DOCTORAATSPROEFSCHRIFT

2005
Faculteit Medische Wetenschappen

Genomic Screening Methodology for Common Diseases and Complex Traits

Multiplicity and Missingness: A Statistical Hurdle?

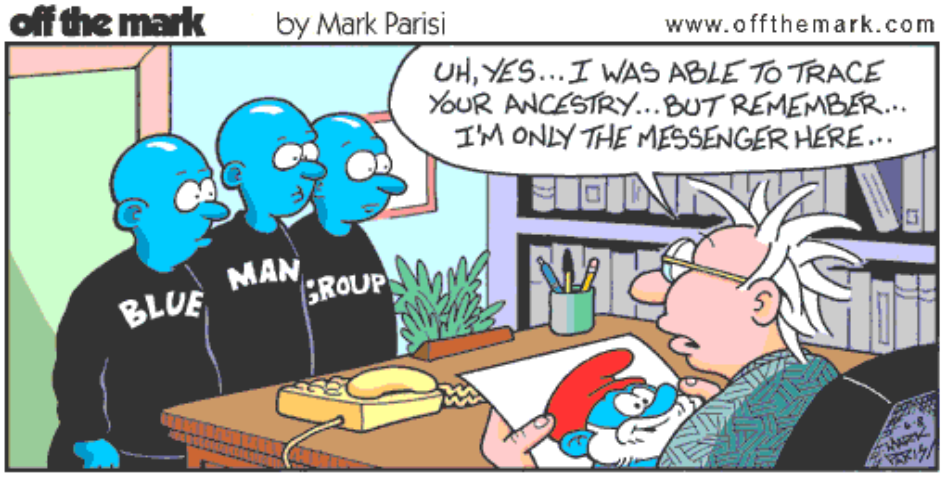
Proefschrift voorgelegd tot het behalen van de graad van Doctor in de Medische Wetenschappen, richting Biomedische Wetenschappen, te verdedigen door

Dr. KRISTEL VAN STEEN

Promotor: Prof. Dr. Geert Molenberghs
Co-Promotor: Prof. Dr. Nan Laird

"It often happens
that the mind of a person
who is learning a new science,
has to pass through all the phases
which the science itself has exhibited
in its historical evolution."

(Stanislao Cannizzaro, 1826-1910)



The Boston Saga

Thanks

I wish to express my gratitude to Prof. Dr. Geert Molenberghs (Hasselt University, Belgium) and Prof. Dr. Nan Laird (Harvard University, USA) for all their help and support during this project.

I would like to thank the scientific staff of the Center for Statistics in the Faculty of Science at the Hasselt University and of the Biostatistics Department at the Harvard School of Public Health (Boston, USA), who I am indebted to for my continuing academic training in the analysis of complex data structures (including missing data) and statistical genetics, for all kinds of practical and technical help, especially Prof. Dr. Christoph Lange. Three years ago, he suggested to come to Harvard for a short research visit, a visit that got prolonged with another 25 months ...

I am also greatly indebted to Prof. Dr. Paul Van Cauwenberge and Prof. Dr. Norbert Fraeyman (Ghent University) for their interest in my scientific work and their encouraging words, when wrapping up this project felt hard to combine with GA^{2}LEN activities, to Benjamin Raby (M.D., Channing Laboratory and Harvard Medical School - Boston, USA), to the IDPBW team of the Ghent University Hospital (in particular Lieve Jorens and Cindy Mettepenningen) and to Mrs. Mia Mortier who spent several weeks reading this thesis, taking care of linguistic aspects.

Finally, ever since I was born, my parents never failed to encourage me and to support me whenever they could. Thank you for the finishing touch and for remarks with respect to the lay-out. To all my friends, whether in the US or in Belgium, thanks for standing by me ... always.

Kristel Van Steen
Gent, 30 november 2005

Contents

List of Tables

List of Figures

Introduction

Chapter 1

How Complex are Complex Diseases?

Most common disorders are complex. Complex diseases are disorders for which the simple Mendelian model or rules of inheritance do not apply. Possible departures from this model can be attributed to allelic heterogeneity (different alleles at a locus are involved in disease susceptibility), locus heterogeneity (different loci increase disease susceptibility), gene-gene interactions, environmental factors, gene-environment interactions.

The genetic analysis of a complex trait involves several steps. The first step in assessing whether a complex disease has a genetic component is detecting and estimating familial aggregation. Familial aggregation of a trait (e.g., higher occurrence rates in siblings or offspring) is a necessary but not a sufficient condition for the presence of disease susceptibility loci. Trait occurrence rates can be elevated by within-family common environmental factors as well. Non-genetic shared family effects are typically separated from genetic shared family effects using twin studies by comparing monozygotic twin pairs (sharing all their genes) with dizygotic twin pairs (sharing 50% of their genes). The larger the similarity between monozygotic twins compared to dizygotic twins, the larger the evidence of genetic determinants for the trait under investigation. Once a genetic component to the trait or disorder is established, the mode of inheritance can be determined via a segregation analysis. Such an analysis provides more information about the inheritance pattern itself, that can be Mendelian (e.g., dominant), non-classical (e.g., with parent-of-origin effect) or non-Mendelian (no pattern at all). For complex traits, even when there

is a locus of major effect, additional genetic and environmental factors may distort Mendelian transmission rates. A segregation analysis in this setting will become more elaborate. The next step is to localise disease genes. This can be accomplished using gene mapping techniques. Statistical gene mapping for complex diseases can be performed in a variety of ways: model-based (parametric) analysis, model-free (non-parametric) analysis of allele sharing and linkage disequilibrium (association) analysis. Once the locations of a susceptibility gene is mapped, the location can be used to clone the gene (a process called positional cloning), after which the gene's function can be determined. Traditionally, a candidate gene approach is adopted. Here, the starting point are genes whose effects are known to be related to biochemical processes that are believed to affect the trait under study in some way. Once the actual causal variant is determined, the road is paved to develop diagnostics and (personalised) treatments (Figure 1.1).

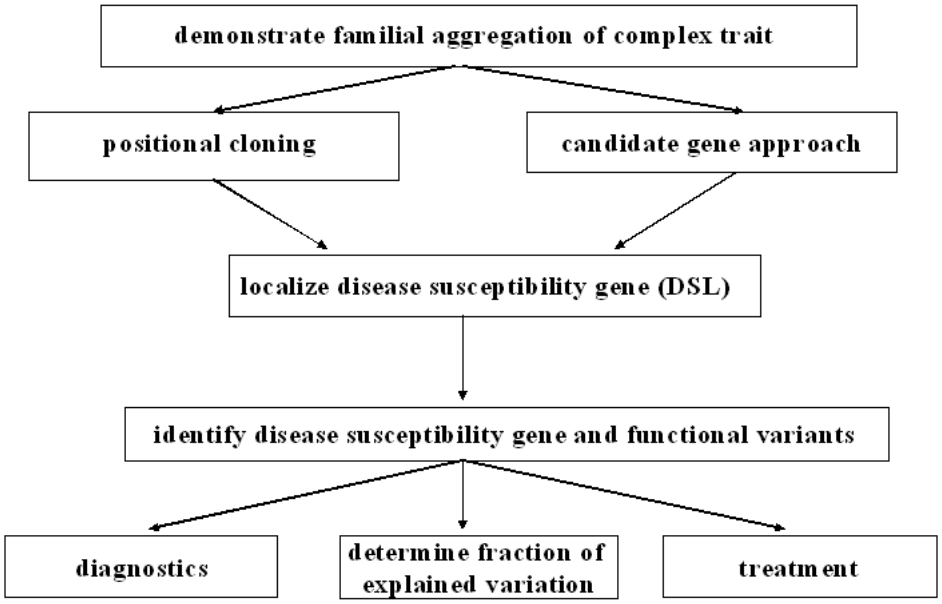


Figure 1.1: Overview of complex trait genetic analysis (Courtesy of Ed Silverman)

Any two unrelated people are the same at about 99.9% of their DNA sequences. Nevertheless, the remaining 0.1% is important because it contains the genetic variants that influence how people differ in their risk of disease or their response to treatment. Sites in the genome where the DNA sequences of many individuals differ by only a single base are called single nucleotide polymorphisms or SNPs. Since there exist about 10 million SNPs in human populations, where the

rarer SNP allele has a frequency of at least 1%, SNPs are particularly attractive in the search for the genetic underpinnings of complex causes for diseases in humans.

In facilitating this search process, an international consortium was founded (in October 2002) to build a map of the common patterns of DNA sequence variation in the human genome, by determining the genotypes and frequencies of sequence variants and their associations between them. To this end, DNA samples were taken from four populations with ancestry from parts of Africa, Asia and Europe (The International HapMap Consortium 2003). As with the completion of the Human Genome Project (April 2003), we are now on the verge of yet another era (Collins *et al.* 2003): The International HapMap Consortium recently released a public data base of common variation in the human genome with more than one million SNPs. This new information source contains accurate and complete genotypes obtained from 269 DNA samples from several populations, including ten 500Kb regions in which practically all information about common DNA variation seems to be extracted (The International HapMap Consortium 2005). The release of this extensive collection of SNPs raises the possibility to use these SNPs as markers in genome-wide association studies.

Whether or not the use of common SNPs in genome-wide association screening are sufficient in understanding most of the genetics underlying common disorders is largely driven by the validity of one of the following hypotheses. The common disease/rare variant hypothesis (CD/RV) holds that a common disease may result from any one of a large number of alleles that occur at a low population frequency (Smith and Lusis 2002). If this is true, then haplotype maps may be of limited use ... The common disease/common variant (CD/CV) hypothesis on the other hand holds that a few alleles at relatively high frequencies ($> 1\%$) represent a significant proportion of susceptibility alleles for common disease (Reich and Lander 2001). The applicability of this hypothesis will largely determine the success of the HapMap Project: the haplotypes being mapped will include only common SNPs, the disease mutations associated with these SNPs will therefore be equally common (Couzin 2002). Empirical evidence indicates that both high and low frequency alleles contribute to common diseases (Wang and Pike 2003).

In Chapter 2 we raise some general issues related to genetic association studies and position the subsequent chapters of this manuscript.

Chapter 2

Disease Association Studies

2.1 The Purpose of a Genetic Association Study

Throughout this chapter, we define genetic association to mean that the variation in the disease trait of interest is explained at least in part by an individual's genotype at a genetic marker. Association analyses are used in a variety of settings, such as looking at the effects of markers in candidate genes, fine mapping under linkage peaks and even whole genome scans. If the marker genotype being tested is a known mutation which influences the trait, values of the trait will be directly associated with presence or absence of the mutation. However, more commonly, we are testing association between a disease trait and a SNP without a known causal relationship to the disease trait. This implies that usually association between the marker and the trait is present if there is allelic association between a mutation at the hypothesised disease locus and the marker being tested.

Allelic association is a population concept, which implies that alleles at one genetic locus are statistically associated with the alleles at a second locus. For two randomly selected markers in a randomly mating population, there should be no allelic association because of the genetic reshuffling that occurs during meiosis, in which case the markers are in linkage equilibrium, otherwise they are in linkage disequilibrium. Alleles at loci on the same chromosome can also display a phenomenon known as linkage. Linkage is a physical concept and making inferences about the relative positions of two or more loci is the subject of a linkage analysis. Linkage analysis refers to a group of methods that analyse the distribution of DNA markers within families to determine if a particular region

of the genome contains a gene related to the phenotype of interest. Whereas association analysis can be family- or population-based, linkage analysis can only use family data. The underlying principle is that for two loci on the same chromosome (also called syntenic loci), separation of the two parental or maternal alleles (generating so-called recombinant haplotypes at the two syntenic loci) can only occur in the presence of a crossover between the loci. The closer the two loci, the less likely a crossover will occur between them during meiosis (the cell division that leads to the formation of egg or sperm cells) and the more likely the parental alleles will be transmitted together to an offspring. An excess of non-recombinant over recombinant gametes implies a recombination fraction between the loci (defined as the probability that a gamete is recombinant) of less than 1/2. The smaller the recombination fraction, the more tightly linked the two loci are.

Linkage mapping has been successful in identifying the genetic basis of many human diseases in which the disease penetrance resembles a simple Mendelian model, for example Huntington's disease, cystic fibrosis and some forms of breast cancer (Risch and Merikangas 1996). There is a growing literature on linkage screens for an array of complex disorders such as schizophrenia, manic depression, autism, type I and II diabetes, multiple sclerosis and lupus, many of them giving rise to conflicting results. One of the reasons for these conflicting findings may be the "multiplicity" of the data. In some settings, multiplicity may give rise to "multicollinearity" issues. We refer to Van Steen *et al.* (2002) for an exploration on multicollinearity in the context of cancer prognostic factor analyses and to Van Steen and Molenberghs (2004a) for some general notes on multicollinearity.

While recombination fraction is one of the most important parameters in linkage studies, linkage disequilibrium (LD) is the basis for association studies. This type of studies refers to methods that determine if a particular form of a DNA polymorphism occurs more frequently in subjects with a phenotype of interest (Figure 2.1). In essence, the methodology is closely related to the methodology used in epidemiology. However, genetic association can be demonstrated at two levels: at the level of the person (via genotype data) or at the chromosomal level (via haplotype data). The controversy of relying on linkage versus association strategies has not been fully resolved. We refer to Mcqueen *et al.* (2004) for a comparison of linkage and association strategies for quantitative traits. In their paper, Risch and Merikangas (1996) have pointed out that association studies for complex diseases may have more power than linkage studies to detect weak genetic

effects exhibited by loci involved in complex diseases. Although their conclusions are based on comparing non-parametric linkage mapping using affected sib pairs and family-based association using the transmission disequilibrium test (TDT, Spielman *et al.* 1993), it remains an important observation since many complex diseases are believed to be operated by multiple genes of small effect.

Risch and Merikangas (1996) also report that it is necessary to collect over 2,000 families and to type at least 1,000,000 SNPs to detect a disease susceptibility allele in a genome-wide association scan, with a genotype relative risk of 1.5 and a frequency of 0.1. Due to difficulties in assessing global statistical significance and controlling false positives, but mainly due to the large costs involved in typing so many markers in several family members, TDT-studies have lost much of their attraction through history. However, many of these obstacles are about to be removed. Advances in genotyping technology (e.g., bead technology) have already driven down the cost of genotyping in the broader marketplace and will continue to do so. Moreover, the newly proposed screening methodology of Chapter 5 (Van Steen *et al.* 2005) provides a clever way to handle the multiple testing problem in genome-wide family-based association testing, identifying associations that achieve genome-wide significance.

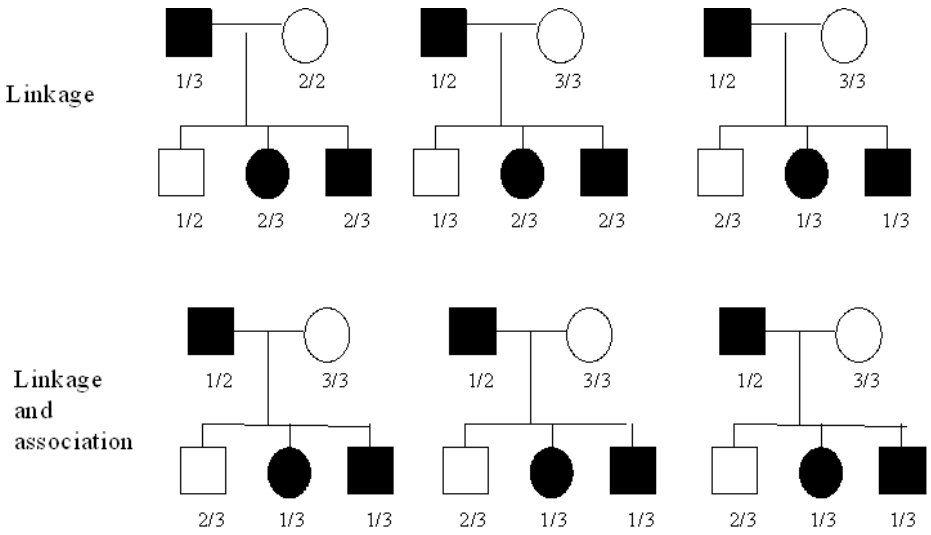


Figure 2.1: Linkage versus association analysis (Courtesy of Ed Silverman)

For a comprehensive view on testing association in genetic studies, we refer to Laird *et al.* (2005).

2.2 Data Designs for Association Studies

The statistical power to detect a genetic association depends upon numerous factors, including sample size, allele frequency differences between trait and marker loci, genetic effect size or the strength of the influence of the gene on the trait, the extent of LD between the selected markers and the disease loci, genotype errors and design (Gordon and Finch 2005). In association studies controls are either population-based or family-based (Figure 2.2). Population controls are easy to collect but are prone to population stratification or admixture. Family-based controls may refer to untransmitted alleles from parents to affected offspring, which are then compared to the transmitted alleles in association testing (Zhao 2000).

Population-based genetic association testing can be non-parametric (χ^2-tests for contingency tables or exact tests) or parametric (regression methods). In family-based genetic association testing, qualitative traits can be handled with the TDT test; both qualitative and quantitative traits can be used in association testing using a score test, such as the FBAT test (acronym for Family-based Association Test in genetic analysis; Rabinowitz and Laird 2000; Laird *et al.* 2000). FBAT builds on the original TDT method (Spielman *et al.* 1993) in which alleles transmitted to affected offspring are compared with the expected distribution of alleles among offspring, but is more general so that tests of different genetic models, tests of different sampling designs, tests involving different disease phenotypes, tests with missing parents, and tests of different null hypotheses, all fit into the same framework.

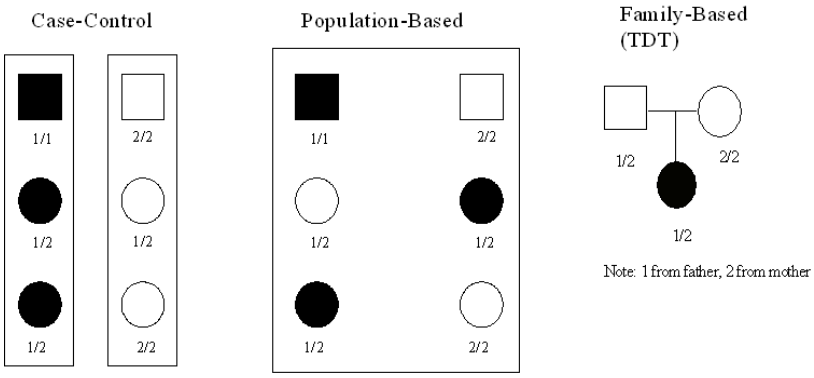


Figure 2.2: Genetic association studies: family-based, case-control, population-based (Courtesy of Ed Silverman)

Although you need larger samples to obtain family-based association tests (Mar-

tin *et al.* 1997), family designs guard against population stratification and allow testing for linkage as well as association.

2.3 Linkage Disequilibrium

Many factors can induce an observed allelic association in a population: mutation, selection, genetic drift, founder effects, and population admixture or stratification. Ordinarily, allelic association will dissipate over time due to meiosis, however it can continue for many generations if the two loci are tightly linked. This is the phenomenon called linkage disequilibrium or LD for short. LD occurs when the genetic material at the two loci is inherited as a single unit. If one locus is a disease susceptibility loci and the other is a nearby marker, the presence of LD between the two loci provides the rational for using association studies to locate disease genes.

The higher the LD, the higher the power, all other factors being equal. A key observation seems to be that the alleles at the marker and disease susceptibility locus are the same, rather than that the quantitative trait locus is common. Hence, in terms of power, not only the allele frequency of the hypothesised disease locus is important, but even more important is the frequency of the marker in LD with it. Conversely, with linkage equilibrium, there is no association between the marker and the disease locus and thus no power to detect association.

There is currently much debate on over how far LD extends to markers around a disease mutation. Mathematical and simulation models do not correspond well in practice to actual data. For sure it is population specific (depending on ancestral demographics), genomic region specific (depending on sequence features) and marker specific (depending on the type of markers considered, for instance SNP or microsatellite). We refer to Pritchard and Przeworski (2001) for a review.

2.4 Haplotype Analysis

There are several motivations to opt for haplotypes instead of single SNPs: (i) they may be more informative than individual SNPs; (ii) they reflect evolutionary history/linkage disequilibrium pattern more accurately; (iii) they may allow identification of key combinations of SNPs.

Especially if the causal mutation falls between two markers, haplotypes may

increase power. Also from a statistical point of view it may be more advantageous to turn to haplotypes, given that SNPs in LD will give rise to correlated test statistics. Instead of testing many markers separately, a smaller number of common haplotypes ($> 1\%$) can be tested.

As haplotype phase is often uncertain and needs to be inferred, procedures that model the distribution of parental haplotypes can lead to substantial bias in parameter estimates when this distribution is misspecified. Allen *et al.* (2005) use a haplotype regression approach to compute robust estimates of haplotype/disease association in family-based studies. A more elaborate view on incomplete data in the context of family-based association testing is given in Chapter 6 (Van Steen *et al.* 2004c).

2.5 Population Stratification

We have already highlighted that although LD implies association, association does not necessarily imply LD. Association can also be "spurious". When studying association between a marker and a disease trait, we take spurious association to indicate that there is actually no disease susceptibility locus in linkage disequilibrium with the marker, even though the trait and the marker are associated.

In practical terms, an observed statistical association between an allele and a phenotypic trait will be due to one of three situations. The allele itself is functional and directly affects the expression of the phenotype. The allele is in linkage disequilibrium with an allele at another locus that directly affects the expression of the phenotype. The finding can be due to chance or factors such as confounding and selection bias. In the presence of confounding factors, the frequencies of the marker allele and the causal variant both vary across populations strata, which then gives rise to (spurious) association. An early example of spurious association due to the presence of population admixture is given by Knowler *et al.* (1988), who studied diabetes mellitus in a population of American Indians with mixed Caucasian heritage.

As mentioned before, case-control and population-based methods are particularly susceptible to spurious associations caused by population stratification. If population stratification can be measured, then testing for association within strata is a good solution. In general, there are two strategies for protecting against population

structure in case-control studies: Structured association (Pritchard *et al.* 2000a,b) and genomic control (Devlin and Roeder 1999). Problems associated with population stratification can be avoided by careful matching of cases and controls (ensuring similar ethnic and genetic background), by genotyping multiple unlinked markers, or by using family-based methods like FBAT.

2.6 Multiple and Multivariate Associations

Most screens are evaluated on a marker-by-marker basis. Ideally all markers are analysed jointly and genetic or biological interactions are accounted for. Evidently, the number of markers will soon exceed the number of observations. Hoh *et al.* (2000, 2001) therefore propose to find a set of SNP loci that is significantly associated with the disease. The idea is to first do a data reduction step and then to model the interactions and make predictions about genetic effects. A reduced set of markers can be obtained in several ways: grouping markers that show individually highly significant association, only using markers in genes that are over-expressed in cases versus controls, etc. The reduced set of markers is then used to develop appropriate models.

One of the major benefits of a classic regression approach in genetic association modelling is the ease of including information from important confounders or predictors and the possibility of acknowledging interactions (such as gene-gene or gene-environment interactions). One of the major drawbacks though is that including too many genetic markers in one model requires too large samples to guarantee adequate power. Problems are compounded when marker data are missing or when marker data are of poor quality (or even erroneous). In addition, in particular for densely spaced markers in candidate gene regions, multicollinearity issues arise from the underlying existing complex LD structure.

In response to these problems a multivariate approach can be adopted in which marker data are modelled conditional on the trait (Chapter 4, Van Steen *et al.* 2004b). In theory, any number of genotypes is allowed. In practice, problems associated with having more parameters than observations remain. However, in this framework missingness and error measurements at the genotype level have become outcome problems. Whereas the first can be integrated in a Rubin framework, the second involves dealing with residual errors. In addition, the problem of multicollinearity is alleviated and as a bonus of the multivariate strategy, we can in-

vestigate the dependence of interactions on disease status and environmental factors.

Apart from a parametric modelling method, a non-parametric approach can be adopted, e.g., Multifactor Dimensionality Reduction (MDR - Ritchie *et al.* 2001, 2003; Hahn *et al.* 2003). Essentially, MDR defines a new variable that incorporates information from several loci that can be divided into high risk and low risk combinations, for a binary trait. Cross-validation (Hastie *et al.* 2001) and permutation testing (Good 2000) can then be used to evaluate the new construct for its ability to classify and predict disease risk status. MDR is particularly attractive since it seems to be able to detect higher-order gene-gene interactions with reasonable to good power in many circumstances, without the pitfall of having to deal with hard to verify model assumptions. This makes the MDR approach, with optimization procedures similar to incorporating machine learning methodology, a promising tool for genome-wide scans (refer to http://liinwww.ira.uka.de/bibliography/Neural/ for a collection of computer science bibliography on neural networks).

2.7 Genome-wide Association Studies

These days we are overwhelmed with data to increase our understanding about the genetic forces that determine disease. SNP arrays of 10K, 100K and 500K SNP are already available. The 1M SNP array is just around the corner. In addition, costs are forecast to come down substantially, coverage is getting better and genotyping error rates can often be reduced to less than 1%.

Also on a larger genomic scale (e.g., genome-wide level) there are two main types of studies to determine the contribution of genes to disease susceptibility: linkage and association. genome-wide linkage studies using SNPs may narrow down linkage peaks, yet after approximately 5-6K SNPs no additional information is gained. Genome-wide association studies using SNPs, potentially use all available data, are more powerful for genes of small to moderate effect and allow for covariate assessment, interactions, effect size, etc. However, there are some major obstacles to overcome, such as how to best select a SNP set and applying a safeguard against an excess of false positives due to multiple testing (Duncan *et al.* 2005).

The proposed methods in Chapter 5 are rather unique in the sense that they use the entire sample and do not require separate screening and validation samples to establish genome-wide significance, as is the case in population-based designs.

Kraft (2005) proposes a multi-stage approach where a portion of the samples are genotyped first with a high-throughput genotyping method, and a small number of the most promising variants are then genotyped in the remaining samples using a lower throughput method. Similar in spirit to controlling the family-wise error rate (as in Satagopan *et al.* 2004), the author aims to limit the false positive report probability (FPRB, Wacholder *et al.* 2004) while maximising the number of expected true positives. The standard strategy for analysing such two-stage data is to view the second stage as a replication study. Instead of focussing on findings that reach statistical significance when stage two data are considered stand-alone, Skol *et al.* (2005) analyse all available data jointly. In many cases, joint analysis seems to be more powerful.

2.8 Multiple Testing

Multiple testing may involve multiple studies, multiple phenotypes, multiple markers, multiple test statistics, and combinations thereof. If we test 500K SNPs for association with a single phenotype at the 5% level, we can expect approximately 25,000 false positives. When multiple phenotypes or haplotypes are considered the number of tests easily exceeds 1M ... Multiple testing is one of the reasons why false positive rates (probability of no association among significant findings) in association analysis can be fairly high. Other reasons may be insufficient power (whether caused by too small sample size or too small genetic effect size), or an inappropriate (too high) critical p-value. Traditional corrections for multiple testing such as Bonferroni-correction (Bonferroni 1936) are far too severe. But also more recently developed strategies such as those controlling the false discovery rate (FDR) seem to break down.

Alternatively, SNPs are selected in such a way that multiple testing becomes less of a problem. For instance, Stram (2004) gives a review of current methods for selecting informative SNPs for association studies, using data from a dense network of SNPs that have been genotyped in a relatively small group of subjects. Several definitions can be given to "optimal" SNP selection, one of them referring to eliminating as much redundancy in the information provided by the SNPs as possible. Tag SNPS are SNPS that uniquely identify a set of haplotypes. The number of tag SNPs that contain most of the information about the patterns of genetic variation is estimated to be about 300,000 to 600,000, which is far fewer than the approximately 10 million common SNPs. Obviously, tag SNPs often greatly reduce genotyping

costs and may reduce test degrees of freedom. We offer an alternative view on the concept of tagging SNPs in Chapter 5.

2.9 Incomplete Data

The applied statistician frequently encounters correlated outcome data. Common situations include multivariate, clustered, and longitudinal data. In such settings, it frequently occurs that not all of the planned measurements of subject i's outcome vector $\boldsymbol{y}_i$ are actually observed, turning the statistical analysis into an incomplete or missing data problem. For example, in a longitudinal study, a subject's response vector may terminate early for a number of reasons outside of the control of the investigator. This feature is referred to as dropout, a special case of missingness. It is almost always necessary to reflect on the nature of the missingness process and its impact on inferences.

When referring to the missing-value, or non-response, process we will use terminology of Little and Rubin (2002). A non-response process is said to be missing completely at random (MCAR) if the missingness is independent of both unobserved and observed data and missing at random (MAR) if, conditional on the observed data, the missingness is independent of the unobserved measurements. A process that is neither MCAR nor MAR is termed non-random (MNAR).

Rubin's taxonomy is particularly relevant when *modelling* data with missing observations, but is less applicable in the context of *testing* hypothesis. Nevertheless, missing observations may have a large impact on test results, especially when test statistics are not appropriately accounted for the incomplete data they are computed from. In what follows, we give a brief historical overview of missing data from a modelling perspective. Incomplete data in the context of family-based association testing will be dealt with in Chapter 6.

Early work on missing values was largely concerned with algorithmic and computational solutions to the induced lack of balance or deviations from the intended study design (Afifi and Elashoff 1966; Hartley and Hocking 1971). In the meantime, a number of applied areas have adopted the practice of analysing incomplete data in relatively simple ways (Little and Rubin 2002; Verbeke and Molenberghs 2000), resorting to, for example, a so-called complete case analysis (CC), where all subjects with incomplete outcome vectors are discarded, or using

single imputation strategies, where values are substituted for missing measurements. Commonly used imputation strategies are mean imputation and last observation carried forward (LOCF). For a detailed criticism of such methods, see Molenberghs *et al.* (2004). The criticism of these methods are to a large extent directed towards the strong assumptions needed for them to be valid. For example, CC requires MCAR while LOCF requires even stronger assumptions.

Over the last three decades, a number of developments have taken place, allowing the use of MAR-based methods. For example, general data augmentation algorithms have been developed, the most famous one undoubtedly being the Expectation-Maximization algorithm (EM; Dempster *et al.* 1977), together with multiple imputation strategies (Rubin 1987). MAR methods are very important, not only because they relax the tight MCAR assumption, but also because their relationship with the concept of ignorability. Indeed, Rubin (1976) and Little and Rubin (2002) have shown that, under MAR and mild regularity conditions (parameters $\boldsymbol{\theta}$ describing the measurement process and $\boldsymbol{\psi}$ describing the missingness process, are functionally independent), likelihood-based and Bayesian inferences are valid, even when the missing data mechanism is ignored (see also Verbeke and Molenberghs 2000). Practically speaking, the likelihood of interest is then based upon the factor $f(\boldsymbol{y}_i|\boldsymbol{\theta})$, where $\boldsymbol{y}_i^o$ refers to the observed portion of the outcome vector and $\boldsymbol{y}_i^m$ likewise stands for the missing part. This is called ignorability. The practical implication is that a software module with likelihood estimation facilities and with the ability to handle incompletely observed subjects manipulates the correct likelihood an hence provides valid inferences (point estimates, standard errors, and likelihood ratio tests). This is at the mild condition that the observed information matrix is used (Kenward and Molenberghs 1998).

The practical implication for likelihood inference is that, as soon as a module is available to handle measurement sequences of unequal length, valid inferences are obtained without any additional work. This type of tools abounds for Gaussian measurements, where a large number of software packages have implemented the linear mixed-effects model (Verbeke and Molenberghs 2000), such as, for example, the SAS procedure MIXED (SAS 2000). When outcomes are of a non-Gaussian type (binary, ordinal, counts, etc.), several options are available. A typical random-effects model in this context is the generalized linear mixed-effects model (Breslow and Clayton 1993; Wolfinger and O'Connell 1993; Fahrmeir and Tutz 2002), implemented, for in the SAS procedure NLMIXED. Since this method is

likelihood-based and the procedure manipulates the correct likelihood (handling the numerical complexity of integrating out the random-effects distribution by numerical integration), the procedure is valid under MAR.

For non-Gaussian outcomes, apart from random-effects models, also marginal models have become popular. Typical models include the Bahadur model (Bahadur 1961) and the multivariate Dale or global odds ratio model (Molenberghs and Lesaffre 1994, 1999). For an overview, see Aerts *et al.* (2002). Since these models specify, in principle, the full likelihood, they can be used to analyse incomplete data as well, under MAR assumptions, and making use of the ignorability property (Kenward *et al.* 1994). However, marginal models for non-Gaussian data imply complex and hard to manipulate likelihoods. In many practical settings involving outcome sequences of moderate to large length, direct likelihood is prohibitive.

As a response to this problem, a number of alternatives have been formulated, the most popular one undoubtedly being generalized estimating equations (GEE; Liang and Zeger 1986; Diggle *et al.* 2002). By transforming the score equations into estimating equations, this method essentially allows confining attention to the specification of the first moments of the outcome sequence only (i.e., the mean structure), thereby circumventing the need to address the association structure while still leading to valid inferences. A number of variations to this theme exist, such as GEE2 (also specifying the second moments; Liang *et al.* 1992) and alternating logistic regressions (Carey *et al.* 1993). When data are incomplete, GEE suffers from its frequentist nature and is in its basic form valid only under MCAR. Therefore, Robins *et al.* (1995b) have developed so-called weighted generalized estimating equations (WGEE), as well as a number of refinements and extensions in subsequent papers, to allow usage of GEE under not only MAR, but even under MNAR settings. The method rests on Horvitz-Thompson ideas (Cochran 1977), weighting contributions by the inverse probability of being observed. The method is elegant and enjoys good properties, but explicitly requires specification of a model for the weights.

More recently, pseudo-likelihood methods (PL; le Cessie and van Houwelingen 1994; Geys *et al.* 1998; Geys *et al.* 1999; Aerts *et al.* 2002) have become popular as an alternative to full-likelihood, and therefore also to GEE and GEE2. Rather than replacing the score equations with alternative functions, the likelihood is replaced by a more tractable function. Computational and statistical performance

(e.g., efficiency) have been shown to range from acceptably good to excellent. A correction is needed to allow the use of pseudo-likelihoods in a MAR setting. This correction seems to be much milder than the one needed for WGEE and, in particular, does not require the construction of any additional model (ongoing research - Van Steen *et al.*).

Of course, whatever MAR developments are made, one can never exclude the operation of a MNAR mechanism. A number of modelling strategies (Diggle and Kenward 1994; Molenberghs *et al.* 1997) have been proposed, but at the same time it has been reported that such strategies are very sensitive to unverifiable modelling assumptions (Kenward 1998). A number of sensitivity analysis tools have been proposed (Molenberghs *et al.* 2001; Verbeke *et al.* 2001; Van Steen *et al.* 2001; Kenward *et al.* 2001; Jansen *et al.* 2003; Robins *et al.* 1998).

In Chapter 3 we introduce three data sets that will be referred to throughout this manuscript: the APOE data in Chapter 4, the CAMP and Affymetrix data in Chapter 5.

Chapter 3

Case Studies

3.1 The CAMP Study

3.1.1 Childhood Asthma Management Programme

Asthma is a complex disease, expressed by several genetic factors, each of which possibly interacting with various environmental stimuli. To date, there has been genome-wide linkage analyses of at least 11 different populations for asthma-related phenotypes. Over twenty chromosomal regions have been identified for asthma and related phenotypes IgE, skin test reactivity, eosinophil count, and airway responsiveness (Hoffjan and Ober 2002). Five asthma genes have been identified by positional cloning, and over 200 genetic association studies of asthma and its associated phenotypes have been reported (Weiss and Raby 2004; Wills-Karp and Ewart 2004; Hoffjan *et al.* 2003).

The Childhood Asthma Management Programme was designed to evaluate whether continuous, long-term treatment with either budesonide (an inhaled corticosteroid) or nedocromil (an inhaled noncorticosteroid drug) safely produces an improvement in lung growth as compared with placebo (i.e., treatment for symptoms only). A total of 1041 asthmatic children, ages 5 to 12 years with mild to moderate asthma, were randomised into the three treatment arms. The mean duration of follow-up was 4.5 years (CAMP 2000). Trial design and methodology have been previously published (CAMP 1999). Appropriate informed consent was obtained from all participating subjects at each of the CAMP centres. DNA samples were collected as part of the ancillary study protocol from approximately 93% of the cohort; parental samples were also collected.

These parent/child trios data from the CAMP Genetics Ancillary Study (CAMP 1999) are used to illustrate and validate our proposed screening technique in candidate gene studies (Van Steen *et al.* 2005, Chapter 5).

3.1.2 Marker and phenotype selection

Within CAMP, DNA samples for complete parent/child trios were available for 651 nuclear families. Inclusion of sib-pairs in CAMP provided 707 parent/child trios within these families. A total of 291 typed SNPs were used.

CAMP not only exhibits an extensive source of phenotypic baseline information before randomization to drug therapy, it also has repeated measures data on a variety of asthma-related phenotypes (Table 3.1). Asthma can be considered a syndrome, with varying contributions of clinical, immunologic and physiologic manifestations, including constellation, pattern and severity of symptoms, markers of atopy and measures of bronchial responsiveness (Clough 1998). The asthma phenotype is heterogeneous and refers to a spectrum of disorders, that result in the common clinical feature of intermittent wheeze. Therefore, instead of using asthma directly as a phenotype, or a classification of different asthma types, we investigated intermediate quantitative phenotypes such as PC20 (i.e., methacholine, 11 measurements between 0-52 months, log-transformed to lnPC20) in a range of studies (e.g., Raby *et al.* 2005a,b,c; Van Steen *et al.* 2005).

Table 3.1: Quantitative phenotypes in CAMP Genetics Ancillary Study participants

Phenotype	Number of Subjects with Phenotype Values	Mean	Standard Deviation
Post-Bronchodilator FEV_1 (% Predicted)	698	102.800	12.700
Post-Bronchodilator FVC_1 (% Predicted)	698	106.300	12.800
Ln PC20 to Methacholine	698	0.042	1.166
Bronchodilator Responsiveness (post-FEV_1 - pre-FEV_1) as Absolute Volume in liters	680	0.159	0.135
Bronchodilator Responsiveness as % of % of Baseline FEV_1	680	10.360	9.420
Bronchodilator Responsiveness as % of % of Predicted FEV_1	680	8.990	7.180
Morning Mean Peak Expiratory Flow Rate	700	245.400	64.700
Evening Mean Peak Expiratory Flow Rate	700	255.300	65.400
Mean Asthma Symptom Score	700	0.611	0.345
Total Eosinophil Count	687	512.900	456.700
Log Total Serum IgE	692	2.630	0.680
Number of Positive Skin Tests	700	3.530	2.660

3.2 The Affymetrix 10K Data Set

Prostate cancer is a group of cancerous cells (a malignant tumor) that begins most often in the outer part of the prostate. It is one of the most common types of cancer in men in the Western world and a leading cause of mortality. Early prostate cancer usually does not cause any symptoms. As the tumor grows, it may spread from the prostate to nearby lymph nodes, bones or other organs.

Each family in the Affymetrix 10K data set was selected through a proband who received treatment for prostate cancer at the Mayo Clinic. A family was eligible if there were at least three men with prostate cancer in the family, of whom at least two were alive for recruitment (Schaid *et al.* 1998; Cunningham *et al.* 2003). In total 160 families were available, which included 437 men affected with prostate cancer and 157 unaffected men and women.

For the genotyping of SNPs, DNA was isolated from peripheral blood lymphocytes by standard methods. The Early Access Affymetrix Mapping 10K array was used in accordance with the manufacturers recommendations (Schaid *et al.* 2004). From the original data set, four men were excluded because of degraded DNA and one family was excluded because it was no longer informative as a result of degraded DNA. This resulted in 433 affected men from 159 families. Because of cost constraints, no genotype data were retrieved for the unaffected members of these original families. Over time, nine new affected subjects and eight new pedigrees were recruited (25 affected men and 17 unaffected men and women). Since all affected men in the original pedigrees were genotyped and all members of the new pedigrees were genotyped, SNP data were available for 167 families with 467 affected men. The research protocol and informed consents were approved by the Mayo Clinic Institutional Review Board.

3.3 ApoE Data for Alzheimer's Disease

There are two types of Alzheimer Disease: early-onset and late-onset AD. Early-onset familial AD is a rare form of AD that usually occurs between the ages of 30 and 60. Late-onset AD is the more common form of AD. It strikes seniors in their late 60s and beyond. Unlike early-onset AD there is no clear inheritance pattern detectable in most families. It is regarded as a complex disorder, triggered by several interacting genes and environmental factors. The best-known identified risk factor for late-onset AD is a gene on chromosome 19 (Figure 3.1) that directs the

manufacture of apolipoprotein E (ApoE) (MIM 107741), a protein that helps carry blood cholesterol throughout the body (Strittmatter *et al.* 1993; Saunders *et al.* 1993). It is found in neurons and other supportive brain cells of healthy brains, but it is also associated in excess amounts with the plaques found in the brains of people with Alzheimer's disease. Three common isoforms of the protein are found in most populations: ApoeE2, ApoeE3 and ApoeE4, determined at the DNA level by the $\epsilon 2$, $\epsilon 3$ and $\epsilon 4$ alleles of the ApoE gene (Weisgraber 1994). For more information on AD, we refer to www.alzinfo.org.

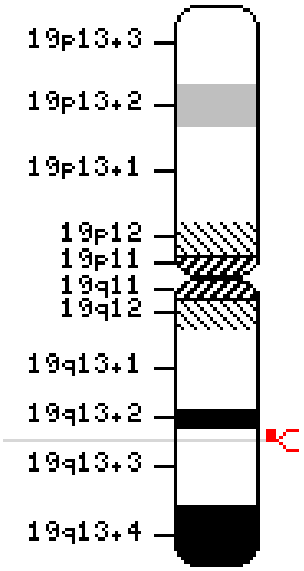


Figure 3.1: Location of the APOE gene in Homo sapiens. The APOE gene is mapped to chromosome 19 in a cluster with APOC1 and APOC2. Defects in apolipoprotein E result in familial dysbetalipoproteinemia, or type III hyperlipoproteinemia (HLP III), in which increased plasma cholesterol and triglycerides are the consequence of impaired clearance of chylomicron and VLDL remnants (Source: NCBI - Entrez Gene).

The ApoE data set refers to a sample of 210 Alzheimer's patients selected from 33 US hospitals and 159 controls taken from a group of non-demented prostate cancer hospital patients recruited in Paris and Nancy, France (Knapp *et al.* 1994). Standard methods were used to carry out blood collection and DNA extraction. A total of 8 SNPs in a 205Kb region of chromosome 19 containing the ApoE gene were

genotyped among the cases and controls (Fallin *et al.* 2001).

Genetic Associations

Chapter 4

Introducing the Multivariate Dale Model in Population-based Genetic Association Studies

4.1 Introduction

Until recently, the most common parametric approaches to study the combined effects of several genetic polymorphisms located within a gene or in a small genomic region are, at the genotype level, logistic regressions and at the haplotype level, haplotype analyses. An alternative modelling approach, based on the case-control principle, is to regard exposures (e.g., genetic data such as derived from Single Nucleotide Polymorphisms - SNPs) as random and disease status as fixed and to use a marginal multivariate model that accounts for inter-relationships between exposures. One such model is the multivariate Dale model. This model is based on multiple logistic regressions. That is why the model, applied in a case-control setting, leads to straightforward interpretations that are similar to those drawn in a classical logistic modelling framework.

Genetic association studies between candidate polymorphisms and the case-control status of unrelated individuals offer a possible approach to identify disease predisposing loci or mutations. Most genomic regions of relevance to a disease

may have many base positions within them that, when upset or mutated, either directly contribute to disease susceptibility or have some related pathological effects (Longmate 2001; Horikawa *et al.* 2000). The association of alleles at these multiple sites with disease may thus be subtle, since affected individuals may possess one or a combination of them. Since it is well-known that, within a small genomic region or a candidate gene, test statistics assessing the association with the trait may be correlated as a result of linkage disequilibrium between alleles at the marker loci, a multi-allelic approach should be advocated. Until recently, the most common parametric approaches to study the combined effects of several alleles at different positions are logistic regressions at the diploid level (Czika *et al.* 2001; Cordell and Clayton 2002) and haplotype analyses at the haploid level (Excoffier and Slatkin 1995).

The first set of methods traditionally incorporates genotypic information at the covariate level of a logistic model. Under a case-control design the probability of disease is functionally related to the genotypes at several loci of interest. Apart from main effects, the models easily accommodate epistatic interactions between the loci. Dominant or recessive properties at particular loci can be assessed via classic statistical tests (see also Cordell and Clayton 2002). However, investigating the effect of several polymorphisms within a small genomic region showing strong linkage disequilibrium between the (some) markers may lead to multicollinearity and sparseness problems in the multi-genotypic tables. Those phenomena may jeopardise substantially extracting clear interpretations of the results obtained. The main problem of multicollinearity lies in the fact that the estimated regression coefficients tend to vary widely from one sample to another when the predictor variables are highly correlated. Many of the estimated regression coefficients individually may be statistically not significant even though a definite statistical relation exists between the response variable (e.g., disease status) and the set of predictor variables (e.g., marker information). In addition, the simple interpretation of the regression coefficients as measuring marginal effects is often unwarranted with highly correlated predictor variables (Neter *et al.* 1996). Moreover, the focus in logistic regression analyses is describing the relationship between covariates (phased or unphased genotypes) and disease status. If the association structure between several loci is of interest, multivariate models are more appropriate.

The second set of methods relies on haplotype frequency estimations that are essentially based on expectation-maximization (EM)-based algorithms (Excoffier and

Slatkin 1995; Hawley and Kidd 1995; Long *et al.* 1995). The profiles of haplotype frequencies between the two groups in a case-control design can be compared by performing an omnibus likelihood ratio (OLR) statistic and the significance of the test can be approximated by an empirical p-value obtained by permutations. The use of haplotypic information can easily accommodate weak linkage disequilibrium, potential allelic heterogeneity and does not require any assumption about the nature of the haplotype frequencies. However, this approach is not without limitations. The various EM algorithms make the limiting assumption of the Hardy-Weinberg equilibrium (i.e., alleles are acting independently of each other within genotypes at the different loci) even though it has been shown to be robust under specific and well-maintained case-control conditions (Stephens *et al.* 2001). Neither recessive nor dominant effects are accounted for which is a real limitation if it is believed that alleles at certain loci may only have an impact on the genotypic level. Furthermore, it is only possible to give a rough positioning of the potential functional disease loci and it remains to be validated how the method performs using genetic variants or markers with smaller effects on the disease status. Finally, the size of the haplotype needs to be determined.

For the sake of completeness, we mention the existence of alternative haplotype approaches to describe or test genetic associations. Haplotype dosage (0, 1 or 2; the count of the number of copies of a haplotype in the pair of haplotypes carried by an individual) is entered in the parameteric model as a linear term, or its expectation is (e.g., Zaykin *et al.* 2002; Schaid *et al.* 2002, Stram *et al.* 2003) hereby accounting for haplotype uncertainty. Tests of association between phenotype and haplotypes can also be constructed using estimating equations methodology and treating unknown or ambiguous haplotypes as latent variables (Zhao *et al.* 2003).

Thus, sensitive statistical methods are needed in the identification of multiple susceptibility genes associated to complex diseases within a small genomic region or a candidate gene. A multivariate model in which disease status is regarded as a covariate of interest and in which the genotype/allelic data are regarded as multiple responses, may serve this purpose. More specifically, we propose the use of the multivariate Dale model (Molenberghs and Lesaffre 1994). This model extends the bivariate global odds ratio model described by Dale (1986) and McCullagh and Nelder (1989). Being not only more intuitive, this approach also enables to investigate existing correlations between exposures. Interpretations are straightforward and similar to those drawn in a classical logistic modelling framework. The latter

is related to the observation that the multivariate Dale model is a marginal model. Hence, all properties that are valid for each of the marginal logistic regressions can be carried over. The only pre-requisite is that cases and controls should be selected independently of exposure status.

The structure of this chapter is as follows: after providing technical details of the multivariate Dale model and showing its utility in genetic association studies, treating marker information at the response level, we present the results obtained in an application of the model to a real-life data set of case-control individuals genotyped on SNPs within a 250Kb region containing the APOE gene. We show that the model can recover both results found by others as well as effects not previously described. For illustrative purposes, we have restricted attention to combinations of at most three markers. The model is applicable to any number of markers, but may require alternative estimation procedures. We finally discuss future directions for development of the model and additional thoughts concerning its advantages and use.

4.2 Materials and Methods

4.2.1 A Marginal Model for Multivariate Ordinal Data

The multivariate Dale model extends the bivariate global odds ratio model described by Dale (1986) and McCullagh and Nelder (1989). As a true multivariate model, it not only accounts for the dependence of the multiple ordinal responses on covariates (which may be time-varying, continuous and/or discrete), but for the dependence between the multiple responses as well. Joint probabilities are decomposed into main effects (described by marginal probabilities) and interactions (described by odds ratios of second and higher orders).

Let $i = 1, \ldots, N$ indicate the covariate or design level, containing n_i subjects. Every subject r at the ith level (group) provides information on T_i distinct markers and for each marker, the subject is scored using a categorical outcome variable. Hence, the outcome for subject r in the ith level, characterised by a vector of covariates $\boldsymbol{x_i}$, is a series of measurements Y_{irt} $(t = 1, \ldots, T_i)$, where Y_{irt} can take on c_t distinct (possibly ordered) values k_t (e.g., $1 \leq k_t \leq c_t$). For instance, in case of a biallelic marker with alleles A and a, the non-missing outcome for a subject r may be aa, Aa or AA (0, 1 or 2 at risk alleles for disease). The number of distinct values for that marker is $c_t = 3$.

Categorical data are typically presented in the form of frequency counts of observations. We therefore summarise the categorical outcomes, measured for subjects with covariate vector $\boldsymbol{x_i}$, in a cross-classification of the outcomes Y_{irt} into a $c_1 \times \cdots \times c_{T_i}$ dimensional contingency table with cell counts

$$Z_i^*(\boldsymbol{k}) \equiv Z_i^*(k_1, \ldots, k_{T_i}). \tag{4.1}$$

We observe that $\sum_{\boldsymbol{k}} Z_i^*(\boldsymbol{k}) = n_i$. At every T_i-dimensional cutpoint, the data table is collapsed into a $2 \times 2 \times \cdots \times 2$ table. Corresponding probabilities are denoted as

$$\mu_i^*(\boldsymbol{k}) = Pr(\boldsymbol{Y}_{ir} = \boldsymbol{k} |\ \boldsymbol{x_i}, \boldsymbol{\theta}).$$

The multivariate Dale model is based on the assumption that every such table arises as a discretization of a multivariate Plackett distribution (Plackett 1965; Molenberghs and Lesaffre 1994). For example, tabular unphased genotype data for two biallelic markers M1 and M2 will give rise to four 2-dimensional cutpoints as in Figure 4.1. Each 2×2 table is considered to be a discretization of a bivariate Plackett distribution.

We note that, given the ordinal nature of the outcomes, a more natural strategy is to work with cumulative counts

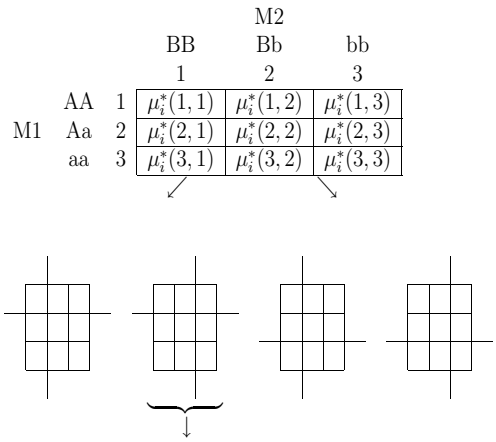
$$Z_i(\boldsymbol{k}) = \sum_{\boldsymbol{\ell} \leq \boldsymbol{k}} Z_i^*(\boldsymbol{\ell}).$$

Here, $\boldsymbol{\ell} \leq \boldsymbol{k}$ is a short-hand notation for $l_j \leq k_j, j = 1, \cdots, T_i$. In other words, $Z_i(\boldsymbol{k})$, where $\boldsymbol{k} = (k_1, \ldots, k_{T_i})$, simply refers to the number of individuals in group i of which the observed response vector is $\boldsymbol{\ell}$ with $\boldsymbol{\ell} \leq \boldsymbol{k}$. The corresponding probabilities are

$$\mu_i(\boldsymbol{k}) = Pr(\boldsymbol{Y}_{ir} \leq \boldsymbol{k} |\ \boldsymbol{x_i}, \boldsymbol{\theta}).$$

Note that $\mu_i(c_1, \ldots, c_{T_i}) = 1$ and $Z_i(c_1, \ldots, c_{T_i}) = n_i$. In addition, marginal counts are given by all counts for which all but one indices are equal to their maximal value: $Z_{itk} \equiv Z_i(c_1, \ldots, c_{t-1}, k, c_{t+1}, \ldots, c_{T_i})$. Bivariate cell counts, i.e., cell counts of a cross-classification of a pair of outcomes, are obtained by setting all but two indices k_s equal to c_s. Trivariate cell counts and counts of higher order are obtained in a similar fashion. In each of these cases, the corresponding probabilities can be derived in a straightforward way: e.g., univariate (cumulative) probabilities referring to the sth marker outcome, are denoted by

$$\mu_{isl} = \mu_i(c_1, \ldots, c_{s-1}, \ell, c_{s+1}, \ldots, c_{T_i}).$$

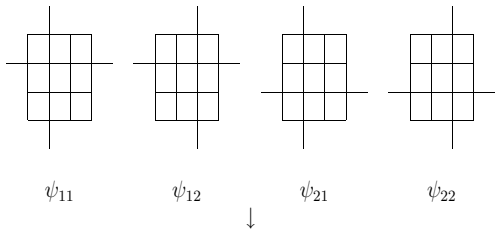


$\mu_i(1,2)$	$\mu_i(1,3) - \mu_i(1,2)$
$\mu_i(3,2) - \mu_i(1,2)$	$\mu_i(3,3) - \mu_i(3,2) - \mu_i(1,3) + \mu_i(1,2)$

or

$\mu_{i,12,12}$	$\mu_{i,11} - \mu_{i,12,12}$
$\mu_{i,22} - \mu_{i,12,12}$	$1 - \mu_{i,22} - \mu_{i,11} + \mu_{i,12,12}$

Figure 4.1: Summarizing a 3×3 table as condensed 2×2 tables



ψ_{11}	ψ_{12}
ψ_{21}	ψ_{22}

Figure 4.2: Global odds ratios for the bivariate case of two diallelic markers M1 and M2

Bivariate (cumulative) probabilities pertaining to the tth and sth marker outcome, are denoted by

$$\mu_{i,ts,k\ell} = \mu_i(c_1, \ldots, c_{t-1}, k, c_{t+1}, \ldots, c_{s-1}, \ell, c_{s+1}, \ldots, c_{T_i})$$

and refer to the probability that $Y_{irt} \leq k$ and $Y_{irs} \leq l$ (Figure 4.1).

The multivariate Dale model involves describing the marginal distributions (T_i in total), $T_i(T_i - 1)/2$ pairs of two-way interactions and three or higher order associations. The latter are often assumed to be constant. The description is completed by specifying link functions and linear predictors for both the univariate margins and the association parameters.

For the univariate marginal links, a convenient choice is the logistic link function:

$$\eta_{itk} = \text{logit}(\mu_{itk}|x_{it}) = \beta_{0\,itk} + \beta_{itk}x_{it}, \;\; (1 \leq t \leq T_i, 1 \leq k < c_t). \tag{4.2}$$

Note that the index k in β_{itk} indicates that the β parameters are allowed to depend on the cutpoints (i.e., marginal non-proportional odds). If evidence is found that the regression parameters are consistent across the cutpoints k, β_{itk} in (4.2) may be replaced by β_{it}, implying a proportional odds model for the response.
Taking up the example of two biallelic markers M1 and M2, we would have to link two cumulative probabilities for marker 1 and two cumulative probabilities for marker 2 to covariates of interest x_{it}. The marginal links under the assumption of non-proportional odds can then be written as:

$$\begin{aligned}
\text{Marker M1:} &\begin{cases} \eta_{i11} &= \text{logit}(\mu_{i11}|x_{i1}) = \beta_{0\,i11} + \beta_{i11}x_{i1}, \\ \eta_{i12} &= \text{logit}(\mu_{i12}|x_{i1}) = \beta_{0\,i12} + \beta_{i12}x_{i1}, \end{cases} \\
\text{Marker M2:} &\begin{cases} \eta_{i21} &= \text{logit}(\mu_{i21}|x_{i2}) = \beta_{0\,i21} + \beta_{i21}x_{i2}, \\ \eta_{i22} &= \text{logit}(\mu_{i22}|x_{i2}) = \beta_{0\,i22} + \beta_{i22}x_{i2}. \end{cases}
\end{aligned} \tag{4.3}$$

Full specification of the association is done in terms of marginal global odds ratios:

$$\psi_{i,ts,k\ell} = \frac{(\mu_{i,ts,k\ell})(1 - \mu_{itk} - \mu_{is\ell} + \mu_{i,ts,k\ell})}{(\mu_{is\ell} - \mu_{i,ts,k\ell})(\mu_{itk} - \mu_{i,ts,k\ell})}. \tag{4.4}$$

For every chosen pair $\{t, s\}$, a set of $(c_t - 1) \times (c_s - 1)$ odds ratios is obtained (Figure 4.2).

Usually, they are modelled on a log odds ratio scale in the following way:

$$\eta_{i,ts,k\ell} = \ln(\mu_{i,ts,k\ell}) - \ln(\mu_{itk} - \mu_{i,ts,k\ell}) - \ln(\mu_{is\ell} - \mu_{i,ts,k\ell}) + \ln(1 - \mu_{itk} - \mu_{is\ell} + \mu_{i,ts,k\ell}).$$

Higher order global odds ratios are defined in a recursive manner. Indeed, if

$$\mu_{it|s}(z_s) = Pr(Z_{irtk_t} = 1 | Z_{irsk_s} = z_s, X_i, \boldsymbol{\theta}) \tag{4.5}$$

is the conditional probability of observing a success at occasion t, given the value z_s is observed at occasion s, and writing the corresponding conditional odds as

$$\psi_{it|s}(z_s) = \mu_{it|s}(z_s)/(1 - \mu_{it|s}(z_s)),$$

the pairwise marginal odds ratio, for occasions t and s, is defined as

$$\psi_{its} = \frac{\{\text{pr}(Z_{irtk_t} = 1, Z_{irsk_s} = 1)\}\{\text{pr}(Z_{irtk_t} = 0, Z_{irsk_s} = 0)\}}{\{\text{pr}(Z_{irtk_t} = 0, Z_{irsk_s} = 1)\}\{\text{pr}(Z_{irtk_t} = 1, Z_{irsk_s} = 0)\}} = \frac{\psi_{it|s}(1)}{\psi_{it|s}(0)},$$

in accordance with (4.4). Multi-way marginal global odds ratios can conveniently be defined in terms of ratios of conditional odds

$$\psi_{it_1 \ldots t_m t_{m+1}} = \frac{\psi_{it_1 \ldots t_m | t_{m+1}}(1)}{\psi_{it_1 \ldots t_m | t_{m+1}}(0)}, \tag{4.6}$$

where $\psi_{it_1 \ldots t_m | t_{m+1}}(z_{m+1})$ is defined by conditioning on e.g., $Z_{irt_{m+1}} = z_{t_{m+1}}$. Of course, they do retain a fully marginal interpretation.

Note that models for the odds ratios may include row-, column- and cell-specific terms, as well as covariate terms. For two biallelic markers M1 (t) and M2 (s), this would imply the following elaborate structure:

$$\eta_{i,ts,k\ell} = \gamma_{0,its} + \gamma_{1,itk} + \gamma_{2,is\ell} + \gamma_{3,its,k\ell} + \gamma_{4,its,k\ell} x_i, \qquad k, \ell = 1, 2,$$

with $\gamma_{0,its}$ an intercept term, and $\gamma_{1,itk}$, $\gamma_{2,is\ell}$ and $\gamma_{3,its,k\ell}$ respectively row-, column- and cell-specific parameters. The parameters $\gamma_{4,its,k\ell}$ reflect potential dependency of the odds ratios on covariates. Note that unicity constraints need to be imposed on the row, column and cell parameters; for instance $\gamma_{1,t1} = 0, \gamma_{2,s1} = 0, \gamma_{3,ts,k1} = 0$ and $\gamma_{3,ts,1\ell} = 0$.

Once the model specification is complete, the choice of estimation procedure needs careful reflection. Indeed, it is common knowledge that likelihood, quasi-likelihood, and GEE-based inferential methods in the analysis of correlated categorical responses tend to give dissimilar numerical results (Prentice 1988; Fitzmaurice

et al. 1993; Fahrmeir and Tutz 1994; Pendergast *et al.* 1996 and Diggle *et al.* 2002). However, with complete data these results should be in close agreement. Also in the presence of many markers a well-considered choice of estimation technique can substantially speed or facilitate convergence.

4.2.2 Estimation Techniques

One of the best-known and most popular approaches after specifying a multivariate probability model is to use maximum likelihood estimation techniques. However, knowing that there exist about 10 million SNPs in human populations, the dimensionality of genetic SNP studies may dramatically increase. It is therefore not surprising that in a full likelihood approach, the computational burden can be quite extensive. Hence, in high-dimensional problems alternative estimation techniques such as a generalized estimating equations approach (GEE) or pseudo-likelihood esimation may be more beneficial. For descriptions and properties of these alternative approaches, we refer to e.g., Liang and Zeger (1986); Zhao and Prentice (1990); Liang *et al.* (1992); Geys *et al.* (1997, 1999); Diggle *et al.* (2002).

Technically, if both multivariate outcomes and familial clustering are present, the pseudo-likelihood-driven methods are to be preferred over GEE2 (Geys *et al.* 1998). Whether or not this statement also holds in practical genetic settings should be investigated. Otherwise, in highly multivariate settings and provided third and higher order associations are not of interest, GEE2 is traditionally more flexible than a full likelihood approach. Generalized estimating equations can easily be derived, using a generalized linear modelling framework to describe the multivariate Dale model. Since in case-control studies, familial links between subjects are typically unknown or non-existing, we will restrict attention to GEEs to estimate the parameters of interest.

As specified above, the multivariate Dale model is designed to model multivariate ordinal data, such as biallelic markers for which the levels are determined by 0, 1 or 2 high-risk alleles. We agree that this limits its use. For instance, a similar approach cannot be adopted with multi-allelic (> 2) markers. However, current research is being carried out to use GEEs with nominal data, whilst mimicking the behaviour of the multivariate Dale model with multivariate ordered categorical outcomes.

4.2.3 Using Genetic Information as Response Data

Realising that in the logistic modelling framework, a pool of potential risk factors can be investigated in a single run, it is not surprising that the logistic model remains a popular tool among (genetic) epidemiologists. However, intuitively it makes more sense to apply multivariate analyses techniques while treating disease status as fixed and the history of risk factors (such as marker information) as random, as in case-control studies. There it *is* known in advance whether or not a given individual has developed the disease, since subjects are selected on the basis of their disease status (being a case or a control). Moreover, there are a few issues that deserve our immediate attention.

First and foremost, incorporating many exposure variables (e.g., SNPs) at the "covariate level", may introduce a problem of multicollinearity. If multicollinearity is present, then one should correct for it. Indeed, the fact that some predictor variables are correlated does not inhibit our ability to obtain a good fit, nor does it affect inferences about mean responses. The main problem lies in the fact that the estimated regression coefficients tend to vary widely from one sample to another when the predictor variables are highly correlated. In addition, many of the estimated regression coefficients individually may be statistically non-significant, whereas there appears to exist a definite statistical relation between the response variable and the set of predictor variables. Moreover, in the presence of highly correlated predictor variables, the simple interpretation of the regression coefficients as measuring marginal effects is often unwarranted. This problem is particularly apparent when searching for primary functional variants. In a classic logistic modelling framework this search can be carried out by constructing a model for the probability of disease, e.g., including only main effects at all variants under investigation. Quantifying the effects of a single locus is achieved by interpreting the corresponding regression coefficients, conditional on the fixed status at the remaining loci. The question remains how much value can be put on this interpretation if the single locus is involved in a complex multicollinearity pattern with other loci included in the model. One of the ways to correct for multicollinearity is then to implement ridge or principal component logistic regression. However, we note that these methods rely on severe assumptions such as the choice of ridge constant or the number of principal components to retain (Ryan 1997; Krzanowski 1988).

Second, most standard statistical software packages can fit regression models

such as logistic models. Despite their ease of use, they suffer from the drawback that *all* information on an individual is discarded whenever at least one predictor variable is missing. This may lead to severe power loss in genetic association studies, in which classically genetic markers are covariates and phenotype the response in a logistic-regression framework.

Third, although storing genetic information at the predictor level does allow investigating gene-gene interactions (epistasis), it is not so flexible for investigating the interplay between polymorphisms and its relation to the multiple facets of a complex disease. This is mainly due to the fact that a linear regression model as for case-control data is not designed to study relationships between predictor variables as such. The main purpose is to investigate the relationship between (combinations of) predictor variables to the probability of disease. If we aim for finding genetic determinants for a disease (e.g., Crohn's disease; Joossens *et al.* 2004), but at the same time suspect that different combinations of variants are associated to different behaviours of the disease (e.g., non-stricturing and non-penetrating, stricturing, penetrating), a multivariate approach (as outlined in Section 2) would be more appropriate. Here, genetic information is stored at the response level. Using a marginal model would still allow drawing conclusions for each response (variant) separately. At the same time, possible associations between the variants are accounted for and can be modelled in detail.

4.2.4 Application to real data: the ApoE region

The relationship between late-onset Alzheimer's disease (AD) and the ApoE ϵ_4 allele is widely recognised (e.g., Farrer *et al.* 1997). Interesting characteristics such as incomplete penetrance and the belief that the ϵ_4 allele is probably only one of several predisposing alleles for AD (Corder *et al.* 1993), has made data on late-onset Alzheimer's disease very popular in studying complex traits and in the process of developing multivariate statistical methodology.

4.3 Results

We implemented several multivariate Dale models to the ApoE data set (Chapter 3, Section 3.3), using the software package GAUSS (1997), as a means to properly account for or to explicitly model associations between markers, under the (for this data plausible) assumption of missing completely at random (MCAR). We first

restricted attention to the bivariate case and a single covariate effect (in particular, disease status). The reason for doing so is twofold: (i) the main purpose is to point out the potential merits of the model and not to develop "the best" descriptive model. (ii) the simplicity of this setting may facilitate attributing analysis features to genetic, biostatistical or numerical characteristics.

In particular, the measurement model in the bivariate case is written in terms of univariate marginal links, as in (4.3):

$$\text{Marker M1:} \begin{cases} \eta_{i11} &= \text{logit}(\mu_{i11}|x_{i1}) = \beta_{0\,i11} + \beta_{i11}x_{i1}, \\ \eta_{i12} &= \text{logit}(\mu_{i12}|x_{i1}) = \beta_{0\,i12} + \beta_{i12}x_{i1}, \end{cases}$$

$$\text{Marker M2:} \begin{cases} \eta_{i21} &= \text{logit}(\mu_{i21}|x_{i2}) = \beta_{0\,i21} + \beta_{i21}x_{i2}, \\ \eta_{i22} &= \text{logit}(\mu_{i22}|x_{i2}) = \beta_{0\,i22} + \beta_{i22}x_{i2}. \end{cases}$$

The possible outcomes for a marker (M1 say) are AA, Aa or aa. Whereas η_{11} relates the odds of having two copies of allele A to disease status, η_{12} links disease status to the odds of having at least one copy of A. For the ApoE data, 'A' represents the following alleles (using the notation of Fallin *et al.* 2001): C (for marker M1), G (M2), T (M3), C (M4), C (M5), G (M6), G (M7) and G (M8).
For the marginal global odds ratios $\psi_{i,ts,k\ell}$ (4.4), we investigate the following models, using a log-scale:

$$\begin{aligned}
(a) \quad & \eta_{i,ts,k\ell} = \gamma_{0,its}, \\
(b) \quad & \eta_{i,ts,k\ell} = \gamma_{0,its,k\ell}, \\
(c) \quad & \eta_{i,ts,k\ell} = \gamma_{0,its} + \gamma_{1,itk} + \gamma_{2,is\ell}, \\
(d) \quad & \eta_{i,ts,k\ell} = \gamma_{0,its} + \gamma_{4,its}\, x_i, \\
(e) \quad & \eta_{i,ts,k\ell} = \gamma_{0,its} + \gamma_{4,its,k\ell}\, x_i, \\
(f) \quad & \eta_{i,ts,k\ell} = \gamma_{0,its,k\ell} + \gamma_{4,its}\, x_i, \\
(g) \quad & \eta_{i,ts,k\ell} = \gamma_{0,its,k\ell} + \gamma_{4,its,k\ell}\, x_i, \\
(h) \quad & \eta_{i,ts,k\ell} = \gamma_{0,its} + \gamma_{1,itk} + \gamma_{2,is\ell} + \gamma_{4,its}\, x_i, \\
(i) \quad & \eta_{i,ts,k\ell} = \gamma_{0,its} + \gamma_{1,itk} + \gamma_{2,is\ell} + \gamma_{4,its,k\ell}\, x_i.
\end{aligned} \tag{4.7}$$

Cell-specific association terms are not considered, since often too little cell-information is available to draw valid conclusions from. The parameter $\gamma_{4,its,k\ell}$ in model (i) incorporates a disease adjusted intercept, row- and column- effect. In other words, $\gamma_{4,its,k\ell} = \gamma_{40,its} + \gamma_{41,itk} + \gamma_{42,is\ell}$ or

$$(i) \quad \eta_{i,ts,k\ell} = (\gamma_{0,its} + \gamma_{40,its}\, x_i) + (\gamma_{1,itk} + \gamma_{41,itk}\, x_i) + (\gamma_{2,is\ell} + \gamma_{42,is\ell}\, x_i).$$

These models can be further reduced by assuming proportional odds or imposing $\beta_{11} = \beta_{12} = \beta_{i1}$ and $\beta_{21} = \beta_{22} = \beta_2$:

$$\text{Marker M1:} \begin{cases} \eta_{i11} &= \text{logit}(\mu_{i11}|x_{i1}) = \beta_{0\,i11} + \beta_{i1} x_{i1}, \\ \eta_{i12} &= \text{logit}(\mu_{i12}|x_{i1}) = \beta_{0\,i12} + \beta_{i1} x_{i1}, \end{cases}$$

$$\text{Marker M2:} \begin{cases} \eta_{i21} &= \text{logit}(\mu_{i21}|x_{i2}) = \beta_{0\,i21} + \beta_{i2} x_{i2}, \\ \eta_{i22} &= \text{logit}(\mu_{i22}|x_{i2}) = \beta_{0\,i22} + \beta_{i2} x_{i2}, \end{cases}$$

with the simple association models $\eta_{i,ts,k\ell} = \gamma_{0,its} + \gamma_{4,its}\, x_i$ or $\eta_{i,ts,k\ell} = \gamma_{0,its}$.

Provided evidence is found for proportional odds, this model is useful to derive a single disease effect quantification per marker, while correcting for the presence of other markers.

In addition, we implemented all possible trivariate combinations of markers, using a single disease quantification per marker, while correcting for the presence of other markers:

$$\text{Marker M1:} \begin{cases} \eta_{i11} &= \text{logit}(\mu_{i11}|x_{i1}) = \beta_{0\,i11} + \beta_{i1} x_{i1}, \\ \eta_{i12} &= \text{logit}(\mu_{i12}|x_{i1}) = \beta_{0\,i12} + \beta_{i1} x_{i1}, \end{cases}$$

$$\text{Marker M2:} \begin{cases} \eta_{i21} &= \text{logit}(\mu_{i21}|x_{i2}) = \beta_{0\,i21} + \beta_{i2} x_{i2}, \\ \eta_{i22} &= \text{logit}(\mu_{i22}|x_{i2}) = \beta_{0\,i22} + \beta_{i2} x_{i2}, \end{cases}$$

$$\text{Marker M3:} \begin{cases} \eta_{i31} &= \text{logit}(\mu_{i31}|x_{i2}) = \beta_{0\,i31} + \beta_{i3} x_{i3}, \\ \eta_{i32} &= \text{logit}(\mu_{i32}|x_{i2}) = \beta_{0\,i32} + \beta_{i3} x_{i3}. \end{cases} \tag{4.8}$$

The pairwise and third order association structures were kept simple:

$$\begin{aligned} \eta_{i,ts,k\ell} &= \gamma_{0,its} + \gamma_{4,its}\, x_i, \\ \eta_{i,123} &= \gamma_{0,i123} + \gamma_{4,i123}\, x_i. \end{aligned} \tag{4.9}$$

4.3.1 All Possible Pairs of Markers

We first restricted attention to the marginal bivariate models (4.3) with association structure specified via expressions (a), (b) and (c) as outlined in the model formulations (4.7). Hence, we were able to quantify differences in the odds of disease when having 2 copies of an allele A versus one copy or none, or when having at least one copy of an allele A versus no copy at all. In general, if in the association model (a) the

parameter $\gamma_{0,ts}$ is statistically significantly different from 0 (i.e., the corresponding odds ratio is statistically significantly different from 1), then significant association parameters are observed in models (b) or (c). This is the case for the combination of markers M1-M2, M3-M4, M3-M5, M3-M6, M4-M5, M4-M6, M5-M7 and M7-M8. We refer to Table 4.1 for a selection of results. The pair M5-M8 is exceptional in this sense. Although a significant association is observed for the marker pair M1-M5, only marginal non-significant results are obtained for models (b) and (c). Note that if significant parameters are observed in models (b) and/or (c), then this does not necessarily lead to a significant overall parameter $\gamma_{0,ts}$ in model (a). This is not surprising, since moving from the more detailed association specifications (b) and (c) towards (a) can be seen as a smoothing process, wiping out sporadic significant associations $\psi_{11}, \psi_{12}, \psi_{21}$ and ψ_{22} by imposing $\psi_{11} = \psi_{12} = \psi_{21} = \psi_{22}$. Examples are given by the marker combinations M2-M8, M3-M7, M3-M8, M4-M7 and M4-M8 (results not shown). Note that the significant parameter $\gamma_{0,ts}$ in model (b) for M3-M8, refers to an odds ratio ψ_{12}, significantly different from 1. The significant parameter $\gamma_{1,t2}$ in model (c) for M3-M8 refers to a significant row effect or a significant difference between the odds ratios ψ_{11} and ψ_{21} (Figure 4.2).

Of main interest is the question whether these marker associations differ between cases and controls. To answer this question, we need to consider models such as models (d) to (i). Significant disease effects are observed for marker combinations M1-M2, M1-M3, M1-M4, M1-M6, M1-M7, M3-M4, M3-M7, M3-M8, M4-M7 and M4-M8. As before, if the association model (d) (respectively (f)) shows evidence for a significant effect of disease status, then significant disease effect parameters are observed in model (e) (respectively (g)). The reverse is not necessarily true (e.g., model (g) and marker pair M3-M4 in Table 4.1). Significant disease effects for either model (d) or (f) are only observed for marker combinations M1-M2, M3-M7, M3-M8, M4-M7 and M4-M8. If row- and column-effects are introduced in the association model (models (h) and (i)) then significant differences between cases and controls are found in the pairs M1-M2, M1-M6, M1-M7, M3-M7 and M3-M8, for particular association features. Performing haplotype analyses on all possible pairs of two markers, M1-M2 and all the combinations involving M3 and/or M4 polymorphisms showed significant omnibus likelihood ratio test results (at the 5% significance level). The multivariate Dale model identified the same set of combinations and also M1-M6 and M1-M7 combinations which harbour M4 responsible for epsilon 4 allele and M3, a neighbouring SNP.

Table 4.1: Selection of analysis results (p-values) after fitting the bivariate Dale model (4.3) under a variety of association models (4.7). The adopted estimation technique is GEE2.

	Association Models								
	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
	$\gamma_{0,its}$	$\gamma_{0,its,11}$	$\gamma_{0,its}$	$\gamma_{4,its}$	$\gamma_{4,i11}$	$\gamma_{4,its}$	$\gamma_{4,i11}$	$\gamma_{4,its}$	$\gamma_{4,its}$
		$\gamma_{0,its,12}$	$\gamma_{1,it2}$		$\gamma_{4,i12}$		$\gamma_{4,i12}$		$\gamma_{4,it2}$
		$\gamma_{0,its,21}$	$\gamma_{2,is2}$		$\gamma_{4,i21}$		$\gamma_{4,i21}$		$\gamma_{4,is2}$
		$\gamma_{0,its,22}$			$\gamma_{4,i22}$		$\gamma_{4,i22}$		
Pair t-s									
M1-M2	0.000	0.001	0.000	0.000	0.320	0.000	0.119	0.000	0.165
		0.000	0.036		0.002		0.006		0.869
		0.000	0.069		0.001		0.002		0.925
		0.000			0.062		0.004		
M1-M3	0.321	0.972	0.644	0.529	0.643	0.481	0.701	0.483	0.887
		0.680	0.452		0.489		0.766		0.570
		0.223	0.780		0.958		0.000		0.968
		0.287			0.587		0.387		
M1-M4	0.898	0.832	0.482	0.382	0.755	0.297	0.000	0.290	0.231
		0.962	0.897		0.332		0.398		0.899
		0.296	0.357		0.701		0.315		0.104
		0.857			0.252		0.264		
M2-M3	0.764	0.942	0.694	0.529	0.802	0.549	0.979	0.550	0.932
		0.808	0.988		0.611		0.746		0.838
		0.584	0.746		0.541		0.983		0.731
		0.878			0.551		0.506		
M2-M4	0.558	0.568	0.976	0.313	0.760	0.288	0.264	0.281	0.255
		0.505	0.906		0.191		0.249		0.681
		0.575	0.755		0.567		0.324		0.029
		0.613			0.293		0.355		
M3-M4	0.000	0.000	0.000	0.220	0.601	0.219	0.516	0.289	0.418
		0.000	0.026		0.303		0.991		0.501
		0.000	0.018		0.003		0.831		0.828
		0.000			0.146		0.085		

As for the marginal response descriptions, only markers M3 and M4 seem to be significantly associated with disease. This is fully in line with expectation. Indeed, marker M4 is a SNP responsible for the $\epsilon 4$ allele, whereas marker M3 is a neighbouring SNP. These two loci have been shown to have alleles in strong linkage disequilibrium (Fallin *et al.* 2001).
Parameter interpretation in the response functions for the pair M3-M4 is as follows (Table 4.2). The odds of disease is increased with a factor $\exp(\beta_{11}) = \exp(1.378)$=3.967 when the individual has two copies of allele A (this is the allele increased among cases) for marker M3 rather than 1 or no copy of allele A. The odds of disease is increased with a factor $\exp(\beta_{12}) = \exp(1.129)$= 3.093 when the individual has at least one copy of A for M3 rather than no copy at all.

In addition, we note that significant links between markers M3 or M4 and disease carry through, no matter the fitted model for the association structure. In fact, in general the β parameter estimates remain fairly stable from model (a) to model (i). This feature is quite attractive and has a simple explanation. Indeed, in the bivariate Dale model, the estimated marginal (β) and association (γ) parameters are orthogonal (Palmgren 1989). As can be seen from Table 4.2 and M3 in M3-M4, most of the fluctuation is observed in the parameter estimate for β_{11}, as it ranges from 1.053 for model (e) to 1.378 for model (a). This leads to an increase in odds of disease for an individual having two copies of allele A versus 1 or 0, by a factor ranging from 2.866 to 3.967.

The property of the bivariate Dale model mentioned before also makes the β parameter estimates for a particular choice of marker (e.g., M3) and model (e.g., model (a)) comparable over all possible marker pair extensions.

Up to now, the marginal models allowed for varying disease effects on the two logits for one particular marker. In other words, we let both parameters β_{11} and β_{12} for a particular marker vary independently from each other. Testing the null hypothesis of proportional odds ($\beta_{11} = \beta_{12}$ and $\beta_{21} = \beta_{22}$) in the bivariate Dale model (4.3) with a single association parameter, using a Wald type statistic and a robust estimate for variance covariance matrices (in particular, a sandwich estimate), yields test values ranging from 0.106 (marker combination M2-M8) to 2.999 (marker combination M3-M4). Hence, based on a chi-square distribution with 2 degrees of freedom and a significance level of 0.05, no evidence was found to reject proportional odds, for none of the pairwise marker combinations.

Table 4.2: GEE2 parameter estimates for disease effects β_{11} and β_{12} (robust standard errors) on marker M3, for a selection of marginal response and association models.

Assoc. Model	Marker Pair	Estimate (Std. Err.) β_{11}	Estimate (Std. Err.) β_{12}
(a)	M3-M1	1.406 (0.507)	1.085 (0.239)
	M3-M2	1.239 (0.472)	1.043 (0.238)
	M3-M4	1.378 (0.485)	1.129 (0.238)
	M3-M5	1.216 (0.472)	1.101 (0.239)
	M3-M6	1.129 (0.454)	1.058 (0.240)
	M3-M7	1.159 (0.471)	1.074 (0.239)
	M3-M8	1.230 (0.474)	1.095 (0.242)
(b)	M3-M4	1.186 (0.439)	1.083 (0.232)
(c)	M3-M4	1.185 (0.437)	1.089 (0.233)
(d)	M3-M4	1.123 (0.423)	1.111 (0.233)
(e)	M3-M4	1.053 (0.427)	1.091 (0.233)
(f)	M3-M4	1.161 (0.458)	1.093 (0.232)
(g)	M3-M4	1.070 (0.432)	1.072 (0.233)
(h)	M3-M4	1.147 (0.439)	1.095 (0.233)
(i)	M3-M4	1.059 (0.432)	1.071 (0.233)

Bearing Palmgren's (1989) result in mind, the particular specification of the association structure does not substantially influence the parameter estimation of the β parameters. For illustrative purposes we completed the model formulation by specifying $\eta_{i,ts,k\ell} = \gamma_{0,its}$, hereby allowing for a single association parameter.

Note that in principle a whole variety of association models can be implemented, such as models (b) to (i) before. Testing (via score-type test statistics based on a chi-squared distribution with 1 degree of freedom) whether the association depends on disease status (this is equivalent with $\gamma_{4,ts} = 0$), yields significant results for marker combinations M1-M2 (χ^2=31.711), M3-M4 (χ^2=4.128), M3-M7 (χ^2= 9.950), M3-M8 (χ^2=7.413), M4-M7 (χ^2=4.231), M4-M8 (χ^2=5.578) and M7-M8 (χ^2=4.896). This coincides with the results from Table 4.1, model (d), with the exception of marker combinations M3-M4 and M7-M8. Table 4.3 shows a selection of the analysis results. Considering marker M4 in the pair M3-M4, the odds of disease have increased with a factor exp(β_{21})=exp(1.395)=4.035 when the individual has two copies of allele A for M3 rather than 1 or 0 copies. The same factor applies when comparing individuals having at least one copy of allele A versus those having no copy.

4.3.2 All Possible 3-Tuples of Markers

Based on the previous results, we consider all possible 3-tuples of markers and fit the marginal response model as specified in (4.8) and a reduced model for the associations by specifying $\eta_{i,ts,kl} = \gamma_{0its}$ and $\gamma_{i,123} = \gamma_{0,i123}$ in each of these cases. Using a score-type test it is possible to test whether the association structure depends on disease status, without fitting the more complex model.

Based on a chi-squared distribution with 4 degrees of freedom, significance at the 0.05 level was attained for all combinations including the pair M1-M2: M1-M2-M3 (χ^2=36.113), M1-M2-M4 (χ^2=35.500), M1-M2-M5 (χ^2=32.332), M1-M2-M6 (χ^2=34.134), M1-M2-M7 (χ^2=41.534), M1-M2-M8 (χ^2=33.400). In addition, significance was attained for the triples M1-M3-M7 (χ^2=11.601), M1-M3-M8 (χ^2=9.459), M1-M6-M7 (χ^2=14.786), M1-M6-M8
(χ^2=16.081), M2-M3-M7 (χ^2=12.415), M2-M3-M8 (χ^2=10.125), M2-M4-M7 (χ^2=9.352, M2-M4-M8 (χ^2=9.352), M3-M4-M7 (χ^2=12.134), M3-M4-M8 (χ^2=11.337), M3-M5-M7 (χ^2=9.690), M3-M6-M7 (χ^2=18.361), M3-M6-M8 (χ^2=11.109), M3-M7-M8 (χ^2=35.996), M4-M6-M8 (χ^2=9.358), M4-M7-M8 (χ^2=11.141).

Table 4.3: Maximum likelihood parameter estimates in a bivariate Dale model (assuming proportional odds and a simple association structure) for the marker combinations M1-M2 and M3-M4.

Parameter	Estimate (Std. Err.)	p-value
Markers M1 and M2:		
$\beta 011$	-0.964 (0.168)	0.000
$\beta 012$	1.067 (0.168)	0.000
$\beta 11 = \beta 12$	-0.108 (0.200)	0.590
$\beta 021$	-1.609 (0.185)	0.000
$\beta 022$	0.421 (0.159)	0.008
$\beta 21 = \beta 22$	0.072 (0.200)	0.720
$\gamma_{0,ts}$	-3.235 (0.248)	0.000
Markers M3 and M4:		
$\beta 011$	-3.089 (0.258)	0.000
$\beta 012$	-1.139 (0.187)	0.000
$\beta 11 = \beta 12$	1.147 (0.235)	0.000
$\beta 021$	-3.344 (0.267)	0.000
$\beta 022$	-1.379 (0.198)	0.000
$\beta 21 = \beta 22$	1.395 (0.243)	0.000
$\gamma_{0,ts}$	5.383 (0.395)	0.000

Fitting the more elaborate association model (4.9) only for these combinations shows that a significant effect of disease status on the association structure between M1 and M2 is responsible for rejecting the null hypothesis $\gamma_{4,i11} = \gamma_{4,i12} = \gamma_{4,i21} = \gamma_{4,i22} = 0$ and $\gamma_{4,123} = 0$ in all cases involving M1-M2. This is in agreement with the highly significant disease effect on the association between M1 and M2, observed in Table 4.1, via model (d). In all other combinations including M3 and/or M4, the significant disease effect on the association between M3 and M7 or M8, and between M4 and M7 or M8, seems to be the determining factor for rejecting the aforementioned null hypothesis. Only for the combinations M1-M6-M7 and M1-M6-M8 we detect a dependence on disease status of the third order association. For the latter triples, disease dependence was also seen for the association between M1 and M6. This was only marginally picked up in the bivariate Dale model using association model (d), with a parameter estimate for $\gamma_{4,its}$ of -0.745 and corresponding standard error 0.387. Parameter interpretations naturally extend to those in the bivariate case.

Performing haplotype analyses on all possible combinations of 3 markers revealed that significance was met for all the combinations containing M1-M2, M3 or M4. The Dale model identified additionally the combinations M1-M6-M7 and M1-M6-M8, harbouring both M4 and M3 (results not shown).

4.4 Discussion

Methods that allow considering several loci simultaneously are particularly attractive since (i) for a given set of polymorphisms within a small genomic region or a candidate gene, the combination of their genotypic information (Corbex *et al.* 2000) or haplotypes (Drysdale *et al.* 2000) can reveal undetected effects by single locus tests; (ii) they provide a profound basis for studying epistasis, one of the most fundamental aspects of genetics (Templeton 2000). Complementary to logistic regression or haplotype-based approaches, multivariate models can be implemented. One such model is the multivariate Dale model, which encompasses a whole family of parametric models, by the choice of different link functions for the margins and/or associations. Model building is facilitated via Wald type or (pseudo-)likelihood driven tests. Estimates of the marker associations are readily available. Associations (expressed by means of odds ratios) between markers can be linked to relevant covariate information apart from disease status. Row-,

column-, and cell-specific terms can be included in a flexible way. Straightforward interpretations for the marginal effects remain, especially when using logistic links. The obtained marginal estimators are "corrected" for possible inter-relationships between the markers.

Applied to a real data set, we showed that the model can recover both results found by others, through haplotype analyses, as well as effects not previously described. The performed analyses showed that a multivariate model for marker data is indeed a handy tool to understand relationships between markers. These relations may depend on covariates of interest, in particular disease status. In that case, the models enable to investigate whether or not there is a different interplay of genetic markers between groups of people with varying disease outcome. Ideally, the model is used in conjunction with other analysis techniques, such as haplotype-based approaches or logistic regression analyses. Since they allow moving through the solution space from different angles, they should be seen as complementary techniques rather than competing ones.

Allelic heterogeneity is allowed, but may require taking additional measures. Indeed, if multiple high-allelic markers are involved, the contingency tables (4.1) are likely to suffer from many sparse cells. One possible solution is to double the data, to augment each zero cell with 1, and to use the correct final standard errors (Agresti 1990). Another solution is to carefully select the estimation procedure. For instance, a GEE2 approach is to be preferred above a likelihood approach when elaborate models (in terms of number of parameters) are in focus. When too many markers are involved, it may be better to split up the full likelihood in several pieces and to patch it up again using pseudo-likelihood methodology.

Neither recessive nor dominant effects can be tested for in a multivariate Dale model, since the latter does not support the analysis of nominal categorical data. This limitation disappears if the model is modified to cover marginal descriptions via generalized logits (instead of cumulative logits in the Dale model). In that setting and assuming a single indicator variable x_i referring to disease status (1=case, 0=control), the first logit (η_{i11}) as in (4.3) would relate disease status to the odds of having genotype AA versus aa, whereas the second logit (η_{i12}) would relate disease status to the odds of having Aa versus aa. The assumption of no dominance is then equivalent to assuming that $\beta_{11} = 2\beta_{12}$. The implementation of this extension will be the subject of future research.

The multivariate Dale model does allow for another type of testing: the proportional odds assumption. In practice, for a biallelic marker M1 as in (4.3), the proportional odds assumption ($\beta_{11} = 2\beta_{12}$) implies that the functional relationship between the odds of having at least one disease allele and the odds of having 2 copies of the disease allele is independent from disease status.

Most standard statistical software packages can fit logistic models. However, when information is missing on one or more covariate measurements for an individual, the individual is most often excluded from the analysis. If we want to take the possibility into account that the missing data mechanism on the collected markers cannot be ignored (Little and Rubin 1987), then this mechanism should be accounted for and/or modelled explicitly. This can be achieved by extending the measurement model (specified as a multivariate Dale model) with a (logistic) model for the missingness process. In this way, the measurement model and the missingness process can be modelled jointly. A sensitivity analysis to assess the impact on conclusions under different assumptions of the missingness process in "responses" is the topic of ongoing research.

Environmental factors as well as interactions of genes and environmental factors can easily be accounted for in the proposed modelling strategy by including them at the response level as well. This increases the dimensionality of the multivariate model. However, the number of association parameters can be substantially reduced by assuming that all associations between an environmental factor and a gene are of the same magnitude. This coincides with the assumption of having a single summarising global association parameter between an environmental factor and a gene. By disconnecting disease status and environmental information in the model (this is: the corresponding variables reside at different sides of the equation sign) it is straightforward to assess the marginal effect of environment on disease as well. Whereas interacting or correlated genes and environmental factors are most intuitively accounted for in multivariate models (e.g., avoiding multicollinearity problems), classical univariate regression models might be more beneficial to describe additive gene-environment effects. Note that the Dale model allows easy assessing of marginal effects. It is less straightforward to assess the additive contribution to disease risk.

Furthermore, we point out that the multivariate Dale model is applicable to

more than three markers and combinations thereof. Practically, it is obvious that as dimensionality increases the number of possible association terms and therefore the number of parameters to estimate increases also. However, Clayton and Jones (1999) argue that higher-order associations fall away at progressively more rapid rates than first-order associations. Moreover, instead of using the full likelihood to obtain parameter estimates, alternative estimation procedures can be considered. The potential of pseudo-likelihood estimation is under investigation and may be the road to travel by, especially in the presence of many markers. Alternatively, first a data dimensionality reduction procedure is applied (e.g., refer to Ritchie *et al.* 2001), followed by a detailed investigation of the nature of the (limited number of) associations using the Dale model.

The question remains what the connection is between biostatistical findings and genetic implications. The analyst has the choice of fitting a broad range of models with respect to the association structure, ranging from a constant to an elaborate model accounting for row-, column-, and cell-specific effects, as well as cell-dependent covariate dependence. Each of these model formulations make it possible to describe a different aspect of the interplay of genetic markers in their relation to disease. The potential use of the multivariate Dale model to pick up markers with small single effects but a strong joint effect will be the subject of further investigation, through the use of extensive simulations and other data sets. Extensions of the model involve (i) accounting for familial relationships and (ii) replacing cumulative logits by generalized logits.
Via the first extension, investigating family-specific relationships between genetic information and facets of a disease becomes within reach. Note that the classical logistic regression methods for case-control data in genomic association studies are indeed easily extended to nuclear family data using conditional rather than unconditional logistic regression models. Analysing family data perfectly fits in a multivariate approach in which the clustering in the data is explicitly acknowledged. Typically, the simplifying assumption of exchangeability among family members is made. This implies that the relation between the responses of any pair of members of the same family can be assumed to be the same. In practice, the response vector in the multivariate Dale approach consists of all markers of all members within the same family. In the simplest setting, it suffices to introduce one additional association parameter that describes the relation between measurements of the same marker within a family. A variable family size may add an additional level of technical complexity to the problem.

The second extension regards the levels of a marker as nominal and clears the way for applying the model to multi-allelic markers.

The authors have developed GAUSS code which is available upon request.

Chapter 5

Genomic Screening in Family-based Association Testing for Quantitative Traits: Validation and Replication Using the same Data Set to correct for Multiple Testing

5.1 Introduction

The Human Genome Project and its spin-offs such as the Allele Frequency/Genotype Project or the HapMap Project are making it increasingly feasible to disentangle the genetic basis of a given complex trait using genome-wide association studies. The statistical challenge in analysing such genome-wide association studies stems from the severe multiple-comparison problem resulting from the analysis of thousands of SNPs. Standard multiple comparison methods are not likely to be successful in finding associations that achieve genome-wide significance. Our proposed method-

ology for family-based association studies successfully deals with this issue in the context of genome-wide association screening, using single SNPs or haplotypes. In relation to developing guidelines for our screening tools, we provide lower bounds for the estimated power to detect the gene harbouring the disease susceptibility locus, which hold regardless of the LD structure present in the data. We assess the power of our approach in the presence of multiple disease susceptibility loci. The proposed screening tools accommodate genomic control and impact the concept of haplotype tagging SNPs. Finally, our methods use the entire sample and do not require separate screening and validation samples to establish genome-wide significance, as population-based designs do.

In humans, single nucleotide polymorphisms are the most common type of genetic variation; eight million SNPs have already been documented and deposited in the dbSNP database (e.g., Sherry *et al.* 1999, International SNP Map Working Group 2001). Their dense distribution across the genome, on average about every 200 base pairs, and their low mutation rate makes them ideal markers for large-scale genome-wide association studies to discover genes in common complex diseases, such as cancer, diabetes or vascular disease (Marnellos 2003). In addition, the recent advances in bioinformatics and array technologies (Chee *et al.* 1996; Wang *et al.* 1998) have made it possible to genotype biological samples for thousands of SNPs.

The success of genome-wide association studies will depend upon whether the increase in numbers of SNPs can be translated into an increase in the overall statistical power, or whether the positive effects of the increase in the number of SNPs are diluted by the multiple-comparison problem. When thousands of SNPs are tested for association with disease-related phenotypes, the p-value needs to be adjusted for the number of tests computed, so as to control type I error rates. Type I error rates include the family-wise error rate and the false discovery rate. Multiple testing procedures such as those proposed by Bonferroni (1936), Holm (1979), Hochberg and Tamhane (1987), Hochberg (1988) and Westfall and Young (1993) adjust p-values to control the family-wise error rate. They often generate unrealistically small significance levels for individual tests, in part because the dependence between test statistics is ignored. Alternative multiple testing approaches control the false discovery rate (e.g., Benjamini and Hochberg 1995; Yekutieli and Benjamini 1999; Benjamini and Yekutieli 2001; Storey and Tibshirani 2003). However, most procedures which aim to control the false positive rate become more conservative as more tests are performed.

Ideally, a SNP data reduction technique would be applied first, so that the number of association tests is diminished and hence the correction for multiple testing is less severe. In order not to bias the test results, the data used to reduce the number of SNPs should differ from the data used for testing. For family data there is a way to create two sources of information using one sample. The basic idea proposed by Lange *et al.* (2003a,b) is to estimate a genetic effect via a regression model that is statistically independent of the family-based analysis, using data from all families. The genetic effect estimate for each SNP is used in a Wald test of no genetic effect, or in power calculations for a family-based test, to screen and select SNPs for association; either the Wald test or power calculations are used to screen SNPs for final testing. The association testing on a much smaller set of SNPs uses family-based tests (FBATs), which are robust to population admixture and/or stratification.

The screening approach of Lange *et al.* (2003a,b) was introduced in the context of testing several sets of related phenotypes. The emphasis of this chapter is the development of new strategies for genomic screening. We derive lower bounds for the estimated power of the screening method to detect a gene harbouring the disease susceptibility locus (DSL), which hold even in the presence of high linkage disequilibrium (LD). In addition, we show that population stratification and admixture have minimal effect on power. The potential of the approach in the context of genomic screening are further illustrated by simulations and application to an asthma study, using the software package PBAT (Lange *et al.* 2004a; Van Steen and Lange 2005).

5.2 Methods

5.2.1 New Tool for Genome-wide Association Screening

The tools for genomic association screening, which we will describe in this section, are implemented in the PBAT software (Lange *et al.* 2004a) and use the unified approach to family-based tests of association (FBAT), introduced by Rabinowitz and Laird (2000) and Laird *et al.* (2000). FBAT builds on the original TDT method (Spielman *et al.* 1993) in which alleles transmitted to affected offspring are compared with the expected distribution of alleles among offspring, and has been generalised to accommodate quantitative phenotypes, missing parental information, use of different genetic models, etc.

In particular, the FBAT statistic is based on a linear combination of offspring genotypes and traits:

$$FBAT = (S - E[S])/\sqrt{V},\ S = \sum_{ij} T_{ij} * X_{ij}, \tag{5.1}$$

where $V = \mathrm{Var}(S)$ and T_{ij} represents the coded phenotype (i.e., the phenotype adjusted for any covariates) of the j-th offspring in family i. The X_{ij} denote the offspring's coded genotype at the locus being tested. It depends on the genetic model under consideration. The FBAT statistic has an approximate standard normal distribution; the null hypothesis being tested is "no linkage and no association". Extension of (6.1) to multiple traits is straightforward (Lange *et al.* 2003c), in which case the distribution of the multivariate FBAT statistic can be approximated by a chi-square distribution (df: rank of V). For studies with quantitative traits that are measured repeatedly, generalized principal component analysis can be used to derive an overall phenotype that maximises the proportion of phenotypic variance explained by the marker. The newly defined trait is used in a univariate FBAT statistic, hereafter referred to as FBAT-PC (Lange *et al.* 2004b).

The screening strategy consists of two steps. The first step is a data reduction technique to select the most promising trait-marker combinations. It involves repeating four components of an algorithm. In particular, the four components of step 1 are:

1. Specify a plausible linear regression model that functionally relates the phenotype(s) of interest to genotypic information. The coding of the marker genotypes is a reflection of the underlying disease model. In linking trait(s) to coded genotype, different selections of covariates with the test locus can be considered as well. For example, the regression model to link an offspring's phenotype Y_{ij} to its genotype X_{ij} and a covariate vector Z_{ij} is

$$\mathrm{E}(Y_{ij}) = a \star X_{ij} + b \star Z_{ij}. \tag{5.2}$$

2. Replace the observed offspring genotypes by their conditional mean given the parental genotypes at the marker (i.e., the between-family component). When the parental genotypes are observed, the conditional mean is computed based on the genotypes (i.e., X_{ij} is replaced by $E[X_{ij}|\text{parental genotypes}]$). When parental genotypes are incomplete, the conditional mean is computed based on the sufficient statistic (Rabinowitz and Laird 2000) (i.e., X_{ij} is replaced by $E[X_{ij}|\text{sufficient statistic}]$). Hence, instead of the regression model (5.2) we use

the conditional mean model

$$\mathrm{E}(Y_{ij}) = a \star E[X_{ij}|\text{parental genotypes}] + b \star Z_{ij}. \tag{5.3}$$

For double homozygous parents,

$$X_{ij} = E[X_{ij}|\text{parental genotypes}].$$

3. Estimate the genetic effect size a in the conditional mean model (5.3) using ordinary least squares estimation ($\hat{a}$).

4. Use one of the two methods to evaluate each combination of trait, marker, covariates and genetic model.

   (i) Method I: Compute the conditional power of the FBAT statistic, given the observed data (Lange 2003a):

   The power of the test statistic is computed, conditional on the offspring's phenotypes and the parental genotypes or the minimal sufficient statistics when parental genotypes are missing. The conditional power depends on the estimated genetic effect $\hat{a}$ from the conditional mean model (5.3).

   (ii) Method II: Calculate the Wald test statistic for the genetic effect in the conditional mean model (Lange 2003b):

$$\frac{\hat{a}^2}{\mathrm{Var}(\hat{a})} \sim \chi^2(1).$$

   Trait-marker combinations can be retained in a variety of ways, according to different criteria. For example, the criterion to retain combinations may be a specific cutoff value for the conditional power of the FBAT statistic (e.g., 80%), or it may be based on a preset number of smallest p-values for the Wald test (e.g., 5 smallest).

The second step in the screening process involves applying the FBAT statistic on the selected combinations of phenotypes and markers. Although population admixture and stratification may bias the estimate of a and thus will affect the power of the proposed testing strategy, step 2 of the screening technique avoids confounding due to model misspecification as well as admixture or population stratification: The final decision on potential marker associations is based on the FBAT test statistic, which guards against these confounding factors. Note that the

null hypothesis being tested is "no association and no linkage between *any* SNP and a disease susceptibility locus".

From a theoretical point of view, screening based on conditional power calculations is preferred, since conditional power calculations are a natural yardstick using both the genetic effect size estimates and the number of informative families in the FBAT statistic, whereas screening based on an overall Wald test does not account for the available number of informative families.

In this chapter, the data reduction in step 1 is achieved by selecting the top K trait-marker combinations for subsequent FBAT testing. Whereas the top K trait/marker combinations in screening method I refer to trait-marker associations with the highest power for the corresponding FBAT test, in screening method II they refer to trait-marker pairs for which the lowest Wald p-value in the conditional mean model is obtained. Guidelines for choosing the optimal value for K are discussed in the next section.

5.3 Results

5.3.1 Simulation Studies

Power

We conducted a series of simulation studies to assess the power of the proposed testing strategies. To account for linkage disequilibrium (LD) between SNPs we used actual genetic data. Genetic data were extracted from Childhood Asthma Management Programme (CAMP) Genetics Ancillary Study (CAMP 1999); 651 trios were ascertained for mild to moderate asthma. We used genotype data on 291 SNPs in selected candidate genes. The phenotypic data will be discussed in the Data Analysis section.

To assess power we selected the interleukin gene IL10 on chromosome 1 as the DSL. This gene is known to be associated with an asthma-related quantitative trait in this data set (Lyon *et al.* 2004). We selected each of the six typed SNPs in IL10 as the causal SNP and, for each offspring, simulated a trait value Y_{ij} from the normal

distribution with unit variance

$$Y_{ij} \sim N(a * X_{ij}, 1), \tag{5.4}$$

where a denotes the genetic effect size and X_{ij} the observed marker score for the selected SNP in IL10, for the jth offspring's in family i. We specified the genetic effect in terms of the heritability h according to

$$h = \mathrm{Var}(a * X_{ij})/\mathrm{Var}(Y_{ij})$$

as in Lange and Laird (2002b). Here, heritability is taken to mean the proportion of phenotypic variance that is explained by the analysed marker. We selected the most promising SNP from the entire pool, in terms of highest conditional power (method I) or smallest p-value for the Wald test (method II), and tested it for association using the FBAT statistic (Laird *et al.* 2000).

We estimated power levels as the proportion of trials successfully identifying the alleged causal mutation in IL10 (i.e., the SNP in IL10 is selected by the screening technique and the associated FBAT-statistic is significant at an α-level of 5%). In general, screening 291 SNPs using method I based on conditional power gives rise to estimated power levels of at least 60%, and power more than 80% with heritability values $h \geq 0.07$ (Table 5.1). Similar results are obtained with screening method II, based on the overall Wald test for no genetic effects, and are also reported in Table 5.1. Screening method II seems to outperform method I. Our methods did not perform well for SNP5 in IL10, which is a rare SNP (estimated minor allele frequency of 0.055) and is not highly correlated with the other typed SNPs in IL10.

To assess the power of our screening technique with thousands of markers, we conducted a simulation study based on genetic data from an Affymetrix GeneChip Mapping 10K Array on prostate cancer. Blood samples were collected from 467 subjects in 167 families, and SNPs were typed using the Affymetrix early access 10,000 SNP mapping array (Kennedy *et al.* 2003). Data on a genome linkage screen were reported in Schaid *et al.* (2004). We selected an LD block of four SNPs as the region carrying the disease susceptibility locus in our simulations. The SNPs in this block were among the few SNPs that were in one gene and in LD with each other. Each of the 4 SNPs was alternatingly chosen as the alleged causal mutation and traits were generated again according to the normal distribution $Y_{ij} \sim N(a * X_{ij}, 1)$, where X_{ij} is the observed marker score for the selected SNP in the 4-block for the jth offspring in family i. Power levels were estimated as the proportion of

trials successfully identifying one of the four SNPs in the selected LD block. We observed a loss in power when increasing the number of SNPs to 10,000 (Table 5.1; size m=10,000). Yet, power levels were still acceptable for the larger heritability values ($h > 0.05$). In contrast to the smaller CAMP data set, screening method I now seems to be more powerful than method II for lower heritability values (≤ 0.05). However, two remarks are in order: (i) The foregoing simulations always assume one gene, with the functional variant being on an observed SNP. Thus the causal variant is always being tested. (ii) Only the top trait-marker combination (for either screening method) is retained for further FBAT testing. In principle, more trait-marker pairs can be pushed forward to the second level of the screening method, but there will be a trade-off between the number of such pairs and the loss in power owing to controlling type I error for FBAT results on those pairs.

To simulate a more realistic situation where the causal SNP is not observed, we chose each SNP as the alleged DSL, simulated trait values as described above according to the normal model (5.4), removed the DSL from the pool of SNPs and determined the most promising SNP from the pool using screening method I or II. We estimated power levels as the proportion of trials successfully identifying the IL10 gene in CAMP or in the Affymetrix LD block of size 4. The identification is successful if one of the SNPs in the gene or block is selected by the screening technique and found significant by the FBAT statistic at the 5% level. Results are listed in Table 5.2.

Overall, the findings of Table 5.2 show that adequate power can be attained to detect the gene harbouring the disease mutation for larger heritability values when only the top screening selection is used. Low heritability values (≤ 0.05) may give rise to poor power, in particular when allele frequencies are relatively low and the number of SNPs tested is large. Even though the SNP pool does not include the actual causal mutation, high correlations between SNPs ensure that the gene containing the mutation is picked up by the screening technique. For example, SNP3 and SNP6 exhibit an r^2 of 0.91 (Table 5.2). The metric r^2 gives an informative description of the degree of LD between alleles at two loci (Wang *et al.* 2005). It is inversely proportional to the sample size that is required for detecting disease association given a fixed genetic risk (Pritchard and Przeworski 2001). When SNP3 is omitted from the pool but was selected in the simulations as the actual causal SNP, SNP6 takes over in the majority of the cases and has the highest selection probability through our screening method. When SNP2 is selected

Table 5.1: Estimated power levels to detect the IL10 gene using SNP data from CAMP or a selected LD-block of 4 SNPs from Affymetrix data for heritabilities in the range 0.05-0.10. The nominal significance level is set to 5%. Screening Method I is based on conditional power calculations; Method II is based on the overall Wald test for genetic effects. Method III uses the Benjamini-Yekutieli (2001) FDR controlling approach to calculate power levels using adjusted p-values. Method IV refers to the Benjamini-Hochberg (1995) FDR corrective approach. Values within parentheses refer to power estimates to detect the simulated causal mutation.

		Causal Mutation in CAMP IL10					
Method I: top 1							
m	h	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6
291	0.05	0.726 (0.368)	0.745 (0.741)	0.673 (0.271)	0.730 (0.548)	0.481 (0.479)	0.695 (0.590)
	0.07	0.871 (0.454)	0.873 (0.872)	0.833 (0.351)	0.865 (0.680)	0.561 (0.561)	0.812 (0.746)
	0.10	0.936 (0.514)	0.943 (0.943)	0.922 (0.403)	0.943 (0.773)	0.687 (0.687)	0.933 (0.874)
Method II: top 1							
m	h	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6
291	0.05	0.842 (0.646)	0.884 (0.864)	0.891 (0.683)	0.843 (0.447)	0.813 (0.813)	0.898 (0.598)
	0.07	0.967 (0.798)	0.976 (0.964)	0.978 (0.800)	0.955 (0.512)	0.967 (0.967)	0.947 (0.620)
	0.10	0.998 (0.881)	0.998 (0.994)	0.999 (0.851)	0.999 (0.559)	0.996 (0.996)	1.000 (0.710)
Method III							
m	h	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6
291	0.05	0.074 (0.059)	0.408 (0.408)	0.203 (0.155)	0.076 (0.063)	0.265 (0.265)	0.235 (0.229)
	0.07	0.172 (0.129)	0.649 (0.647)	0.364 (0.294)	0.167 (0.154)	0.389 (0.389)	0.443 (0.437)
	0.10	0.309 (0.257)	0.882 (0.882)	0.622 (0.556)	0.332 (0.315)	0.584 (0.582)	0.735 (0.724)
		Causal Mutation in AFFY block					
Method I: top 1							
m	h	SNP1	SNP2	SNP3	SNP4		
10,000	0.05	0.5870 (0.2681)	0.6897 (0.2644)	0.4545 (0.0909)	0.5270 (0.0541)		
	0.07	0.7714 (0.4000)	0.8406 (0.3333)	0.7826 (0.1159)	0.7937 (0.0664)		
	0.10	0.9496 (0.5108)	0.9643 (0.3786)	0.9583 (0.1250)	0.9670 (0.0693)		
Method II: top 1							
m	h	SNP1	SNP2	SNP3	SNP4		
10,000	0.05	0.4058 (0.1522)	0.4598 (0.0920)	0.3182 (0.0455)	0.3649 (0.1216)		
	0.07	0.6857 (0.2929)	0.7391 (0.1159)	0.6884 (0.1304)	0.7203 (0.2413)		
	0.10	0.9568 (0.3453)	0.9500 (0.1786)	0.9583 (0.1667)	0.9373 (0.3729)		
Method III							
m	h	SNP1	SNP2	SNP3	SNP4		
10,000	0.05	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)		
	0.07	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.008 (0.000)		
	0.10	0.032 (0.032)	0.040 (0.032)	0.057 (0.057)	0.049 (0.041)		
Method IV							
m	h	SNP1	SNP2	SNP3	SNP4		
10,000	0.05	0.024 (0.008)	0.008 (0.008)	0.008 (0.008)	0.000 (0.000)		
	0.07	0.041 (0.041)	0.008 (0.008)	0.033 (0.033)	0.024 (0.016)		
	0.10	0.153 (0.153)	0.113 (0.105)	0.098 (0.098)	0.146 (0.138)		

Table 5.2: Estimated power levels (in %) via simulations based on the CAMP ($m = 291$) or Affymetrix ($m = 10,000$) data. The values in column "SNPi" refer to estimated power levels when the causal SNPi is removed from the SNP set tested. We list either Pr(IL10 is selected, via one of the 6 available SNPs, by first level screening, and found significant in terms of the FBAT statistic at the 5% level) or list Pr(one of 4 SNPs in a fixed block is selected by first level screening, and found significant in terms of the FBAT statistic at the 5% level), using screening Method I based on conditional power, screening Method II based on the overall Wald test for genetic effects or controlling FDR in Method III (Benjamini-Yuketieli 2001) and Method IV (Benjamini-Hochberg 1995). Different heritabilities are considered in the range 0.05-0.10.

		Causal Mutation in CAMP IL10					
Method I: top 1							
m	h	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6
291	0.05	0.694	0.167	0.653	0.664	0.005	0.557
	0.07	0.838	0.324	0.809	0.817	0.007	0.679
	0.10	0.923	0.569	0.913	0.918	0.012	0.882
Method II: top 1							
m	h	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6
291	0.05	0.750	0.258	0.849	0.788	0.004	0.837
	0.07	0.918	0.456	0.955	0.934	0.006	0.909
	0.10	0.989	0.709	0.998	0.991	0.011	0.995
Method III							
m	h	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6
291	0.05	0.059	0.017	0.161	0.041	0.000	0.109
	0.07	0.133	0.039	0.309	0.085	0.006	0.229
	0.10	0.253	0.069	0.558	0.186	0.002	0.460
		Causal Mutation in AFFY block					
Method I: top 1							
m	h	SNP1	SNP2	SNP3	SNP4		
10,000	0.05	0.5435	0.6782	0.4545	0.5270		
	0.07	0.7571	0.8333	0.7899	0.8077		
	0.10	0.9353	0.9571	0.9583	0.9736		
Method II: top 1							
m	h	SNP1	SNP2	SNP3	SNP4		
10,000	0.05	0.3696	0.4598	0.3182	0.3514		
	0.07	0.6500	0.7391	0.6957	0.7308		
	0.10	0.9353	0.9500	0.9583	0.9439		
Method III							
m	h	SNP1	SNP2	SNP3	SNP4		
10,000	0.05	0.000	0.000	0.000	0.000		
	0.07	0.000	0.000	0.000	0.008		
	0.10	0.024	0.016	0.057	0.049		
Method IV							
m	h	SNP1	SNP2	SNP3	SNP4		
10,000	0.05	0.016	0.008	0.000	0.000		
	0.07	0.041	0.008	0.033	0.024		
	0.10	0.065	0.105	0.082	0.130		

The r^2 measure of LD for CAMP SNP pair (i, j)=(1,2) is 0.11; (1,3): 0.32; (1,4): 0.94; (1,5): 0.02; (1,6): 0.31; (2,3): 0.33; (2,4): 0.11; (2,5): 0.02; (2,6): 0.33; (3,4): 0.31; (3,5): 0.07; (3,6): 0.91; (4,5): 0.02; (4,6): 0.30; (5,6): 0.07. The r^2 measure of LD for AFFYMETRIX SNP pair (i, j)=(1,2) is 0.92; (1,3): 0.92 ; (1,4): 0.88 ;(2,3): 0.96; (2,4): 0.92; (3,4): 0.92.

as the actual causal variant, SNPs 3 and 6 are significant. These are the SNPs in IL10 that are most correlated with SNP2. When we selected SNP5 in IL10 as the unobserved DSL, neither screening method (I and II) had power to identify IL10; SNP5 is poorly correlated with the other five SNPs in IL10 (Table 5.2). The high correlations among the four Affymetrix SNPs in the LD block resulted in similar power levels (Tables 5.1 and 5.2).

Figure 5.1 (total number of SNPs is 291) shows the IL10 SNP selection distribution in CAMP, apart from the selection probabilities of IL10 itself, based on one out of 6 SNPs in IL10 being selected. Even though the SNP pool does not include the actual causal mutation, high correlations between SNPs (rather than high values of Lewontin's linkage disequilibrium measure D'; Lewontin 1988) ensure that the gene hosting the mutation is picked up by the screening technique. SNP selection probabilities for the Affymetrix data are displayed in Figure 5.2.

We also compared PBAT's screening-tools with procedures to control false discovery rate (FDR) at the 5% level based on Benjamini and Yekutieli (2001; Method III) and Benjamini and Hochberg (1995; Method IV). To account for general dependencies such as those arising from LD patterns between SNPs in candidate genes, we included only results for Benjamini-Yekutieli's procedure (2001) for the CAMP data in Tables 5.1 and 5.2 (Method III). The results were derived via the Bioconductor R multtest package (Rv2.0.0) and gave similar results as other FDR controlling procedures such as the one described in Benjamini-Hochberg (1995). The power to detect the gene carrying the DSL, or the SNP itself, was always higher for PBAT's screening methods I and II (based on power or the overall Wald test for genetic effects). For the Affymetrix data and large heritability values ($\geq h = 0.07$) our screening methodology outperformed Method III up to a 30-fold increase in power (Table 5.1). We came to similar conclusions regarding Method IV, the FDR-procedure by Benjamini-Hochberg (1995). Differences in power performance between selection based on FDR controlling, or selection based on our screening techniques, were even more apparent when the actual disease causing mutation was not being tested (Table 5.2). Although Method IV is expected to be less conservative than Method III, they both fail to detect a signifcant association in the Affymetrix data. Because our results (Tables 5.1 and 5.2) refer to probabilities of both selecting a gene or SNP (using methods I or II) and obtaining a significant FBAT statistic at the 5% level, power estimates can be smaller than the nominal level.

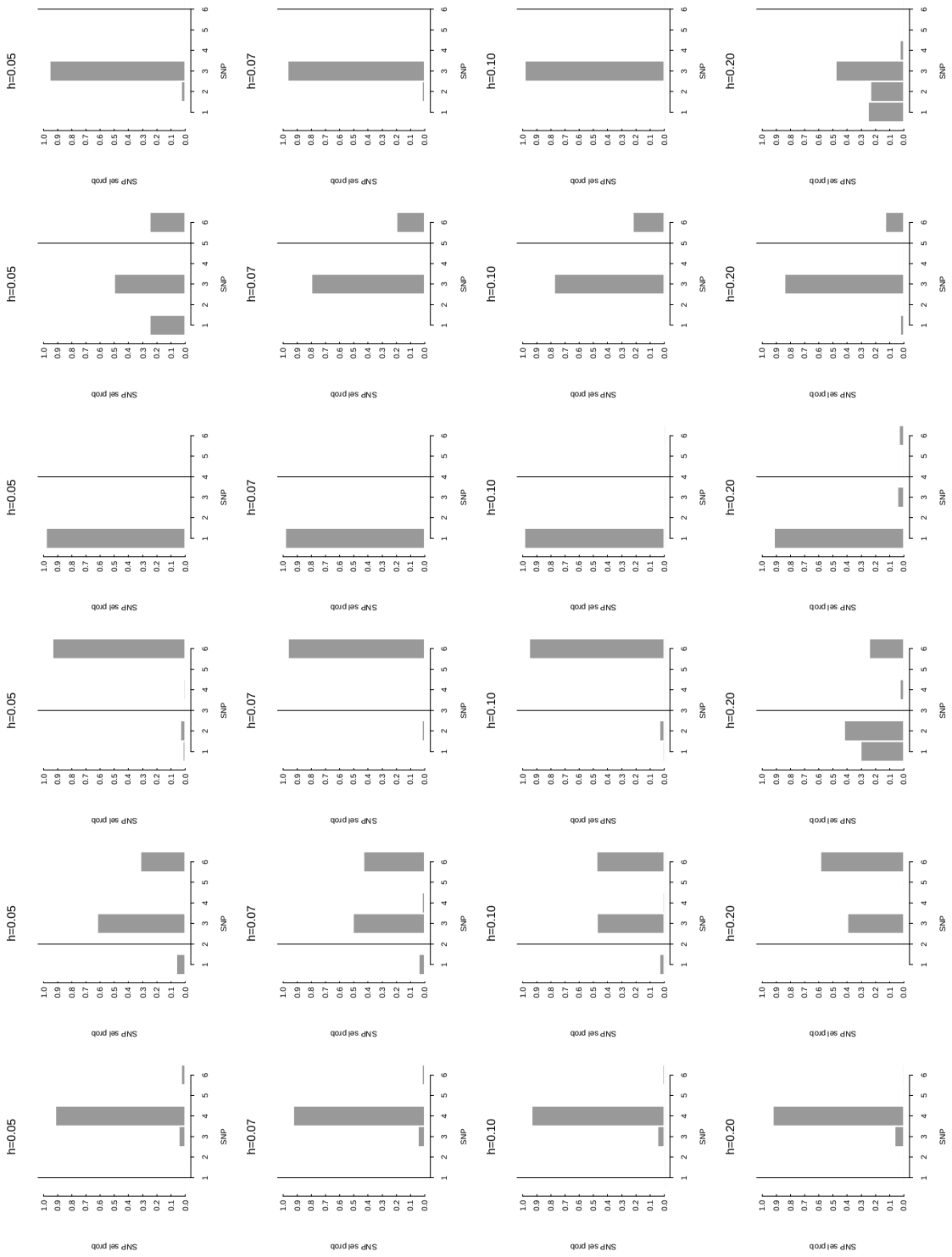


Figure 5.1: Probability of SNP selection within IL10 when screening 291 SNPs minus one. The vertical line indicates the omitted SNP. The leave-one-out SNP for column i is SNPi. This is also indicated by the vertical line in each plot. Rows 1 to 4 pertain to settings with heritabilities 0.05, 0.07, 0.10 and 0.20, respectively.

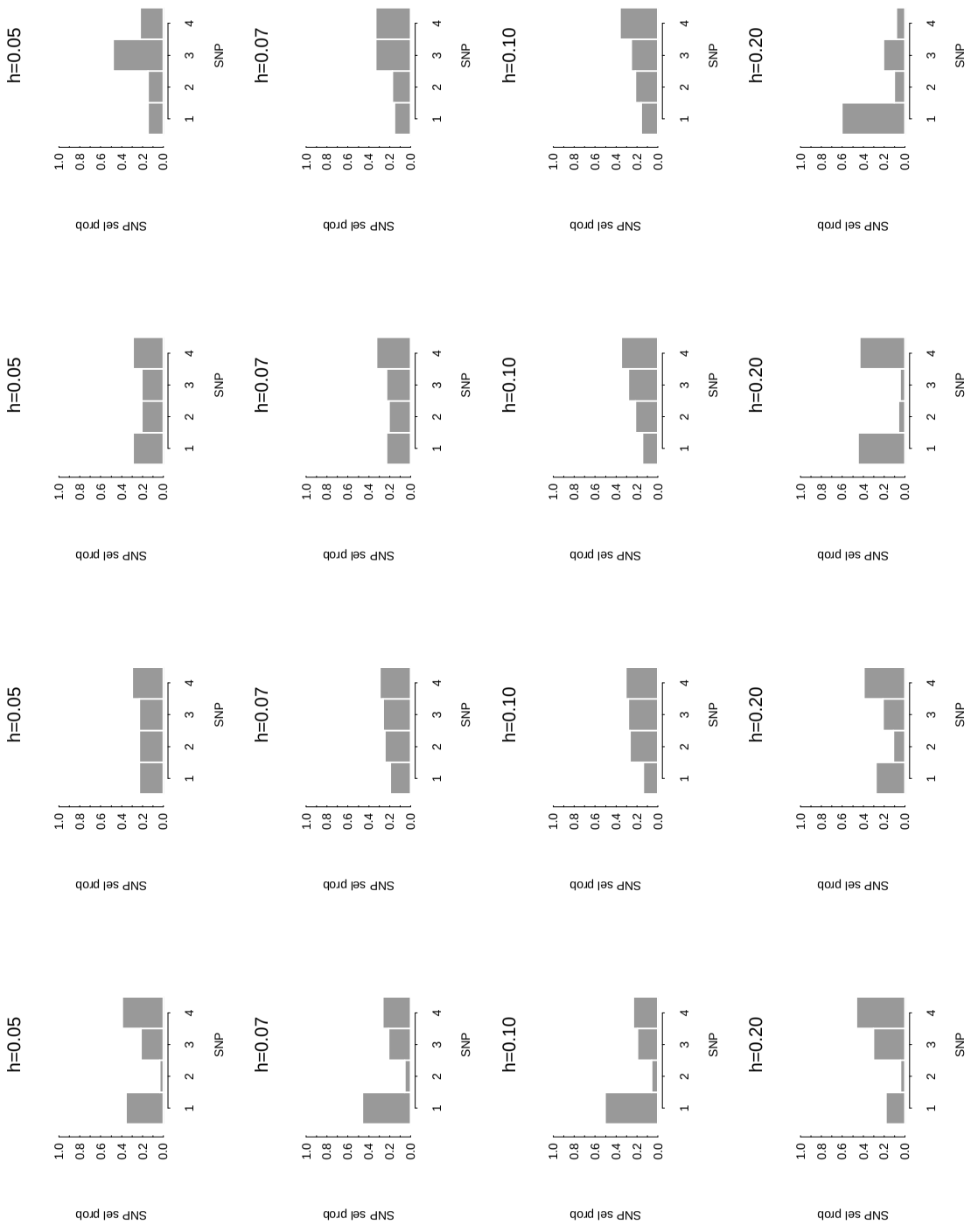


Figure 5.2: Probability of SNP selection within a 4-SNP block when screening 10,000 SNPs. Rows 1 to 4 assume the heritabilities 0.05, 0.07, 0.10 and 0.20, respectively.

Guidelines for the number of top selections

To address the issue of selecting a single trait-marker combination versus multiple combinations, we repeated the simulations, defining the power to identify the disease susceptibility locus as the probability that it is observed in the top K ranking of the chosen screening method and results in a significant FBAT statistics at the 5% significance level, Bonferroni-corrected for K comparisons made. We defined the power to find the gene carrying the DSL in a similar fashion. The simulation results using genotype data from the CAMP study and the Affymetrix 10,000 SNP platform are illustrated in Figures 5.3.a and 5.3.b, respectively. Only results for SNP1 are shown. Dashed lines refer to PBATs screening power to identify the actual causal variant. Full lines refer to power levels to pick up the gene carrying the DSL.

With respect to SNP selection, clear benefits can be gained by including more than 1 top selection with either screening method (Figure 5.3.a, 5.3.b). At some point, these benefits are offset by the required correction for multiple testing. This became apparent when we used relatively few SNPs for screening (Figure 5.3.a). In the case of large SNP pools, selecting the top five combinations gave an acceptable balance between detecting (multiple) associations and reducing power (Figure 5.3.b).

When attention is restricted to "gene" selection, results for large samples such as the Affymetrix data favour picking the top trait-marker selection. The power to detect the gene carrying the DSL increases with increasing heritability and decreases with increasing number of top selections. We observed a steeper decrease in power when we removed the DSL from the data set. For the smaller CAMP sample (Figure 5.3.a), retaining a single trait-marker combination did well for screening on the basis of Wald tests (method II) but was less powerful for screening on the basis of power (method I). Figures 5.3.a and 5.3.b indicate similar trends in case the actual causal variant is removed from the total set of SNPs being screened, yet power was generally lower. Overall, screening on the basis of power and retaining the five most promising trait-marker combinations seemed to yield excellent results.

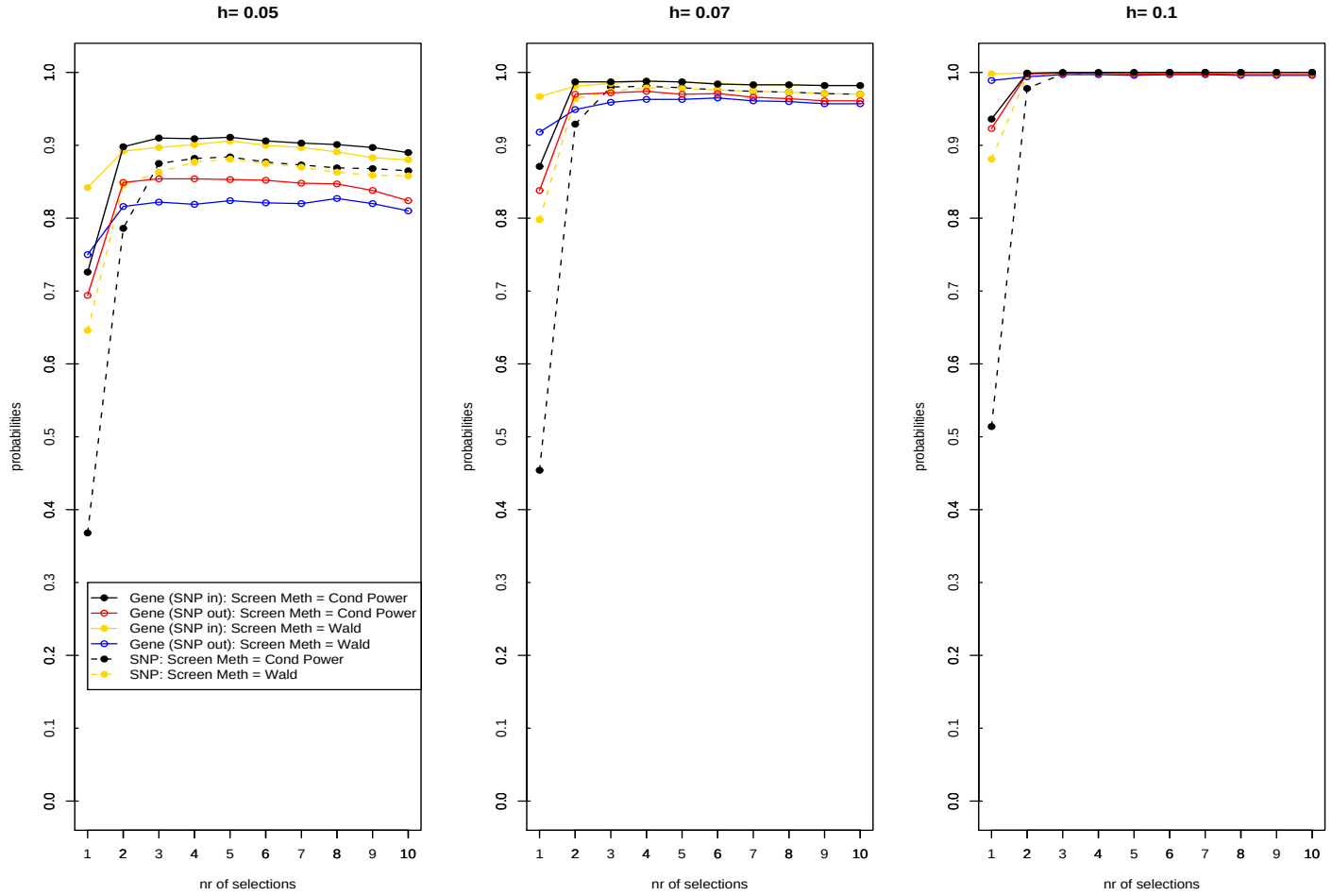


Figure 5.3.a: Power plots versus the number of top trait-marker combination retained after first-level screening under different screening scenarios, using 291 SNPs from CAMP: Method I is based on conditional power sorting (high to low) and method II is based on ranking the p-values (low to high) obtained by the Wald test for genetic effects. Full (dashed) lines refer to probabilities to pick up the gene harbouring the disease mutation (the actual causal variant). Closed bullets indicate that the alleged disease mutation SNP1 is included in the SNP search set. Heritabilities range from 0.05-0.10.

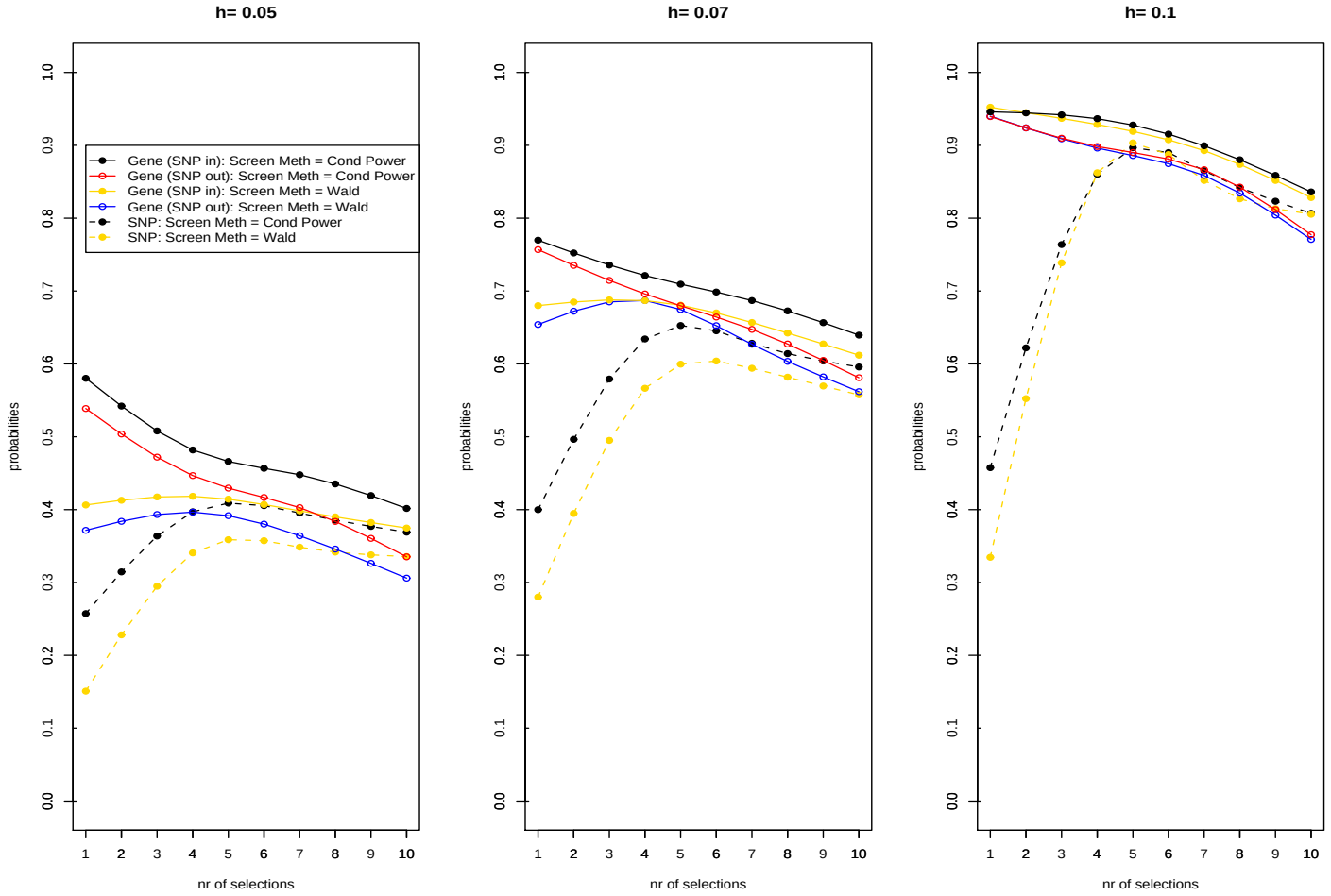


Figure 5.3.b: Power plots versus the number of top trait-marker combination retained after first-level screening under different screening scenarios, using the 10K Affymetrix setting: Method I is based on conditional power sorting (high to low) and method II is based on ranking the p-values (low to high) obtained by the Wald test for genetic effects. Full (dashed) lines refer to probabilities to pick up the gene harbouring the disease mutation (the actual causal variant). Unless open bullets are used, the disease mutation SNP1 is included in the SNP search set. Heritabilities range from 0.05-0.10.

Our simulations reconfirm the belief that existing LD patterns are important in terms of power to detect SNPs or genes. High correlations exist between SNP1 and the other three SNPs in the gene selected for the Affymetrix-based simulations. If the DSL was removed from the search SNP pool, the gene or region of interest can still be picked up by a single SNP that is in high correlation with the removed SNP. In the CAMP data set, retaining more than one trait-marker combination boosted the power to detect the gene. We could increase the signal by adding more trait-marker pairs for further testing, but adding too many combinations could just as well worsen the signal.

Type I Error

By simulating the quantitative trait from a normal distribution with mean 0 and variance 1, we assessed the actual significance levels attained by our two-level testing strategy. As before, both screening techniques (method I and II) were considered. For each combination of SNP and trait, we either estimated the power or computed the Wald test and then selected the most promising SNP to be tested for association with the FBAT-statistic. We estimated the empirical significance levels as the proportion of SNPs that are selected by the screening technique and then found significant by the FBAT statistic at the 5% level.

For 100,000 replicates, the estimated significance levels for the 291 CAMP SNPs were 4.92% and 4.87% for methods I and II, respectively. Thus, both screening methods well maintain the pre-specified significance level. This is not surprising since under the null-hypothesis of no association between any SNP and the trait, the screening methods select the SNP showing the strongest association by chance. For the selection of the SNP, no information is used that would bias the significance level of the subsequently computed FBAT statistic. In other words, under the null hypothesis of no association, the screening techniques and the FBAT statistic are statistically independent (Lange *et al.* 2003a,b) and the significance level of the FBAT statistic is expected to be the significance level of the overall procedure. We obtained similar using the 10K Affymetrix SNP data. For α-level of 5%, the estimated significance levels were 4.86% and 5.14% for methods I and II, respectively.

5.3.2 Analytical Power Considerations for Genomic Association Screening

In this section, we give an analytical expression for the power of PBAT's screening techniques for a single quantitative trait and one disease locus. The derivations naturally extend to include multiple disease loci as well. Hence, here the null hypothesis assumes no association and no linkage between any SNP and a disease susceptibility locus. Generalisations to a non-quantitative trait or multiple traits are straightforward.

Throughout, we assume that n trios are genotyped and that the model for the quantitative trait is given by the additive model

$$Y_i \quad = \quad \mu_i + a_k * X_{ik}$$

where Y_i denotes the quantitative trait for the ith proband, a_k the genetic effect size for the kth SNP and X_{ik} the marker score for the kth SNP in the ith offspring (Falconer and Mackay 1996). A single subscript i is used in Y_{ij} before, since every family i is assumed to have a single offspring ($j = 1$). The number of genotyped SNPs is denoted by m. For simplicity, we take $\mu_i = 0$ and $\mathrm{Var}(Y_i) = 1$. In addition, we assume that the first SNP is the disease susceptibility locus, i.e. $a_1 > 0$ and $a_k = 0$ for $k = 2, .., m$.

The power of the screening technique is then defined as

$$\begin{aligned} Power^{PBAT} &= P(\text{true SNP is selected and FBAT statistic is significant}) \\ &= P(\text{FBAT statistic is significant} \,|\text{true SNP is selected}\,) * \\ &\quad P(\text{true SNP is selected}) \end{aligned} \tag{5.5}$$

Whereas the probability $P(\text{FBAT statistic is significant} \,|\text{true SNP is selected}\,)$ can be computed directly, using the unconditional power calculations described by Lange *et al.* (2002) and Lange and Laird (2002a) and implemented in the PBAT software, the best we can do for the probability $P(\text{true SNP is selected})$ is to approximate it by providing a lower bound.

Probability $P(\text{true SNP is selected})$ can be approximated by

$$P(\text{true SNP is selected}) \quad = \quad P(\hat{a}_1 > \hat{a}_k, k = 2, .., m), \tag{5.6}$$

where the $\hat{a}_k$ are the least-squares estimators for the genetic effect size a_k of the k-th SNP in the expression for the conditional mean model (Methods Section, step 2 of algorithm), and

$$\hat{a}_k = \frac{\sum_i E(X_{ik}|P_{ik1}, P_{ik2})Y_i}{\sum_i E(X_{ik}|P_{ik1}, P_{ik2})^2}.$$

When SNPs are grouped into LD blocks, we can assume independence between blocks and can rewrite expression (5.6) as:

$$\begin{aligned} & P(\hat{a}_1 > \hat{a}_k, k = 2, .., m) \\ & = \prod_{l=1}^{\tilde{m}} P\Big(\bigcap_{k \in M_l \setminus \{1\}} \{\hat{a}_1 > \hat{a}_k\} \Big). \end{aligned} \tag{5.7}$$

Here, $\tilde{m}$ denotes the number of LD blocks and $M_l \setminus \{1\}$ represents the index set for the SNPs in the l-th block, the first SNP excluded. Note that equation (5.7) equals $\prod_{k=2}^{m} P(\hat{a}_k - \hat{a}_1 < 0)$ under independence of the SNPs.

Furthermore, using the Bonferroni inequality,

$$\begin{aligned} & \prod_{l=1}^{\tilde{m}} P\Big(\bigcap_{k \in M_l \setminus \{1\}} \{\hat{a}_1 > \hat{a}_k\} \Big) \\ & \geq \prod_{l=1}^{\tilde{m}} \{1 - \sum_{k \in M_l \setminus \{1\}} P(\hat{a}_1 < \hat{a}_k)\} \\ & = \prod_{l=1}^{\tilde{m}} \{1 - \sum_{k \in M_l \setminus \{1\}} [1 - P(\hat{a}_1 > \hat{a}_k)]\} \\ & = \prod_{l=1}^{\tilde{m}} \{ \sum_{k \in M_l \setminus \{1\}} [c_{M_l} + P(\hat{a}_1 > \hat{a}_k)]\}, \end{aligned} \tag{5.8}$$

where c_{M_l} is a constant depending on the number of elements in the index set M_l. A lower bound of (5.8) is given by

$$\prod_{l=1}^{\tilde{m}} \{ \prod_{k \in M_l \setminus \{1\}} P(\hat{a}_1 > \hat{a}_k) \},$$

since $\sum_{k \in M_l \setminus \{1\}} \{c_{M_l} + P(\hat{a}_1 > \hat{a}_k)\} = O(\sum_{k \in M_l \setminus \{1\}} P(\hat{a}_1 > \hat{a}_k)) \geq O(\prod_{k \in M_l \setminus \{1\}} P(\hat{a}_1 > \hat{a}_k))$.

Combining results from (5.6), (5.7) and (5.8), we obtain the inequality

$$P(\text{true SNP is selected}) \quad \geq \quad \prod_{k=1}^{m} P(\hat{a}_1 > \hat{a}_k). \tag{5.9}$$

Moreover, assuming Hardy-Weinberg equilibrium for parental genotypes and using the results in Lange and Laird (2002b), it can be shown that the variance of $\hat{a}_k$ is given by

$$Var(\hat{a}_k) \quad = \quad \frac{1}{np_k\,(3\,p_k + 1)} \quad \text{for } n \to \infty,$$

where p_k denotes the target allele frequency of the kth SNP. For a large number of trios n, the distribution of $\hat{a}_k - \hat{a}_1$ can be approximated by

$$(\hat{a}_k - \hat{a}_1) \quad \sim \quad N\left(-a_1, \frac{2}{np_k\,(3\,p_k + 1)}\right)$$

Therefore, inequality (5.9) can be re-written as

$$P(\text{true SNP is selected}) \quad \geq \quad \prod_{k>1} P_{N\left(-a_1, \frac{2}{np_k(3\,p_k+1)}\right)}\,(Z < 0)$$

This further simplifies to

$$\begin{aligned} P(\text{true SNP is selected}) \quad &\geq \quad \left[P_{N\left(-a_1, \frac{2}{np_{min}(3\,p_{min}+1)}\right)}\,(Z < 0)\right]^{(m-1)} \\ &= \quad \pi_{h,p_{min},n,m}, \end{aligned} \tag{5.10}$$

since

$$P_{N\left(-a_1, \frac{2}{np_k(3\,p_k+1)}\right)}\,(Z < 0) \geq P_{N\left(-a_1, \frac{2}{np_{min}(3\,p_{min}+1)}\right)}\,(Z < 0)\,,$$

with $p_{min} = \min_{1<k\leq m}(p_k)$.

These power calculations are derived under the assumption that one of the SNPs is the disease susceptibility locus and that genetic effects of other SNPs in the pool are zero (i.e., "$a_k = 0, k > 1$"-assumption), even for SNPs that are in strong LD with the disease susceptibility locus. This is not a realistic assumption to make if the study aims to identify the causal mutation. However, if the goal of the study is to identify the causal mutation or SNPs that are in strong LD with it, this assumption seems justifiable. Furthermore, the far too pessimistic inequality (5.10) should largely compensate for the "$a_k = 0, k > 1$"-assumption. So despite the fact that the resulting bound for the overall power is likely to be too conservative, promising

unconditional power levels P(FBAT statistic is significant|true SNP is selected) in expression (5.6) with high values for $\pi_{h,p_{min},n,m}$ are indicative for adequate actual power.

Values for $\pi_{h,p_{min},n,m}$ under a variety of settings (i.e., different sample size n, disease allele frequency p_{min}, heritability h and number of SNPs m), are shown in Table 5.3. The factor $\pi_{h,p_{min},n,m}$ seems only modestly influenced by the disease allele frequency p_{min}. Moreover, the influence of the number of SNPs on $\pi_{h,p_{min},n,m}$ is almost completely diminished by the level of heritability h and sample size n. However, the unconditional power levels obtained for a single SNP analysis (Table 5.3, column "power FBAT") are highly dictated by allele frequencies, minimum frequencies p_{min} (minimum of all allele frequencies in the data) as low as 0.001 giving rise to poor power.

Assuming 1500 trios, power levels for heritability values in the range of 0.07-0.20 exceed 70% with five million SNPs. Power may drop below 50% when more than 13.6 million SNPs are analysed at once, but only for moderately low heritability values (~ 0.07). For heritability values larger than 10%, power is well-maintained, even with 3 billion SNPs (the approximate genome size), in which case power levels still exceed 85%. These are exciting results given that the most likely scenario, at least for the next few years, is the 500K chip from Affymetrix. For low heritability values (≤ 0.05) and a low disease allele frequency (< 0.001), power levels drop below 50% when more than 162,500 SNPs are screened.

Our screening technique uses the conditional mean model in which the estimated effects might be distorted by population stratification. The extent on PBAT's screening results of the potential distortion by this phenomenon is investigated in the next section.

5.3.3 Population Stratification and/or Admixture

A problem with cohort or case-control association studies is the presence of undetected population structure. This can lead to both false positive results and lack of power to detect true genetic associations. Family-based designs for genetic association do not suffer from this problem. However, PBATs screening tools rely on the conditional mean model and the estimated effects from the screening might be distorted by population stratification. Its impact becomes more severe as sample size and number of markers increases (Pritchard and Donnelly 2001).

Table 5.3: Estimated lower bounds $\pi_{h,p,n,m}$ for PBAT's genome-wide association screening technique and a selection of different sample-sizes n, minimum allele frequency p_{min} (over all available SNPs in the data), heritabilities h and number of SNPs m. Column "power FBAT" gives power levels for a single SNP analysis, obtained via unconditional power calculations by approximation (Lange *et al.* 2002). Lines for heritability settings (one of 0.05, 0.07, 0.10) where both the estimated power for FBAT and $\pi_{h,p,n,m}$ is one, are omitted.

n	p_{min}	h	power FBAT	m=250	m=500	m=1,000	m=10,000	m=100,000	m=500,000
700	0.001	0.05	0.050	0.7454	0.5550	0.3077	0.0000	0.0000	0.0000
		0.07	0.050	0.9662	0.9334	0.8711	0.2512	0.0000	0.0000
		0.10	0.050	0.9988	0.9975	0.9950	0.9515	0.6083	0.0833
700	0.010	0.05	0.874	0.7830	0.6124	0.3747	0.0001	0.0000	0.0000
		0.07	0.925	0.9737	0.9480	0.8986	0.3431	0.0000	0.0000
		0.10	0.962	0.9991	0.9983	0.9966	0.9663	0.7094	0.1797
	0.100	0.05	0.981	0.9676	0.9361	0.8761	0.2660	0.0000	0.0000
		0.07	0.997	0.9984	0.9968	0.9936	0.9376	0.5248	0.0398
		0.10	1.000	1.0000	1.0000	0.9999	0.9994	0.9942	0.9714
n	p_{min}	h	power FBAT	m=250	m=500	m=1,000	m=10,000	m=100,000	m=500,000
1500	0.001	0.05	0.270	0.9989	0.9979	0.9957	0.9583	0.6529	0.1186
		0.07	0.275	1.0000	1.0000	0.9999	0.9995	0.9949	0.9749
		0.10	0.279	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	0.010	0.05	0.997	0.9993	0.9985	0.9971	0.9711	0.7461	0.2312
		0.07	0.999	1.0000	1.0000	1.0000	0.9997	0.9970	0.9851
	0.100	0.05	1.000	1.0000	1.0000	1.0000	0.9995	0.9953	0.9770
n	p_{min}	h	power FBAT	m=250	m=500	m=1,000	m=10,000	m=100,000	m=500,000
2000	0.001	0.05	0.455	1.0000	0.9999	0.9999	0.9986	0.9864	0.9337
		0.07	0.466	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998
		0.10	0.481	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	0.010	0.05	1.000	1.0000	1.0000	0.9999	0.9992	0.9917	0.9591
		0.07	1.000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
	0.100	0.05	1.000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998

Case-control tests of association that are valid in the presence of population structure can be classified as genomic control methods (GC) or structured association methods (SA). Whereas the first uses random markers to estimate the proper null distribution of an association test statistic in the presence of population structure (Devlin and Roeder 1999), the latter uses random markers to infer the details of the population structure itself, prior to testing for association conditional on the estimated individuals' ancestries (Pritchard *et al.* 2000a,b). The statistical advantage of SA methods depends on the degree of informativeness of the available marker data to make correct inferences about the true structure, and the potential of the method needs further exploration. GC methods will perform well when a sufficient number of loci are used in estimating the correcting factor, but using too many loci will lead to a substantial loss in power (Marchini *et al.* 2004). For a thorough discussion on pros and cons between both approaches, refer to Pritchard and Donnelly (2001).

We examine the effect on PBATs screening tools of one method for correcting for population structure, namely the GC method proposed by Devlin and Roeder (1999). Following their ideas and having the conditional mean model (Methods Section, step 2 of the algorithm) in mind, the usual Wald chi-square statistic for no genetic effect may no longer have a chi-square distribution and needs to be inflated by a factor λ. According to Bacanu *et al.* (2002), all available markers may be used to estimate the correction factor in a study of hundreds or thousands of loci, in which only a few genuine trait-marker associations are expected. Since ranking the m (this is total number of SNPs in the data) corrected Wald test values

$$\frac{\hat{a}_k^2}{\lambda \mathrm{Var}(\hat{a}_k)} \sim \chi^2(1),\ 1 \leq k \leq m$$

is then the same as ranking the uncorrected test values $\hat{a}_k^2/\mathrm{Var}(\hat{a}_k)$, genomic control does not affect PBAT's screening method II. In contrast, screening based on conditional power (method I) will not be invariant to genomic control. Since the impact factor λ under genomic control inflates the variance of the estimated genetic effect $\hat{a}_k$, p_k the target allele frequency of the kth SNP, it is given by:

$$\mathrm{Var}_{\mathrm{GC}}(\hat{a}_k) \quad = \quad \frac{\lambda}{np_k\,(3\,p_k + 1)} \quad \text{for } n \to \infty.$$

Hence, this will alter the lower bounds given in Table 5.3.

Apart from the lower bounds $\pi_{h,p,n,m}$ reported in Table 5.3, also the values in column "Power FBAT" are affected by genomic control, since the latter depends on the proportion of phenotypic variance explained by the analysed marker. The loss of power with increasing degrees of population stratification, is most severe for small heritability values ≤ 0.05. With heritability values ≥ 0.07, more SNPs can be added for genomic screening before detecting similar drops in power levels (Van Steen *et al.* 2005).

5.3.4 Multiple disease susceptibility loci

When considering passing multiple SNPs from stage one to stage two, all power studies conducted so far have assumed that there is only one gene causing the disease in the sample. This is the most commonly used underlying assumption in most statistical genetic analyses, even though it is believed that most common diseases are governed by multiple causative loci. In this section, we use a similar set-up for power studies as before, but extend our simulations to allow for the presence of up to five disease susceptibility loci. Using the CAMP and Affymetrix genotypic data as a platform we selected five regions, including the IL10 region for CAMP and the SNP-block of size 4 for the Affymetrix data used to generate Tables 5.1 and 5.2. To represent the LD structure in the data, one causal mutation was selected in each of the 5 regions for the Affymetrix data and one gene was selected in each of the 5 regions for the CAMP data. Traits were generated according to a normal distribution, now including as many as five genetic contributions. For each replicate of the simulation study, we generated heritability values for each locus from a uniform distribution with either mean h value of either 0.03 (for all loci considered) or 0.05 (assuming that when multiple loci elevate disease risk, their singular effects are small).

The screening results using either Method I (based on power) or Method II (based on an overall Wald test for genetic effect) are reported in Tables 5.4.a and 5.4.b. Although we explored the effects on power of selecting the top K SNPs, using various values for K, we show results only for selecting the top five or ten SNPs. Overall, selecting the ten most promising combinations for subsequent FBAT testing works best in the presence of multiple disease susceptibility loci. Comparing Method I with Method II in the presence of multiple disease loci, screening based on conditional power well outperforms screening Method II. This observation holds for both the CAMP data and the Affymetrix data. For the CAMP data, excellent power was achieved to detect as many as three DSLs with Method I,

even for heritability values as low as $h = 0.03$. For Method II, this observation held for detecting as many as two DSLs. Using the Affymetrix data, the results in Table 5.4.b show that screening based on conditional power (Method I) in the presence of multiple trait influencing loci gives good to excellent power to detect two of the loci. Given four DSLs, the power to detect two of them was 0.754 and to detect three of them was 0.246 (Table 5.4.b; column $h = 0.05$; top 10).

Results for the FDR methods of Benjamini and Yekutieli (2001)and Benjamini Hochberg (1995) are included in Tables 5.4.a and 5.4.b as Methods III and IV. The following patterns can be observed: (i) The methods have modest power to detect one locus; (ii) There is virtually no power to detect any DSL using the 10K Affymetrix platform; (iii) In general, the power to detect any trait influencing locus with Methods III and IV is much lower than the corresponding power estimates using screening Method I (based on conditional power). For instance, for $h = 0.03$ and using the CAMP data set, the power estimates with screening Method I and top ten most promising combinations are on average approximately a 20-fold of the corresponding non-zero power estimates obtained with Method III, depending on whether respectively two or five DSLs are considered. The effect is even amplified when the number of SNPs is increased (Table 5.4.b).

5.3.5 Data Analysis: Childhood Asthma Management Programme

Asthma is a complex genetic disorder, with increasing prevalence (Mannino *et al.* 2002). Significant heritability has been reported (Duffy *et al.* 1990). A total of eleven groups have reported linkage for asthma and related phenotypes IgE, skin test reactivity, eosinophil count, and airway responsiveness (PC20) in over 30 genomic regions (Hoffjan and Ober 2002). To date five genes involved in asthma have been identified by positional cloning. More than 200 positive genetic association studies of asthma and its phenotypes have been reported (Weiss and Raby 2004; Wills-Karp and Ewart 2004; Hoffjan *et al.* 2003). Many of these associations have not been replicated. Potential reasons for the conflicting findings of studies of genetic association in asthma include: small sample size, genotyping error, failure to correct for multiple comparisons, genetic and environmental heterogeneity and (for case-control studies) population stratification. Forty associations have been replicated in at least two populations and fifteen associations have been replicated five or more times.

Table 5.4.a: Estimated power levels to detect multiple disease susceptibility loci, based on the CAMP genetic data set. The average heritability of a DSL, in simulating trait values, is either 0.03 or 0.05 for all loci considered. The nominal significance level is set to 5%. Screening Method I is based on conditional power calculations, Method II on the overall Wald test for genetic effects, Method III/IV on controlling FDR (Benjamini and Yekutieli 2001 / Benjamini-Hochberg 1995).

	h=0.03					h=0.05				
CAMP (m=291)										
	Nr of Identified Genes					Nr of Identified Genes				
	1	2	3	4	5	1	2	3	4	5
Method I: top 5										
Nr of DSL 2	0.732	0.229	-	-	-	0.764	0.250	-	-	-
top 10	0.971	0.668				0.981	0.682			
Nr of DSL 3	0.927	0.553	0.128	-	-	0.939	0.642	0.199	-	-
	0.997	0.913	0.460			1.000	0.979	0.700		
Nr of DSL 4	0.890	0.487	0.121	0.005	-	0.948	0.676	0.266	0.042	-
	0.994	0.842	0.394	0.039		0.998	0.976	0.742	0.278	
Nr of DSL 5	0.844	0.417	0.075	0.003	0.000	0.925	0.618	0.230	0.032	0.000
	0.979	0.727	0.282	0.047	0.003	0.995	0.926	0.672	0.243	0.033
Method II: top 5										
Nr of DSL 2	0.545	0.086	-	-	-	0.514	0.085	-	-	-
top 10	0.804	0.274				0.766	0.249			
Nr of DSL 3	0.740	0.289	0.031	-	-	0.720	0.270	0.039	-	-
	0.919	0.562	0.116			0.899	0.551	0.128		
Nr of DSL 4	0.790	0.336	0.049	0.003	-	0.825	0.414	0.111	0.009	-
	0.961	0.675	0.227	0.024		0.963	0.751	0.363	0.082	
Nr of DSL 5	0.753	0.278	0.038	0.001	0.000	0.804	0.383	0.089	0.011	0.001
	0.946	0.562	0.178	0.016	0.001	0.959	0.706	0.321	0.071	0.004
Method III	Nr of Identified Genes					Nr of Identified Genes				
	1	2	3	4	5	1	2	3	4	5
Nr of DSL 2	0.153	0.014	-	-	-	0.471	0.098	-	-	-
Nr of DSL 3	0.358	0.098	0.005	-	-	0.804	0.458	0.116	-	-
Nr of DSL 4	0.416	0.142	0.035	0.002	-	0.855	0.615	0.298	0.058	-
Nr of DSL 5	0.439	0.150	0.025	0.000	0.000	0.861	0.519	0.221	0.027	0.002

Table 5.4.b: Estimated power levels to detect multiple disease susceptibility loci, based on the Affymetrix genetic data set. The average heritability of a DSL, in simulating trait values, is either 0.03 or 0.05 for all loci considered. The nominal significance level is set to 5%. Screening Method I is based on conditional power calculations, Method II on the overall Wald test for genetic effects, Method III/IV on controlling FDR (Benjamini and Yekutieli 2001 / Benjamini-Hochberg 1995).

	h=0.03	h=0.05
	Affymetrix (m=10,000)	

Method I: top 5 / top 10

h=0.03

Nr of DSL \ Nr of Identified Genes	1	2	3	4	5
2	0.646	0.000	-	-	-
	0.665	0.000	-	-	-
3	0.776	0.079	0.000	-	-
	0.823	0.063	0.000	-	-
4	0.846	0.247	0.010	0.000	-
	0.914	0.255	0.025	0.000	-
5	0.730	0.205	0.005	0.000	0.000
	0.822	0.222	0.025	0.000	0.000

h=0.05

Nr of DSL \ Nr of Identified Genes	1	2	3	4	5
2	0.969	0.000	-	-	-
	0.984	0.000	-	-	-
3	0.984	0.390	0.000	-	-
	0.996	0.287	0.000	-	-
4	0.972	0.643	0.116	0.003	-
	0.997	0.754	0.246	0.015	-
5	0.947	0.534	0.051	0.000	0.000
	0.987	0.696	0.185	0.005	0.000

Method II: top 5 / top 10

h=0.03

Nr of DSL \ Nr of Identified Genes	1	2	3	4	5
2	0.098	0.000	-	-	-
	0.236	0.000	-	-	-
3	0.118	0.000	0.000	-	-
	0.307	0.004	0.000	-	-
4	0.106	0.003	0.000	0.000	-
	0.424	0.008	0.000	0.000	-
5	0.072	0.000	0.000	0.000	0.000
	0.345	0.008	0.000	0.000	0.000

h=0.05

Nr of DSL \ Nr of Identified Genes	1	2	3	4	5
2	0.323	0.000	-	-	-
	0.504	0.000	-	-	-
3	0.260	0.008	0.000	-	-
	0.472	0.024	0.000	-	-
4	0.268	0.005	0.000	0.000	-
	0.711	0.099	0.003	0.000	-
5	0.182	0.005	0.000	0.000	0.000
	0.648	0.089	0.000	0.000	0.000

Method: III / IV

h=0.03

Nr of DSL \ Nr of Identified Genes	1	2	3	4	5
2	0.059	0.000	-	-	-
	0.201	0.000			
3	0.138	0.000	0.000	-	-
	0.303	0.000	0.000		
4	0.258	0.000	0.000	0.000	-
	0.485	0.010	0.000	0.000	
5	0.368	0.000	0.000	0.000	0.000
	0.563	0.008	0.000	0.000	0.000

h=0.05

Nr of DSL \ Nr of Identified Genes	1	2	3	4	5
2	0.413	0.000	-	-	-
	0.587	0.000			
3	0.630	0.004	0.000	-	-
	0.799	0.004	0.000		
4	0.770	0.018	0.000	0.000	-
	0.909	0.071	0.000	0.000	
5	0.833	0.003	0.000	0.000	0.000
	0.937	0.033	0.003	0.000	0.000

We illustrate the use of our proposed screening technique to a genomic data set, using parent/child trios in the Childhood Asthma Management Programme (CAMP) Genetics Ancillary Study. A total of 1041 asthmatic children were randomised into three different treatment groups (CAMP 1999). Appropriate informed consent was obtained from all participating subjects at each of the CAMP centres. Blood samples were collected as part of the ancillary study protocol; parental samples were also collected. Genotype information was used on a pool of 291 SNPs for 701 children in 651 pedigrees. As a quantitative phenotype we used the log of PC20 scores (lnPC20), measured repeatedly over time. It is a measure of airways responsiveness, a primary phenotype of asthma.

We first selected the baseline values at randomization and adjusted for age, age of onset, weight, height (first and second order terms) and gender. Attention was restricted to Caucasians only. In particular, screening was performed on 291 SNPs using methods I and II, with a significance level of 5%. We used recessive genetic models, motivated by the success of this model in several genetic association studies for asthma (Randolph *et al.* 2004, Lazarus *et al.* - unpublished data). Furthermore, to ensure the asymptotic validity of the FBAT statistic, we did not calculate it when fewer than 20 families were informative. When only a few families are informative for the FBAT statistic (i.e., most of the families are double homozygous and the offspring's genotype codes show only little variation), the estimate for the genetic effect will be unstable and the asymptotic properties of the FBAT statistic will no longer be valid. This may lead to unreliable screening results using these estimates.

Table 5.5 (first panel) shows the selected trait-marker combinations, from a first-level screening by selecting the five highest conditional power levels (column 6). It can be seen that, after having reduced the total number of SNP-trait combinations to 5, only IL10 g.-627 A>C reaches significance for the asthma-related trait LNPC20 at randomization with a p-value of 0.0058 (compared to 0.05/5=0.01 using a Bonferroni correction for multiple comparisons). Screening results after selecting the five smallest p-values for the Wald test statistic are also shown in Table 5.5 (column 7). This screening method failed to identify any significant associations. Not accounting for multiple testing, there were only four FBAT p-values in the entire data set <0.01; two of these were smaller than the one highlighted in the screening method I process. The smallest FBAT p-value was 0.0022 (Power: 0.0003; WALD p-value: 0.6944).

We next consider a family-based generalized principle component analysis to

Table 5.5: Data analysis results from screening a moderate number of SNPs using the CAMP data. The reported FBAT p-values are not corrected for multiple testing. Method I is based on conditional power calculations; Method II is based on the Wald test for genetic effects. Panel 1 (last column) shows the estimated proportions of phenotypic variance explained by the analysed SNP (h).

Univariate FBAT

					Meth I		FBAT	
Gene	SNP	rs ♯	Allele	♯ Info Fam	Power	WALD p-val	p-val	h
IKBKAP	Leu1023Leu	rs11791783	C	220	0.2831	1.5381E-02	0.1921	0.0306
IL10	g.-627 A>C	rs1800872	A	67	0.2138	2.4775E-02	0.0058	0.0656
IL10	g.-854 T>C	rs1800871	T	78	0.2019	8.1205E-02	0.0355	0.0596
Tbx21	g.-16115 A>G	rs1808192	T	142	0.1381	6.3913E-02	0.7327	0.0264
VDR	Ile352Ile	rs731236	G	179	0.1364	1.4245E-02	0.4800	0.0364
						Meth II	FBAT	
Gene	SNP	rs ♯	Allele	♯ Info Fam	Power	WALD p-val	p-val	h
VDR	Ile352Ile	rs731236	G	179	0.1364	1.4245E-02	0.4800	0.0364
IKBKAP	Leu1023Leu	rs11791783	C	220	0.2831	1.5381E-02	0.1921	0.0306
IKBKAP	Ile816Leu	rs10759326	A	233	0.1087	1.6590E-02	0.0172	0.0347
ADRB2	Ile164Thr	rs1800888	C	22	0.0147	2.1569E-02	0.9755	0.0217
IL10	g.-1117 A>G	rs1800896	A	194	0.1120	2.1745E-02	0.4073	0.0426

Multivariate FBAT

					Meth I		FBAT
Gene	SNP	rs ♯	Allele	♯ Info Fam	Power	WALD p-val	p-val
VDR	g.34059 A>C	rs7975232	A	221	0.7814	2.0505E-03	0.9122
TLR4	g. -6143 A>G	rs1927914	A	276	0.7407	2.9748E-04	0.0086
CRHBP	g.-8093C>T	rs1700676	T	135	0.6970	5.0940E-02	0.4878
VDR	Ile352Ile	rs731236	G	180	0.6922	6.4160E-03	0.5728
ADRB2	g.45702 C>T	rs1036173	T	242	0.6035	1.0130E-02	0.5434
						Meth II	FBAT
Gene	SNP	rs ♯	Allele	♯ Info Fam	Power	WALD p-val	p-val
PPARG	g.30,132 C>G	rs709150	G	184	0.0999	2.7953E-08	0.9890
LOX	g.-2,241 T>G	rs840466	T	58	0.2989	3.9514E-08	0.3382
CHCR1	g.49823 C>A	rs242949	C	218	0.0663	7.6767E-08	0.7501
IL12B	g.11776 C>A	rs1368439	A	187	0.0478	8.8965E-08	0.5671
IL13	Arg130Gln	rs20541	G	219	0.0330	1.6390E-07	0.1063

evaluate which SNPs impact LNPC20 measurements made repeatedly over time. Due to dropout, only the first five time points (including randomization time) were considered. In particular, an overall phenotype with maximal heritability was derived and a univariate test statistic (FBAT-PC) was computed (Table 5.5). For studies with quantitative traits that are measured repeatedly this statistic is often more powerful than single time-point analyses (Lange *et al.* 2004b). Whereas screening method II again does not identify significant associations, screening based on conditional power does highlight a SNP, not residing in the IL10 gene (TLR4 g.-6143 A>G, allele A, p=0.0086), with marginal significance using a Bonferroni type correction for drawing conclusions on five tests jointly. Not accounting for multiple testing, there were only five FBAT p-values in the entire data set <0.01; four of these were smaller than the one highlighted via screening method I. The smallest FBAT p-value was 0.0025 (Power: 0.0092; WALD p-value: 0.3171).

The number of comparisons with the screening technique is far less than we would have to deal with when looking at all possible tests (in our data 440, restricting attention to Caucasians and recessive models) leading to a Bonferroni benchmark significance level of $0.05/440 = 0.0001$. We computed FBAT adjusted p-values using the procedures of Bonferroni (1936), Holm (1979), Hochberg (1988), Benjamini and Hochberg (1995), Sidak (1967, 1971), Benjamini and Yekutieli (2001). None of these methods for controlling type I error rates identified significant associations.

Sometimes large discrepancies between the population-based p-values (via Wald tests) and the family-based p-values (via the FBAT statistic) are observed. These can be explained by the fact that the Wald statistic referred to in Table 5.5 uses between-family information in all families, whereas the FBAT statistic uses only within-family information from informative families (i.e., at least one heterozygote parent).

5.4 Discussion

With currently over 10 million SNPs available in public data bases, advances in automation and parallel genotyping will substantially reduce the costs involved in carrying out genome-wide association studies with a continuing growing SNP pool (Jiang *et al.* 2003). The number of tests for association between each marker and the trait of interest will increase vastly and some correction for statistical

significance is warranted. A similar problem occurs when a genomic region of interest is saturated with markers for follow-up of a linkage peak. As markers in such a region tend to be more correlated than markers across the genome in a genome scan, controlling for multiplicity becomes even more difficult. In this chapter, we have addressed the multiple testing problem, one of the most important statistical hurdles involved in candidate gene or genome-wide association studies.

There are 2 types of association tests: population-based (either case-control or cohort) and family-based (related individuals). Typically, in the first design cases and controls (or the cohort) are cross-classified by genotype and analysed via a non-parametric χ^2 test or a parametric logistic regression model. In the second design, the analysis is often based on allele transmission rates, using TDT-type of statistics. The methods proposed in this manuscript are only applicable to family-based studies. One of the problems with recruiting family data is that family members may not be available (e.g., for late-onset diseases) and that it usually involves an increased genotyping cost compared to the case-control setting. Genotyping errors may bias the analysis results, but unlike in a case-control design, genetic information on available family members may help to detect and correct this type of errors. In addition, family-based data can be used in assessing haplotypes, to resolve phase, and is not susceptible to confounding due to population substructure. Finally, family data allow testing of both association and linkage, while population-based data only allows testing of association.

Our studies show that genome-wide association studies have the best power when

- family data with at least 1,500 probands are available,
- the minor allele frequency for SNPs in the analysis is at least 0.01,
- heritability is moderate (> 0.03).

Under these conditions, genome-wide association studies with as many as one million SNPs can be successful. The information gained from the increased number of SNPs are not diluted by the multiple-comparison problem. The number of SNPs has only a modest effect on the overall power of our screening techniques; its effect seems to be bounded for higher heritability values. The screening technique maintains its protective character for extended data sets with a few hundred-thousand SNPs. In addition, our screening methods are robust against effects of population

stratification and admixture.

An adequate rule of thumb to identify one DSL is to screen on conditional power (Method I) and to consider the five most promising combinations for FBAT testing. To identify multiple loci of small effects, a substantial gain in power can be achieved when considering the ten most promising combinations in stage I for further FBAT testing. When extended pedigrees have to be analysed or when the phenotypic vectors in a multivariate version of the FBAT statistic are highly dimensional, a balance needs to be found between loss of power (Method II) and increased computational burden (Method I). In either case, the better performance of method I for large numbers of SNPs may alter views on current strategies (Satagopan *et al.* 2004) for genome-wide association studies in case-control designs.

Studies of nucleotide diversity estimate that a common SNP with minor allele frequency >0.100 roughly occurs once every 600 base pairs (Kruglyak and Nickerson 2001). Thus on average 50 common polymorphisms are to be expected in a gene, assuming that the average gene in the human genome spans about 27Kb (e.g., Venter *et al.* 2001). This provides an important source of information in tracking down the genetic basis of complex diseases. We have illustrated that our screening methods are a useful tool, to detect common disease susceptibility loci, even with small effects. Applying our screening tools using the haplotype features of PBAT in this setting may increase power even further. This is currently under investigation.

Because of the computational and statistical issues involved in analysing thousands of SNPs in single SNP analysis, investigators have searched for methods to reduce the sheer amount of data, using LD patterns, with a minimum loss of information (e.g., the International HapMap Consortium 2003, Carlson *et al.* 2003). Leaving aside whether SNP selection should be based on haplotype blocks or not (Zhai *et al.* 2004), the end product is a reduced set of tagging SNPs, often with low LD between them. Using simulations, we showed that our proposed screening techniques do not require a priori identification of causal variants to identify disease-associated regions; their success relies on the assumption that untyped risk-related SNPs are correlated with one or more typed SNPs. From this perspective, adding more SNPs is beneficial, since it increases potential LD with (the) actual causal mutation(s). In general, adding more SNPs comes at the cost of power loss when corrections for multiple testing need to be applied (e.g., Bonferroni-type corrections to control type I error), but our screening methods are

only moderately affected by adding "non-causal" SNPs.

In theory, Type I error rates can also be controlled by permutation-based methods, similar in nature to those proposed by Churchill and Doerge (1994, 1996): If there is a real association between a trait and some marker(s), that association can be destroyed by randomly shuffling the trait values. Such permutation procedures are numerically challenging for large data sets with thousands of SNPs. As the field is moving towards genotyping arrays with more than 500,000 SNPs, small p-values are expected, and the number of permutations needs to be large to guarantee acceptable accuracy levels. In PBAT, a genomic screen of 2,000 trios genotyped on 300,000 SNPs takes 1 day on a single processor. Alternatively, the transmitted and untransmitted status of alleles from parents to offspring can be randomised. Such an approach was adopted by Lin *et al.* (2004) in the context of genome-wide association studies using the classical TDT test (Spielman *et al.* 1993). This procedure is more challenging, since reshuffled genotypes with respect to phenotypes should still resemble the original LD structure that is present in the data.

Our screening tools increase the detection power, provide an analytical method to use genotype data in full and can be used with multivariate quantitative traits, time-to-onset traits, covariate adjustment, multiple alleles, haplotypes, extended pedigrees and missing parents. Genomic control (Devlin and Roeder 1999) does not affect PBAT screening method II. Power is reduced with increasing degrees of population stratification using method I but remains acceptable. Compared with FDR methods, our screening tools are particularly attractive to detect multiple trait influencing loci in data sets with thousands of SNPs, even for low heritability values. Unlike FDR-methods our screening-technique does not necessarily identify the SNP with the smallest FBAT p-value, but combines p-value information with an estimate for genetic effect. Furthermore, our screening-technique does not require a screening and replication sample. Both screening and replication steps can be accomplished in a single data set.

Missing Data

Chapter 6

Approaches to Handle Incomplete Data in Family-based Association Testing

6.1 Introduction

The high throughput of data that has arisen with the complete sequence of the human genome has left statistical genetics with a rich information source. The wide availability of software and the increased capacity for computational power has improved the potential to access and process the data. One of the problems though is the incompleteness of the data: unobserved or partially observed data points due to technical deficiency, reasons associated with the patient's status, erroneous measurements of phenotype or genotype, to name a few. These sources of incompleteness seriously jeopardise the credibility of analytical results when not properly addressed.

In this chapter, we aim to give a perspective view on the occurrence and analysis of different forms of incomplete data in family-based genetic association testing.

The concern about how to deal with incomplete data in general and missing data specifically is not new. The history of accounting for missing observations in statistical analysis is characterised by many casual events. However, during

the last 25 years missing data issues have gained momentum, mainly due to the work of Rubin (1976). While much effort has focussed on conceptual work and developing a taxonomy, considerable attention has been dedicated to the practical implementation of proposed strategies. Putting these developments outside their frame of thought has shown to be very dangerous in for example clinical trials (Mallinckrodt *et al.* 2003). Whereas in clinical trials the variation in missing data aspects is fairly limited, a large diversity exists in statistical genetics. Speaking the "language" fluently is mandatory to understand the nature of the problem and to propose adequate solutions.

Keeping in mind that the primary goal of genetic analysis is to specify the correct effect of genotype on phenotype, we note two remarkably different scenarios of genetic analysis that engender different aspects of the missing data problem. First, we can measure phenotypes and subsequently use this source of data to draw inferences about the genotype structure. Second, we can measure both phenotype and genotype variables. In the first scenario, missingness issues are confined to the phenotype level. For the second scenario the genotype level may be contaminated with (partially) unobserved data as well. On top of the classical missingness problem, measurement uncertainty or measurement error may be superimposed.

For genetic association studies, a general paradigm is a regression analysis, in which the disease trait is the response variable and the coded genotype the predictor variable. Here, statistical models are developed and the validity of conclusions crucially depends on the underlying model assumptions. One such set of assumptions pertains to unobserved or "missing" measurements. This is particularly relevant when modelling genetic associations. Obviously, in the context of testing genetic associations, different approaches need to be adopted to properly account for incomplete data.

The remainder of this chapter is organised as follows. Section 6.2 gives an overview of family-based association testing. Incomplete data taxonomy is re-introduced in Section 6.3. In Section 6.4 we focus on incomplete data issues (e.g., detection, impact assessment, currently used remedial measures), targeted to family-based association tests. Less commonly used approaches dealing with incomplete data in this context are also discussed (Section 6.5). We do not differentiate between the type of markers; the discussions naturally extend from single marker analysis to the analysis of multiple tightly linked markers (haplotype analysis). The missing

data problem is compounded in the haplotype setting in that unknown phases need to be resolved as well.

6.2 Family-based Association Testing

Family-based association tests (Thomson 1995; Gauderman *et al.* 1999) use the genetic information from the family members to construct the distribution of the test statistic under the null hypothesis of no linkage and no association, conditioning upon phenotypes and parental genotypes. For qualitative traits, family-based association tests can be traced back to the Transmission Disequilibrium Test (TDT) of Spielman *et al.* (1993, 1994) and Ewens and Spielman (1995) which, using affected individuals, compares observed and expected marker scores computed using parental genotypes and Mendel's law of segregation. The technique uses family-based controls and therefore addresses a major concern in genetic association analyses: spurious associations caused by population substructure. Since its conception, many have extended or modified the TDT to include multi-allelic markers (Sham and Curtis 1995 - ETDT; Zhao *et al.* 2000), haplotypes (Bourgain *et al.* 2000 - MILC; Clayton and Jones 1999), family designs other than the two parent/one affected child setting (Spielman and Ewens 1998; Martin *et al.* 2000 - PDT), quantitative traits (Abecassis *et al.* 2000; Lunetta *et al.* 2000), an evolutionary-based haplotype analysis approach (Seltman *et al.* 2001 - ET-TDT), covariates or important confounders (Rice *et al.* 1995; Whittemore *et al.* 2005 - COVTDT), missing parental genotype information (Sun *et al.* 1999; Weinberg 1999; Lee 2002), or to allow for genotyping errors (Gordon *et al.* 2001).

Self *et al.* (1991) developed a conditional logistic regression model, modelling the marker relative risk for trio data. Each case is matched to three controls, such that the controls are the remaining three marker genotypes that the parents of each case could have passed on. The regression framework lends itself easily to test the null hypothesis of no association of the genetic marker alleles with disease.
Schaid (1996) developed a more general score statistic using the conditional likelihoods developed by Self *et al.* (1991), extended to multi-allelic markers and log-risk models for dichotomous phenotypes. The template for Schaid's test statistic are Rao's efficient score statistics. These require evaluation only under the null hypothesis, unlike likelihood ratio test statistics that require full likelihood evaluation. The major drawback of Schaid's conditional test is the necessity to have observed parental genotypes for all cases included in the test, a requirement that is

particularly problematic, for example, in the genetic analysis of late-onset diseases. Cordell and Clayton (2002) have analysed case-parent trios via conditional logistic regression as well, using the case and three pseudocontrols derived from the untransmitted parental alleles. They have described a unified approach for genetic association analysis with nuclear families or case-control data in a (stepwise) regression context. For nuclear family data, the approach can be seen as an extension of the genotype relative risk method of Schaid and Sommer (1993) and Schaid (1996). The practicality of their regression method is that it can be performed using standard statistical software and that additional effects such as parentoforigin effects can be included. Apart from the misspecification problem in regression modelling, a major drawback is that the technique has not been adapted yet to include extended pedigrees without splitting them up into nuclear families.

Likelihood methods that deal with missing parental genotype information are presented in Schaid and Sommer (1993), but assume a homogeneous population in Hardy-Weinberg equilibrium. As outlined in Schaid (1996) this restriction can be avoided by inferring parental genotypes using information from available members in the pedigree. The likelihood-based TDT approach of Clayton and Jones (1999) and Clayton (1999) handles missing parental genotypes but is, to date, restricted to dichotomous outcomes. Because missing parental genotypes are handled by estimating parameters of the missing genotype distribution from the data, without conditioning on founder genotypes, it is not robust against population admixture.

In summary, Self *et al.* (1991) and Schaid (1996) first model genotype to disease and subsequently apply a correction for admixture before computing the score statistic. Lunetta *et al.* (2000) first model disease versus genotype, which naturally extends the association models to measured phenotypes as well. While treating the phenotype as an outcome for modelling purposes, it is treated as fixed in calculating the distribution of the test statistic. The distribution of the derived score statistic is calculated as a function of offspring genotypes, conditional on parental genotypes and offspring trait values. As a consequence, the conditional test is unbiased even when the association model or phenotype distribution is misspecified; a correction for admixture is implemented after the computation of the score statistic. Although this method can be extended to time-to-onset data, categorical, ordinal, continuous and multivariate phenotypes, and inclusion of important predictors, adequate estimates of nuisance parameters (e.g., when traits vary among subjects in the sample) are not always available, in which case ad-

ditional steps are taken, such as computing the score statistic with minimal variance.

In response to the need of a flexible methodology that addresses the aforementioned issues, Rabinowitz and Laird (2000) and Laird *et al.* (2000) introduced a unified approach to family-based tests of association referred to as FBAT. While this general approach in family-based association testing also builds on score tests computing the conditional distribution of offspring genotype given their phenotypes and parental genotypes, it allows tests of different genetic models, tests of different sampling designs, tests involving various types of phenotypic data, tests with missing parents, and tests of different null hypotheses, all in the same framework. In particular, the FBAT statistic is based on a linear combination of offspring genotypes and traits:

$$FBAT = (S - E[S])/\sqrt{V},\ S = \sum_{ij} T_{ij} * X_{ij}, \tag{6.1}$$

where $V = \text{Var}(S)$ and T_{ij} represents the coded phenotype (i.e., the phenotype adjusted for any covariates) of the j-th offspring in family i. The X_{ij} denote the offspring's coded genotype at the locus being tested and depends on the genetic model under consideration. The test adjusts for ascertainment and avoids bias due to population admixture or stratification and mis-specification of the trait distribution. The hedging against bias is achieved by calculating the null distribution of the test statistic, conditional on the sufficient statistics for any nuisance parameter under the null. For instance, if parental genotypes are missing, then the FBAT approach also conditions on the sufficient statistics for parental genotypes.

In what follows, we will focus on the FBAT statistic of Rabinowitz and Laird (2000) and Laird *et al.* (2000). We will organise the discussion on incomplete data in family-based association tests around missingness in (i) parental genotypes, (ii) offspring-coded genotypes X_{ij}, (iii) traits T_{ij}, and (iv) in covariates to recode observed offspring's phenotypes. Genotyping errors (v) and incompleteness in haplotype analysis (vi) are discussed separately.

6.3 Incomplete Data: What's in a Name?

Incomplete data issues cover a much broader variety of applications than missing data. The term "incomplete data" is generally used when measurements are observed but recorded on a "coarser" scale than may actually be possible. "Missing data" simply involve situations in which certain planned measurements are

unobserved, but can be viewed as one particular aspect of such "coarsened data". Heitjan and Rubin (1991) offer an alternative view on handling incomplete data when conducting statistical inference. In this chapter, we not only handle the missing data problem in FBAT's, we extend the discussion to include measurements with error as well. Such data can be thought of as "incomplete" in that the true value, and hence the discrepancy between true and observed values, is unobserved.

The taxonomy on "missing data" goes back to Rubin's work of 1976, in which he describes three missing data mechanisms. Following his classification system, the data are said to be missing completely at random (MCAR) if the probability that a measurement is missing is independent of the other outcomes, observed or not. Data are said to be missing at random (MAR) when the cause of missingness is unrelated to the missing values, but may be related to observed ones. The latter may be either observed covariates or response variables. In all other cases, the missing data mechanism is said to be missing not at random (MNAR). Missing or erroneous genotypes, even when technical failures are responsible for them, do not necessarily follow a MCAR process.

The aforementioned taxonomy is particularly useful in the context of parameter estimation. It is therefore usually presented in the selection modelling framework (e.g., Glynn *et al.* 1986; Little 1993, 1995), where the joint distribution f of measurements and missingness process is factorised into the marginal measurement distribution and the conditional distribution of the missingness indicators, given the outcomes. Nevertheless, whenever a trait-marker model is specified and estimates thereof are used in the computation of the test statistic, missingness in the components of the model are a point of concern and deserves special attention. This may also apply to the FBAT statistic, when recoding phenotypes as traits T_{ij} in $S = \sum_{ij} T_{ij} * X_{ij}$ using the residuals from a regression model. However, even when the regression model is not valid, the FBAT test remains valid, even though it may have reduced power.

Genetic analysis usually assumes that an individual's actual marker genotype is coded correctly, i.e., error-free. Issues with heterozygotes, such as whether the marker alleles at the locus really different, are hard to pick up when checking for Mendelian errors using standard genetic software. As long as one parent has one variant allele, there will not seem to be an incompatibility. Genotyping errors, whether random or systematic, are a major concern in statistical genetics. Apart

from affecting the accuracy of marker maps, mistyping also affects the localization of traits (Sobel *et al.* 2002). Genotyping errors may lead to a substantial loss of power in gene-mapping studies and may severely underestimate the strength of correlations between trait- and marker-locus genotypes. Several authors have shown that even a small error rate of $1-2\%$ can have a serious impact on linkage results (e.g., Abecasis 2001 for family-based analysis of quantitative traits). When testing for association and linkage, Mitchell *et al.* (2003) have shown that undetected genotyping errors can inflate the type I error rate of the transmission disequilibrium test. The inflated error rate is explained by over transmission of common alleles in the presence of undetected genotyping errors, resulting in a systematic bias in the test statistic.

There is a vast literature on error detection (refer to Gordon *et al.* 2001 for examples) but only limited sources are available that offer methodology to actually deal with the problem in linkage and or linkage disequilibrium (LD) analysis; removal of the error by treating the erroneous measurement as missing is usually not a good idea. Sobel *et al.* (2002) address many aspects of detecting and integrating genotyping errors in statistical genetics. In the context of linkage analysis, Göring and Terwilliger (2000a,b,c,d) in a series of papers introduce the concept of hypercomplex-valued recombination fraction, confirming that crossing field boundaries opens up new perspectives in problem solving. In the context of LD analysis, Gordon *et al.* (2001) developed a TDT test that allows for random genotyping errors (TDT_{ae}).

6.4 Incomplete Data in FBAT-testing

In this section, we elaborate on the nature and extent of missingness and genotyping errors in family-based association testing using the FBAT statistic. This is important because TDT-type tests are known to inflate type I error rate where there is missing parental genotype information (Curtis and Sham 1995) or undetected genotype errors (Gordon *et al.* 2001, Mitchell *et al.* 2003).

6.4.1 Missing Parental Genotypes

Problems frequently encountered in genetic family studies include, for example, the availability of family members for late-onset disease and an increased genotyping cost compared to the case-control setting. In their characterization, Rabinowitz and Laird (2000) derive the minimal sufficient statistics in the presence of missing

constituent information. When parental data are complete, the observed traits in all family members and the parental marker genotypes are the minimal sufficient statistics. For incomplete parental data, the partially observed parental genotypes and the offspring genotype configuration are sufficient statistics for the missing parental genotypes.

Other approaches that can deal with missing parental genotype information include the Sibship Disequilibrium Test (SDT) of Horvath and Laird (1998), the Sib Transmission/Disequilibrium Test (S-TDT, Spielman and Ewens 1998), the discordant alleles test (DAT, Boehnke and Langefeld 1998), a class of transmission/disequilibrium test-like statistical tests based on the difference between the estimated allele frequencies in the affected and control populations (Risch and Teng 1998, Teng and Risch 1999), the quantitative trait transmission disequilibrium test of Schaid and Rowland (1999), methods developed by Weinberg (1999), Allison *et al.* (1999) and Clayton (1999), and the Reconstruction-Combined Transmission/Disequilibrium Test (RC-TDT) of Knapp (1999).

Similar to the conditioning arguments used by Rabinowitz and Laird (2000), Laird *et al.* (2000), Horvath *et al.* (2001), Horvath *et al.* (2004) suggest extending their conditioning strategy to tightly linked markers for haplotype analysis by conditioning on the sufficient statistic for resolving phase in phase-unknown parental genotypes as well.

6.4.2 Missing Offspring Genotypes

Missing genotype data can lead to an increased number of false positives if samples with a particular genotype are more likely not to be classified during genotyping, for instance when the adopted genotyping method has a lower success rate for heterozygotes (Hirschhorn and Daly 2005). The missingness process in this situation is MNAR because the missing genotype depends on what would have been observed when the genotyping method works well. Removing records with incomplete information may lead to biased results under MCAR or when the incompleteness can be explained by observed factors (MAR).

The current practice in FBAT analyses is to ignore individuals with incomplete genotype information or to treat a missing allele as any other allele. Family data may help to resolve genotype issues but the available data are often not sufficient. Also when haplotypes rather than SNPs are considered, missing or erroneous geno-

types matter in the construction of the haplotypes and estimation of the haplotype frequencies. Sebastiani *et al.* (2004) proposed an extension to the TDT, called robust TDT (rTDT), that is able to handle incomplete genotypes on both parents and children and does not rest on any assumption about the missing data mechanism.

Sometimes it may be beneficial to deliberately introduce missingness, an example of missingness by design. In family-based studies, the multiple testing problem as a consequence of increased dimensionality can be handled by creating two sources of information, so that data used to reduce the number of SNPs is different from the data used for FBAT testing. Using the conditional mean model approach (Lange *et al.* 2003a,b; Van Steen and Lange 2005), the data are first analysed in a screening step.

The basic idea is to estimate a genetic effect via a regression model that utilises data not used in the association test statistic; observed marker scores for offspring in informative families (a family is informative for FBAT testing when it gives a non-zero contribution to the FBAT statistic) are set to missing and replaced by their conditional expected marker scores. In this screening step, the scientist can look at all possible associations between the markers and traits, and select a subset of promising marker-trait combinations, typically between $5-10$ combinations (Van Steen *et al.* 2005) without biasing the significance level of subsequently computed tests. Only the selected subset is then put forward to the hypothesis testing step.

6.4.3 Missing Traits

Little research has been done in the area of missing phenotype information within a pedigree analysis. Murphy *et al.* (2004) present a new testing strategy for family-based association tests when missing multivariate phenotype data are present. The standard methodology used by FBAT involves the list-wise deletion of missing observations (FBAT-GEE; complete removal of a subject). Alternatively, subjects' data are still utilized for phenotypic outcomes where their information are not missing (case-wise deletion). The latter is the "observed FBAT-GEE" described in Lange *et al.* (2003c). In addition, Murphy *et al.* (2004) discuss several imputation techniques for missing phenotypes: e.g., the Expectation-Maximization (EM) algorithm (Dempster *et al.* 1977) and the Data Augmentation (DA) algorithm (Tanner and Wong 1987) in strata according to parental mating types or unstratified, and imputation using the conditional mean model (Lange *et al.* 2003a,b; Chapter 5: Model (5.3)). For the latter, all model parameters of the multivariate mean model are first esti-

mated using the observed data. Secondly, the missing phenotypes are imputed based on the effect size estimates and the expected marker score $E[X_{ij}|\text{parental genotypes}]$ (X_{ij} denotes the coded genotype of the j-th offspring in family i) under the null hypothesis of no linkage and no association.

When the data are MCAR, the "observed FBAT-GEE" outperformed the discussed imputation methods in terms of power of the multivariate FBAT test. When the data are MAR and strong environmental correlations exist among the phenotypes, both EM- and DA-based imputation methods, after stratification of the data by mating type, are recommended. When the data are MAR and moderate environmental correlations exist, imputation based on the conditional mean model seems to achieve better power.

Data augmentation refers to methods for constructing iterative algorithms via the introduction of unobserved data or latent variables. Whereas Dempster *et al.* (1977) popularised the method for deterministic algorithms, Tanner and Wong (1987) did so for stochastic algorithms. A valuable overview of data augmentation schemes is given by van Dyck and Meng (2001).

6.4.4 Missing Covariates

Environmental factors that are known to influence disease risk in population designs may also play a role in family-based studies. If this is the case, incorporating them as confounding factors in the genetic analysis can improve the power to detect disease-associated genes. There are two scenarios: (i) all measurements are missing, e.g., because the importance is unknown to date, or (ii) some measurements are missing e.g., because information has been collected from different data bases. Provided that the missingness in the covariates is unrelated to the outcome in the regression model, one can always obtain unbiased estimates of regression parameters, by using only subjects with complete data on all covariates in the model (Jones 1996). Missing covariates may affect PBAT's screening tools in that first level screening is based on genetic effect estimates from a parametric regression model, as explained before. Missing covariates may also affect the power of the FBAT test when recoding phenotypes, using residuals derived from modelling strategies, so as to create new constructs that exhibit reduced variation.

Often in studies that explicitly differentiate between responses and covariates, the conditional distribution of the incomplete covariate is specified, given (a subset of) all other covariates and the response, as a means to predict incomplete entries.

In many cases there is no obvious reason to believe many of the distributional assumptions made. Instead of adopting a full parametric approach (Little 1992; Blackhurst and Schluchter 1989; Ibrahim 1990; Ibrahim and Weisberg 1992) a semi-parametric estimation procedure may be advocated (Robins and Rotnitzky 1995). The latter occasionally suffers from the drawback of severe efficiency loss. Didelez (2002) discussed pros and cons that focussed on alternative strategies using non-Bayesian ideas while Zhao *et al.* (1996) examined regression analyses with missing covariate data using estimating equations. In addition, Lipsitz and Ibrahim (1998) proposed a set of estimating equations for Cox' proportional hazard models (Cox and Oakes 1984) and under ignorable missing covariate data. Parameter estimates are obtained via a Monte Carlo EM algorithm. Leong *et al.* (2001) generalise the technique to non-ignorable missing covariates in the Cox model using biological marker data. Horton and Laird (1998, 2001) discuss the method of weights as an implementation of the EM algorithm for general maximum likelihood analysis of regression models. Bayesian methods such as multiple imputation (MI) can also be used to deal with missing covariate information.

In contrast to single imputation, multiple imputation, first described by Rubin (1977) and later covered in more detail by Rubin (1987) and Schafer (1997), refers to replacing each missing value by more than one imputed values. By imputing several values for a single missing component, it is explicitly acknowledged that imputed values and truly observed values are entirely different. Not only does multiple imputation provide an adequate way to account for sampling uncertainty, it can also be regarded as a component of a sensitivity analysis to investigate assumptions about relevant parts of models under consideration. Nevertheless, some remarks on this approach are in order. First the imputation model should be at least approximately compatible with the analysis to be performed on the imputed data sets. Second, it should be rich enough to preserve the associations or relationships among variables that will be the focus of later investigation. Third, convergence of algorithms involved are related to rates of missing information. Fourth, in a particular study the imputer and the analyst may be two different persons, not necessarily having the same model in mind.

Several publications, such as Horton and Lipsitz (2001), set out practical guidelines for multiple imputation. These authors list several imputation models, encompassing settings with monotone or non-monotone missingness patterns, missing outcomes and/or predictors, and continuous or discrete variables. In addition, the

dual problems of missingness and error (Section 4.5) can be handled in one analysis. Ghosh-Dastidar and Schafer (2003) describe how to replace an observed data set containing both missing values and errors with multiple simulated versions of an ideal data set that is complete and error-free. The ideal data sets are then analysed separately and combined using the same rules as for MI. Xie and Paik (1997) discuss multiple imputation methods for missing covariates in generalized estimating equation models.

6.4.5 Genotyping Errors

Errors in predictors is yet another issue distinct from the concern of incomplete covariate information. In general, conventional parametric and non-parametric regression techniques are no longer valid when errors in the predictors are expected. This jeopardises family-based association tests that rely on functionally relating trait and marker data which are important, for example, when estimating genetic effect size in the screening procedures implemented in the PBAT software (Van Steen and Lange 2005). Also the comparison between transmitted and non-transmitted alleles itself, as in TDT tests and derivatives thereof, is not guarded against genotyping errors, with biased analysis results as a consequence.

Genetic information on available family members may help to detect and correct Mendelian genotyping errors, i.e., those that are inconsistent with Mendelian patterns inheritance. It is estimated that Mendelian errors underestimate genotyping errors by a factor 4. Thus, if the data show 20 Mendelian errors in 10 markers typed in 400 trios, the Mendelian error rate of 20/4000=0.5% translates roughly to a genotype error rate of 2% (Douglas *et al.* 2002).

Estimation of genetic effect is often based on regression models relating trait to marker data, for which a vast amount of remedial strategies exist to deal with error in the predictors. Kuechenhoff and Carroll (1997) discuss the estimation of parameters in a particular segmented generalized linear model with additive measurement error in predictors, with a focus on linear and logistic regression. Two viewpoints, regression calibration and simulation extrapolation, are considered. Regression calibration involves the calculation of expected values, and hence an underlying distribution for the error-prone predictors. One possible assumption is a distribution of a mixture of normals with an unknown number of components, and to use regression splines. Simulation-extrapolation (SIMEX) does not require any assumptions about the distribution of the unobserved error-prone predictor. The SIMEX method also serves

its purpose in variance component analyses while testing for zero variance components, or correlation within clusters and heterogeneity across clusters, in the setting of generalized linear mixed measurement error models (Lin and Carroll 1999).

6.4.6 Haplotype Analysis

All missing data aspects previously mentioned also apply when haplotype markers formed by several adjacent loci are considered. Haplotypes may reduce the dimension of association tests. Another advantage of using haplotypes in association analysis is often improved power to detect disease-associated loci. The major disadvantage of haplotypes is that they are seldom known, even when the genotypes are observed. Although haplotyping can be done at the molecular level (Douglas *et al.* 2001 - whole genome derived haplotypes) it remains rather expensive. The costs can be substantially reduced by applying haplotype-reconstructive algorithms. There are several methods to infer haplotypes from available data. Schaid *et al.* (2002) mention the following broad classes to account for ambiguous haplotypes: (i) parsimony algorithms (Clark 1990), (ii) Bayesian population genetic model based on coalescent theory (Stephens *et al.* 2001; Zhang *et al.* 2001; Stephens and Donelly 2003; Stephens and Scheet 2005) and (iii) Maximum Likelihood-based methods (Terwilliger and Ott 1994; Excoffier and Slatkin 1995; Hawley and Kidd 1995; Long *et al.* 1995).

Another approach for testing haplotypes with missing phase uses conditioning arguments. The conditioning principle as applied to family-based tests, for complete parental genotype data and phased genotype data, is described in Lazzeroni and Lange (1998). When both parents' genotypes are known, Dudbridge *et al.* (2000) propose a test that is based on a conditioning approach similar to one developed for a single marker in the general case of missing parents (Knapp 1999, Rabinowitz and Laird 2000). For the FBAT statistic applied to haplotype markers, haplotype frequencies are estimated using the parents in nuclear families. An EM algorithm is used to include founders with missing haplotype information or to resolve phase. Phase-unknown subjects are included in the evaluation of the FBAT test statistic using a set of weights assigned to the possible phased genotypes that are consistent with any ambiguous unphased genotype (Horvath *et al.* 2004). Hence, this method combines a conditioning argument with the weighting approach of Zhao *et al.* (2000).

The EM algorithm of Dempster *et al.* (1977) indeed remains an attractive means to estimate parameters under an incomplete data model, whenever a likelihood

approach is feasible. Since the method was first mentioned, many actions have been undertaken to improve the EM algorithm, such as the Louis method (Louis 1982) to obtain consistent estimators of the standard errors of the maximum likelihood estimates. As Meilijson (1989) reports, gradient methods do provide estimates, but suffer the drawback that an explicit expression of the score function is needed. Therefore, Meilijson combines aspects of both the EM algoritm and Newton-type methods to obtain a fast improvement to the classic EM. Additional modifications of the classic EM algorithm include the ECM algorithm of Meng and Rubin (1993) and ECME algorithm of Liu and Rubin (1994), or evolving from reversing the E and M step to stochastic EM algorithms (e.g., Marschner 2001). Marschner shows that the proposed modification of the stochastic EM algorithm outperforms the usual stochastic EM algorithm and the ML estimator in small sample simulation studies of standard censoring and mixture problems.

For a thorough review of both population-based and family-based haplotype inference methods and handling genotype score uncertainty in inferring haplotype inference, refer to Niu (2004). Alternatively, a multiple imputation technique can be used to allow multiple imputation of individual haplotypes, such as the one implemented in the SNPHAP programme (written by David Clayton).

6.5 Non-Classical Methods to Account for Missing Data

The issue of MNAR mechanism causes considerable controversy, even among missing data researchers. The main reason is that MNAR is a severe assumption regarding incomplete data which cannot be verified on a purely statistical basis. One school advocates MNAR models of ever increasing complexity ("supermodel fallacy") while another school strongly supports sensitivity analyses. However, there is an erroneous assumption that the more complex a model is the less problematic sensitivity becomes, frequently referred to as the supermodel fallacy. The mitigation of the sensitivity issue is fundamentally impossible to attain in this manner mainly due to the lack of parameter identification (Molenberghs *et al.* 1999). In a sensitivity analysis, several models are proposed and the impact of assumptions on conclusions are investigated (Verbeke and Molenberghs 2000; Thijs *et al.* 2001; Verbeke *et al* 2001; Molenberghs *et al.* 2001; Kenward *et al.* 2001; Scharfstein *et al.* 1999; Vach and Blettner 1995; Raab and Donnelly 1999). Multiple settings are possible, such as parametric versus semi-parametric or selection models versus pattern mixture

models (Thijs *et al.* 2001).

An overview of currently available approaches in biostatistics to deal with missing data is given in Table 6.1. Many good textbooks exist that provide more insight into these strategies.

Obviously, family-based association testing also benefits from sensitivity analysis, e.g., it is important to investigate the sensitivity of the haplotype-based analysis for different haplotype reconstruction methods. While a general awareness of the need for sensitivity analysis has grown and a few proposals have been made, many of these are considered useful but are still ad hoc procedures. Any approach should ideally identify and incorporate both sources of uncertainty: imprecision due to finite sampling and ignorance due to incompleteness.

6.6 I can't see the Black Hole ... It's Missing

In this chapter, we identified several occurrences of incomplete data in family-based association testing. Many techniques developed in biostatistics to handle incomplete data carry through to statistical genetics. For instance, in the context of regression-based score tests it is possible to use standard theory for likelihoods in missing data problems (Little and Rubin 1987) to extend the expression for the components of the score statistic (e.g., as in Schaid *et al.* 2002). Nevertheless, it is important to realise that missing data problems generally differ from those encountered in the rest of biostatistics in the sense that data do not simply split into complete and incomplete records. Moreover, a clear distinction should be made between genetic association modelling and genetic association testing.

Some would advocate that just about *all* of genetic statistics is about missing data problems: The actual genotypes of the putative locus are always unknown, even when the phenotypes and marker genotypes are known. It may be clear that the choice of analysis technique in the presence of incomplete data requires careful reflection. Clearly, there is room for further assessment of the effects of missingness or measurement error on the test results. One such way is with a sensitivity analysis. The existing approaches to deal with missing data or erroneous measurements in statistical genetics in general and family-based association testing in specific (such as EM-based algorithms or multiple imputation techniques) are constantly being improved. The future will reveal to which extent more novel approaches, such as causal inference, can have an impact on the field of statistical genetics.

Table 6.1: Standard and less commonly used approaches to deal with missingness in biostatistics.

Classical Methods under MCAR (Little and Rubin 2002)
- Complete case analysis
- Simple imputation: unconditional mean imputation,
conditional mean imputation (e.g., Buck 1960),
hot deck imputation (e.g., Ford 1983) and
cold deck imputation (e.g., Shao 2000)
- Incomplete data analysis: frequentist available case analysis
Classical Methods under non-MCAR (Little and Rubin 2002)
- Multiple imputation (Rubin 1987; Rubin and Schenker 1991; Schafer 1997;
Horton and Lipsitz 2001)
- Data augmentation (overview: van Dyck and Meng 2001)
- Incomplete data analysis:
likelihood-based available case analysis,
pseudo-likelihood (Le Cessie and Van Houwelingen 1994) and
estimating equations (Liang and Zeger 1986)
Non-classical Methods
- Sufficient statistics (Dynkin 1951)
- Supermodel fallacy and sensitivity analysis (Verbeke and Molenberghs 2000)
- Causal inference (Pearl 2000)
- Local influence (Cook 1986; Verbeke and Molenberghs 2000; Verbeke *et al.* 2001;
Molenberghs and Verbeke 2005)
- Interval of ignorance (Molenberghs *et al.* 2001; Vansteelandt *et al.* 2005)
- Modelling framework specific analysis:
selection (Glynn *et al.* 1986),
pattern mixture (Thijs *et al.* 2001) and
shared parameter models (Wu and Carroll 1988)

Conclusion

Chapter 7

Through the Looking Glass

7.1 Gene-Gene Interactions

A central problem in modelling interactions is the curse of dimensionality (Bellman 1961), which refers to the exponential growth of hypervolume as a function of dimensionality. Translated to the aim of disentangling the complex mix of high-order gene-gene interactions, modelling these interactions will involve many multilocus genotype combinations without any observed data. The analysis of sparse tables may give rise to problems related to goodness-of-fit testing, since the asymptotic approximations of standard chi-square statistics are no longer optimal. Another type of problems relates to parameters and their associated standard errors that are non-existent ($\pm\infty$) or highly biased. Although these problems can be solved (e.g., by adopting a Bayesian approach as in Smith and Queen 1996), we still need to encompass in our strategy an extensive space of epistatic models of which the contributing loci display no (or very small) main effects. A traditional parametric approach that computes estimates of population parameters may therefore not be the road to travel by (Moore and Ritchie 2004).

Because of the large number of potential genotypes when looking at multiple loci at once, analysing sets of SNPs is far more tedious than analysing each SNP separately. Ideally, a more manageable data set is obtained first. Perhaps the most relevant question in this context is, how can we optimally reduce the data so that a maximal amount of information is retained? One way to do so, is by implementing a data-mining technique such as the MDR approach of Ritchie *et al.* (2001, 2003) for population-based designs, as mentioned in Chapter 2. To date, the technique

has not been generalised to quantitative trait analysis, nor elaborate family-based designs. Alternatively, Bayesian networks are used to describe the dependencies of genotype on phenotype (Sebastiani *et al.* 2005). It appears that conditioning on phenotype reduces the complexity of SNP/phenotype associations and can identify larger clusters of such associations (Hoh and Ott 2003). These clusters can than be targeted in a subsequent family-based association analysis, modifying the screening tools of Chapter 5 to target epistatic effects on the trait(s) of interest.

7.2 Gene-Environment Interactions

The study of gene-environment interactions not only provides useful information on biological pathways, it is also an important aid when developing new therapeutic strategies. One important environmental agent in the aetiology of a disease is drug therapy. Many drugs are currently believed to be affected by polymorphic genes. Studying the effect of genetic variability on treatment response or risk of serious adverse reactions to drugs is the subject of pharmacogenetics. One of the major potentials of pharmacogenetics is providing the basis for matching a patient to a particular treatment therapy. For a review on gene-environment interactions in human diseases, we refer to Hunter (2005).

Asthma is a disease of chronic airway inflammation characterised by reversible airway obstruction and increased airway responsiveness (NIH 1995, 1997; Warner and Naspitz 1998). According to the World Health Organization the number of asthma patients is growing by 50% every decade. Estimates suggest that over 150 million people worldwide have asthma. Asthma has become an epidemic .... Early events in life seem to be crucial for the development of the disease (Weiss 1998; Cookson 1999). It is believed that up to $60 - 80\%$ of the variations in treatment responses are due to genetic factors (Drazen *et al.* 2000). The list of pharmacogenetic traits that lead to a differential response to drug treatment has expanded dramatically as a result of the Human Genome Project and its spin-offs such as the identification of single nucleotide polymorphisms (Murphy 2002).

The issue of asthmatic non-responders is a problem that is becoming increasingly recognised in the treatment of asthma. It is therefore important to develop methods for understanding the genetic components that are responsible for (non-)responder asthma phenotypes (Palmer *et al.* 2002). In general, there are 3 classes of drugs used to treat asthma (Sayers and Hall 2005): Leukotriene modifiers (reducing

leukotriene effects in the lungs), β_2 agonists (inhalers for quick relief of symptoms or long term bronchodilator therapy) and inhaled steroids (are used daily to reduce inflammation and symptoms). Fench and Hall (2002) and Hall (2002) give an overview of the pharmacogenetics involved in airway treatment targets. In summary, genetic variants may alter response to drugs in three main ways: (i) variation in metabolism of a drug among individuals, (ii) variation among population members regarding drug adverse effects that are not based on the drug's action, (iii) genetic variation in the drug treatment target or target pathways (Palmer *et al.* 2002).

From a clinical perspective, individuals are best tested repeatedly over time, since asthma symptoms are likely to evolve over time, potentially through different patterns (Clough 1998). From a statistical point of view, when data are collected on many related phenotypes, testing all phenotypes individually for association with the marker loci of interest and subsequently adjusting the obtained p-values for multiple testing may fail to detect true associations. Although permutation of the observed data under the null hypothesis of no linkage and no association eliminates the need to adjust p-values after multiple testing, it often remains highly computer-intensive. Multivariate tests, such as the multivariate family-based association test using generalized estimating equations (FBAT-GEE; Lange *et al.* 2003c) provide an alternative. The FBAT-GEE test can be used for repeatedly measured phenotypes without making distributional assumptions for the phenotypic observations. Using generalized principal component analysis and amplifying the genetic effects of each measurement by constructing a new overall phenotype with maximal heritability, FBAT-PC has been shown to be of practical relevance when analysing repeatedly measured quantitative traits (Lange *et al.* 2004b). The test is particularly useful when covariates are partially or completely unknown and therefore cannot be modelled adequately.

Genome-wide association studies using single nucleotide polymorphisms have been proposed as a more powerful strategy for detecting genetic effects than linkage mapping (Risch and Merikangas 1996). Genetic association studies can aid in identifying pathway candidate genes. Once the genes that influence asthma treatment response have been determined, important sequence variants can be identified and the underlying molecular mechanisms of their effects can be characterised.

In ongoing research, we combine the advantages of performing family-based association screening with the advantages of using phenotypic assessments over time. We

extend the existing screening methodology of Lange *et al.* (2003a,b) and Van Steen *et al.* (2005) and tailor it to the field of pharmacogenetics of asthma. Moreover, by incorporating treatment data into a family-based design we can directly estimate gene-drug interactions.

Software

A large number of computer programmes are available for family-based association tests, including AFBAC (Thomson 1995), QTDT (Abecasis *et al.* 2000), FBAT (Horvath and Laird 1998; Horvath *et al.* 2000, 2001; Laird *et al.* 2000; Lake *et al.* 2000), TRANSMIT (Clayton 1999) and PDT (Martin *et al.* 2000). These software packages primarily focus on the computation of various test statistics. PBAT provides methods for a wide range of situations that arise in family-based association studies, using FBAT statistics. More specifically, there are two main components: tools for the planning of family-based association studies and data analysis tools. The latter include the screening tools of Chapter 5. The PBAT-software can be downloaded via the URL

http://www.biostat.harvard.edu/∼clange/default.htm.

PBAT is an interactive software package that is available for Windows XP, Linux and UNIX operating system (Lange *et al.* 2004a; Van Steen and Lange 2005). PBAT's newest version (v2.5) includes many features that were not available in earlier versions (Lange *et al.* 2004a), such as haplotype analysis tools that can be invoked using batch-mode or user-interface, more flexible specifications in power calculations, allowance for discrete trait distribution when applicable. In particular, PBAT incorporates the features of the FBAT package (http://www.biostat.harvard.edu/ fbat/fbat.htm) but provides many additional options for designing association/linkage studies and analysing data with multiple continuous traits. Perhaps the most striking feature, which gives PBAT a unique advantage over most available software in the field, is its implementation of the screening techniques, i.e. the conditional mean model approach (Lange *et al.* 2003a,b), that allow the user to handle the multiple comparison problem at a genome-wide level (Van Steen *et al.* 2005). Further advantages of PBAT are the analytical power and sample size calculations for family-based association tests (Lange and Laird 2002a; Lange *et al.* 2002). PBAT is especially well-suited for quantitative traits while accounting for important predictors.

Acknowledgements

Most of this work was carried out within the framework of the Belgian IUAP/PAI network "Statistical Techniques and Modelling for Complex Substantive Questions with Complex Data" (Chapters 4, 6), and supported in parts by grant MH59532 of the National Institutes of Health (Chapter 5). In addition, we thank Steve Gracon of Pfizer for the use of the ApoE data (Chapter 4). Nadia Tahri (co-author of Van Steen *et al.* 2004b) would like to thank the Association Nationale de la Recherche Technique (ANRT) for financial support.

Our publication in Nature Genetics (Van Steen *et al.* 2005; Chapter 5) would have been impossible without all the CAMP families for their enthusiastic participation in the CAMP Genetics Ancillary Study, supported by the National Heart, Lung and Blood Institute, N01-HR-16049. We also acknowledge the CAMP investigators and research team, supported by NHLBI, for collecting the CAMP Genetic Ancillary Study data. Additional support for this research came from P01 HL67664 (STW), R01 HL66386-01 and U01 HL66795-01 from the National Heart Lung and Blood Institute. Support was obtained by grant MH 59532 of the National Institutes of Health. All work undertaken from the CAMP Genetics Ancillary Study was conducted at the Channing Laboratory at the Brigham and Women's Hospital under appropriate CAMP policies and human subjects protections. M.B.M. was supported by the National Research Service Award, Training Programme in Psychiatric Epidemiology and Biostatistics (T32 MH17119).

Samenvatting

De erfelijke informatie bij de mens is in code opgeslagen in het DNA (desoxyribo nucleïnezuur), een lineaire molecule die georganiseerd is in 22 paar chromosomen en 1 paar geslachtschromosomen. Chromosomen zijn ketens van de basenparen A (adenine), C (cytosine), T (thymine) en G (guanine). In totaal bestaat het menselijk genoom uit zo'n 3 miljard van deze bouwstenen. Een deel van het erfelijk materiaal bestaat uit genen (zo'n 30.000), die de instructies bevatten om eiwitten te maken. De DNA sequenties van twee mensen, die willekeurig uit de populatie worden gekozen, zijn voor 99.9% identiek. De overige 0.1%, hoe klein ook, bevatten de genetische variaties die niet alleen bepalen hoe het individu reageert op een therapeutische interventie, maar die ook bepalen hoe groot het risico is om een medische aandoening te ontwikkelen. Posities in het menselijk genoom waar de DNA sequenties van verschillende individuen slechts 1 basepaar verschillen, worden SNPs (single nucleotide polymorphisms) genoemd. Omdat SNPs zo frequent voorkomen in menselijke populaties (ongeveer 10 miljoen, waarbij de MAF (minor allele frequency) $> 1\%$) worden ze vaak als hét instrument bij uitstek gezien om de genetische onderbouw van complexe ziekten te ontrafelen. Er bestaan verschillende technieken om genen voor complexe ziekten in kaart brengen: parametrische analyse, niet-parametrische analyse van "allele-sharing" via een koppelingsanalyse (linkage analysis), linkage disequilibrium (associatie-) analyse.

Complexe ziekten zijn aandoeningen waarvoor het eenvoudige overervingsmodel van Mendel niet van toepassing is. Mogelijke redenen hiervoor zijn (i) dat er meerdere varianten op een locus betrokken zijn bij het ziektebeeld, (ii) dat er verschillende posities op het genoom aanwezig zijn die de kans verhogen op ontwikkeling van de ziekte, (iii) gen-gen interacties, (iv) omgevingsfactoren, (v) gen-omgeving interacties. Astma is een voorbeeld van een complexe ziekte, waarin zowel genetische factoren als omgevingsfactoren een belangrijke rol spelen. Het wordt veroorzaakt door een ontsteking van de luchtwegen, die leidt tot een

overgevoeligheid voor externe prikkels zoals rook of koude. Deze prikkels kunnen dan een vernauwing van de luchtwegen veroorzaken, waardoor de astmapatient last krijgt van een piepende ademhaling of het benauwd krijgt.

In hoofdstuk 1 geven we een kort overzicht van de verschillende fasen die betrokken zijn bij de analyse van een complex fenotype en wijzen we op de rol van het Internationale HapMap Project op SNP-gebruik in genetische associatiescreenings op grote schaal. In hoofdstuk 2 beschrijven we de factoren die een rol spelen bij genetische associatiestudies in het algemeen en deze op genoomschaal in het bijzonder. Dit hoofdstuk is de bindende factor tussen hoofdstukken 4, 5, 6 en 7. Een beknopte beschrijving van de gebruikte data sets in dit proefschrift, met een motivering van hun gebruik is te vinden in hoofdstuk 3.

Er zijn verschillende kaders mogelijk waarbinnen een genetische associatiescreening van een complexe ziekte kan uitgewerkt worden. Voor welk design men ook opteert (populatie of families), de vraag blijft hoe men bij mogelijks honderdduizenden testen een globaal significantieniveau kan garanderen. Een bijkomend probleem is dat de teststatistieken die SNPs met fenotype(s) associeren veelal gecorreleerd zijn op een subtiele manier, door de complexe koppelingscorrelaties (LD) tussen verschillende posities in het menselijk genoom. Het debat over hoe ver LD kan reiken rond een mutatie is nog verre van afgelopen. Wiskundige en simulatiemodellen in overeenstemming brengen met de praktijk blijkt een huzarenstukje. In elk geval spelen de aard van de populatie, de plaats in het genoom en het type merker dat gebruikt wordt een niet onbelangrijke rol (Pritchard en Przeworski 2001).

Multicollineariteit in logistische regressie (waarbij ziektestatus versus genotype(s) worden gemodelleerd) is het resultaat van sterke correlaties tussen onafhankelijke variabelen. Multicollineariteit doet varianties van parameterschattingen exploderen, en dit kan leiden tot statistische niet-significante resultaten op individuele parameters terwijl het model in zijn geheel, met alle variabelen, toch sterk significant is. Om dit probleem het hoofd te bieden, introduceren we het Dalemodel in genetische associatiestudies met een beperkt aantal merkers (hoofdstuk 4) en beschouwen we genotype informatie (de verschillende merker datapunten) als uitkomstmaten. Dit laat ons ondermeer toe de genetische correlatiestructuur in detail te bestuderen.

Een associatie van SNP varianten tussen twee loci kan totstandkomen door puur toeval, door populatie stratificatie, of door koppeling van de loci. In wezen

is men voornamelijk geïnteresseerd in de laatste vorm van associatie. Een kritiek op case-controle studies in genetische epidemiologie is, dat ze kunnen leiden tot eerder aangehaalde onechte associaties als er populatiestratificatie aanwezig is. Om een dergelijke stratificatie op het spoor te komen, suggereerden Pritchard en Rosenberg (1999) om een groot aantal merkers in the studie mee op te nemen en deze merkers, die verondersteld worden niet gekoppeld te zijn aan merkers in de kandidaat-genregio's, te controleren op merker-fenotype associatie. Eens het probleem gedetecteerd kan men het controleren door cases met controles overeen te stemmen op bijvoorbeeld etniciteit. Blijven werken met ongerelateerde individuen kan ook. In het laatste geval zijn verschillende methodes uitgewerkt door bijvoorbeeld Devlin en Roeder (1999) en Pritchard *et al.* (2000a,b). Een alternatieve oplossing om de negatieve effecten van populatiestratificatie te vermijden, is over te stappen op familiedesigns.

De eerste gemeenschappelijke test voor koppeling en associatie was de Transmission Disequilibrium Test (TDT) van Spielman *et al.* (1993), uitgewerkt voor het trio-design (een case met ouders). Deze test, initieel geïntroduceerd als test voor koppeling alleen, gaat na of de transmissiekansen van een SNP van ouders naar geaffecteerd kind groter zijn dan 0.5. In principe gebruikt deze test dus controles gebaseerd op familie, eerder dan controles uit de algemene populatie. Hierdoor ondervangt men het probleem van ongemeten populatiestratificatie dat aanleiding kan geven tot onechte of "spurious" associaties.

Sinds zijn conceptie zijn er heel wat extensies aan de klassieke TDT toegevoegd. FBAT (Rabinowitz and Laird 2000; Laird *et al.* 2000) staat voor een ganse methodologie van familiegebaseerde associatietesten en omkadert standard genetische modellen, binaire en niet-bineaire fenotypes, multivariate fenotype-analyses met merkers die > 2 varianten kunnen hebben, ontbrekende genotype informatie voor ouders, haplotypes, uitgebreide stambomen, etc.

Bij een associatiescreening van 500.000 SNPs met een enkel fenotype, onder een significantieniveau van 1%, verwachten we 5.000 positieve testresultaten (i.e., er is koppeling en associatie) die in wezen vals zijn. Bekijken we meerdere fenotypes of merkers met meerdere varianten, dan stijgt dit aantal alleen nog maar. Traditionele correcties zoals Bonferroni-aanpassingen zijn vaak te drastisch en conservatief; Meer recent ontwikkelde aanpakken zoals deze, gebaseerd op het controleren van FDR (false discovery rate), lijken evenmin stand te houden in de context van

genome-wide genetische associatiestudies.

In Hoofdstuk 5 bekijken we alle beschikbare data, zonder een afwijking te introduceren in het significantieniveau, en selecteren we de meest optimale SNP-fenotype combinaties. De uitgewerkte methodologie voor familiegebaseerde associatiestudies is succesvol in het opsporen van genetische associaties die genoomwijde significantie garanderen. Naast het uitwerken van praktische richtlijnen bij het gebruik van de voorgestelde screeningsmethode, gaan we eveneens de bruikbaarheid na van de techniek voor het ontdekken van > 1 ziektegerelateerde locus. Een belangrijk gegeven is, dat alle data betrokken worden in het screeningsproces. Met ander woorden, de aanpak vereist geen afzonderlijke screenings- en valideringsdata, zoals het geval is voor populatiegebaseerde designs.

Zodra data worden verzameld, bestaat de kans om met ontbrekende gegevens (of foutieve gegevens) geconfronteerd te worden. In hoofdstuk 6 geven we een beschouwende visie over onvolledige gegevens in de context van familiegebaseerde genetische associatietesten. Niettegenstaande testen en modelleren een verschillende benadering vereisen, zijn Rubins strategie en taxonomie toch van toepassing op de FBAT statistieken en screeningsmethode van hoofdstuk 5 omdat eerst een datareductie via een (conditioneel) regressiemodel wordt beoogd.

Tot slot richten we onze blik op toekomstgericht onderzoek in hoofdstuk 7, in het bijzonder op gen-gen en gen-omgevings interacties. Beschouwen we een therapeutische interventie als omgevingsfactor, dan belanden we bij de farmacogenetica die de rol van erfelijkheid onderzoekt in de reactie op medicijnen: Welke genetische varianten kunnen verantwoordelijk geacht worden voor ernstige bijwerkingen op een medicijn? Welke genetische varianten zijn gekoppeld aan een verminderde therapierespons?

“Genomic screening methodology for common diseases and complex traits - Multiplicity and missingness: a statistical hurdle?” Multipliciteit en ontbrekende gegevens: niet altijd een onoverkomelijk probleem in genetische associatiestudies.

Bibliography

Aerts, M., Geys, H., Molenberghs, G., Ryan, L.M. (2002). *Topics in modelling of clustered data*, London: Chapman & Hall.

Abecasis, G.R., Cardon, L.R., Cookson, W.O. (2000) A general test of association for quantitative traits in nuclear families. *Am J Hum Genet*, **66**, 279-292.

Abecasis, G.R., Cherny, S.S., Cardon, L.R. (2001) The impact of genotyping error on family-based analysis of quantitative traits. *Eur J Hum Genet*, **9**, 130-134.

Afifi, A., Elashoff, R. (1966) Missing observations in multivariate statistics I: Review of the literature. *J Am Stat Assoc*, **61**, 595-604.

Agresti, A. (1990) *Categorical data analysis.* New York: John Wiley & Sons.

Allen, A.S., Satten, G.A., Tsiatis, A.A. (2005) Locally-efficient robust estimation of haplotype-disease association in family-based studies. *Biometrika*, **92**, 559-571.

Allison, D.B., Heo, M., Kaplan, N., Martin, E.R. (1999) Sibling-based tests of linkage and association for quantitative traits. *Am J Hum Genet*, **64**, 1754-1764.

Bacanu, S.-A., Devlin, B., Roeder, K. (2002) Association studies for quantitative traits in structured populations. *Genet Epidemiol*, **22**, 78-93.

Bahadur, R.R. (1961) A representation of the joint distribution of responses of n dichotomous items. In: *Studies in item analysis and prediction*, (Ed. H. Solomon), Stanford Mathematical Studies in the Social Sciences VI. Stanford, CA: Stanford University Press.

Bellman, R. (1961), *Adaptive control processes: a guided tour.* Princeton: Princeton University Press.

Benjamini, Y., Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Stat Soc B*, **57**, 289-300.

Benjamini, Y., Yekutieli, D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann Stat*, **29**, 1165-1188.

Blackhurst, D.W., Schluchter, M.D. (1989) Logistic regression with a partially observed covariate. *Commun Stat Simulat*, **18**, 163-177.

Boehnke, M., Langefeld, C.D. (1998) Genetic association mapping based on discordant sib pairs: the discordant alleles test (DAT). *Am J Hum Genet*, **62**, 950-961.

Bonferroni, C.E. (1936) Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, **8**, 3-62.

Bourgain, C., Genin, E., Quesneville, H., Clerget-Darpoux, F. (2000) Search for multifactorial disease susceptibility genes in founder populations. *Ann Hum Genet*, **64**, 255-265.

Breslow, N.E., Day, N.E. (1980) Statistical methods in cancer research, Vol. 1 - *The analysis of case-control studies*, IARC Sci Publ, Lyon, France.

Buck, S.F. (1960) A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *J Roy Stat Soc B*, **22**, 302-306.

Carey, V.C., Zeger, S.L., Diggle, P.J. (1993) Modelling multivariate binary data with alternating logistic regressions. *Biometrika*, **80**, 517-526.

Carlson, C., Eberle, M., Rieder, M., Yi, Q., Kruglyak, L., Nickerson, D. (2003) Haplotypes and informative SNP selection algorithms: Don't block out information. *Proceedings of the Seventh Annual International Conference on Computational Molecular Biology*, 19-27.

Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X.C., Stern, D., Winkler, J., Lockhart, D.J., Morris, M.S., Fodor, S.P.A. (1996) Accessing genetic information with high density DNA arrays. *Science*, **274**, 610-614.

Childhood Asthma Management Program Research Group (1999) The childhood asthma management program (CAMP): design, rationale, and methods. *Control Clin Trials*, **20**, 91-120.

Childhood Asthma Management Program Research Group (2000) Long-term effects of budesonide or nedocromil in children with asthma. *New Engl J Med*, **343**, 1054-1063.

Churchill, G.A., Doerge, R.W. (1994) Empirical threshold values for quantitative trait mapping. *Genetics*, **138**, 963-971.

Churchill, G.A., Doerge, R.W. (1996) Permutation tests for multiple loci affecting a quantitative character. *Genetics*,**142**, 285-294.

Clark, A.G. (1990) Inference of haplotypes from PCR-amplified samples of diploid populations. *Mol Biol Evol*, **7**, 111-122.

Clayton, D. (1999) A Generalization of the Transmission/disequilibrium test for uncertain-haplotype transmission. *Am J Hum Genet*, **65**, 1170-1177.

Clayton, D., Jones, H. (1999) Transmission/disequilibrium tests for extended marker haplotypes. *Am J Hum Genet*, **65**, 1161-1169.

Clough JB (1998) Phenotype stability in asthma and atopy in childhood. *Clin Exp Allergy*, **28**, 22-25.

Cochran, W.G. (1977) *Sampling techniques*, New York: Wiley & Sons.

Collins, F.S., Green, E.D., Guttmacher, A.E., Guyer, M.S. (2003) A Vision for the future of genomics research. *Nature*, **24**, 835.

Cook, R.D. (1986) Assessment of local influence. *J Roy Stat Soc B*, **48**, 133-169.

Cookson, W. (1999) The alliance of genes and environment in asthma and allergy. *Nature*, **402**, Suppl 25, B5-B11.

Corbex, M., Poirier, O., Fumeron, F., Betoulle, D., Evans, A., Ruidavets, J.B., Arveiler, D., Luc, G., Tiret, L., Cambien, F. (2000) Extensive association analysis between the CETP gene and coronary heart disease phenotypes reveals several putative functional polymorphisms and gene-environment interaction. *Genet Epidemiol*, **19**, 64-80.

Cordell, H.J., Clayton, D.G. (2002) A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to HLA in type 1 diabetes. *Am J Hum Genet*, **70**, 124-141.

Corder, E.H., Saunders, A.M., Strittmatter, W.J., Schmechel, D.E., Gaskell, P.C., Small, G.W., Roses, A.D., Haines, J.L., Pericak-Vance, M.A. (1993) Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science*, **261**, 921-923.

Couzin, J. (2002) New mapping project splits the community. *Genomics*, **296**, 1391-1392.

Cox, D.R., Oaks, D. (1984) *Analysis of survival data*, London: Chapman & Hall.

Cunningham, J.M., McDonnell, S.K., Marks, A., Hebbring, S., Anderson, S.A., Peterson, B.J., Slager, S., French, A., Blute, M.L., Schaid, D.J., Thibodeau, S.N. (2003) Genome linkage screen for prostate cancer susceptibility loci: results from the Mayo Clinic Familial Prostate Cancer Study. *Prostate*, **57**, 335346L.

Curtis, D., Sham, P.C. (1995) A note on the application of the transmission/disequilibrium test when a parent is missing. *Am J Hum Genet*, **56**, 811-812.

Czika, W.A., Weir, B.C., Edwards, S.R., Thompson, R.W., Nielsen, D.M., Brocklebank, J.C., Zinkus, C., Martin, E.R., Hobler, K.E. (2001) Applying data mining techniques to the mapping of complex disease genes. *Genet Epidemiol* **21**, Suppl 1, S435-S440.

Dale, J.R. (1986) Global odds ratio models for bivariate, discrete, ordered responses. *Biometrics*, **42**, 909-917.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J Roy Stat Soc B*, **39**, 1-38.

Devlin, B., Roeder, K. (1999) Genomic control for association studies. *Biometrics*, **55**, 997-1004.

Didelez, V. (2002) ML- and semiparametric estimation in logistic models with incomplete covariate data. *Stat Neerl*, **56**, 330-345.

Diggle, P.J., Kenward, M.G. (1994) Informative drop-out in longitudinal data analysis (with discussion). *Appl Stat*, **43**, 49-93.

Diggle, P.J., Heagerty, P., Liang, K.Y., Zeger, S.L. (2002) *Analysis of longitudinal data*, Oxford: Clarendon Press.

Douglas, J.A., Boehnke, M., Gillanders, E., Trent, J.M., Gruber, S.B. (2001) Experimentally-derived haplotypes substantially increase the efficiency of linkage disequilibrium studies. *Nat Genet*, **28**, 361-364.

Douglas, J.A., Skol, A.D., Boehnke, M. (2002) Probability of detection of genotyping errors and mutations as inheritance inconsistencies in nuclear-family data. *Am J Hum Genet*, **70**, 487-495.

Drazen, J.M., Silverman, E.K., Lee, T.H. (2000) Heterogeneity of therapeutic responses in asthma. *Brit Med Bull*, **56**, 1045-1070.

Drysdale, C.M., McGraw, D.W., Stack, C.B., Stephens, J.C., Judson, R.S., Nandabalan K., Arnold, K., Ruano, G., Liggett, S.B. (2000) Complex promotor and coding region 2-adrenergic receptor haplotypes alter receptor expression and predict in vivo responsiveness. *Proc Natl Acad Sci USA*, **97**, 10483-10488.

Dudbridge, F., Koeleman, B.P.C., Todd, J.A., Clayton, D.G. (2000) Unbiased application of the transmission/disequilibrium test to multilocus haplotypes. *Am J Hum Genet*, **66**, 2009-2012.

Duffy, D.L., Martin, N.G., Battistutta, D., Hopper, J.L., Mathews, J.D. (1990) Genetics of asthma and hay fever in Australian twins. *Am Rev Respir Dis*, **142**, 1351-1358.

Dynkin, E.B., (1951) Necessary and sufficient statistics for a family of probability distributions. English translation in *Selected Translations in Mathematical Statistics and Probability*, **1**, 23-41.

Ewens, W.J., Spielman, R.S. (1995) The transmission/disequilibrium test: history, subdivision, and admixture. *Am J Hum Genet*, **57**, 455-564.

Excoffier, L., Slatkin, M. (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol*, **12**, 921-927.

Fahrmeir, L., Tutz, G. (1994, 2002) *Multivariate statistical modelling based on generalized linear models.* Heidelberg: Springer-Verlag.

Falconer, D.S., Mackay, T.F.C. (1996) *Introduction to quantitative genetics.* London: Longmann & Co.

Fallin, D., Cohen, A., Essioux, L., Chumakov, I., Blumenfeld, M., Cohen, D., Schork, N. (2001) Genetic analysis of case/control data using estimated haplotype frequencies: application to APOE locus variation and Alzheimer's disease. *Genome Res*, **11**, 143-151.

Farrer, L.A., Cupples, L.A., Haines, J.L., Hyman, B., Kukull, W.A., Mayeux, R. Myers, R.H., Pericak-Vance, M.A., Risch, N., van Duijn, C.M. (1997) Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease. A meta-analysis. APOE and Alzheimer Disease Meta Analysis Consortium. *J Amer Med Assoc*, **278**, 1349-56.

Fench, A., Hall, I.P. (2002) Pharmacogenetics of asthma. *J Clin Pharmacol*, **53**, 3-15.

Fitzmaurice, G.M., Laird, N.M., Rotnitzky, A. (1993) Regression models for discrete longitudinal responses. *Stat Sci*, **8**, 284-309.

Ford, B.L. (1983) An overview of hot-deck procedures. In : *Incomplete data in sample surveys* (Eds. Madow, W.G., Olkin, I., Rubin, D.B.), New York: Academic Press, 185-207.

Gauderman, W.J., Witte, J.S., Thomas, D.C. (1999) Family-based association studies. *J Natl Cancer I* , **26**, 31-37.

GAUSS for Windows NT/95 Version 3.2.32 (Dec 19, 1997). Copyright 1984-1997 Aptech Systems, Inc. Maple Valley, WA.

Geys, H., Molenberghs, G., Ryan, L.M. (1997) Pseudo-likelihood inference for clustered binary data. *Comm Stat Theory*, **26**, 2743-2767.

Geys, H., Molenberghs, G., Lipsitz, S.R. (1998) A note on the comparison of pseudo-likelihood adn generalized estimating equations for marginal odds ratio models. *J Am Stat Assoc*, **94**, 734-745.

Geys, H., Molenberghs, G., Ryan, L.M. (1999) Pseudo-likelihood modeling of multivariate outcomes in developmental toxicology. *J Am Stat Assoc*, **94**, 734-745.

Ghosh-Dastidar, B., Schafer, J.L. (2003) Multiple edit/multiple imputation for continuous multivariate survey data. *J Am Stat Assoc*, **98**(464).

Glynn, R.J., Laird, N.M., Rubin, D.B. (1986) Selection modeling versus mixture modeling with nonignorable nonresponse. In: *Drawing Inferences from Self-Selected Samples* (Ed. Wainer, H.), New York: Springer-Verlag, 115-142.

Good, P.I. (2000) *Permutation test: a practical guide to resampling methods for testing hypotheses.* Berlin: Springer-Verlag.

Gordon, D., Heath, S.C., Liu, X., Ott, J. (2001) A transmission/disequilibrium test that allows for genotyping errors in the analysis of single-nucleotide polymorphism data. *Am J Hum Genet*, **69(2)**, 371-380.

Gordon, D., Finch, S.J. (2005) Factors affecting statistical power in the detection of genetic association. *J Clin Invest*, **115**, 1408-1418.

Göring, Terwilliger (2000a) Linkage analysis in the presence of errors I: complex-valued recombination fractions and complex phenotypes. *Am J Hum Genet*, **66**, 1095-1106.

Göring, Terwilliger (2000b) Linkage analysis in the presence of errors II: marker-locus genotyping errors modeled with hypercomplex recombination fractions. *Am J Hum Genet*, **66**, 1107-1118.

Göring, Terwilliger (2000c) Linkage analysis in the presence of errors III: marker-loci and their map as nuisance parameters. *Am J Hum Genet*, **66**, 1298-1309.

Göring, Terwilliger (2000d) Linkage analysis in the presence of errors IV: joint pseudomarker analysis of linkage and/or linkage disequilibrium on a mixture of pedigrees and singletons when the mode of inheritance cannot be accurately specified. *Am J Hum Genet*, **66**, 1310-1327.

Hahn, L.W., Ritchie, M.D., Moore, J.H. (2003) Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. *Bioinformatics*, **19**, 376-382.

Hall, I.P. (2002) Pharmacogenetics, pharmacogenomics and airway disease. *Respir Res*, **3**, 10-16.

Hartley, H.O., Hocking, R. (1971) The analysis of incomplete data. *Biometrics*, **27**, 7783-7808.

Hastie, T., Tibshirani, R., Friedman, J.H. (2001) *The elements of statistical learning: data mining, inference, and prediction.* New-York: Springer-Verlag.

Hawley, M.E., Kidd, K.K. (1995) HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. *J Hered*, **86**, 409-411.

Heitjan, D.F., Rubin, D.B. (1991) Ignorability and coarse data. *Ann Stat*, **19**, 2244-2253.

Hirschhorn, J.N., Daly, M.J. (2005) Genome-wide association studies for common diseases and complex traits. *Nat rev genet*, **6**, 95-108.

Hochberg, Y., Tamhane, A.C. (1987) *Multiple comparison procedures.* New York: John Wiley & Sons.

Hochberg, Y. (1988) A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, **75**, 800-802.

Hoffjan, S., Ober, C. (2002) Present status on the genetic studies of asthma. *Curr Opin Immunol*, **14**, 709-717.

Hoffjan, S., Nicolae, D., Ober, C. (2003) Association studies for asthma and atopic diseases: a comprehensive review of the literature. *Respir Res*, **4**, 14.

Hoh, J., Wille, A., Zee, R., Lindpaintner, K., Ott, J. (2000) Selecting SNPs in two-stage analysis of disease association data: A model-free approach. *Ann Hum Genet*, **64**, 413-417.

Hoh, J., Wille, A., Ott, J. (2001) Trimming, weighting and grouping SNPs in human case-control association studies. *Genome Res*, **11**, 2115-2119.

Hoh, J., Ott, J. (2003) Mathematical multi-locus approaches to localizing complex human trait genes. *Nat Rev Genet*, **4**, 701-709.

Holm, S. (1979) A simple sequentially rejective multiple testing procedure. *Scand J Stat*, **6**, 65-70.

Horikawa, Y., Oda, N., Cox, N.J., Li, X., Orho-Melander, M., Hara, M., Hinokio, Y., Lindner, T.H., Mashima, H., Schwarz, P.E., del Bosque-Plata, L., Horikawa, Y., Oda, Y., Yoshiuchi, I., Colilla, S., Polonsky, K.S., Wei, S., Concannon, P., Iwasaki, N., Schulze, J., Baier, L.J., Bogardus, C., Groop, L., Boerwinkle, E., Hanis, C.L., Bell, G.I. (2000) Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. *Nat Genet*, **26**, 163-175.

Horton, N.J., Laird, N.M. (1998) Maximum likelihood analysis of generalized linear models with missing covariates. *Stat Methods Med Res*, **8**, 37-50.

Horton, N.J., Laird, N.M. (2001) Maximum likelihood analysis of logistic regression models with incomplete covariate data and auxiliary information. *Biometrics*, **57**, 34-42.

Horton, N.J., Lipsitz, S.R. (2001) Multiple imputation in practice: comparison of software packages for regression models with missing variables. *Am Stat Assoc*, **35**, 244-254.

Horvath, S., Laird, N.M. (1998) A discordant-sibship test for disequilibrium and linkage: no need for parental data. *Am J Hum Genet*, **63**, 1886-1897.

Horvath, S., Laird, N.M., Knapp, M. (2000) The transmission/desequilibrium test (TDT) and parental genotype reconstruction for X-chromosomal markers. *Am J Hum Genet*, **66**, 1161-1167.

Horvath, S., Xu, X., Laird, N.M. (2001) The family-based association test method: strategies for studying general genotype-phenotype associations. *Eur J Hum Genet*, **9**, 301-306.

Horvath, S., Xin, X., Lake, S.L., Silverman, E.K., Weiss, S.T., Laird, N.M. (2004) Family-based tests for associating haplotypes with general phenotype data: application to asthma genetics. *Genet Epidemiol*, **26**, 61-69.

Hunter, D.J. (2005) Gene-environment interactions in human diseases. *Nat Rev Genet*, **6**, 287-298.

Ibrahim, J.G. (1990) Incomplete data in generalized linear models. *J Am Stat Assoc*, **85**, 765-769.

Ibrahim, J.G., Weisberg, S. (1992) Incomplete data in generalized linear models with continuous covariates. *Aust J Stat*, **34**, 461-470.

The International HapMap Consortium. (2003) *Nature*, **426**, 789-796.

The International HapMap Consortium. (2005) *Nature*, **437** 1299 - 1320.

International SNP Map Working Group (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, **409**, 928-933.

Jansen, I., Molenberghs, G., Aerts, M., Thijs, H., Van Steen, K. (2003) A local influence approach applied to binary data from a psychiatric study. *Biometrics*, **59**, 410-419.

Jiang, R., Duan, J., Windemuth, A., Stephens, J.C., Judson, R., Xu, C. (2003) Genome-wide evaluation of public SNP databases. *Pharmacogenomics*, **6**, 779-789.

Jones, M.P. (1996) Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Jam Stat Assoc*, **91**, 222-230.

Joossens, S., Vermeire, S., Van Steen, K., Godefridis, L., Claessens, G., Pierik, M., Vlietinck, R., Aerts, R., Rutgeerts, P., Bossuyt, X. (2004) Pancreatic autoantibodies in inflammatory bowel disease. *Inflamm Bowel Dis*, **10**, 771-777.

Kennedy, G.C., Matsuzaki, H., Dong, S., Liu, W.M., Huang, J., Liu, G., Su, X., Cao, M., Chen, W., Zhang, J., Liu, W., Yang, G., Di, X., Ryder, T., He, Z., Surti, U., Phillips, M.S., Boyce-Jacino, M.T., Fodor, S.P., Jones, K.W. (2003) Large-scale genotyping of complex DNA. *Nat Biotechnol*, **21**, 1233-1237.

Kenward, M.G., Lesaffre, E., Molenberghs, G. (1994) An application of maximum likelihood and estimating equations to the analysis of ordinal data from a longitudinal study with cases missing at random. *Biometrics*, **50**, 945–953.

Kenward, M.G. (1998) Selection models for repeated measurements with nonrandom dropout: an illustration of sensitivity. *Stat Med*, **17**, 2723-2732.

Kenward, M.G., Molenberghs, G. (1998) Likelihood-based frequentist inference when data are missing at random. *Stat Sci*, **12**, 236-247.

Kenward, M.G., Goetghebeur, E.J.T., Molenberghs, G. (2001) Sensitivity analysis of incomplete categorical data. *Stat Model*, **1**, 31-48.

Knapp, M., Knopman, D., Solomon, P., Pendlebury, W., Davis, C., Gracon, S., Tacrine Study Group (1994) A 30-week randomized controlled trial of high-dose tacrine in patients with Alzheimer's disease. *J Am Med Assoc*, **271**, 985-991.

Knapp, M. (1999) The transmission/disequilibrium test and parental-genotype reconstruction: the reconstruction-combined transmission/disequilibrium test. *Am J Hum Genet*, **64**, 861-870.

Knowler, W.C., Williams, R.C., Pettitt, D.J., Steinberg, A.G. (1988) Gm3-5,13,14 and type 2 diabetes mellitus - an association in American Indians with genetic admixture. *Am J Hum Genet*, **43**, 520-526.

Kraft, P. (2005) Efficient two-stage genome-wide association designs based on false positive report probabilities. (work in progress)

Kruglyak, L., Nickerson, D.A. (2001) Variation is the spice of life. *Nat Genet*, **27**, 234-236.

Krzanowski, W.J. (1988) *Principles of multivariate analysis - A users's perspective.* Oxford: Clarendon Press.

Kuechenhoff, H., Carroll, R.J. (1997) Segmented regression with errors in predictors: semi-parametric and parametric methods. *Stat Med*, **16**, 169-188.

Laird, N.M., Horvath, S., Xu, X. (2000) Implementing a unified approach to family-based tests of association. *Genet Epidemiol*, **19**, Suppl 1, S36-S42.

Laird, N.M., Kraft, P., Lange, C., Van Steen, K. (2005) Testing for association in genetic studies. In: *Respiratory Genetics*, (Eds. E.K. Silverman, S.D. Shapiro, D.A. Lomas, S.T. Weiss), Oxford: Oxford University Press.

Lake, S.L., Blacker, D., Laird, N.M. (2000) Family-based tests of association in the presence of linkage. *Am J Hum Genet*, **67**, 1515-1525.

Lange, C., DeMeo, D.L., Laird, N.M. (2002) Power calculations for a general class of family-based association tests: quantitative traits. *Am J Hum Genet*, **71**, 1330-1341.

Lange, C., Laird, N.M. (2002a) Power calculations for a general class of family-based association tests: Dichotomous traits. *Am J Hum Genet*, **71**, 575-584.

Lange, C., Laird, N.M. (2002b) On a general class of conditional tests for family-based association studies in genetics: the asymptotic distribution, the conditional power, and optimality considerations. *Genet Epidemiol*, **23**, 165-180.

Lange, C., DeMeo, D.L., Silverman, E., Weiss, S., Laird, N.M. (2003a) Using the noninformative families in family-based association tests: a powerful new testing strategy. *Am J Hum Genet*, **73**, 801-811.

Lange, C., Lyon, H., DeMeo, D.L., Raby, B.A., Silverman, E., Weiss, S. (2003b) A new powerful non-parametric two-stage approach for testing multiple phenotypes in family-based association studies. *Hum Hered*, **56**, 10-17.

Lange, C., Silverman, E.K., Xu, X., Weiss, S.T., Laird, N.M. (2003c) A multivariate family-based association test using generalized estimating equations: FBAT-GEE. *Biostatistics*, **4**, 195-206.

Lange, C., DeMeo, D.L., Silverman, E.K., Weiss, S.T., Laird, N.M. (2004a) PBAT: Tools for family-based association studies. *Am J Hum Genet*, **74**, 367-369.

Lange, C., Van Steen, K., Andrew, T., Lyon, H., DeMeo, D.L., Raby, B., Murphy, A., Silverman, E.K., MacGregor, A., Weiss, S.T., Laird, N.M. (2004b) A family-based association test for repeatedly measured quantitative traits adjusting for unknown environmental and/or polygenic effects. *Statistical Applications in Genetics and Molecular Biology [online]*, **3**, Article 17.

Lazzeroni, L.C., Lange, K. (1998) A conditional inference framework for extending the transmission/disequilibrium test. *Hum Hered*, **48**, 67-81.

Le Cessie, S., Van Houwelingen, J.C. (1994) Logistic regression for correlated binary data, *Appl Stat*, **43**, 95-108.

Lee, W.C. (2002) Transmission/disequilibrium test when neither parent is available in some families: a non-iterative approach. *J Cancer Epidemiol Prev*, **7**, 97-103.

Leong, T., Ibrahim, J.G., Lipsitz, S.R. (1999) Using missing data methods in genetic studies. *Stat Med*, **18**, 473-485.

Leong, T., Lipsitz, S.R. and Ibrahim, J.G. (2001) Incomplete covariates in the Cox model with applications to biological marker data. *Appl Stat*, **50**, 467-484.

Lewontin, R. (1988) On measures of gametic disequilibrium. *Genetics*, **120**, 849-852.

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

Liang, K.-Y., Zeger, S.L., Qaqish, B. (1992) Multivariate regression analyses for categorical data (with discussion). *J Roy Stat Soc B*, **54**, 3-40.

Lin, S., Chakravarti, A., Cutler, D.J. (2004) Exhaustive allelic transmission disequilibrium tests as a new approach to genome-wide association studies. *Nat Genet*, **36**, 1181-1188.

Lin, X., Carroll, R.J. (1999) SIMEX variance component tests in generalized linear mixed measurement error models. *Biometrics*, **55**, 613-619.

Lipsitz, S.R., Ibrahim, J.G. (1998) Estimating equation with incomplete categorical covariates in the Cox model. *Biometrika*, **54**, 1002-1014.

Little, R.J.A., Rubin, D.B. (1987, 2002) *Statistical analysis with missing data.*, New York: John Wiley & Sons.

Little, R.J.A. (1992) Regression with missing X's: A review. *J Am Stat Assoc*, **87**, 1227-1237.

Little, R.J.A. (1993) Pattern-mixture models for multivariate incomplete data. *J Am Stat Assoc*, **88**, 125-134.

Little, R.J.A. (1995) Modeling the drop-out mechanism in repeated-measures studies. *J Am Stat Assoc*, **90**, 1112-1121.

Liu, C.H., Rubin, D.B. (1994) The ECME algorithm: a simple extension to EM and ECM with faster monotone convergence. *Biometrika*, **81**, 633-648.

Long, J.C., Williams, R.C., Urbanek, M. (1995) An E-M algorithm and testing strategy for multiple-locus haplotypes. *Am J Hum Genet*, **56**, 799-810.

Longmate, J.A. (2001) Complexity and power in case-control studies. *Am J Hum Genet*, **68**, 1229-1237.

Louis, T.A. (1982) Finding the observed information matrix when using the EM algorithm. *Ann Stat*, **22**, 326-339.

Lunetta, K.L., Faraone, S.V., Biederman, J., Laird, N.M. (2000) Family-based tests of association and linkage that use unaffected sibs, covariates, and interactions. *Am J Hum Genet*, **66**, 605-614.

Lyon, H., Lange, C., Lake, S., Silverman, E.K., Randolph, A.G., Kwiatkowski, D., Raby, B.A., Lazarus, R., Weiland, K.M., Laird, N., Weiss, S.T. (2004). IL10 gene polymorphisms are associated with asthma phenotypes in children. *Genet Epidemiol*, **26**, 155-165.

Mallinckrodt, C.H., Scott Clark, W., Carroll, R.J., Molenberghs G. (2003) Assessing response profiles from incomplete longitudinal clinical trial data with subject dropout under regulatory conditions. *J Biopharm Stat*, **13**, 179-190.

Mannino, D.M., Homa, D.M., Akinbami, L.J., Moorman, J.E., Gwynn, C., Redd, S.C. (2002) Surveillance for asthma - United States, 1980-1999. *MMWR Surveill Summ*, **51**, 1-13.

Marchini, J., Cardon, L.R., Phillips, M.S., Donnelly, P. (2004) The effects of human population structure on large genetic association structures. *Nat Genet*, **36**, 512-517.

Marnellos, G. (2003) High-throughput SNP analysis for genetic association studies. *Curr Opin Drug Di De*, **6**, 317-321.

Marschner, I.C. (2001) Miscellanea on stochastic versions of the EM algorithm. *Biometrika*, **88**, 281-286.

Martin, E.R., Kaplan, N.L., Weir, B.S. (1997) Tests for linkage and association in nuclear families. *Am J Hum Genet*, **61**, 439-448.

Martin, E.R., Monks, S.A., Warren, L.L., Kaplan, N.L. (2000) A test for linkage and association in general pedigrees: the pedigree disequilirium test. *Am J Hum Genet*, **67**, 146-154.

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

Mcqueen, M.B., Murphy, A. , Kraft, P., Su, J., Lazarus, R., Laird, N.M., Lange, C., Van Steen, K. (2006) Comparison of linkage and association strategies for quantitative traits using the COGA data set. *BMC Genet*, **0**, 000-000.

Meilijson, I. (1989) A fast improvement to the EM algorithm on its own terms. *J Roy Stat Soc B*, **51**, 127-138.

Meng, X.L., Rubin, D.B. (1993) Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, **80**, 267-278.

Mitchell, A.A., Cutler, D.J., Chakravanti, A. (2003) Undetected genotyping errors cause apparent overtransmission of common alleles in the transmission/disequilibrium test. *Am J Hum Genet*, **72(3)**, 598-610.

Molenberghs, G., Lesaffre, E. (1994) Marginal modelling of correlated ordinal data using a multivariate Plackett distribution. *J Am Stat Assoc*, **89**, 633-644.

Molenberghs, G., Kenward, M. G., Lesaffre, E. (1997) The analysis of longitudinal ordinal data with non-random dropout. *Biometrika*, **84**, 33-44.

Molenberghs, G., Goetghebeur, E.J.T., Lipsitz, S.R., Kenward, M.G. (1999) Nonrandom missingness in categorical data: strengths and limitations. *Amer Statist*, **53**, 110-118.

Molenberghs, G. and Lesaffre, E. (1999) Marginal modelling of multivariate categorical data. *Stat Med*, **18**, 2237-2255.

Molenberghs, G., Kenward, M.G., and Goetghebeur, E. (2001) Sensitivity analysis for incomplete contingency tables: the Slovenian plebiscite case. *Appl Stat*, **50**, 15-29.

Molenberghs, G., Thijs, H., Jansen, I., Beunckens, C., Kenward, M.G., Mallinckrodt, C., Carroll, R.J. (2004) Analyzing incomplete longitudinal clinical trial data. *Biostatistics*, **5**, 445-464.

Molenberghs, G., Verbeke, G. (2005) *Models for discrete longitudinal data.* New York: Springer, 683 p.: ill..- (Springer series in statistics).- ISBN 0-387-25144-8

Moore, J.H., Ritchie, M.D. (2004) The challenges of whole-genome approaches to common diseases. *JAMA*, **291**, 1642-1643.

Murphy, M.P. (2002) Pharmacogenomics: a critical component of patient stratification during drug development. *Expert Rev Mol Diagn*, **2**, 1-4.

Murhpy, A., Van Steen, K., Lange, C. (2004) On missing phenotype data in multivariate family-based association tests: FBAT-GEE-IMP and imputation strategies based on the EM-algorithm, the DA-algorithm and the conditional mean model. *Far East J Theor Stat*, **13**, 175-188.

Neter, N., Kutner, M.H., Nachtsheim, C.J., Wasserman, W. (1996) *Applied linear statistical models.* Chicago: IRWIN.

NIH publication no. 95-3659: Global strategy for asthma management and prevention: NHLBI/WHO workshop report. Bethesda, Md.: National Heart, Lung, and Blood Institute, 1995.

NIH publication no. 97-4051: National Asthma Education and Prevention Program. Expert panel report 2: guidelines for the diagnosis and management of asthma. Bethesda, Md.: National Heart, Lung, and Blood Institute, 1997.

Niu, T. (2004) Algorithms for inferring haplotypes. *Genet Epidemiol*, **27**, 334-347.

Palmer, L.J., Silverman, E.S., Weiss, S.T., Drazen, J.M. (2002) Pharmacogenetics of asthma. *Am J Resp Crit Care*, **165**, 861-866.

Palmgren, J. (1989) *Regression models for bivariate responses.* Technical Report 101, School of Public Health and Community Medicine, Dept of Biostatistics, University of Washington, Seattle.

Pearl, J. (2000) *Causality: models, reasoning, and inference*. New York: Cambridge University Press.

Pendergast, J.F., Gange, S.J., Newton, M.A., Lindstrom, M.J., Palta, M., Fisher, M.R. (1996) A survey of methods for analyzing clustered binary response data. *Int Stat Rev*, **64**, 89-118.

Plackett, R.L. (1965) A class of bivariate distributions. *J Am Stat Assoc*, **60**, 516-522.

Prentice, R.L. (1988) Correlated binary regression with covariates specific to each binary observation. *Biometrics*, **44**, 1033-1048.

Pritchard, J.K., Stephens, M., Donnelly, P. (2000a) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945-959.

Pritchard, J.K., Stephens, M., Rosenberg, N.A., Donnelly, P. (2000b) Association mapping in structured populations. *Am J Hum Genet*, **67**, 170-181.

Pritchard, J.K., Donnelly, P. (2001) Case-control studies of association in structured or admixed populations. *Theor Popul Biol*, **60(3)**, 227-237.

Pritchard, J.K., Przeworski, M. (2001). Linkage disequilibrium in humans: models and data. *Am J Hum Genet*, **69**, 1-14.

Raab, G.M., Donnelly, C.A. (1999) Information on sexual behaviour when some data are missing. *Appl Stat*, **48**, 117-133.

Rabinowitz, D., Laird, N.M. (2000) A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Hum Hered*, **50**, 211-223.

Raby, B.A., Hwang, E.-S., Van Steen, K., Tantisara, K., Peng, S., Litonjua, A, Azarus, R., Gaillourakis, C., Rioux, J., Silverman, E.K., Glimcher, L., Weiss, S.T. (2005a) T-bet polymorphisms are associated with asthma and airways hyperresponsiveness. *Am J Respir Crit Care Med*, in print.

Raby, B.A., Van Steen, K., Celedón, J.C., Litonjua, A.A., Lange, C., Weiss, S.T. for the CAMP Research Group (2005b) Paternal history of asthma and airway responsiveness in children with asthma. *Am J Respir Crit Care Med*, **172**, 552-558.

Raby, B.A., Van Steen, K., Lazarus, R., Celedón, J.C., Silverman, E., Weiss, S.T. (2005c) Eotaxin polymorphisms and serum total Immunoglobin E levels in children with asthma. *J Allergy Clin Immun*, **0**, 000-000.

Randolph, A.G., Lange, C., Silverman, E.K., Lazarus, R., Silverman, E.S., Raby, B.A., Brown, A., Ozonoff, A., Richter, B., Weiss, S.T. (2004) IL12B gene is associated with asthma. *Am J Hum Genet*, **75**, 709-715.

Reich, D.A., Lander, E.S. (2001) On the allelic spectrum of human disease. *Trends Genet*, **17**, 502-510.

Rice, J.P., Neuman, R.J., Hoshaw, S.L., Daw, E.W., Gu, C. (1995) TDT with covariates and genomic screens with mod scores: their behavior on simulated data. *Genet Epidemiol*, **12**, 659-664.

Risch, N., Merikangas, K. (1996) The future of genetic studies of complex human diseases. *Science*, **273**, 1516-1517.

Risch, N., Teng, J. (1998) The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases I. DNA pooling. *Genome Res*, **8**, 1273-1288.

Ritchie, M.D., Hahn, L.W., Roodi, N., Bailey, L.R., Dupont, W.D., Parl, F.F., Moore, J.H. (2001) Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet*, **69**, 138-147.

Ritchie, M.D., Hahn, L.W., Moore, J.H. (2003) Power of multifactor dimensionality reduction for detecting gene-gene and gene-environment interactions. *Genet Epidemiol*, **24**, 150-157.

Robins, J.M., Rotnitzky, A. (1995) Semiparametric efficiency in multivariate regression models with missing data. *J Am Stat Assoc*, **90**, 122-129.

Robins, J.M., Rotnitzky, A., and Zhao, L.P. (1995) Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J Am Stat Assoc*, **90**, 106-121.

Robins, J.M., Rotnitzky, A., and Scharfstein, D.O. (1998) Semiparametric regression for repeated outcomes with non-ignorable non-response. *J Am Stat Assoc*, **93**, 1321-1339.

Rubin, D.B. (1976) Inference and missing data. *Biometrika*, **63**, 581-592.

Rubin, D.B. (1977) Formalizing subjective notions about the effect of nonrespondents in sample surveys. *J Am Stat Assoc*, **72**, 538-543.

Rubin, D.B. (1987) *Multiple imputation for nonresponse in surveys.* New York: John Wiley & Sons.

Rubin, D.B., Schenker, N. (1991) Multiple imputation in health-care data bases: an overview and some applications, *Stat Med*, **10**, 585-598.

Ryan, T.P. (1997) *Modern regression methods.* New York: John Wiley & Sons.

SAS/STAT®*User's Guide, Version 8* (2000). Cary, NC: SAS Institute Inc.

Satagopan, J.M., Venkatraman, E.S., Begg, C.B. (2004) Two-stage designs for gene-disease association studies with sample size constraints. *Biometrics*, **60**, 589.

Saunders, A.M, Strittmatter, W.J., Schmechel, D., George-Hyslop, P.H., Pericak-Vance, M.A., Joo, S.H., Rosi, B.L., Gusella, J.F., Crapper-MacLachlan, D.R., Alberts, M.J., *et al.* (1993) Association of apolipoprotein E allele epsilon 4 with late-onset familial and sporadic Alzheimer's disease. *Neurology*, **43**, 1467-1472.

Sayers, I., Hall, I.P. (2005) Pharmacogenetic approaches in the treatment of asthma. *Curr Allergy Asthm R*, **5**, 101-108.

Schafer, J.L. (1997) *Analysis of incomplete multivariate data.* London: Chapman & Hall.

Schaid, D.J., Sommer, S.S. (1993) Genotype relative risks: methods for design and analysis of candidate-gene association studies. *Am J Hum Genet*, **53**, 1114-1126.

Schaid, D.J. (1996) General score tests for associations of genetic markers with disease using cased and their parents. *Genet Epidemiol*, **13**, 423-449.

Schaid, D.J., McDonnell, S.K., Blute, M.L., Thibodeau, S.N. (1998) Evidence for autosomal dominant inheritance of prostate cancer. *Am J Hum Genet*, **62**, 14251438.

Schaid, D.J., Rowland, C.M. (1999) Quantitative trait transmission disequilibrium test: allowance for missing parents. *Genet Epidemiol*, **17**, Suppl, S307-S312.

Schaid, D.J., Rowland, C.M., Tines, D.E., Jacobson, R.M., Poland, G.A. (2002) Score test for association between traits and haplotypes when linkage phase is ambiguous. *Am J Hum Genet*, **70**, 425-434.

Schaid, D.J., Guenther, J.C., Christensen, G.B., Hebbring, S., Rosenow, C., Cunningham, J.M., McDonnell, S.K., Slager, S., Blute, M.L., Thibodeau, S.N. (2004) Comparison of microsatellites versus single nucleotide polymorphisms by a genome linkage screen for prostate cancer susceptibility loci. *Am J Hum Genet*, **75**, 948-965.

Scharfstein, D.O., Rotnitzky, A., Robins, J.M. (1999) Adjusting for non-ignorable drop-out using semiparametric nonresponde models (with discussion). *J Am Stat Assoc*, **94**, 1096-1146.

Sebastiani, P., Abad, M.M., Alpargu, G., Ramoni, M.F. (2004) Robust transmission/disequilibrium test for incomplete family genotypes. *Genetics*, **168**, 2329-2337.

Sebastiani, P., Ramoni, M.F., Nolan, V., Baldwin, C.T., Steinberg, M.H. (2005) Genetic dissection and prognostic modeling of overt stroke in sickle cell anemia. *Nature Genetics*, **37**, 435-440.

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

Seltman, H., Roeder, K., Devlin, B. (2001) Transmission/disequilibrium test meets measured haplotype analysis: family-based association analysis guided by evolution of haplotypes. *Am J Hum Genet*, **68**, 1250-1263.

Sham, P.C., Curtis, D. (1995) An extended transmission/disequilibrium test (TDT) for multi-allele marker loci. *Ann Hum Genet*, **1995**, 323-336.

Shao, J. (2000) Cold deck and ratio imputation. *Survey Methodology*, **26**, 79-85.

Sherry, S.T., Ward, M., Sirotkin, K. (1999) dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Res*, **9**, 677-679.

Sidak, Z. (1967) Rectangular confidence regions for the means of multivariate normal distributions. *Am Stat Assoc*, **62**, 626-633.

Sidak, Z. (1971) On probabilities of rectangles in multivariate Student distributions: their dependence on correlations. *Ann Math Stat*, **42**, 169-175.

Skol, A.D., Scott, L.J., Abecasis, G.R., Boehnke, M. (2005) Forget replication: Joint analysis is more efficient for whole genome association studies. Abstract 180: ASHG 55th Annual Meeting, Salt Lake City, USA.

Smith, D.J., Lusis, A.J. (2002) The allelic structure of common disease. *Hum Mol Genet*, **11**, 2455-2461.

Smith, J.Q., Queen, C.M. (1996) Bayesian models for sparse probability tables. *Ann Stat*, **24**, 2178-2198.

Sobel, E., Papp, J.C., Lange, K. (2002) Detection and integration of genotyping errors in statistical genetics. *Am J Hum Genet*, **70**, 496-508.

Spielman, R.S., Mc Ginnis, R.E., Ewens, W.J. (1993) Transmission test for linkage disequilibrium: the insuline gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet*, **65**, 578-580.

Spielman, R.S., McGinnis, R.E., Ewens, W.J. (1994) The transmission/disequilibrium test detects cosegregation and linkage. *Am J Hum Genet*, **54**, 559-560; author reply 560-563.

Spielman, R.S. and Ewens, W.J. (1998) A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *Am J Hum Genet*, **62**, 450-458.

Stephens, M., Smith, N.J., Donnelly, P. (2001) A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet*, **69**, 906-914.

Stephens, M., Donnelly, P. (2003) A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet*, **73**, 1162-1169.

Stephens, M., Scheet, P. (2005) Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am J Hum Genet*, **76**, 449-462.

Storey, J.D., Tibshirani, R. (2003) Statistical significance for genome-wide studies. *Proc Natl Acad Sci*, **100**, 9440-9445.

Stram, D.O., Pearce, C.L., Bretsky, P., Freedman, M., Hirschhorn, J.N., Altshuler, D., Kolonel, L.N., Henderson, B.E., Thomas, D.C. (2003) Modeling and E-M estimation of haplotype-specific relative risks from genotype data for a case-control study of unrelated individuals. *Hum Hered*, **55**, 179-190.

Stram, D.O. (2004) Tag SNP selection for association studies. *Genet Epidemiol*, **27**, 365-374.

Strittmatter, W.J., Saunders, A.M., Schmechel, D., Pericak-Vance, M., Enghild, J., Roses, A.D. (1993) Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. *Proc Natl Acad Sci*, **90**, 1977-1981.

Sun, F., Flanders, W.D., Yang, Q., Khoury, M.J. (1999). Transmission disequilibrium test (TDT) when only one parent is available: the 1-TDT. *Am J Epidemiol*, **150**, 97-104.

Tanner, M.A., Wong, W.H. (1987) The calculation of posterior distributions by data augmentation. *J Am Stat Assoc*, **82**, 528-550.

Templeton, A.R. (2000) Epistasis and complex traits. In: *Epistasis and the Evolutionary Process*(Eds. Wade, M.J., Brodie, E.D., III, Wolf, J.B.), Oxford: Oxford University Press.

Teng, J., Risch, N. (1999) The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases. II. Individual genotyping. *Genome Res*, **9**, 234-241.

Terwilliger, J.D., Ott, J. (1994) *Handbook of human genetic linkage.* Baltimore: Johns Hopkins University Press.

Thijs, H., Molenberghs, G., Michiels, B., Verbeke, G., Curran, D. (2001) Strategies to fit pattern-mixture models. *Biostatistics*, **3**, 245-265.

Thomas, D.C., Haile, R.W., Duggan, D. (2005) Recent developments in genomewide association scans: a workshop summary and review. *Am J Hum Genet*, **77**, 337-345.

Thomson, G. (1995) Mapping disease genes: family-based association studies. *Am J Hum Genet*, **57**, 487-498.

Vach, W., Blettner, M. (1995) Logistic regression with incompletely observed categorical covariates - Investigating the sensitivity against violation of the missing at random assumption. *Stat Med*, **14**, 1315-1329.

van Dyck, D.A., Meng, X.-L. (2001) The art of data augmentation. *JCGS* , **10**, 1-111.

Vansteelandt, S., Goetghebeur, E., Kenward, M. G., Molenberghs, G. (2005) Ignorance and uncertainty as inferential tools in a sensitivity analysis. *Stat Sinica*, in print.

Van Steen, K., Molenberghs, G., Verbeke, G., Thijs, H. (2001) A local influence approach to sensitivity analysis of incomplete longitudinal ordinal data. *Stat Model*, **1**, 125-142.

Van Steen, K., Curran, D., Kramer, J., Molenberghs, G., Van Vreckem, A., Bottomley, A., Sylvester, R. (2002) Multicollinearity in prognostic factor analyses using the EORTC QLQ-C30: Identification and impact on model selection. *Stat Med*, **21**, 3865-3884.

Van Steen, K., Molenberghs, G. (2004a) Multicollinearity. In: *Encyclopedia of Biopharmaceutical Statistics* (Ed. Shein-Chung Chow), ISBN: 0-8247-4263-X

Van Steen, K., Tahri, N., Molenberghs, G. (2004b) Introducing the multivariate Dale model in population-based genetic association studies. *Biometrical J*, **46**, 187-202.

Van Steen, K., Markel, P., Vlietinck, R., Molenberghs, G., Laird, N.M. (2004c) Approaches to handle missingness in family-based association testing, to be sumitted to *Ann Hum Genet*.

Van Steen, K., Lange, C. (2005) PBAT: a comprehensive software package for genome-wide association analysis of complex familybased studies. *Human Genomics*, **2**, 67-69.

Van Steen, K., Mcqueen, M.B., Herbert, A., Raby, B., Lyon, H., Demeo, D.L., Murphy, A., Su, J., Datta, S., Rosenow, C., Christman, M., Silverman, E.K., Laird, N.M., Weiss, S.T., Lange, C. (2005) Screening and replication in family-based association testing using the same data set. *Nat Genet*, **37**, 683-691.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., et al (2001) The sequence of the human genome. *Science*, **291**, 1304-1351.

Verbeke, G., Molenberghs, G. (2000) *Linear mixed models for longitudinal data.* New York: Springer-Verlag.

Verbeke, G., Molenberghs, G., Thijs, H., Lesaffre, E., Kenward, M.G. (2001) Sensitivity analysis for non-random dropout: a local influence approach. *Biometrics*, **57**, 43-50.

Wacholder, S., Chanock, S., Garcia-Closas, M., El Ghormli, L., Rothman, N. (2004) Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J Natl Cancer I*, **96**, 434-442.

Wang, D.G., Fan, J.B., Siao, C.J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., Kruglyak, L., Stein, L., Hsie, L.,

Topaloglou, T., Hubbell, E., Robinson, E., Mittmann, M., Morris, M.S., Shen, N.P., Kilburn, D., Rioux, J., Nusbaum, C., Rozen, S., Hudson, T.J., Lipshutz, R., Chee, M., Lander, E.S. (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*, **282**, 1077-1082.

Wang, W.Y.S., Pike, N. (2003) The allelic spectra of common diseases may resemble the allelic spectrum of the full genome. *Med Hypotheses*, **63(4)**, 748-751.

Wang, W.Y.S., Barratt, B.J., Clayton, D.G., Todd, J.A. (2005) Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet*, **6**, 109-118.

Warner, J.O., Naspitz, C.K. (1998) Third International Pediatric Consensus statement on the management of childhood asthma. *Pediatr Pulm*, **25**, 1-17.

Weinberg, C.R. (1999) Allowing for missing parents in genetic studies of case-parent triads. *Am J Hum Genet*, **64**, 1186-1193.

Weisgraber, K.H. (1994) Apolipoprotein E: structure-function relationships. *Adv Protein Chem*, **45**, 249-302.

Weiss, S.T. (1998) Environmental risk factors in childhood asthma. *Clin Exp Allergy*, **28**, Suppl 5, 29-34.

Weiss, S.T., Raby, B.A. (2004) Asthma genetics 2003. *Hum Mol Genet*, **13**, R83-R89.

Westfall, P.H., Young, S.S. (1993) *Resampling-based multiple testing: Examples and methods for p-value adjustment.* New York: John Wiley & Sons.

Whittemore, A.S., Halpern, J., Ahsan H. (2005) Covariate adjustment in family-based association studies. *Genet Epidemiol*, **28(3)**, 244-255.

Wills-Karp, M., Ewart, S.L. (2004) Time to draw breath: asthma-susceptibility genes are identified. *Nat Rev Genet*, **5**, 376-387.

Wolfinger, R., O'Connell, M. (1993) Generalized linear mixed models: a pseudo-likelihood approach. *J Stat Comput Sim*, **48**, 233-243.

Wu, M.C., Carroll, R.J. (1988) Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, **44**, 175-188.

Xie, F., Paik, M.C. (1997) Multiple imputation methods for the missing covariates in generalized estimating equation. *Biometrics*, **53**, 1538-1546.

Yekutieli, D., Benjamini, Y. (1999) Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *J Stat Plan Infer*, **82**, 171-196.

Zaykin, D.V., Westfall, P.H., Young, S.S., Karnoub, M.A., Wagner, M.J., Ehm, M.G. (2002) Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Hum Hered*, **53**, 79-91.

Zhai, W., Todd, M.J., Nielsen, R. (2004) Is haplotype block identification useful for association mapping studies? *Genet Epidemiol*, **27**, 80-83.

Zhang, S., Pakstis, A.J., Kidd, K.K., Zhao, H. (2001) Comparisons of two methods for haplotype frequency estimation from population data. *Am J Hum Genet*, **69**, 906-912.

Zhao, L.P., Prentice, R.L. (1990) Correlated binary regression using a quadratic exponential model. *Biometrika*, **77**, 642-648.

Zhao, L.P., Lipsitz, S., Lew, D. (1996) Regression analysis with missing covariate data using estimating equations. *Biometrics*, **52**, 1165-1182.

Zhao, H., Zhang, S., Merikangas, K., Trixler, M., Wildenauer, D., Sun, F., Kidd, K. (2000) Transmission/disequilibrium tests using multiple tightly linked markers. *Am J Hum Genet*, **67**, 939-946.

Zhao, H. (2000) Family-based association studies. *Stat Methods Med Res*, **9**, 563-587.

Zhao, H., Zhang, S., Merikangas, K.R., Trixler, M., Wildenauer, D.B., Sun, F., Kidd, K.K. (2000) Transmission/disequilibrium tests using multiple tightly linked markers. *Am J Hum Genet*, **67**, 936-946.

Zhao, L.P., Li, S.S., Khalid, N. (2003) A method for the assessment of disease associations with single-nucleotide polymorphism haplotypes and environmental variables in case-control studies. *Am J Hum Genet*, **72**, 1231-1250.