J. H. Newcorn, and J. M. Halperin (2000). Parent-teacher concordance for dsm-iv attention-deficit/hyperactivity disorder in a clinic-referred sample. *Journal of the American Academy of Child and Adolescent Psychiatry 39*(3), 308–313.

Morton, N. E. (1959). Genetic tests under incomplete ascertainment. *American Journal of Human Genetics 11*, 1–16.

Neuman, R. J., A. Heath, W. Reich, K. K. Bucholz, P. A. F. Madden, L. W. Sun, R. D. Todd, and J. J. Hudziak (2001). Latent class analysis of adhd and comorbid symptoms in a population sample of adolescent female twins. *Journal of Child Psychology and Psychiatry and Allied Disciplines 42*(7), 933–942.

Niu, T. H. (2004). Algorithms for inferring haplotypes. *Genetic Epidemiology 27*, 334–347.

Ott, J. (1999). *Analysis of human genetic linkage*. Third edition, Johns Hopkins University Press, Baltimore.

Pfeiffer, R. M., M. H. Gail, and D. Pee (2001). Inference for covariates that accounts for ascertainment and random genetic effects in family studies. *Biometrika 88*, 933–948.

Putter, H., L. A. Sandkuijl, and J. C. van Houwelingen (2002). Score test for detecting linkage to quantitative traits. *Genetic Epidemiology 22*, 345–355.

Rao, D. C. and R. Wette (1987). Nonrandom sampling in genetic epidemiologymaximum-likelihood methods for multifactorial analysis of quantitative data ascertained through truncation. *Genetic Epidemiology 4*, 357–376.

Rao, D. C., R. Wette, and W. J. Ewens (1988). Multifactorial analysis of family data ascertained through truncation - a comparative-evaluation of 2 methods of statistical-inference. *American Journal of Human Genetics 42*, 506–515.

Reich, D. E. and E. S. Lander (2001). On the allelic spectrum of human disease. *Trends Genet. 17*, 502–10.

Risch, N. (1990a). Linkage strategies for genetically complex traits .I. multilocus models. *American Journal of Human Genetics 46*, 222–228.

Risch, N. (1990b). Linkage strategies for genetically complex traits .II. the power of affected relative pairs. *American Journal of Human Genetics 46*, 229–241.

Risch, N. (1990c). linkage strategies for genetically complex traits .III. the effect of marker polymorphism on analysis of affected relative pairs. *American Journal of Human Genetics 46*, 242–253.

Risch, N. (2000). Searching for genes in complex diseases: lessons from systemic lupus erythematosus. *Journal of Clinical Investigation 105*, 1503–1506.

Risch, N. and H. P. Zhang (1995). Extreme discordant sib pairs for mapping quantitative trait loci in humans. *Science 268*, 1584–1589.

Roks, G. P. and, v. S. J., M. Cruts, E. Boeren, L. Sandkuijl, P. Snijders, C. Van Broeckhoven, B. Oostra, and C. van Duijn (2001). *Alzheimer's disease in a Dutch recent genetically isolated population. The GRIP Study, in Alzheimer's disease. A genetic epidemiologic approach.* Ph. D. thesis, p. 83-90. Erasmus MC: Rotterdam.

Sasieni, P. D. (1997). From genotypes to genes: doubling the sample size. *Biometrics 53*, 1253–61.

Satten, G. A. and M. P. Epstein (2004). Comparison of prospective and retrospective methods for haplotype inference in case-control studies. *Genetic Epidemiology 27*, 192–201.

Schaid, D. J. (2004). Evaluating associations of haplotypes with traits. *Genetic Epidemiology 27*, 348–364.

Schaid, D. J., C. M. Rowland, D. E. Tines, R. M. Jacobson, and G. A. Poland (2002). Score tests for association between traits and haplotypes when linkage phase is ambiguous. *American Journal of Human Genetics 70*, 425–434.

Schaid, D. J., J. P. Sinnwell, and S. N. Thibodeau (2005). Robust multipoint identical-by-descent mapping for affected relative pairs. *American Journal of Human Genetics 76*, 128–138.

Sebastiani, P., R. Lazarus, S. T. Weiss, L. M. Kunkel, I. S. Kohane, and M. F. Ramoni (2003). Minimal haplotype tagging. *Proceedings of the National Academy of Sciences of the United States of America 100*, 9900–9905.

Self, S. and K. Liang (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under non-standard conditions. *JASA 82*, 605–610.

Shaffer, D., P. Fisher, C. P. Lucas, M. K. Dulcan, and M. E. Schwab-Stone (2000). Nimh diagnostic interview schedule for children version iv (nimh disc-iv): Description, differences from previous versions, and reliability of some common diagnoses. *Journal of the American Academy of Child and Adolescent Psychiatry 39*(1), 28–38.

Sham, P. C. (1998). *Statistics in human genetics*. New York: Oxford University Press Inc.

Sham, P. C. and D. Curtis (1995). Monte Carlo tests for associations between disease and alleles at highly polymorphic loci. *Annals of Human Genetics 59*, 97–105.

Sham, P. C., D. Curtis, and C. J. MacLean (1996). Likelihood ratio tests for linkage and linkage disequilibrium: asymptotic distribution and power. *American Journal of Human Genetics 58*, 1093–5.

Shih, M. C. and A. S. Whittemore (2001). Allele-sharing among affected relatives: non-parametric methods for identifying genes. *Statistical Methods in Medical Research 10*, 27–55.

Siegmund, D. (2001). Is peak height sufficient? *Genetic Epidemiology 20*, 403–408.

Slager, S. L. and D. J. Schaid (2001). Case-control studies of genetic markers: Power and sample size approximations for armitage's test for trend. *Human Heredity 52*, 149–153.

Stram, D. O. (2004). Tag snp selection for association studies. *Genetic Epidemiology 27*, 365–374.

Stram, D. O., H. Lee, and D. C. Thomas (1993). Use of generalized estimating equations in segregation analysis of continuous outcomes. *Genetic Epidemiology 10*, 575–579.

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

Terwilliger, J. D. (1995). A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. *American Journal of Human Genetics 56*, 777–87.

Thapar, A., R. Harrington, K. Ross, and P. McGuffin (2000). Does the definition of adhd affect heritability? *J Am Acad Child Adolesc Psychiatry 39*(6), 1528–1536.

Thapar, A., A. Hervas, and P. McGuffin (1995). Childhood hyperactivity scores are highly heritable and show sibling competition effects: Twin study evidence. *Behaviour Genetics 5*, 537–544.

Thapar, A., J. Holmes, K. Poulton, and R. Harrington (1999). Genetic basis of attention deficit and hyperactivity. *British Journal of Psychiatry 174*, 105–111.

Thompson, E. A. (1975). Estimation of pairwise relationships. *Annals of Human Genetics 39*, 173–188.

Todd, R. D., E. R. Rasmussen, R. J. Neuman, W. Reich, J. J. Hudziak, K. K. Bucholz,

P. A. F. Madden, and A. Heath (2001). Familiality and heritability of subtypes of attention deficit hyperactivity disorder in a population sample of adolescent female twins. *American Journal of Psychiatry 158*(11), 1891–1898.

Tregouet, D. A. and L. Tiret (2000). Applications of the estimating equations theory to genetic epidemiology: a review. *Annals of Human Genetics 64*, 1–14.

Tritchler, D., Y. Liu, and S. Fallah (2003). A test of linkage for complex discrete and continuous traits in nuclear families. *Biometrics 59*, 382–92.

Uh, H. W., J. J. Houwing-Duistermaat, H. Putter, and van Houwelingen H. C. (2005). How to quantify information loss due to phase ambiguity in haplotype case-control studies. *BMC Genet. 6 Suppl 1*, S108.

Uitte de Willige, S., M. C. de Visser, J. J. Houwing-Duistermaat, F. R. Rosendaal, H. L. Vos, and R. M. Bertina (2005). Genetic variation in the fibrinogen gamma gene increases the risk of deep venous thrombosis by reducing plasma fibrinogen gamma' levels. *Blood 106*, 4176–83.

Vaessen, N., P. Heutink, J. J. Houwing-Duistermaat, P. J. L. M. Snijders, T. Rademaker, L. Testers, M. R. Batstra, L. A. Sandkuijl, C. M. van Duijn, and B. A. Oostra (2002). A genome-wide search for linkage-disequilibrium with type 1 diabetes in a recent genetically isolated population from the netherlands. *Diabetes 51*, 856–859.

van der Ende, J. (1999). *Multiple informants: Multiple views*. In: Child psychiatric epidemiology. Accomplishments and future directions., Koot HM CA, Ferdinand RF, ed. Assen (The Netherlands): Van Gorcum & Company, pp 39-52.

van der Meer, F. J. M., T. Koster, J. P. Vandenbroucke, E. Briet, and F. R. Rosendaal (1997). The leiden thrombophilia study (lets). *Thrombosis and Haemostasis 78*, 631–635.

van Duijn, C., M. Dekker, V. Bonifati, R. Galjaard, J. Houwing-Duistermaat, P. Snijders, L. Testers, G. Breedveld, M. Horstink, L. Sandkuijl, J. van Swieten, B. Oostra, and P. Heutink (2001). Park7, a novel locus for autosomal recessive early-onset parkinsonism, on chromosome 1p36. *American journal of Human Genetics 69*, 629–34.

Verbeke, G. and G. Molenberghs (1997). *Linear mixed models in practice: a sas-oriented approach*. New York: Springer-Verlag.

Verhulst, F. C., H. M. Koot, and J. van der Ende (1994). Differential predictive value of parents and teachers reports of childrens problem behaviorsa longitudinal-study. *Journal of Abnormal Child Psychology 22*(5), 531–546.

Verhulst, F. C., J. van der Ende, R. F. Ferdinand, and M. C. Kasius (1997). The prevalence of dsm-iii-r diagnoses in a national sample of dutch adolescents. *Archives of General Psychiatry 54*(4), 329–336.

Wang, W. Y., B. J. Barratt, D. G. Clayton, and J. A. Todd (2005). Genome-wide association studies: theoretical and practical concerns. *Nature Reviews of Genetics 6*, 109–118.

Whittemore, A. S. (1995). Logistic-regression of family data from case-control studies. *Biometrika 82*, 57–67.

Wright, F. A. (1997). The phenotypic difference discards sib-pair qtl linkage information. *American Journal of Human Genetics 60*, 740–742.

Zhang, H. P. and N. Risch (1996). Mapping quantitative-trait loci in humans by use of extreme concordant sib pairs: Selected sampling by parental phenotypes. *American Journal of Human Genetics 59*, 951–957.

Zhao, H. Y. (2000). Family-based association studies. *Statistical Methods in Medical Research 9*, 563–587.

Zhao, L. P., L. Hsu, S. Holte, Y. Chen, F. Quiaoit, and R. L. Prentice (1998). Combined association and aggregation analysis of data from case-control family studies. *Biometrika 85*, 299–315.

Ziegler, A., C. Kastner, and M. Blettner (1998). The generalised estimating equations: an annotated bibliography. *Biometrical Journal 40*, 115–139.

Zondervan, K. T. and L. R. Cardon (2004). The complex interplay among factors that influence allelic association. *Nat. Rev. Genet. 5*, 89–101.

Zwetselaar, M. (2003). Pedkin, software to calculate kinship and inbreeding coefficients.

Curriculum Vitae

Rachid el Galta was born on 15 August 1972 in Elkbab, Morocco. In 1991, he obtained his Baccalaureate degree from Lycée technique, Beni Mellal, Morocco. In 1998 he graduated in mathematics at the University of Rabat, Morocco. In the same year he moved to the Netherlands. After he followed a Dutch language course for about eight months he started his master study in statistics at the University of Amsterdam. He also obtained a certificate for Dutch as second language (NT2). In the end of 2001 he received the MSc. degree in Business- and Industrial Statistics. In January 2002 he started the project described in this thesis at the department of Epidemiology and Biostatistics, Erasmus Medical Center, Rotterdam. During this period he followed several courses in statistical genetics. In December 2003 he moved along with the project to the department of Medical Statistics, Leiden University Medical Center, Leiden, where he completed this thesis. The content of this thesis has been presented in different international conferences, e.g. International Genetic Epidemiology Society Conference, New Orleans (2002), USA; Conference of the Royal Statistical Society (RSS 2003) - Statistical Genetics and Bioinformatics, Diepenbeek, Belgium; International Genetic Epidemiology Society Conference (2004), Noordwijkerhoud, the Netherlands; IBS Multi-Regional Conference (2005), Leicester, UK; 25th meeting for statisticians (2005), Oslo, Norway. He also received the ANed Biometry Award 2006. In February 2006 he moved to London to join the Institute of Cancer Research, as scientific researcher.