Statistical Methods for Genetic Association Studies With Response - Selective Sampling Designs

Brunilda Balliu

Cover design: Ermal Tahiraj, Athens, Greece
Printed by: Off Page

ISBN: 978-94-6182-584-1

Research leading to this thesis was supported by the Netherlands Organization for Scientific Research Grant (917.66.344), the Dutch Arthritis Foundation (Reumafonds), European Union's Seventh Framework Program for research under grant agreement no. 305280(MIMOmics), and two grants from the German Research Foundation; BO 1955/2-3 and WU 314/6-2.

Statistical Methods for Genetic Association Studies With Response - Selective Sampling Designs

Proefschrift

ter verkrijging van de graad van Doctor aan de Universiteit Leiden, op gezag van Rector Magnificus prof.mr. C.J.J.M. Stolker, volgens besluit van het College voor Promoties te verdedigen op donderdag 10 september 2015 klokke 16:15 uur

door

Brunilda Balliu
geboren te Vlorë, Albania
in 1987

PROMOTIECOMMISSIE

Promotor:
Prof.dr. J.J. Houwing-Dustermaat

Co-Promotor:
Dr. S. Boehringer

Overige leden:
Prof.dr. H. Cordell, Institute of Genetic Medicine, Newcastle University, Newcastle, United Kingdom
Prof.dr. F.R. Rosendaal
Prof.dr. A.H. Zwinderman, Department of Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center , Amsterdam, The Netherlands

Dedicated to my friends and family for their enduring love and support.

Διά τὸ θαυμάζειν ἡ σοφία.
Wisdom begins in wonder.

– Edith Hamilton, *The Greek Way*, 1930
(paraphrase from Plato's *Theaetetus*, ca. 368 BC).

Table of Contents

Acknowledgments

The research presented in this thesis is the result of my work in the Department of Medical Statistics and Bioinformatics of the Leiden University Medical Center. Thanks are owed to many fantastic people. First and foremost, my promotor Prof.dr. Jeanine Houwing-Duistermaat and my co-promotor Dr. Stefan Boehringer for encouraging my research and for allowing me to grow as a scientist. I am very thankful for the excellent example Prof.dr. Jeanine Houwing-Duistermaat has provided as a successful woman biostatistician and professor and for the endless guidance and encouragement of Dr. Boehringer when I was stuck. It has been an honor to be Dr. Boehringer's first Ph.D. student. I would also like to thank my reading committee members: Prof.dr. Heather Cordell, Prof.dr. Frits R. Rosendaal, and Prof.dr. Koos Zwinderman for their time, interest, and helpful comments.

The current and past members of the Statistical Genetics group have contributed immensely to my personal and professional time at the LUMC. I am especially grateful to my colleagues and officemates Hae Won Uh, Fabrice Colas, Ivonne Martin, and Renaud Tissier for being a source of friendship as well as good advice, even during tough times in the Ph.D. pursuit. Many thanks also go to the other, past and present, group members that I have had the pleasure to work with or alongside: Marcus de Jong, Roula Tsonaka and Mar Rodriguez. My time at the LUMC was made enjoyable in large part due to the many PhD fellows and young researcher that became good friends and a part of my life here: Alina Nicolaie, Alexia Kakourou, Dimitris Ziagkos, Mia Klinten Grand, Roberta Rovito, Rosa Meijer, Zhenia Aizenberg, and of course Theodor Balan. I would also like to thank the rest of my colleagues from the LUMC from whom I benefited a lot: Bart Mertens, Erik van Zwet, Henk Jan van der Wijk, Hein Putter, Jelle Goeman, Liesbeth de Wreede, Lies de Kler-van der Poel, Marta Fiocco, Ramin Monajemi, Ron Wolterbeek, Ronald Brand, Szymon Kiełbasa, Saskia le Cessie, Theo Stijnen, and Watze Hoekstra, as well as many friends and collegues outside LUMC: Angie Markou, Carolina Medina, Doug Speed, Ermal Tahiraj, Katerina and Georgia Papadimitropoulou, Marta Mansi, Suzette Matthijsse and Stavros Nikolakopoulos.

Several people have contributed, both consciously and unconsciously, to my decision to pursue a Ph.D. and to continue academic research, by introducing me to the amazing world of statistics and teaching me how good science is done. I am thankful to Prof.dr. Athanasios Yannacopoulos, Prof.dr. Dimitris Karlis, Prof.dr. Ioannis Ntzoufras, Prof.dr. Petros Dellaportas, Prof.dr. Eleni Kandilorou, Prof.dr. Richard Gill, and Prof.dr. Henk Kelderman.

At the end I would like to express appreciation to my family and three very special people for all their love and encouragement. To Maarten Kampert for his support, nourishment, and much much more. To Reinald Shyti who has been my fellow

traveler during this Ph.D. journey. To Noah Zaitlen who spent sleepless nights with me and was always my support in the moments when there was no one to answer my queries. And most importantly, to my parents Todi and Lefteria, my sister Blerina, and my brother Bledar të dashur babi, mami, motra dhe vëllai fjalët nuk mund të shprehin se sa mirënjohës jam për të gjithë sakrificat që keni bërë per mua.

1

Introduction to Genetic Association Studies

1.1 Introduction

Before outlining the specific novel contributions of this work, some background is given to lend them context and show their relevance to the field. The human genome consists of 23 pairs of chromosomes comprised of 2.3 billion base pairs of DNA in the haploid genome. If we examine the DNA of two individuals, the differences in their genome will include individual nucleotide changes called *single nucleotide polymorphisms* (SNPs), changes in the number of copies of a segment of DNA called copy number variations (CNVs), and other structural changes such as inversions, translocations, and VNTR-polymorphisms. It is believed that *heritability*, the proportion of the variability in a phenotype explained by genetic factors, is mostly due to changes such as these, with some growing evidence for epigenetic effects [Koch, 2014].

In genetic epidemiology, genetic association studies aim to assess the association between genetic variants and complex traits like common diseases. Often in such studies, individuals are collected from two groups, the cases who have the trait of interest, and the controls that are members of the same population but do not have the disease. The individuals are genotyped and differences in the allele frequencies of the genetic variants between the cases and controls are assessed. The diseases of interest in such studies have in many cases low prevalence, e.g. the prevalence of rheumatoid arthritis and multiple sclerosis, two of the diseases we study here, ranges from .5-1.0% [Silman and Hochberg, 2001] and from $.005 - .08\%$ [World Health Organization, 2008], respectively. The putative high-risk alleles can also be rare, with frequencies even below 1%. This means that traditional population-based case-control and cohort studies will generally be inefficient, since most subjects will never develop the disease of interest or have the exposure of interest [Kraft and Thomas,

2000]. Some of the strategies to deal with this problem involve *response-selective sampling* strategies.

Case-control studies of unrelated individuals or family members constitute a very efficient design for collecting covariate information in epidemiological studies and they are the most widely used designs for genetic association studies. Each study design has its advantages and disadvantages. In studies of cases and unrelated controls sufficiently large study populations can be readily assembled without the need to enroll also family members of the recruited participants [Evangelou et al., 2006]. However, such studies are susceptible to confounding due to unaccounted population admixture [Cardon and Palmer, 2003; Hattersley and McCarthy, 2005; Wang et al., 2005], an issue usually addressed by using principal component analysis [Price et al., 2006], they can be under-powered to detect low frequency variants, and they cannot be used for estimating more complex disease generating mechanisms, such as ones arising only from a specific parent-offspring genotype combinations [Weinberg, C. R., 1999; Sinsheimer et al., 2003; Spinka et al., 2005; Hsieh et al., 2007; Ainsworth et al., 2011].

On the other hand, family-based study designs have the advantage that there is a common genetic background among the family members. Thus, the problem of population stratification is mitigated. Methods for family data can take advantage of the ability to model the dependence of genotypes within families. This can increase efficiency of parameter estimates by making more effective use, not only of subjects for whom we have both trait and genotype data, but also of subjects for whom we only have trait data, since subjects who are not genotyped can also contribute information about the relationship between trait and the genetic variant being studied [Kraft and Thomas, 2000]. Furthermore, family-based studies can be more powerful to detect rare variants that aggregate in families [Evangelou et al., 2006]. Moreover, families tend to be more homogeneous regarding exposure to environmental factors possibly associated to the disease etiology. The main disadvantage of family-based studies, however, is that it is usually more difficult to accumulate large enough samples of well-characterized families. Sample sizes need to be large enough to avoid type I error inflation both in the screening process, as well as in the validation of the modest genetic effects that genome-wide association studies target [Ioannidis, 2003].

It is well known that in studies with response-selective sampling designs, the distribution of the covariates contains information about the parameters of interest, i.e. the effect of the covariates on the trait [Scott and Wild, 2001]. Such studies enable us to increase the efficiency of parameter estimates by taking advantage of the dependence among the parameters of interest and the parameters needed to characterize the distribution of the covariates. Thus, accounting for ascertainment in studies with response-selective sampling can increase power to detect associations [Chatterjee and Carroll, 2005; Zaitlen et al., 2012a]. Moreover, when a secondary phenotype is of interest, other than the primary phenotype used to ascertain the samples, modelling the ascertainment is necessary to avoid bias and false positive results regarding the association of the covariates with the secondary phenotype [Lin and Zeng, 2009].

Marginal tests based on individual SNPs have dominated association analyses in the past decade. However, most common complex diseases do not arise from a single genetic cause, but rather a combination of multiple genetic and environmental factors

[Fisher, 1930]. Alternative approaches, which more closely model the underlying biological mechanisms, such as jointly modelling multiple genetic variants, or jointly modelling genetic variants with intermediate cellular phenotypes, might have the potential to discover novel genetic marker associated with disease which would have been missed in standard single SNP association studies [Chen et al., 2008; Li, 2013; Zhao et al., 2014; Huang et al., 2014].

The rest of the introduction is structured as follows. First, we describe different approaches for modelling the ascertainment in case-control or family-based association studies. Next, we present different models for the relation between the genetic variants and the disease. Last, we give an outline of the next chapters of the thesis and a brief explanation of the main novel contributions of each work.

1.2 Accounting for response-selective sampling

Suppose that a process leads to realization of data according to a model

$$f(\boldsymbol{Y}, \boldsymbol{X}; \boldsymbol{\alpha}, \boldsymbol{\beta}) = f(\boldsymbol{Y}|\boldsymbol{X}; \boldsymbol{\alpha}) f(\boldsymbol{X}; \boldsymbol{\beta}).$$

Here, $\boldsymbol{Y}$ is a binary response variable, $\boldsymbol{X}$ is a vector of covariates, $\boldsymbol{\alpha}$ are the parameters needed to characterize $f(\boldsymbol{Y}|\boldsymbol{X})$, and $\boldsymbol{\beta}$ are the parameters needed to characterize $f(\boldsymbol{X})$. $\boldsymbol{X}$ can be multivariate and any elements of $\boldsymbol{X}$ can be either discrete or continuous. The first term, $f(\boldsymbol{Y}|\boldsymbol{X}; \boldsymbol{\alpha})$, is a logistic regression model and $f(\boldsymbol{X}; \boldsymbol{\beta})$ is the density of $\boldsymbol{X}$. The purpose of $\boldsymbol{\alpha}$ is to characterize the conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{X}$ so that $f(\boldsymbol{X}; \boldsymbol{\beta})$ does not involve $\boldsymbol{\alpha}$. Our goal is the estimation of $\boldsymbol{\alpha}$.

When N observations are sampled from the joint distribution of $(\boldsymbol{Y}; \boldsymbol{X})$, i.e. $f(\boldsymbol{Y}, \boldsymbol{X})$, or sampled conditionally on some or all of the variables in $\boldsymbol{X}$, $f(\boldsymbol{X})$ is ancillary and it is standard to base inferences about $\boldsymbol{\alpha}$ on the likelihood made up of conditional terms,

$$L(\boldsymbol{\alpha}; \boldsymbol{Y}, \boldsymbol{X}) = \prod_{i=1}^{N} f(\boldsymbol{Y}_i|\boldsymbol{X}_i, \boldsymbol{\alpha}). \tag{1.1}$$

No modelling of $f(\boldsymbol{X})$ is required. This is very convenient because $\boldsymbol{X}$ often contains many covariates and is too complicated for modelling to be feasible, unless parametric assumptions are made about the nature of $f(\boldsymbol{X})$.

When the probability that a unit with $(\boldsymbol{Y}; \boldsymbol{X})$ will be observed involves $\boldsymbol{Y}$ (response-selective sampling), that is observations are sampled from the distribution $f(\boldsymbol{X}|\boldsymbol{Y})$, $f(\boldsymbol{X})$ is no longer ancillary and (1.1) no longer applies. Nevertheless, Prentice and Pyke [1979] showed that fitting a standard prospective logistic regression that ignores the retrospective sampling nature of the design yields the maximum likelihood estimates of the regression parameters under a *semi-parametric* model $f(\boldsymbol{X}|\boldsymbol{Y}) = f(\boldsymbol{Y}|\boldsymbol{X}) f(\boldsymbol{X}) / f(\boldsymbol{Y})$ that allows $f(\boldsymbol{X})$ to be non-parametric. More recently, Rabinowitz [1997] and Breslow et al. [2000] used modern semi-parametric theory to show that the prospective logistic regression analysis of case-control data is efficient in the sense that it achieves the variance lower bound of the underlying semi-parametric model. However, under the case-control design, the variance lower bound for estimators of the regression parameters under particular constraints for $f(\boldsymbol{X})$,

e.g. independence between elements of $\boldsymbol{X}$, or under particular models for $f(\boldsymbol{X})$, e.g. parametric assumptions, will be lower than that of the more general model that allows a completely non-parametric covariate distribution, and equivalently of the prospective logistic regression approaches [Chatterjee and Carroll, 2005].

In the next sections we present three likelihoods for the analysis of family-based case-control data: the prospective, joint, and retrospective likelihoods. The later is also appropriate for the analysis of case-control data of unrelated individuals.

1.2.1 Ascertainment-Corrected Prospective Likelihood

Let $\mathcal{A}$ be the event that a unit was ascertained in the sample. In the case of family-based case-control studies the whole family is a unit. The prospective likelihood is based on modelling a unit's disease risk given the covariates. The *ascertainment-corrected prospective likelihood* has the form

$$L^p(\boldsymbol{\alpha}) = P(\boldsymbol{Y}|\boldsymbol{X}, \mathcal{A}) = \frac{P(\boldsymbol{Y}, \boldsymbol{X}, \mathcal{A})}{P(\boldsymbol{X}, \mathcal{A})} = \frac{P(\mathcal{A}|\boldsymbol{Y}, \boldsymbol{X})P(\boldsymbol{Y}|\boldsymbol{X})}{P(\mathcal{A}|\boldsymbol{X})}.$$

Notice here that the prospective likelihood only involves the regression parameters $\boldsymbol{\alpha}$. If we assume that subjects selection directly depend only upon potential subjects disease status, not on their covariates, the term $P(\mathcal{A}|\boldsymbol{Y}, \boldsymbol{X})$ simplifies to $P(\mathcal{A}|\boldsymbol{Y})$ in the above likelihood. An additional assumption typically made in studies with response-selective sampling is the assumption of *complete ascertainment*, i.e. for all the units included in the sample $P(\mathcal{A}|\boldsymbol{Y}) = 1$. Then the likelihood is expressed as follows

$$L^p(\boldsymbol{\alpha}) = \frac{P(\boldsymbol{Y}|\boldsymbol{X})}{P(\mathcal{A}|\boldsymbol{X})}. \tag{1.2}$$

The numerator of the likelihood is the *penetrance function*, which models the disease probability of a unit conditional on the unit's covariates. The penetrance function could include only the genotypes of the individuals or genotypes and additional clinical or environmental covariates. In the next section we present several such functions. The denominator models the ascertainment probability of a unit conditional on the unit's covariates. For case-control studies of unrelated individuals this information is more difficult to obtain and the prospective logistic regression without the ascertainment correction is typically used. On the other hand, for family-based studies modelling the probability of ascertainment given the covariates is possible. Consider for example a study which includes families in a study if at least K offspring in the families present the disease. Then, the denominator in (1.2) can be written as follows

$$P(\mathcal{A}|\boldsymbol{X}) = \prod_{i=1}^{N} P\left(\sum_{j=1}^{n_i} Y_{ij} \geq K \middle| \boldsymbol{X}\right) = \prod_{i=1}^{N} \left[1 - \sum_{k=0}^{K-1} P\left(\sum_{j=1}^{n_i} Y_{ij} = k \middle| \boldsymbol{X}\right)\right],$$

where N is the total number of families in the sample, i is the index that runs through all the families, n_i is the size of family i, and j is the index that runs through the family members in each family.

1.2.2 Ascertainment Assumption Free Retrospective Likelihood

The retrospective likelihood is based on modelling the distribution of covariates conditional on the outcome and the ascertainment and is given as follows

$$L^r(\boldsymbol{\alpha}, \boldsymbol{\beta}) = P(\boldsymbol{X}|\boldsymbol{Y}, \mathcal{A}) = P(\boldsymbol{X}|\boldsymbol{Y}).$$

Prentice and Pyke (1979) showed that this likelihood can further be factored into two components, the first identical to the standard prospective likelihood, and the second depending upon the distribution of covariates.

$$L^r(\boldsymbol{\alpha}, \boldsymbol{\beta}) = P(\boldsymbol{X}|\boldsymbol{Y}) = \frac{P(\boldsymbol{Y}|\boldsymbol{X})P(\boldsymbol{X})}{P(\boldsymbol{Y})}.$$

This enables us to estimate again the regression parameters $\boldsymbol{\alpha}$ from the first component of the likelihood. The maximization of the first component leads to the maximum likelihood estimates of the entire likelihood, subject to a constraint based on the marginal population disease rate $P(\boldsymbol{Y})$. For discrete covariates $\boldsymbol{X}$ the retrospective likelihood can further be expressed as follows

$$L^r(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{P(\boldsymbol{Y}|\boldsymbol{X})P(\boldsymbol{X})}{\sum_{\boldsymbol{X}^*} P(\boldsymbol{Y}|\boldsymbol{X}^*)P(\boldsymbol{X}^*)}, \tag{1.3}$$

where the denominator sums over all possible values of $\boldsymbol{X}$, i.e. $\boldsymbol{X}^*$. For continuous covariates $\boldsymbol{X}$, the denominator will involve integrals instead of summations.

An additional challenge for modelling and maximizing the retrospective likelihood comes from the need to model both the population distribution of the covariates $\boldsymbol{X}$ and the marginal distribution of the outcome $\boldsymbol{Y}$ (by integrating over the population distribution of covariates). In the genetics context, there is a strong basis for modelling the distribution of genotypes of unrelated individuals, using the Hardy Weinberg equilibrium (HWE) assumption, or the distribution of genotypes within families, using the HWE assumption, the random mating assumption and the Mendelian laws of inheritance. Thereby, it becomes feasible to directly maximize the retrospective likelihood. On the other hand, when $\boldsymbol{X}$ involves continuous or discrete covariates, other than genotypes, e.g. age and gender of the individuals or intermediate cellular phenotypes, modelling and maximizing the retrospective likelihood is not straightforward. In this case, specific assumptions about the nature of $P(\boldsymbol{X})$ need to be made, in order for $P(\boldsymbol{X})$ to be identifiable from case-control data. Such assumptions include for example parametric assumptions about the distribution of covariates in $\boldsymbol{X}$ or independence assumptions among the covariates in $\boldsymbol{X}$. When these assumptions do not hold (model misspecification), the retrospective likelihood can provide biased parameter estimates and thus flexible modelling strategies should be employed for a good trade-off between efficiency and robustness.

The retrospective likelihoods is *ascertainment-assumption free* - that is, if the probability of a unit being ascertained depends only on the unit's phenotypes, then we do not have to explicitly model how ascertainment depends on phenotypes. The advantage of this approach is that by conditioning on the disease outcomes, one automatically conditions on ascertainment, thereby making this approach relevant to case-control analyses of unrelated individuals or families sampled in an ad hoc

manner, for whom ascertainment correction with the usual prospective likelihood would be impossible. The disadvantage is, of course, that by conditioning on all the phenotypes, rather than just the ascertainment event, one may 'over-condition', thereby perhaps leading to some loss of efficiency relative to the analysis that would be possible if the ascertainment event could be defined.

1.2.3 Ascertainment-Corrected Joint Likelihood

The ascertainment-corrected joint likelihood is based on the joint probability of covariates and phenotypes and is given as follows

$$L^j(\boldsymbol{\alpha}, \boldsymbol{\beta}) = P(\boldsymbol{Y}, \boldsymbol{X}|\mathcal{A}) = \frac{P(\mathcal{A}|\boldsymbol{Y}, \boldsymbol{X})P(\boldsymbol{Y}|\boldsymbol{X})P(\boldsymbol{X})}{P(\mathcal{A})} = \frac{P(\boldsymbol{Y}|\boldsymbol{X})P(\boldsymbol{X})}{P(\mathcal{A})}.$$

The denominator here is the probability of ascertainment. Similarly to the ascertainment - corrected prospective likelihood, modelling the ascertainment is not feasible for studies with ad hoc sampling. However, continuing the example of the previous section, when families are included in the sample if at least K offspring are affected, the denominator can be expressed as

$$\begin{aligned} P(\mathcal{A}) &= \prod_{i=1}^{N} P\left(\sum_{j=1}^{n_i} Y_{ij} \geq K\right) = \prod_{i=1}^{N}\left[1 - \sum_{k=0}^{K-1} P\left(\sum_{j=1}^{n_i} Y_{ij} = k\right)\right] \\ &= \prod_{i=1}^{N}\left\{1 - \sum_{\boldsymbol{X}^*}\left[\sum_{k=0}^{K-1} P\left(\sum_{j=1}^{n_i} Y_{ij} = k|\boldsymbol{X}^*\right) P(\boldsymbol{X}^*)\right]\right\}. \end{aligned}$$

Here, the sum is over all possible covariate values and all family phenotype vectors with no case, at least one case until at least $K-1$ cases. The joint likelihood entails the weakest conditioning of all three likelihoods, $P(\mathcal{A})$, rather than $P(\mathcal{A}|\boldsymbol{X})$ for the prospective likelihood or $P(\boldsymbol{Y})$ for the retrospective likelihood, and thus should be more efficient than either [Kraft and Thomas, 2000].

1.3 Models of disease mechanisms

In this section we will explore different sets of covariates $\boldsymbol{X}$ that can be available in association studies. The standard analysis of genome wide association study data individually evaluates the relationship between each SNP ($\boldsymbol{G}$) and disease. In this case, one may a fit a logistic regression model to assess the association between each SNP and disease:

$$P(\boldsymbol{Y}|\boldsymbol{G}) = logit^{-1}\left(\alpha_0 + \alpha_1 \boldsymbol{G}\right), \tag{1.4}$$

where $logit^{-1}$ is the inverse $logit$ link function; α_0 is the intercept; $\boldsymbol{G}$ is coded in a log additive manner to reflect the number of alleles an individual carries at this SNP (i.e., 0, 1, or 2) and α_1 is the parameter of interest: the log odds ratio reflecting the impact of one additional allele of a SNP on disease risk.

Most common complex diseases do not arise from a single genetic cause, but rather a combination of multiple genetic and environmental factors (i.e., they are polygenic) [Fisher, 1930; Risch and Merikangas, 1996; Witte, 2010]. To assess such joint effects on disease, model (1.4) can be extended to include multiple SNPs, as well as non-genetic exposures. An alternative to single SNP methods are methods based on haplotypes. Haplotypes, tuples of alleles, play key roles in the study of the genetic basis of disease. These roles vary from biologic function to providing information about ancient ancestral chromosome segments that harbor alleles that influence human traits. Haplotype-based association studies compare the frequencies of haplotypes between cases and controls or model the penetrance function depending on haplotypes.

Assume that we are studying the potential association between a genetic variant ($\boldsymbol{G}$) and a binary trait. Furthermore, assume we have also measured environmental or clinical covariate ($\boldsymbol{C}$) associated with the trait but independent of the variant of interest in the source population, so it is not a confounder (Figure 1.1). In this case $\boldsymbol{X} = (\boldsymbol{G}, \boldsymbol{C})$. If we ascertain a random sample of study subjects, then the variant of interest and covariate will remain independent (Figure 1.1.a). Thus, the most powerful model for assessing association between the genetic variant and the binary trait includes the environmental covariate in a logistic regression model [Robinson and Jewell, 1991; Neuhaus and Jewell, 1993; Neuhaus, 1998; Pirinen et al., 2012], that is

$$P(\boldsymbol{Y}|\boldsymbol{G}, \boldsymbol{C}) = logit^{-1}\left(\alpha_0 + \alpha_1 \boldsymbol{G} + \alpha_2 \boldsymbol{C}\right),$$

where $\boldsymbol{G}$ is the genetic variant, $\boldsymbol{C}$ is the environmental covariate, α_1 is the log odds ratio reflecting the impact of one additional allele of a SNP on disease risk and α_2 is the log odds ratio reflecting the impact of one additional unit of C on disease risk.

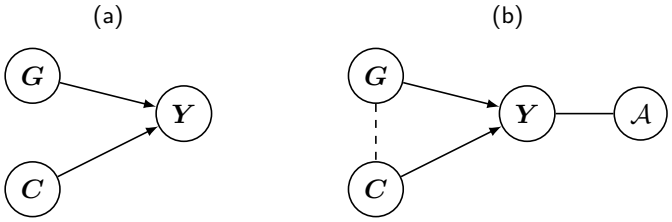


Figure 1.1: **Example to illustrate possible correlation structures among risk factors and a trait in (a) a random sample and (b) a case-control sample**. G: SNP, C: clinical or environmental covariate, Y: binary disease trait, A: ascertainment. Continuous arrows between two nodes connect variables that could be correlated in the population while dashed lines represent induced correlations due to ascertainment.

In the presence of ascertainment, cases will be enriched for both risk genotypes and high-risk covariate levels. As a result, the genetic variant and covariate might end up being correlated in the sample (dashed line in Figure 1.1.b). Including both covariates in a logistic regression model could substantially increase the standard error of the genetic variant association (i.e., due to the induced correlation), resulting in a larger power loss than might arise from omitting the covariate [Mefford and Witte,

2012]. Fortunately, using the retrospective likelihood approach in (1.3) one can address this problem by explicitly imposing the independence assumption between the genetic variant and the covariate [Umbach and Weinberg, 1997; Chatterjee and Carroll, 2005], that is

$$L^r(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{P(\boldsymbol{Y}|\boldsymbol{G}, \boldsymbol{E})P(\boldsymbol{G})P(\boldsymbol{E})}{\sum_{\boldsymbol{G}^*, \boldsymbol{E}^*} P(\boldsymbol{Y}|\boldsymbol{G}^*, \boldsymbol{E}^*)P(\boldsymbol{G}^*)P(\boldsymbol{E}^*)}.$$

It is known that the phenotype of an organism is sometimes determined, not only by its own genotype and environment, but also by the environment and genotype of its parents. Examples of such situation are maternal effects, i.e. when an organism shows the phenotype expected from the genotype of the mother, irrespective of its own genotype. Other examples of such situations are the non-inherited maternal antigen effects (NIMA), i.e. antigens passed from the mother to the offspring during pregnancy, which increase or decrease the disease risk of an offspring. To capture such effects, model (1.4) can be extended to incorporate maternal genotype information,

$$g\{\mathbb{E}(Y|G^c, G^m)\} = \alpha_0 + \alpha_1 G^c + \alpha_2 G^m + \alpha_3 f(G^c, G^m),$$

where G^c and G^m are the genotypes of the child and mother; α_1 and α_2 are their effects on disease risk of the child; $f(G^c, G^m)$ is a function that takes into account the different offspring-mother genotype combinations that can result in a NIMA effect with

$$f(G^c, G^m) = \begin{cases} 1 & \text{if } G^m \text{, but not } G^c \text{ , increases or decreases disease risk,} \\ 0 & \text{if } G^m \text{ , does not increase or decrease disease risk.} \end{cases},$$

and α_3 is the NIMA effect.

The two factors we try to bridge in genetic association studies are SNPs and disease risk. While this approach has successfully identified many associations, the biological mechanisms underpinning the change in risk remain often unknown. Intermediate cellular phenotypes, such as gene expression and DNA methylation, which are now being collected in addition to genetic data, provide an opportunity to address this issue. Performing joint analysis over these multiple data types (i.e. *integrative omics*) has advantages for both biological and statistical reasons. For example, gene expression and DNA methylation can help explain variability of the effect of the SNP on disease when the effect of the SNP on disease is mediated via gene expression and/or DNA methylation, illustrated in Figure 1.2.a, or they can help remove unwanted variation from the phenotype when each variable has an independent effect on disease risk, illustrated in Figure 1.2.b. In both cases this will increase the power of detecting the overall effect of SNPs on disease risk.

1.4 This thesis

This dissertation is primarily concerned with a new set of methods, resources, tools, and techniques designed to address some of the problems mentioned above and improve the power of genetic association studies. The core motivation behind the

Figure 1.2: **Example to illustrate possible correlation structures among a binary disease trait (Y) and the omics risk factors.** The omics risk factors are a SNP (G), a gene expression measurement (E), and a DNA methylation measurement (M). (a) The effect of G on Y is mediated via E and/or M and (b) Each of E, M, and G have an independent effect on Y. Continuous arrows between two nodes connect variables that could be correlated in the population while dashed lines represent mediation effect.

thesis is to construct statistical methods that use "richer" models for the relationship between the genetic variants and the phenotype, compared to models used in standard genetic association studies, incorporate information from both family and case-control based studies; different types of data; genetic, genomic, epigenomic and environmental information; and allow the genetics community to answer more complicated questions about the genetic architecture behind complex traits. Each Chapter is based on a paper, already published, submitted or prepared for submission, that addresses different issues of genetic association studies and current studies of the genetic basis of human disease. In the next section we present these problems and the solutions we propose.

Chapter 2 describes a novel method to improve the power of GWAS by combining data from multi-case family studies and twin studies. To maximise efficiency in parameter estimation we base the inference about the parameters of interest on an ascertainment-corrected joint likelihood. To take into account the correlation of disease risks among family members, due to shared but unmeasured genetic or environmental factors, we use a family-specific random term. We show in both simulated and real data that this families and twins combined ascertainment-corrected joint likelihood approach is more efficient for estimating the parameters of interest, as compared to a families-only approach or a prospective approach which ignores the ascertainment.

Chapter 3 covers a novel method we developed for improving the power of GWAS by performing haplotype-based association studies. A limitation of haplotype-based methods is that the number of parameters increases exponentially with the number of SNPs, inducing a commensurate increase in the degrees of freedom and weakening the power to detect associations. To address this limitation, we introduce a hierarchical linkage disequilibrium model for disease mapping, based on a re-parameterization of the multinomial haplotype distribution. The hierarchy in our parameters enables flexible testing over a range of parameter sets: from joint single SNP analyses through the full haplotype distribution tests. We show via extensive simulations that our approach maintains the type I error at nominal level and has increased power under

many realistic scenarios, as compared to single SNP-based and standard haplotype-based studies.

Chapter 4 investigates the contributions that linkage-based methods, such as identical-by-descent mapping, can make to association mapping to identify rare variants in next-generation sequencing data. Linkage mapping methods are more powerful for identifying highly penetrant variants with low frequencies while association mapping methods are more suitable for identifying more common variants with moderate effect sizes. The hope is that, by combining both methods, we would be able to identify variants with moderate effect sizes and moderate to low frequencies. We apply the method to next-generation sequencing longitudinal family data from Genetic Association Workshop 18.

Chapter 5 introduces a novel statistical method to improve the power of GWAS and further characterize genetic mechanism behind complex diseases by using integrative omics. Recent works on integrative omics use prospective approaches, modelling case-control status conditional on omics and non omics risk factors. In this chapter, we propose a novel statistical method for integrating multiple omics and non-omics factors in case-control association studies based on a retrospective likelihood function, which accounts for the ascertainment present in the case-control data. The new method has increased efficiency over prospective approaches in both simulated and real data.

In addition to methods related to the analysis of GWAS, which focus mainly on phenotype-genotype-related questions, I include research that focuses on phenotype-only-related questions. Here, diseases of interest are Mendelian disorders, such as Fragile X and Cornelia de Lange and the objective is, not to identify the genes related to the disease, but to identify special facial features that would help in the discrimination between different syndromes. In a second stage, such features could be used as intermediate phenotypes in a GWAS. Chapter 6 of this thesis presents a method for automated syndrome classification and visualization based on data transformations prior to analysis. These transformations are low-variance in the sense that each involves only a fixed small number of input features. We show that classification accuracy can be improved when penalized regression techniques are employed, as compared to a principal component analysis pre-processing step. In order to visualize the resulting classifiers, we develop importance plots highlighting the influence of coordinates in the original 2D space. These plots assist in assessing plausibility of classifiers, interpretation of classifiers, and determination of the relative importance of different features.

2

Combining Family and Twin Data in Association Studies [1]

Summary

It is hypothesized that certain alleles can have a protective effect not only when inherited by the offspring but also as non-inherited maternal antigens (NIMA). To estimate the NIMA effect, large samples of families are needed. When large samples are not available, we propose a combined approach to estimate the NIMA effect from ascertained nuclear families and twin pairs. We develop a likelihood-based approach allowing for several ascertainment schemes, to accommodate for the outcome-dependent sampling scheme, and a family-specific random term, to take into account the correlation between family members. Simulations show that the combined likelihood is more efficient for estimating the NIMA odds ratios as compared to a families-only approach. To illustrate our approach, we used data from a family and a twin study from the United Kingdom on rheumatoid arthritis, and confirmed the protective NIMA effect, with an odds ratio of .477 (95% CI .264-.864). The method is publicly available at `https://github.com/BrunildaBalliu/NIMA`.

2.1 Introduction

Genetic studies typically focus on testing whether a genetic variant is associated with disease risk directly through the genotype of the offspring, i.e. offspring allelic effect, to identify susceptibility genes involved in complex disorders. However, many genes influence disease susceptibility through more complex biological mechanisms, such as conditions during embryonic or fetal life. One such mechanism, the non-inherited

[1] Published in *Genetic Epidemiology*.

maternal antigens (NIMA) effect, may be involved in the pathogenesis of certain autoimmune diseases, such as rheumatoid arthritis (RA) [Hsieh et al., 2007; Feitsma et al., 2007], renal graft survival [Smits et al., 1998], and scleroderma [Nelson et al., 1998; Azzouz et al., 2011]. The NIMA effect affects disease susceptibility through a specific maternal-offspring genotype combination, i.e. the mother carries the allele of interest but the offspring does not. When the NIMA effect is present and not correctly modeled it can result in biased estimates of the offspring allelic effect [Weinberg, C. R., 1999; Sinsheimer et al., 2003].

In order to investigate such mechanisms, ascertained multi-case family designs are typically used. They are known to improve efficiency when studying the association of a disease with low prevalence and a low frequency variant, as compared to case-control studies of unrelated individuals [Kraft and Thomas, 2000]. To accommodate for potential residual correlation in disease risks among family members, due to shared but unmeasured genetic or environmental factors, mixed models with family specific random terms are used. An ascertainment correction is needed to account for the outcome-dependent sampling schemes, often used to increase efficiency when studying a disease with low prevalence.

Several methods have been developed to model and/or test for the NIMA effect [Hsieh et al., 2006; Feitsma et al., 2007]. However, these methods are not appropriate for families that contain both multiple cases and healthy siblings. Feitsma et al. [2007] use information only from one affected offspring per family. Hsieh et al. [2006] take into account information from multiple affected siblings, but the correlation between disease outcomes among family members, is ignored. Ignoring this correlation may have an effect on the ascertainment correction, resulting in biased results for both standard errors and effect sizes [Kraft et al., 2005; Hsieh et al., 2006]. Both methods ignore the information available from healthy siblings by excluding them from the analysis.

Recruiting, genotyping, and interviewing members of multi-case families can be difficult due to the lack of clear sampling definition and the high cost, resulting in data sets with small sample size, thus low power to detect the effect of interest. To enhance the statistical power to identify disease susceptibility genes, Pfeiffer et al. [2008] and Zheng et al. [2010] proposed to combine family-based studies with case-control studies using a prospective likelihood (PL) approach, modelling the distribution of the phenotypes of family members conditional on their genotypes. These methods focus on direct effects, and as expected, due to the larger sample size, they increase the power to detect the direct offspring allelic effect [Pfeiffer et al., 2008; Zheng et al., 2010]. Typically, studies with multi-case families lack power to estimate the effects of rare protective factors, such as the NIMA effect. To address this problem, we propose to combine the multi-case family study with a twin-based study and use the joint likelihood (JL), which models the joint genotype and phenotype distribution, instead of the PL. The JL can be more efficient for estimating the genetic odds ratios since it only conditions on the ascertainment event, and uses information from the modelling the genotype distribution of the parents and offspring [Kraft and Thomas, 2000].

The parental genotypes of twins are not at hand thus the twin likelihood itself contains no information about the NIMA effect. However, we can include the NIMA parameter in the model as a nuisance parameter and marginalize the likelihood by summing over all possible parental genotypes combinations. We can then estimate

the direct protective effect from both family and twin likelihood and the indirect NIMA effect from the family likelihood. In a similar way, Chen et al. [2012] use a semi-parametric likelihood where the environmental effect is treated as a nuisance parameter. By combining families with a twin study, as compared to a case-control study, we have more information on familial genotypes distribution, by assuming Mendelian inheritance, random mating and HWE.

The disease of interest in this article is RA, a genetic disorder in which alleles of the HLA-DRB1 gene contribute most to the genetic risk. A group of alleles in this gene, called DERAA alleles, are known to have a protective effect against RA, when present in the genotype of the offspring. Recent observations suggest that biologically relevant exposure to HLA-antigens may occur during fetal development and subsequently through the persistence, of maternal cells in the offspring. This phenomenon is called micro-chimerism. It has been proposed that not only inherited but also non-inherited maternal HLA-antigens can influence RA susceptibility [Feitsma et al., 2007]. This implies that the exposure of DERAA-negative offspring to maternal DERAA-positive HLA-DRB1 antigens during fetal development might have a protective effect on the offspring. We applied the combined joint likelihood (CJL) to 94 multi-case RA nuclear families [Hay et al., 1993; Worthington et al., 1994] and 78 dizygotic twin pairs [Silman et al., 1993], both collected from the National Repository of Family Material of the Arthritis and Rheumatism Council's.

Our method is a general framework for family-based association analysis, incorporating the advantages of several previously proposed methods such as combining different data sets, likelihood-based modelling, ascertainment correction and modeling correlation between disease outcome of siblings.This novel method models the joint genotype and phenotype distribution, taking into account the ascertainment and correlation present in the data, and combines families and twins studies to increase information to estimate the NIMA effect. In the next sections we introduce the general idea of the CJL for family-based and twin-based studies; we provide detailed estimation procedures for the family study and generalize the method to the twin study. The performance of our proposed method is assessed via an extensive simulation study and different approaches are compared for several scenarios, on the efficiency to estimate genetic odds ratios. The proposed method is illustrated with an analysis of the Arthritis and Rheumatism Council data.

2.2 Material And Methods

2.2.1 Notation and Data

Consider a study where information is available from two different data sets, a family-based and a twin-based study. For every family, genotype and phenotype information is available for the offspring, affected and/or healthy, and most of their parents. Families were ascertained on the event of at least two affected offspring per family. Genotypic and phenotypic information is also available for each twin, but not for their parents. Twin pairs were ascertained such that each pair contains at least one affected member.

Let $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, ..., Y_{in_i})$ denote phenotypes or disease status of n_i offspring in family i, where Y_{ij}=1 if offspring j is affected and Y_{ij}=0 if j is unaffected, i=1,...,N_f

and j=1,...,n_i. Similarly, let $\mathbf{G}_i^c = (G_{i1}^c, G_{i2}^c, ..., G_{in_i}^c)$ denote the genotypes of the n_i offspring and $\mathbf{G}_i^p = (G_i^m, G_i^f)$ their maternal and paternal genotypes. We denote by N_f and N_t the total number of families and twin pairs respectively. Last, let A_i be the ascertainment event for a family or twin pair.

2.2.2 Statistical Models

A commonly used approach for family data is the conditional logistic regression [Breslow and Day, 1980]. It conditions on the number of observed cases in each family, to accommodate for the outcome-dependent sampling scheme, and uses a family specific random term, to account for dependencies in disease risk among siblings. When twins are also available, we propose to estimate the genetic odds ratios by maximizing the combined likelihood for families and twins, instead of a families-only approach. Under the assumptions that the data sets are sampled separately from the same population, with no overlap between them and with comparable data collection methods, the combined likelihood can be obtained by the product of the likelihoods for each independent study.

Likelihood for family-based study

To model the association between genotypes and phenotypes of family members we use the JL. This approach is based on the joint probability of phenotypes and genotypes, that is $P\left(\mathbf{Y}_i, \mathbf{G}_i^c, \mathbf{G}_i^p \mid A_i\right)$ and is given by:

$$JL_f\left(\boldsymbol{\theta}\right) = \prod_{i=1}^{N_f} P\left(\mathbf{Y}_i, \mathbf{G}_i^c, \mathbf{G}_i^p \mid A_i\right), \tag{2.1}$$

where $\boldsymbol{\theta}$ is the parameter vector. $P\left(\mathbf{Y}_i, \mathbf{G}_i^c, \mathbf{G}_i^p \mid A_i\right)$ for family i is defined as follows:

$$\begin{aligned} P\left(\mathbf{Y}_i, \mathbf{G}_i^c, \mathbf{G}_i^p \mid \sum_{j=1}^{n_i} Y_{ij} \geq 2\right) &= \frac{P\left(\mathbf{Y}_i, \mathbf{G}_i^c, \mathbf{G}_i^p, \sum_{j=1}^{n_i} Y_{ij} \geq 2\right)}{P\left(\sum_{j=1}^{n_i} Y_{ij} \geq 2\right)} \\ &= \frac{P\left(\mathbf{Y}_i \mid \mathbf{G}_i^c, \mathbf{G}_i^p\right) \times P\left(\mathbf{G}_i^c \mid \mathbf{G}_i^p\right) \times P\left(\mathbf{G}_i^p\right)}{P\left(\sum_{j=1}^{n_i} Y_{ij} \geq 2\right)}. \end{aligned} \tag{2.2}$$

The second identity of (2.2) requires two assumptions. First, subjects selection should depend only upon potential subjects' disease status, not on their covariates, that is

$$P\left(\sum_{j=1}^{n_i} Y_{ij} \geq 2 \mid \mathbf{Y}_i, \mathbf{G}_i^c, \mathbf{G}_i^p\right) = P\left(\sum_{j=1}^{n_i} Y_{ij} \geq 2 \mid \mathbf{Y}_i\right).$$

Secondly, families should be selected under complete ascertainment, that is $P\left(\sum_{j=1}^{n_i} Y_{ij} \geq 2 \mid \mathbf{Y}_i\right) = 1$ for a family with at least two affected offspring, and 0 otherwise.

The numerator of (2.2) is a product of the *disease penetrance function* $P(\mathbf{Y}_i \mid \mathbf{G}_i^c, \mathbf{G}_i^p)$, the *transmission probabilities* $P\left(\mathbf{G}_i^c \mid \mathbf{G}_i^p\right)$ and the *parental genotype probabilities* $P\left(\mathbf{G}_i^p\right)$. The disease penetrance function models the disease probability of n_i offspring given the genotypes of the family. We will explain how we model the penetrance function in the next section. We assume Mendelian inheritance for the transmission probability $P\left(\mathbf{G}_i^c \mid \mathbf{G}_i^p\right)$, random mating for the parents and HWE for the genotype distribution. Thus, the parental genotype probability $P\left(\mathbf{G}_i^p\right)$ is characterised by a single parameter, the allele frequency q.

The denominator is the *ascertainment correction* and models the probability that at least two offspring in the family are affected. This probability can be expressed in terms of the marginal distribution by summing the joint distribution of phenotype and genotypes over all possible genotype combinations in a family, that is :

$$P\left(\sum_{j=1}^{n_i} Y_{ij} \geq 2\right) = 1 - \sum_{\mathbf{G}_*^c, \mathbf{G}_*^p} P\left(\mathbf{G}_*^c \mid \mathbf{G}_*^p\right) \times P\left(\mathbf{G}_*^p\right) \tag{2.3}$$
$$\times \left\{ P\left(\sum_{j=1}^{n_i} Y_{ij} = 1 \mid \mathbf{G}_*^c, \mathbf{G}_*^p\right) + P\left(\sum_{j=1}^{n_i} Y_{ij} = 0 \mid \mathbf{G}_*^c, \mathbf{G}_*^p\right) \right\}.$$

Disease penetrance function

In this section we present the penetrance function for a family in the data set. Given a set of family-specific random effects u_i, we assume that $(Y_{i1}, Y_{i2}, ..., Y_{in_i})$ are conditionally independent. Thus, the penetrance function for one family can be expressed as the product of the penetrance functions for each offspring in the family:

$$P\left(\mathbf{Y}_i \mid \mathbf{G}_i^c, \mathbf{G}_i^p, u_i\right) = \prod_{j=1}^{n_i} P\left(Y_{ij} = y_{ij} \mid G_{ij}^c, \mathbf{G}_i^p, u_i\right).$$

In order to estimate the parameters of interest, we use the marginal probability of the disease outcome of the ith family, given by:

$$P\left(\mathbf{Y}_i \mid \mathbf{G}_i^c, \mathbf{G}_i^p\right) = \int_{u_i} P\left(\mathbf{Y}_i \mid \mathbf{G}_i^c, \mathbf{G}_i^p, u_i\right) f\left(u_i\right) du_i. \tag{2.4}$$

We assume that the random intercept is normally distributed, $u_i \sim N\left(0, \tau_u^2\right)$. The integral is analytically intractable and we resort to numerical integration. To evaluate the integral we used the Gauss - Hermite Quadrature rule.

Last, we specify the individual penetrance function. We consider here the case where a direct offspring allelic effect and an indirect NIMA effect affect the disease probability for each offspring. We assume no direct maternal or paternal allelic effect. The disease probability for each offspring is a function of offspring genotype, combination of maternal and offspring genotype and the random effect u_i:

$$P\left(Y_{ij} = 1 \mid G_{ij}^c, \mathbf{G}_i^m, u_i\right) = logit^{-1}(\beta_0 + \beta_1 \times I[OAE_{ij}] \tag{2.5}$$
$$+\beta_2 \times I[NIMA_{ij}] + u_i),$$

Table 2.1: Possible genotype combination of mother-offspring pair and resulting protective effects. PA: protective allele, NIMA: non-inherited maternal antigens.

Offspring genotype	Maternal genotype	Resulting effect
0 copies of PA	0 copies of PA	Reference Category
0 copies of PA	1 copy of PA	NIMA effect
1/2 copies of PA	0/1/2 copies of PA	Offspring allelic effect

where $logit^{-1}$ is the inverse logit function, $logit^{-1}(x) = \frac{exp(x)}{1+exp(x)}$. Parameter β_0 is the intercept of the logistic model. Let $I[.]$ denote an indicator function. OAE_{ij} denotes an event of offspring allelic effect. We assume a dominant model, where $OAE_{ij} = 1$ when one or two copies of the protective allele are present in the offspring's genotype and zero otherwise. Parameter β_1 represents the log odds ratio of disease probability for the offspring allelic effect. Let $NIMA_{ij}$ denote an event of NIMA, where $NIMA_{ij} = 1$ if a copy of the protective allele is present in the maternal genotype but not present in the offspring's genotype and zero otherwise. Parameter β_2 represents the log odds ratio of the NIMA effect. The interpretation of parameters is conditional on the family specific random effects. In Table 2.1 all possible genotype combination of mother-offspring pair and resulting effects are reported.

Likelihood for twin-based study

In this section we modify the JL presented in the previous section to model data from twin-based studies. Since no parental genotypes are available in the twin study, it is not possible to estimate the indirect NIMA effect. Namely, the twin likelihood contains no information about NIMA. However, we need to include the NIMA parameter in the twin likelihood to ensure that the parameters of the family and twin likelihood have the same interpretation. Missing data is dealt with by marginalizing over all possible parental genotypes combinations, treating β_2 as a nuisance parameter. Following the notation used in (2.1), the JL for the twin data set is given by:

$$JL_t(\boldsymbol{\theta}) = \prod_{i=1}^{N_t} P(\mathbf{Y}_i, \mathbf{G}_i^c, \mid A_i), \tag{2.6}$$

where $P(\mathbf{Y}_i, \mathbf{G}_i^c, \mid A_i)$ for twin pair i is given as follows:

$$P\left(\mathbf{Y}_i, \mathbf{G}_i^c \mid \sum_{j=1}^{2} Y_{ij} \geq 1\right) = \sum_{\mathbf{G}_*^p} P\left(\mathbf{Y}_i, \mathbf{G}_i^c, \mathbf{G}_*^p \mid \sum_{j=1}^{2} Y_{ij} \geq 1\right)$$
$$= \sum_{\mathbf{G}_*^p} \frac{P(\mathbf{Y}_i \mid \mathbf{G}_i^c, \mathbf{G}_*^m) \times P(\mathbf{G}_i^c \mid \mathbf{G}_*^p) \times P(\mathbf{G}_*^p)}{1 - \sum_{\mathbf{G}_*^c, \mathbf{G}_*^p} P\left(\sum_{j=1}^{2} Y_{ij} = 0 \mid \mathbf{G}_*^c, \mathbf{G}_*^m\right) \times P(\mathbf{G}_*^c \mid \mathbf{G}_*^p) \times P(\mathbf{G}_*^p)}.$$

Combined likelihood for the family and twin studies

To obtain joint estimates for the NIMA and direct offspring allelic effect we maximize the combined likelihood for both data sets, given by the product of the likelihood contribution from family study (2.1), and the likelihood contribution from twin study (2.6):

$$CJL\left(\tau_u, \beta_0, \beta_1, \beta_2\right) = JL_f\left(\tau_u, \beta_0, \beta_1, \beta_2\right) \times JL_t\left(\tau_u, \beta_0, \beta_1, \beta_2\right). \tag{2.7}$$

Information to estimate the direct allelic effect, the baseline risk and the variance of the random effect comes both from twins and families. On the other hand, the family likelihood allows us to estimate also the NIMA effect. By adding the twins to the families, we borrow information to better estimate the direct allelic effect, which will also improve the estimate of the NIMA parameter through the family likelihood.

2.3 Simulation Study

The primary goal of the simulation study was to test efficiency gain for estimating effects that depend on parental genotype, such as NIMA, when a twin data set, with missing parental information, is combined with a data set comprised of nuclear families. In addition, we wanted to study the finite sample properties of the JL itself and relative to the PL. In particular, we investigated the impact of family size, variance of random effects and ascertainment scheme on the parameter estimates, and compared our method with the PL used in previous studies, in terms of efficiency and bias of estimates of NIMA effect.

In each scenario, genotype frequencies were selected to mimic the frequency of DERAA alleles in the English population, i.e. .15 [Ann Morgan, personal communication]. To generate genotypes of family members, maternal and paternal genotypes were generated assuming random mating and HWE. Offspring genotypes were generated assuming Mendelian transmission. Disease outcomes of offspring were generated according to the random effects model (2.5). The family-specific random intercept was assumed to be normally distributed with mean zero and variance either 1.5 or 2.5, resembling results from previous literature on heritability of RA [van der Woude et al., 2009]. Two different ascertainment schemes were used, that is, families were included in the study if at least one or two offspring were affected. Twins were generated as families with two offspring, ascertained such that at least one twin per pair is affected. Parental genotype and phenotype information was ignored to mimic the real data set. We set β_0 to -3, representing a common disease with population prevalence approximately 5%. The true parameter values for offspring allelic and NIMA effect, β_1 and β_2, were fixed at -.5 and -1, corresponding to an odds ratio of .6 and .4 respectively. In total, 16 scenarios were generated, each consisting of 10^3 simulated data sets, with corresponding family and sample size, ascertainment scheme and variance of the random effect as indicated in Table 2.2.

To study the finite sample properties of the JL, we applied the likelihood to all scenarios of Table 2.2. Results are summarised in Table 2.3. Effect of different family and sample size on the parameter estimates is reflected by comparing scenarios 1-4. When both sample and family size are small, e.g. scenario 1, τ_u^2 is overestimated

Table 2.2: Simulation scenarios with varying sample and family size, ascertainment scheme and variance of the random effects. fam: family, of: offpsring, $\sum_j Y_{ij} \geq 1$: at least one affected offspring, $\sum_j Y_{ij} \geq 2$: at least two affected offspring, τ_u^2: variance of random effect.

Scenario	Nr. fam	Nr. of	Ascertainment	τ_u^2
1	100	3	$\sum_j Y_{ij} \geq 1$	1.5
2	100	5	$\sum_j Y_{ij} \geq 1$	1.5
3	500	3	$\sum_j Y_{ij} \geq 1$	1.5
4	500	5	$\sum_j Y_{ij} \geq 1$	1.5
5	100	3	$\sum_j Y_{ij} \geq 1$	2.5
6	100	5	$\sum_j Y_{ij} \geq 1$	2.5
7	500	3	$\sum_j Y_{ij} \geq 1$	2.5
8	500	5	$\sum_j Y_{ij} \geq 1$	2.5
9	100	3	$\sum_j Y_{ij} \geq 2$	1.5
10	100	5	$\sum_j Y_{ij} \geq 2$	1.5
11	500	3	$\sum_j Y_{ij} \geq 2$	1.5
12	500	5	$\sum_j Y_{ij} \geq 2$	1.5
13	100	3	$\sum_j Y_{ij} \geq 2$	2.5
14	100	5	$\sum_j Y_{ij} \geq 2$	2.5
15	500	3	$\sum_j Y_{ij} \geq 2$	2.5
16	500	5	$\sum_j Y_{ij} \geq 2$	2.5

resulting in an underestimated β_0. However, estimates of the log odds ratios for the offspring allelic and NIMA effect are nearly unbiased, -2.3% and 3.4% respectively. Increasing family size from 3 to 5, scenario 2, reduces the bias of both effects to .1% and 2.4% and their standard deviations by 8.5% and 11.43% respectively. On the other hand, increasing the number of families from 100 to 500, scenario 3, reduces the bias of both effects to -1.4% and -1.0% and their standard deviations by 55.6% and 58.4% respectively. To study the effect of different τ_u^2 on the parameter estimates, we compared scenarios 1-4 with scenarios 5-8 or/and scenario's 9-12 with scenarios 13-14. When τ_u^2 increases from 1.5 to 2.5, from scenario 1 to scenario 5, bias on the estimate of β_0 and τ_u^2 itself increases. However, this does not introduce much bias in the estimation of the offspring allelic and NIMA parameters. Different ascertainment schemes were compared by contrasting scenarios 1-4 with scenarios 9-12. Bias in τ_u^2 and β_0 estimates increases when ascertainment is $\sum_j Y_{ij} \geq 2$, as compared to $\sum_j Y_{ij} \geq 1$ while estimates of the offspring allelic and NIMA parameters remain unbiased, e.g. bias in scenario 9, for β_1 and β_2, is 1.9% and 5.7% respectively.

Next, we compare the two different likelihoods to model family/twin data in terms of efficiency, the PL used in existing methods, with the approach we use in this article, the JL. We define the percentage of efficiency improvement of likelihood A over B, for estimating a parameter β, as EI$= (1 - \frac{Var(\beta_A)}{Var(\beta_B)}) \times 100$. Positive values mean that likelihood A performs better. In Figure 2.1.a we plot the EI of the JL over the PL, for estimating the log odds ratios of the offspring allelic and NIMA effect. All values are positive, thus the JL is always more efficient. Improvement mainly depends on sample size and less on family size, e.g. EI is approximately the same in scenario 1 and 3 as compared to scenario 2. Moreover, improvement, due to JL, is higher when information is limited, i.e. when families are small and ascertainment is $\sum_j Y_{ij} \geq 2$.

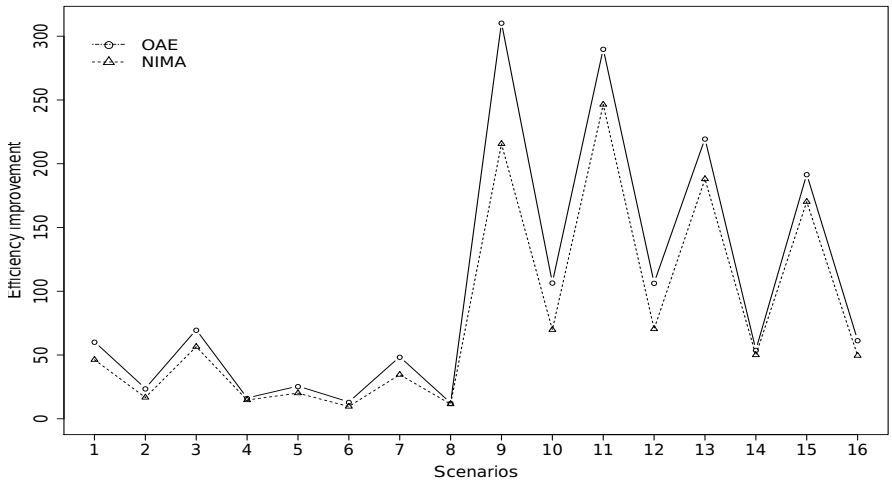
Last, we compared the performance of the JL when different data sources are available: ascertained families-only versus ascertained families and twins. In terms of likelihoods, we compare the JL in (2.1) with the CJL in (2.7). Efficiency improvement of the families-only against the combined approach, with families and 100 twin pairs, is plotted in Figure 2.1.b. The CJL approach is more efficient under all scenarios studied. The percentage of improvement is similar across different values of variance of the random effects or ascertainment scheme. Nonetheless, improvement is noticeably high when the sample size of the nuclear family data is small. When the twin data set was added, we expected efficiency improvement for the offspring allelic effect, due to increased sample size. Interestingly, there was also efficiency improvement for the NIMA effect, which depends on the maternal genotype. The parameter estimates and their standard deviations, using the CJL, are listed in Table A.2.1 of the Appendix.

In order to asses the performance of our method when both direct offspring and NIMA effects are under the null, $\beta_1 = \beta_2 = 0$, and cases in which there only exists a direct offspring, $\beta_2 = 0$, or only a NIMA effect, $\beta_1 = 0$, we simulated the scenarios presented in Table 2.2 with the corresponding effect sizes. We first estimated the effects optimising the JL using only the families. Later, we added 100 twin pairs and optimized the CJL. The estimated effect sizes remain unbiased. The results are listed in Tables A.2.2 and A.2.3 of the Appendix for the JL and in Tables A.2.4, A.2.5 and A.2.6 of the Appendix for the CJL.

Table 2.3: Summary statistics for parameter estimates of the JL (2.1) under the penetrance model (2.5) for each scenario described in Table 2.2. Each entry lists the mean estimates (standard deviation of estimates) over 1000 simulated data sets. JL: joint likelihood.

	True Values							
	$\tau_u^2 = 1.5$		$\beta_0 = -3$		$\beta_1 = -.5$		$\beta_2 = -1$	
1	2.148	(2.538)	-3.365	(1.460)	-.477	(.349)	-1.034	(.507)
2	1.584	(1.038)	-3.049	(.579)	-.501	(.319)	-1.024	(.449)
3	1.571	(.771)	-3.052	(.466)	-.486	(.155)	-.990	(.211)
4	1.543	(.411)	-3.022	(.226)	-.497	(.140)	-1.000	(.197)
	$\tau_u^2 = 2.5$		$\beta_0 = -3$		$\beta_1 = -.5$		$\beta_2 = -1$	
5	3.517	(3.382)	-3.476	(1.612)	-.478	(.397)	-1.016	(.541)
6	2.724	(1.662)	-3.104	(.766)	-.503	(.344)	-1.022	(.469)
7	2.587	(1.152)	-3.045	(.572)	-.492	(.169)	-1.001	(.231)
8	2.577	(.629)	-3.033	(.290)	-.504	(.149)	-1.003	(.196)
	$\tau_u^2 = 1.5$		$\beta_0 = -3$		$\beta_1 = -.5$		$\beta_2 = -1$	
9	2.827	(3.001)	-4.236	(2.864)	-.519	(.265)	-1.057	(.407)
10	1.980	(1.765)	-3.408	(1.442)	-.501	(.258)	-1.020	(.386)
11	2.472	(2.290)	-3.929	(2.194)	-.499	(.120)	-.999	(.173)
12	1.607	(.672)	-3.091	(.560)	-.497	(.112)	-.994	(.167)
	$\tau_u^2 = 2.5$		$\beta_0 = -3$		$\beta_1 = -.5$		$\beta_2 = -1$	
13	3.468	(3.347)	-3.673	(2.465)	-.540	(.316)	-1.056	(.459)
14	2.944	(2.032)	-3.308	(1.341)	-.501	(.299)	-1.024	(.423)
15	3.524	(2.852)	-3.778	(2.148)	-.500	(.144)	-1.016	(.205)
16	2.716	(1.084)	-3.141	(.721)	-.501	(.129)	-.999	(.175)

(a)



(b)

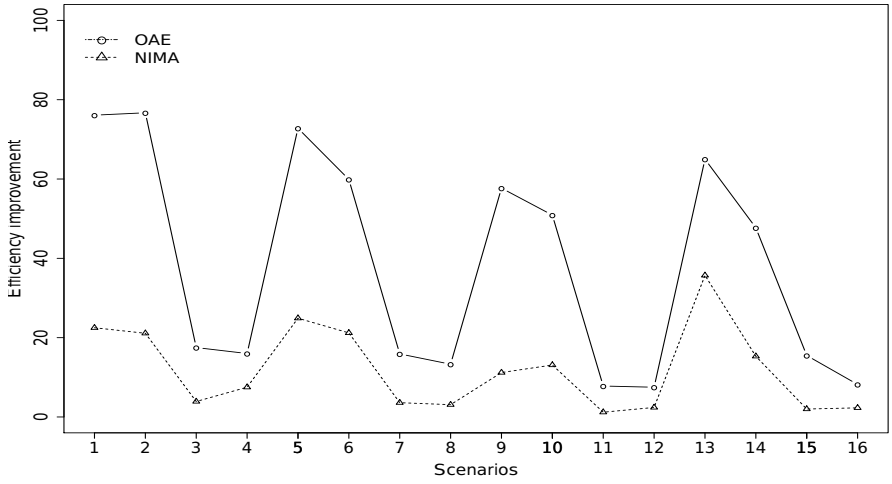


Figure 2.1: Efficiency improvement (EI) of (a) JL against PL and (b) CJL of families and twins, against the JL for families-only, compared for different family/sample size, ascertainment schemes and variance of random effect. (a) Values below zero represent no EI by using the JL and values above zero represent EI of the JL against the PL. (b) Values below zero represent no EI by using the CJL and values above zero represent EI of the CJL against the JL. Each point represents the EI in each of the sixteen scenarios presented in Table 2.2. JL: joint likelihood, PL: prospective likelihood, and CJL: combined joint likelihood, OAE: offspring allelic effect, NIMA: non-inherited maternal antigens effect.

The performance of our approach will vary across different frequencies of the protective allele. All the results presented above concern an allele frequency of .15, in order to mimic the allele frequency in the population we are studying. To study the performance of the method when allele frequency is lower, we also applied the CJL to samples generated with a protective allele frequency of .05. As expected, the parameter estimates are more biased for small sample sizes. Larger samples are needed to obtain unbiased estimates. Results are listed in Table A.2.7 of the Appendix.

2.4 Data Example

This study was motivated by a data set consisting of 94 ascertained nuclear families, collected from the Arthritis and Rheumatism Council. Our goal is to study the effect of NIMA in RA susceptibility. In 51 families the genotype of one of the parents, mainly the father, was missing. In 34 families, of which 8 had a missing mother and 26 a missing father, we were able to construct the genotypes using the genotypes of the offspring and the genotype of the other parent. Namely, we reconstructed the missing genotype in accordance with Mendelian transmission law. For the remaining 17 families, of which 9 were mothers and 8 were fathers, we were able to reconstruct only one of the alleles using this approach. In order to impute the second allele, we made use of the initial 4-digit allele coding of the HLA-DRB1 gene. There are 26 possible 4-digit sequences in the HLA-DRB1 gene, six of which express this DERAA allele, see van der Woude et al. [2010]. We imputed the second allele based on sampling from control 4-digit allele distribution. For 6 out of 9 mothers we had only the first 2 digits of the 4-digit genotyping and for the rest 3 we had no information about the second allele.

Families mainly contain two, three and four offspring. There are also three large families with five, eight and ten offspring. 86 families out of 94 contain exactly two affected offspring and 8 families contain three affected offspring. The maternal-offspring genotype combination that leads to the potential NIMA effect occurs only in 8 families. In these 8 families, 4 have one child, 2 have two children and 2 have three children under potential NIMA effect. In addition, 20 offspring belonging to 13 families are under offspring allelic effect. Since there is so little information in the family data set, we decided to combine it with a data set of 78 ascertained twin pairs, also collected from the Arthritis and Rheumatism Council in the same period. Pairs mainly contain one affected member and only in 3 pairs both members are affected. In 4 pairs both twins carry the DERAA allele, DERAA-concordant, while in 10 pairs only one twin has the allele, DERAA-discordant. In total, 18 twins are under offspring allelic effect. Information on parental genotype of twins is not available, thus the exact number of twins under a possible NIMA effect cannot be determined.

Initially, we only analyzed the family data, using both the JL and the PL approach. Results are listed in the first two lines of Table 2.4. None of the likelihoods gave statistically significant results for the NIMA effect, estimated odds ratios .176 (95% C.I. .010-3.066) and .607 (95% C.I. .348 1.058) for the PL and the JL approach respectively. Concerning the offspring allelic effect, only the JL resulted in a statistically significant result, odds ratios .194 (95% C.I. .023-1.622) for the prospective and .297 (95% C.I. .179 .493) for the joint likelihood approach. Then

Table 2.4: Parameter estimates (95% C.I.) of the disease penetrance model (2.5) by types of likelihood approaches used, prospective (PL), joint (JL), or combined joint likelihood (CJL), and type of data included, families only or families and twins.

Design	τ_u^2	β_0	β_1	β_2
		Families Only		
PL	1.570	.005	.190	.180
	(1.160-2.130)	(.001-.025)	(.020-1.620)	(.010-3.070)
JL	2.130	.001	.300	.610
	(1.630-2.790)	(.000-.006)	(.180-.490)	(.390-1.060)
		Families and Twins		
CJL	2.420	.002	.240	.480
	(1.710-3.420)	(.000-.010)	(.160-.380)	(.270-.870)

we combined the families with the twins and applied the CJL. The odds ratio of the NIMA effect was statistically significant, .477 (95% C.I. .264-.864) and the confidence intervals of the odds ratios of the offspring allelic effect became narrower; .241 (95% C.I. .152-.380).

To conclude, we estimated a significant protective effect of the DERAA allele, coming directly from the genotype of the offspring and indirectly from the maternal genotype. That is, individuals carrying the DERAA allele have a decrease in risk of RA compared to individuals who do not carry it. Furthermore, individuals who do not carry the protective allele DERAA, but their mother does, have a decrease in risk to develop RA as compared to non-DERAA carriers whose mother also does not carry the protective allele.

2.5 Discussion

In this article, we have presented a likelihood-based method for association studies combining family with twin data. Our method is appropriate for testing and estimating effects of genes that act directly through the individual's genotype but also for genes that act through complex biological mechanisms. We overcome the problem of small sample size by combining the family data set with a twin data set and using a JL approach to model the association between genotypes and phenotypes. By using a JL approach, we exploit the information coming from Mendelian transmission law, HWE, random mating and modeling of parental genotype distribution, to increase the efficiency to estimate the genetic odds ratios. The combined approach, not only enhances the statistical power to detect direct allelic effects, but also effects depending on maternal-offspring genotype combinations, such as NIMA effects. Namely, we use information from both data sets to better estimate the direct allelic effect, which gives us increased efficiency to estimate also the indirect NIMA effect. The method takes into account both the sampling scheme of the data and residual correlation be-

tween phenotype of siblings using an ascertainment correction and a family-specific random effects model.

Our approach extends existing methods for combining data sets [Pfeiffer et al., 2008; Zheng et al., 2010] to include indirect effects, using a JL, instead of a PL approach and adding twins, instead of a case-control data set. We compared the proposed JL method with the traditionally used PL approach and showed that our method is more efficient for estimating the genetic odds ratios, especially for small families with stringent selection schemes. For prospective or joint likelihood methods, including ours, ascertainment correction is essential to obtain unbiased parameter estimates. Here, we considered cases for which subjects' selection depends only upon potential subjects' disease status and not on their covariates. When ascertainment is also based on covariates, here genotypes, another model for ascertainment correction should be considered.

Using the JL, power can considerably increased, however at the cost of greater computational intensity, in the presence of large families. In our data set, the families where relatively small and numerical optimization of the JL was possible on a single computer. However, in the presence of large families, the computational burden rises exponentially with the family size. For given parameter values and allele frequency, the denominator (2.3) for family i sums over maximum 3^{n_i} possible familial genotype combinations. If all the families in the data set have a fixed size, the denominator needs to be calculated only p times for each maximization iteration, where p is the number of sample points to use for the Gauss-Hermite Quadrature approximation of the integral (2.4). Unfortunately, this is rarely the case in real data sets where the family size varies but the computation burden can be essentially reduced by using a grid search.

Here, we combine a family data set with a twin data set. However, the method can be extended to include other types of readily available data, such as sibling-pairs, monozygotic twins, or case-parent trios data sets. Nowadays, the combination of already available data is facilitated from existing nationwide registries of families and twins at high risk for particular traits. Extension of the likelihood-based analysis described here, to accommodate multi-allelic marker, is trivial, if HWE and random mating assumptions are made. Although we have focused on association of single SNPs, the approach can be extended to allow for the analysis of haplotypes. Since haplotypes combine linkage disequilibrium information from multiple markers simultaneously, this approach could be more powerful than our current approach. Direct extension to accommodate haplotypes is not straightforward, due to the increase in the number of parameters needed to model the haplotypes, and is beyond the scope of this article. The proposed method can be extended to other complex biological mechanisms, such as maternal effects or imprinting, by adding the appropriate covariates in the logistic regression (2.5). Last, by incorporating our method to methodology applied in Houwing-Duistermaat et al. [2000], we could study whether genetic NIMA effects of RA could create a protection for diseases associated with RA, such as cardiovascular disease or anaemia.

We employed fully parametric models for the random effects distribution. Since no straightforward diagnostics are available to evaluate the validity of the random effects model assumptions, there is a potential for model misspecification. Nevertheless, the estimates of the fixed effects are robust to moderate misspecifications of the

underlying random effects distribution [Heagerty and Kurland, 2001; Pfeiffer et al., 2003]. One could also analyze the data simply by using a GEE approach [Liang and Zeger, 1986]. However, since the GEE estimates do not take into account the sampling design, the resulting covariate effect estimates might be biased, because the family and twin data sets are not a random sample of the families and twins in the population. While the random effects model allows one to accommodate ascertainment of the families as well as residual familial correlation, the interpretation of the parameters is conditional on the random effects [Fitzmaurice et al., 1993]. Marginal parameter estimates can be obtained using the approximate formula of Diggle et al. [1994]. This approximation uses the variance of the random effects. In the simulation study we observed that the estimate of the variance, needed for the marginalization, might be biased when sample size is small. Thus we recommend to use the approximation formula only when the sample size and/or family size are large, e.g. 500 families with 3 offspring when ascertainment is at least one affected offspring.

To conclude, we confirmed the protective effect of the inherited DERAA alleles, offspring allelic effect, and the non-inherited maternal DERAA alleles, NIMA effect. The simulation study and the result of the real data analysis suggest that a combined approach can be more powerful, as compared to a families-only approach, when the information on the initial family data set is restricted.

2.6 Appendix

Table A.2.1: Summary statistics for parameter estimates of the CJL when both direct genetic and NIMA effect are present, for each scenario in Table 2.2. Each entry lists the mean estimates (standard deviation of estimates) over 1000 simulated data sets.

	True Values							
	$\tau_u^2 = 1.5$		$\beta_0 = -3$		$\beta_1 = -.5$		$\beta_2 = -1$	
1	2.211	(2.511)	-1.391	(1.467)	-.508	(.263)	-1.104	(.458)
2	1.613	(.966)	-1.052	(.552)	-.517	(.240)	-1.080	(.408)
3	1.588	(.779)	-1.057	(.472)	-.492	(.143)	-1.007	(.207)
4	1.547	(.403)	-1.021	(.223)	-.502	(.130)	-1.013	(.190)
	$\tau_u^2 = 2.5$		$\beta_0 = -3$		$\beta_1 = -.5$		$\beta_2 = -1$	
5	3.558	(3.193)	-1.503	(1.596)	-.514	(.302)	-1.091	(.484)
6	2.749	(1.559)	-1.103	(.742)	-.525	(.272)	-1.083	(.426)
7	2.608	(1.151)	-1.051	(.574)	-.499	(.157)	-1.018	(.227)
8	2.581	(.617)	-1.032	(.287)	-.509	(.140)	-1.017	(.193)
	$\tau_u^2 = 1.5$		$\beta_0 = -3$		$\beta_1 = -.5$		$\beta_2 = -1$	
9	2.394	(2.403)	-1.645	(1.770)	-.509	(.211)	-1.088	(.386)
10	1.934	(1.594)	-1.320	(1.177)	-.511	(.210)	-1.065	(.363)
11	1.937	(1.495)	-1.386	(1.315)	-.499	(.112)	-1.009	(.172)
12	1.606	(.634)	-1.084	(.511)	-.500	(.108)	-1.005	(.165)
	$\tau_u^2 = 2.5$		$\beta_0 = -3$		$\beta_1 = -.5$		$\beta_2 = -1$	
13	3.893	(3.382)	-1.820	(1.312)	-.526	(.246)	-1.089	(.394)
14	3.073	(2.147)	-1.342	(1.312)	-.520	(.246)	-1.077	(.394)
15	3.252	(2.311)	-1.522	(1.609)	-.499	(.134)	-1.025	(.203)
16	2.741	(1.084)	-1.151	(.706)	-.505	(.124)	-1.010	(.173)

Table A.2.2: Summary statistics for parameter estimates of the JL under the null hypothesis of no direct genetic or NIMA effects, for each scenario in Table 2.2. Each entry lists the mean estimates (standard deviation of estimates) over 1000 simulated data sets.

	True Values							
	$\tau_u^2 = 1.5$		$\beta_0 = -3$		$\beta_1 = 0$		$\beta_2 = 0$	
1	2.559	(2.842)	-3.882	(.295)	-.013	(.419)	-.033	(3.189)
2	1.842	(1.552)	-3.280	(.283)	-.015	(.397)	-.027	(1.792)
3	1.717	(1.257)	-3.179	(.125)	-.007	(.178)	-.019	(1.368)
4	1.539	(.521)	-3.036	(.123)	.001	(.168)	-.007	(.583)
	$\tau_u^2 = 2.5$		$\beta_0 = -3$		$\beta_1 = 0$		$\beta_2 = 0$	
5	3.752	(2.734)	-3.870	(.318)	-.010	(.456)	-.032	(3.667)
6	3.004	(1.588)	-3.319	(.310)	-.025	(.432)	-.044	(2.326)
7	2.927	(1.355)	-3.314	(.134)	-.004	(.192)	-.015	(1.812)
8	2.621	(.620)	-3.085	(.136)	.001	(.186)	-.003	(.866)
	$\tau_u^2 = 1.5$		$\beta_0 = -3$		$\beta_1 = 0$		$\beta_2 = 0$	
9	2.644	(4.027)	-4.567	(.210)	-.009	(.326)	-.034	(2.803)
10	1.949	(2.483)	-3.568	(.202)	-.002	(.287)	-.029	(1.881)
11	2.674	(4.067)	-4.613	(.086)	-.001	(.133)	-.003	(2.856)
12	1.725	(1.537)	-3.291	(.090)	-.001	(.127)	-.010	(1.154)
	$\tau_u^2 = 2.5$		$\beta_0 = -3$		$\beta_1 = 0$		$\beta_2 = 0$	
13	3.006	(3.402)	-3.569	(.265)	-.009	(.392)	-.030	(3.006)
14	2.949	(2.149)	-3.424	(.243)	-.007	(.333)	-.021	(2.139)
15	3.094	(3.446)	-3.647	(.112)	-.001	(.166)	-.006	(3.109)
16	2.786	(1.431)	-3.287	(.106)	-.001	(.151)	-.001	(1.422)

Table A.2.3: Summary statistics for parameter estimates of the JL under the hypothesis of no NIMA effect, for each scenario in Table 2.2. Each entry lists the mean estimates (standard deviation of estimates) over 1000 simulated data sets.

	True Values							
	$\tau_u^2 = 1.5$		$\beta_0 = -3$		$\beta_1 = -.5$		$\beta_2 = 0$	
1	2.523	(2.872)	-3.860	(.317)	-.524	(.402)	-.027	(3.146)
2	1.792	(1.519)	-3.235	(.303)	-.512	(.380)	-.023	(1.732)
3	1.716	(1.279)	-3.186	(.135)	-.504	(.169)	-.014	(1.361)
4	1.538	(.536)	-3.032	(.131)	-.498	(.162)	-.012	(.588)
	$\tau_u^2 = 2.5$		$\beta_0 = -3$		$\beta_1 = -.5$		$\beta_2 = 0$	
5	3.760	(2.859)	-3.899	(.325)	-.519	(.442)	-.031	(3.742)
6	3.045	(1.632)	-3.359	(.320)	-.527	(.430)	-.037	(2.314)
7	2.901	(1.363)	-3.300	(.144)	-.504	(.188)	-.018	(1.791)
8	2.621	(.651)	-3.084	(.140)	-.500	(.183)	-.005	(.898)
	$\tau_u^2 = 1.5$		$\beta_0 = -3$		$\beta_1 = -.5$		$\beta_2 = 0$	
9	2.601	(4.038)	-4.535	(.224)	-.522	(.315)	-.034	(2.757)
10	2.013	(2.626)	-3.685	(.216)	-.502	(.281)	-.037	(1.926)
11	2.575	(3.623)	-4.500	(.090)	-.505	(.127)	-.004	(2.518)
12	1.680	(1.471)	-3.232	(.097)	-.500	(.122)	-.009	(1.092)
	$\tau_u^2 = 2.5$		$\beta_0 = -3$		$\beta_1 = -.5$		$\beta_2 = 0$	
13	3.052	(3.395)	-3.628	(.258)	-.526	(.387)	-.042	(2.937)
14	2.967	(2.229)	-3.446	(.250)	-.504	(.331)	-.027	(2.173)
15	3.094	(3.057)	-3.626	(.112)	-.508	(.156)	-.012	(2.738)
16	2.820	(1.525)	-3.322	(.112)	-.499	(.149)	-.005	(1.481)

Table A.2.4: Summary statistics for parameter estimates of the CJL under the null hypothesis of no direct genetic or NIMA effects, for each scenario in Table 2.2. Each entry lists the mean estimates (standard deviation of estimates) over 1000 simulated data sets.

	True Values							
	$\tau_u^2 = 1.5$		$\beta_0 = -3$		$\beta_1 = 0$		$\beta_2 = 0$	
1	2.586	(2.787)	-3.874	(.213)	-.068	(.359)	-.331	(2.079)
2	1.882	(1.525)	-3.272	(.215)	-.074	(.344)	-.312	(1.726)
3	1.736	(1.268)	-3.182	(.115)	-.024	(.171)	-.095	(1.381)
4	1.545	(.522)	-3.026	(.116)	-.019	(.162)	-.081	(.584)
	$\tau_u^2 = 2.5$		$\beta_0 = -3$		$\beta_1 = 0$		$\beta_2 = 0$	
5	3.839	(2.775)	-3.898	(.229)	-.075	(.394)	-.339	(2.689)
6	3.047	(1.557)	-3.315	(.230)	-.090	(.365)	-.326	(1.911)
7	2.943	(1.357)	-3.307	(.125)	-.023	(.189)	-.093	(1.018)
8	2.628	(.610)	-3.074	(.126)	-.022	(.178)	-.077	(.490)
	$\tau_u^2 = 1.5$		$\beta_0 = -3$		$\beta_1 = 0$		$\beta_2 = 0$	
9	2.288	(2.574)	-3.855	(.160)	-.025	(.293)	-.208	(2.244)
10	1.946	(1.966)	-3.480	(.173)	-.031	(.270)	-.201	(1.651)
11	1.904	(2.019)	-3.519	(.079)	-.003	(.124)	-.043	(1.510)
12	1.683	(1.313)	-3.223	(.086)	-.006	(.126)	-.049	(1.015)
	$\tau_u^2 = 2.5$		$\beta_0 = -3$		$\beta_1 = 0$		$\beta_2 = 0$	
13	3.462	(2.701)	-3.815	(.200)	-.037	(.333)	-.238	(1.966)
14	3.080	(1.976)	-3.467	(.195)	-.048	(.305)	-.218	(1.145)
15	2.978	(2.162)	-3.482	(.095)	-.005	(.151)	-.058	(1.058)
16	2.740	(1.347)	-3.227	(.101)	-.010	(.148)	-.045	(.354)

Table A.2.5: Summary statistics for parameter estimates of the CJL under the hypotheisis of no NIMA effect, for each scenario in Table 2.2. Each entry lists the mean estimates (standard deviation of estimates) over 1000 simulated data sets.

	True Values							
	$\tau_u^2 = 1.5$		$\beta_0 = -3$		$\beta_1 = -.5$		$\beta_2 = 0$	
1	2.599	(2.885)	-3.897	(.223)	-.575	(.344)	-.329	(2.130)
2	1.874	(1.580)	-3.261	(.226)	-.576	(.327)	-.317	(1.777)
3	1.749	(1.330)	-3.202	(.124)	-.522	(.164)	-.092	(1.419)
4	1.546	(.536)	-3.024	(.122)	-.520	(.154)	-.088	(.586)
	$\tau_u^2 = 2.5$		$\beta_0 = -3$		$\beta_1 = -.5$		$\beta_2 = 0$	
5	3.771	(2.797)	-3.869	(.243)	-.584	(.379)	-.346	(2.648)
6	3.037	(1.580)	-3.321	(.241)	-.597	(.359)	-.333	(2.159)
7	2.918	(1.367)	-3.297	(.132)	-.525	(.184)	-.098	(1.804)
8	2.629	(.668)	-3.074	(.132)	-.525	(.175)	-.082	(.923)
	$\tau_u^2 = 1.5$		$\beta_0 = -3$		$\beta_1 = -.5$		$\beta_2 = 0$	
9	2.262	(2.485)	-3.828	(.178)	-.526	(.294)	-.215	(2.174)
10	1.923	(1.934)	-3.465	(.184)	-.529	(.264)	-.212	(1.605)
11	1.850	(1.876)	-3.462	(.087)	-.504	(.125)	-.044	(1.371)
12	1.643	(1.230)	-3.169	(.093)	-.507	(.121)	-.050	(.945)
	$\tau_u^2 = 2.5$		$\beta_0 = -3$		$\beta_1 = -.5$		$\beta_2 = 0$	
13	3.525	(2.652)	-3.893	(.207)	-.537	(.339)	-.256	(2.867)
14	3.099	(2.007)	-3.496	(.201)	-.547	(.303)	-.235	(2.131)
15	3.086	(2.168)	-3.591	(.106)	-.507	(.150)	-.061	(2.057)
16	2.755	(1.322)	-3.240	(.108)	-.510	(.147)	-.053	(1.312)

Table A.2.6: Summary statistics for parameter estimates of the CJL under the hypotheisis of no direct genetic effect, for each scenario in Table 2.2. Each entry lists the mean estimates (standard deviation of estimates) over 1000 simulated data sets.

	True Values							
	$\tau_u^2 = 1.5$		$\beta_0 = -3$		$\beta_1 = 0$		$\beta_2 = -1$	
1	2.641	(2.971)	-3.959	(.205)	-.038	(.954)	-1.386	(2.257)
2	1.863	(1.575)	-3.283	(.206)	-.042	(.744)	-1.343	(1.778)
3	1.746	(1.354)	-3.209	(.111)	-.014	(.225)	-1.088	(1.460)
4	1.537	(.537)	-3.025	(.112)	-.011	(.219)	-1.088	(.594)
	$\tau_u^2 = 2.5$		$\beta_0 = -3$		$\beta_1 = 0$		$\beta_2 = -1$	
5	3.834	(2.835)	-3.936	(.227)	-.040	(.782)	-1.367	(2.718)
6	3.028	(1.549)	-3.332	(.223)	-.050	(.485)	-1.329	(1.760)
7	2.881	(1.334)	-3.281	(.121)	-.012	(.226)	-1.090	(1.077)
8	2.611	(.600)	-3.072	(.122)	-.012	(.220)	-1.081	(.826)
	$\tau_u^2 = 1.5$		$\beta_0 = -3$		$\beta_1 = 0$		$\beta_2 = -1$	
9	2.444	(2.638)	-4.061	(.155)	-.013	(.463)	-1.281	(2.304)
10	2.050	(2.022)	-3.627	(.169)	-.017	(.405)	-1.239	(1.694)
11	2.066	(2.102)	-3.744	(.077)	-.001	(.179)	-1.055	(1.064)
12	1.739	(1.361)	-3.303	(.083)	-.002	(.173)	-1.054	(.510)
	$\tau_u^2 = 2.5$		$\beta_0 = -3$		$\beta_1 = 0$		$\beta_2 = -1$	
13	3.750	(2.816)	-4.108	(.193)	-.018	(.445)	-1.275	(2.093)
14	3.209	(1.996)	-3.614	(.190)	-.023	(.392)	-1.229	(1.165)
15	3.254	(2.242)	-3.778	(.094)	-.001	(.192)	-1.060	(1.137)
16	2.843	(1.337)	-3.338	(.099)	-.006	(.179)	-1.050	(.839)

Table A.2.7: Summary statistics for parameter estimates of the CJL when both direct genetic and NIMA effect are present, for each scenario in Table 2.2. Each entry lists the mean estimates (standard deviation of estimates) over 1000 simulated data sets. The data are simulated with a frequency of the protective allele of .05.

	True Values							
	$\tau_u^2 = 1.5$		$\beta_0 = -3$		$\beta_1 = -.5$		$\beta_2 = -1$	
1	2.627	(2.966)	-3.956	(.355)	-.549	(12.533)	-4.059	(3.228)
2	1.858	(1.535)	-3.288	(.347)	-.565	(5.313)	-2.729	(1.721)
3	1.758	(1.321)	-3.220	(.187)	-.513	(.355)	-1.164	(1.427)
4	1.546	(.532)	-3.03	(.179)	-.518	(.347)	-1.136	(.584)
	$\tau_u^2 = 2.5$		$\beta_0 = -3$		$\beta_1 = -.5$		$\beta_2 = -1$	
5	3.731	(2.826)	-3.860	(.385)	-.571	(7.746)	-3.004	(3.726)
6	2.989	(1.508)	-3.322	(.371)	-.592	(3.835)	-2.251	(2.082)
7	2.892	(1.299)	-3.288	(.199)	-.520	(.347)	-1.158	(1.717)
8	2.606	(.604)	-3.070	(.192)	-.525	(.345)	-1.131	(.829)
	$\tau_u^2 = 1.5$		$\beta_0 = -3$		$\beta_1 = -.5$		$\beta_2 = -1$	
9	2.297	(2.562)	-3.894	(.292)	-.526	(6.810)	-3.143	(2.233)
11	1.946	(1.978)	-3.506	(.283)	-.514	(4.629)	-2.434	(1.662)
12	1.971	(2.015)	-3.619	(.145)	-.501	(.300)	-1.104	(1.502)
13	1.720	(1.301)	-3.277	(.153)	-.511	(.291)	-1.096	(1.005)
	$\tau_u^2 = 2.5$		$\beta_0 = -3$		$\beta_1 = -.5$		$\beta_2 = -1$	
14	3.528	(2.657)	-3.906	(.334)	-.524	(4.330)	-2.230	(2.912)
15	3.089	(1.965)	-3.509	(.320)	-.537	(3.234)	-1.888	(2.107)
16	3.181	(2.261)	-3.688	(.173)	-.506	(.311)	-1.119	(2.179)
17	2.812	(1.377)	-3.303	(.169)	-.513	(.297)	-1.096	(1.385)

3

Powerful Testing via Hierarchical Linkage Disequilibrium in Haplotype Association Studies [1]

Summary

Marginal tests based on individual SNPs are routinely used in genetic association studies. Studies have shown that haplotype-based methods may provide more power in disease mapping than methods based on single markers when, for example, multiple disease-susceptibility variants occur within the same gene. A limitation of haplotype-based methods is that the number of parameters increases exponentially with the number of SNPs, inducing a commensurate increase in the degrees of freedom and weakening the power to detect associations. To address this limitation, we introduce a hierarchical linkage disequilibrium model for disease mapping, based on a re-parametrization of the multinomial haplotype distribution, where every parameter corresponds to the cumulant of each possible subset of a set of loci. This hierarchy present in the parameters enables us to employ flexible testing strategies over a range of parameter sets: from standard single SNP analyses through the full haplotype distribution tests, reducing degrees of freedom and increasing the power to detect associations. We show via extensive simulations that our approach maintains the type I error at nominal level and has increased power under many realistic scenarios, as compared to single SNP and standard haplotype-based studies. To evaluate the performance of our proposed methodology in real data, we analyze genome-wide data on rheumatoid arthritis from the Wellcome Trust Case-Control Consortium. The method is publicly available at `https://github.com/BrunildaBalliu/HierarchicalLD`.

[1] Submitted for publication.

3.1 Introduction

Marginal tests based on individual single nucleotide polymorphisms (SNPs) have dominated association analyses in the past decade. Although single SNP analyses have led to the identification of hundreds of genetic variants associated with many complex diseases [Hindorff et al., 2009], greater power might be achieved by using haplotype-based approaches, analyzing multiple markers simultaneously. Haplotype-based association methods incorporate linkage disequilibrium (LD) information from multiple markers and can be more powerful for gene mapping than methods based on single SNPs [Akey et al., 2001; Zaykin et al., 2002; Epstein and Satten, 2003]. For example, haplotype-based methods will be more powerful when multiple disease-susceptibility variants, each with an independent effect, occur within the same gene [Morris and Kaplan, 2002]. Moreover, haplotype-based methods could be preferable to single SNP-based association methods when diseases arise from the interaction of multiple cis-acting susceptibility variants found within a gene, forming a 'super-allele' [Joosten et al., 2001; Tavtigian et al., 2001; Hollox et al., 2001; Clark et al., 1998; Drysdale et al., 2000], since haplotype based methods allow for super-additivity of multiple genetic variants, whereas marginal tests do not [Epstein and Satten, 2003].

Standard haplotype association methods test for differences in haplotype distributions between cases and controls or perform regression analyses in which haplotypes are treated as categorical variables [Schaid et al., 2002; Zaykin et al., 2002; Epstein and Satten, 2003; Spinka et al., 2005; Lin and Zeng, 2006; Boehringer and Pfeiffer, 2009]. Two detailed reviews on existing methods for haplotype-based association analysis are provided by Schaid [2004] and Liu et al. [2008]. Moving from single-SNP to haplotype-based analyses results in a considerable increase in polymorphism and in a commensurate increase in the number of association parameters and therefore the degrees of freedom (d.f.) of the association tests. As a result, the global score or likelihood ratio test statistics will be weakly powered. Moreover, when the haplotype data is sparse, the χ^2 approximation of the distribution of the test statistics might be invalid. An additional difficulty is the ambiguity in haplotype phase when only genotype data are observed. Ambiguity can be handled using an expectation-maximization (EM) algorithm [Dempster et al., 1977; Excoffier and Slatkin, 1995], however, the additional assumption of Hardy-Weinberg equilibrium (HWE) is needed. The d.f. problem and the problem due to many rare haplotypes remain a limitation and force to employ heuristic methods, such as grouping of rare haplotypes [Schaid, 2004]. Due to these limitations of the haplotype-based methods and the myriad possible genetic architectures of complex human diseases, the relative efficiency of using haplotypes versus single markers remains largely unexplored and is often decided by practical considerations.

In this work, we introduce a hierarchical LD model for trait mapping that enables us to employ flexible testing strategies over a range of parameter sets: from standard single SNP analyses through the comparison of full haplotype distributions, thereby allowing to reduce d.f. and increase the power to detect associations. Our model is based on a re-parametrization of the multinomial haplotype distribution, where every parameter corresponds to the joint cumulant of each possible subset of a set of loci [Thiele, 1899; Brillinger, 1991]. For M SNPs, the new parametrization consists of allele frequencies of each SNP, standard pairwise LD parameters (i.e. D'), and higher-

order $(3, \ldots, M)$ LD parameters, corresponding to generalization of the pairwise LD to multiple SNPs. The proposed method is applicable to phased and unphased data and is particularly useful for detecting SNP-SNP interaction effects, long range differences in LD, the presence of 'super-alleles', and all situations where standard haplotype analysis would be considered. Moreover, due to properties of the hierarchy, direct optimization procedures can be constructed, rather than EM-based estimation. Higher order LD among alleles at more than two loci has been suggested in the past by Bennett [1952] and described in Weir [1990] for the case of three and four SNPs. However, to the best of our knowledge, a full parametrization of the haplotype distribution in terms of LD parameters, for an arbitrary number of SNPs, has not yet been provided.

In the following sections, we develop the re-parametrization of the multinomial haplotype distribution, describe estimation procedures and statistical tests with reduced d.f. for inference, and provide guidelines on how our method can be used. A simulation study, based on realistic haplotype distribution from the Wellcome Trust Case Control Consortium (WTCCC) [Burton et al., 2007] and different disease generating models show that the procedure maintains the type I error rate at nominal level and has increased power over the standard single SNP or haplotype based association methods for a variety of realistic scenarios. We apply our method to unphased SNP genotype data from the WTCCC data on rheumatoid arthritis (RA) and identify several new associations.

3.2 Material and methods

3.2.1 Basic notation and assumptions

Consider the case of genotype measurements of M bi-allelic loci. Let $h \in H$ be a haplotype at these loci, with $H = \{0,1\}^M$ the set of possible haplotypes, $|H| = 2^M$. We assume that $h \sim Mult(1, \boldsymbol{\theta})$ with $\boldsymbol{\theta} = (\theta_h)_{h \in H}$ the parameter vector of the haplotype frequencies, $\boldsymbol{\theta} \in \Theta$ and $\Theta = \{\boldsymbol{\theta} \mid \boldsymbol{\theta} \in (0,1)^{2^M}, \sum_{h \in H} \theta_h = 1\}$.

For the situation when genotypes instead of haplotypes are observed, let $\mathbf{G} = (G_1, \ldots, G_N)$ denote genotypes of N individuals; $D = (h_1, h_2)$ denotes a diplotype, *i.e.* an ordered haplotype pair, and $S(g)$ denotes the set of diplotypes that are consistent with genotype g. By assuming HWE, we can model the diplotype distribution using the product distribution. Then, the likelihood of the data can be expressed as [Schaid, 2004]

$$L_0(\mathbf{G}; \boldsymbol{\theta}) = \prod_{i=1}^{N} \sum_{(h_1,h_2) \in S(G_i)} \theta_{h_1} \times \theta_{h_2}.$$

In the following, we consider case-control studies, with N_1 controls, N_2 cases and sample size $N = N_1 + N_2$. For genotypes $G = (G^{ca}, G^{co})$ the likelihood becomes

$$L(G, \theta) = L_0(G^{ca}, \boldsymbol{\theta}^{ca}) L_0(G^{co}, \boldsymbol{\theta}^{co}),$$

where $\boldsymbol{\theta}^{ca}$ and $\boldsymbol{\theta}^{co}$ are haplotype frequencies for cases and controls, respectively. Standard haplotype testing compares haplotype frequencies of cases and controls as

follows:

$$H_0 : \Theta^0 = \left\{(\boldsymbol{\theta}^{ca}, \boldsymbol{\theta}^{co}) \in \Theta^2 \mid \boldsymbol{\theta}^{ca} = \boldsymbol{\theta}^{co}\right\}, H_1 : \Theta^1 = \left\{(\boldsymbol{\theta}^{ca}, \boldsymbol{\theta}^{co}) \in \Theta^2\right\}. \quad (3.1)$$

Under the null hypothesis, parameters for cases and controls are constrained to be equal, while under the alternative any parameter component can differ between the groups. The EM algorithm can be used to maximize the log-likelihood and compute the maximum likelihood estimates under both the null and alternative hypothesis. The LR-statistic is then

$$LR = 2\left[\log L\left(\mathbf{G}; \hat{\boldsymbol{\theta}}^{\mathbf{1}}\right) - \log L\left(\mathbf{G}; \hat{\boldsymbol{\theta}}^{\mathbf{0}}\right)\right],$$

where $\hat{\boldsymbol{\theta}}^{\mathbf{0}} = \operatorname{argmax}_{\theta \in \Theta^0} L(\mathbf{G}; \boldsymbol{\theta})$ and $\hat{\boldsymbol{\theta}}^{\mathbf{1}} = \operatorname{argmax}_{\theta \in \Theta^1} L(\mathbf{G}; \boldsymbol{\theta})$. It follows from standard likelihood theory that LR is asymptotically $\chi^2_{2^M-1}$ distributed.

3.2.2 Re-parametrization of the multinomial haplotype distribution

In order to achieve our goal of reducing the d.f., we present a hierarchical model of LD. To this end, Lemma 1 establishes a re-parametrization $\boldsymbol{\delta}$ of the multinomial haplotype frequencies $\boldsymbol{\theta}$, where every parameter corresponds to the joint cumulant of each possible subset of a set of M loci. We start by defining the joint cumulant.

Definition. *Let $A = \{A_1, A_2, \ldots, A_M\}$ be a set of random variables. Let P_A refer to the set of partitions of set A into nonempty subsets (blocks). So, for $p \in P_A$, each $b \in p$ is a block. Then, the joint cumulant of the set of random variables A is given as*

$$\kappa(A) = \kappa(A_1, A_2, \ldots, A_M) = \sum_{p \in P_A} (-1)^{|p|-1}(|p|-1)! \prod_{b \in p} E\left(\prod_{A \in b} A\right),$$

where $|p|$ denotes the cardinality of set p.

We also use M-th order cumulant to denote $\kappa(A)$. The joint cumulant is a measure of how far random variables are from independence [Ahlbach et al., 2012]. Notice that if M = 1 or M = 2, the joint cumulant reduces to the expected value and covariance, namely $\kappa(A_1) = \mathsf{E}(A_1), \kappa(A_1, A_2) = \mathsf{E}(A_1 A_2) - \mathsf{E}(A_1)\mathsf{E}(A_2)$.

Lemma 1. *Let $A = \{A_1, A_2, \ldots, A_M\}$ a set of M random variables with $A_j \in \{0, 1\}$. For each $s \in S = 2^A \setminus \emptyset$, let $\delta_s = \kappa(s)$, i.e. the joint cumulant of random variables s. Then $\boldsymbol{\delta} = (\delta_s)_{s \in S}$ is a re-parametrization of $\boldsymbol{\theta}$.*

Here 2^A denotes the power set of A. We interpret A_i as a bi-allelic locus and get that the haplotype distribution can be described by a set of cumulants for which each cumulant uniquely corresponds to a subset of the M loci. Note that first order cumulants correspond to allele frequencies and second order cumulants correspond to standard pairwise LD. Thus, in cases of two SNPs, the re-parametrization reduces to the standard decomposition into allele frequencies and pairwise LD parameters [Weir, 1990]. A proof of Lemma 1 is given in appendix A.1. For a set $\{A_1, A_2, A_3\}$

of random variables, we will write δ_{123} as a shorthand of $\delta_{\{A_1,A_2,A_3\}}$ and η_{123} for $\mathsf{E}(A_1A_2A_3)$. η_{123} is the haplotype frequency for loci 1,2, and 3 with allele 1 chosen at each locus.

As an example to illustrate the lemma, consider the case of three loci. The eight haplotype frequencies $\boldsymbol{\theta} = (\theta_{000}, \theta_{100}, \theta_{010}, \theta_{001}, \theta_{110}, \theta_{101}, \theta_{011}, \theta_{111})^\mathsf{T}$ can be re-parametrized into three allele frequencies, denoted by δ_1, δ_2, and δ_3, three pairwise LD parameters, denoted by δ_{12}, δ_{13}, and δ_{23}, and one third order LD parameter, denoted by δ_{123}, that is $\boldsymbol{\delta} = (\delta_1, \delta_2, \delta_3, \delta_{12}, \delta_{13}, \delta_{23}, \delta_{123})^\mathsf{T}$. The pairwise LD parameters for all pair (j, k) of SNPs are given as

$$\delta_{jk} = \mathsf{E}(A_jA_k) - \mathsf{E}(A_j) \times \mathsf{E}(A_k) = \eta_{jk} - \delta_j \times \delta_k. \tag{3.2}$$

As in the case of pairwise LD, higher order LD parameters express the difference between observed and expected haplotype frequencies, when expected frequencies are computed under the assumption of independence, with a value of zero indicating that at least two disjoint subsets of SNPs are independent of each other, and any cumulant involving two (or more) independent SNPs will be zero [Ahlbach et al., 2012]. This becomes apparent from the third order LD parameter:

$$\delta_{123} = \eta_{123} - \delta_1\eta_{23} - \delta_2\eta_{13} - \delta_3\eta_{12} + 2\delta_1\delta_2\delta_3. \tag{3.3}$$

3.2.3 Parameter estimation

The re-parametrization of the haplotype frequencies into allele frequencies and different order LD parameters introduces a hierarchy in the parameters. Specifically, higher order parameters (corresponding to singletons, pairs, triples, etc) only depend on lower order parameters and are independent of same or higher order parameters, given the lower order ones. This hierarchical structure enables us to construct direct optimization procedures avoiding the need for an EM algorithm.

As an example, consider again the case of three SNPs. In the first step we estimate the allele frequencies $\hat{\delta}_j$, $j = 1, 2, 3$. In the second step we estimate the pairwise LD parameters, denoted by $\hat{\delta}_{jk}, j \neq k$, for all pairs j, k of SNPs. Notice that in (3.2) each δ_{jk} depends only on allele frequencies δ_j and δ_k, which we have estimated in the first step, and a single parameter η_{jk} involving a one-dimensional optimization. Similarly, δ_{123} is estimated by a one-dimensional optimization over η_{123} as all other terms in (3.3) can be recovered by applying Lemma 1 from the parameters already estimated. The whole algorithm starts with allele frequencies and performs $2^M - 1 - M$ ensuing single-parameter optimizations.

3.2.4 Standardized LD parameters

LD parameters have the disadvantage of depending on allele frequencies [Hedrick, 1987]. For the two locus case, Lewontin [1964] suggested normalizing the pairwise LD parameter by dividing it by achievable extremes for fixed allele frequencies:

$$\delta_{jk}^{max} = \begin{cases} \min(\delta_j, \delta_k) - \delta_j\delta_k, & \text{if } \delta_{jk} \geq 0 \text{ and} \\ \left|\max(0, \delta_j + \delta_k - 1) - \delta_j\delta_k\right|, & \text{if } \delta_{jk} < 0. \end{cases} \tag{3.4}$$

We suggest to generalize this concept to establish a standardized LD measure for an arbitrary number of loci. Recall that δ_A can be written as

$$\delta_A = \eta_A - \sum_{p \in P_A \setminus A} (-1)^{|p|}(|p|-1)! \prod_{b \in p} \eta_b = \eta_A - \sum_{p \in P_A \setminus A} R_\delta(p),$$

where $R_\delta(p)$ are terms depending on loci $b \in p$ with $|b| < M$. These rest terms $R_\delta(p)$ are considered fixed and bounds for η_A are to be determined completely analogous to the two locus case. Then

$$\delta_A^{\max} = \begin{cases} \eta_A^{\max} - R_\delta, & \text{if } \delta_A \geq 0 \text{ and} \\ \left|\eta_A^{\min} - R_\delta\right|, & \text{if } \delta_A < 0, \end{cases} \tag{3.5}$$

where $R_\delta = \sum_{p \in P_A \setminus A} R_\delta(p)$, and $\eta^{\max}$ and $\eta^{\min}$ are the upper and lower bound for η_A and are defined in appendix A.2. The standardized version of δ_A is then given as follows

$$\delta_A^{'} = \frac{\delta_A}{\delta_A^{\max}}$$

A value of 1 or -1 indicates that the examined loci have not been exposed to all possible recombinations and at least one of all possible haplotype is not present in the population. $\eta_A^{\min}$ and $\eta_A^{\max}$ can be used to define the parameter space in the LD-parametrization which we denote with Δ in the following.

3.2.5 Parameter testing

The hierarchy present in our parametrization enables us to focus on certain orders in the the hierarchy, thus sparing d.f. as compared to testing the full distribution. We start by re-formulating the global haplotype test in terms of LD parameters. Let $\boldsymbol{\delta}^{ca} = (\delta_s^{ca})_{s \in S}$ and $\boldsymbol{\delta}^{co} = (\delta_s^{co})_{s \in S}$ be parameter vectors for cases and controls, respectively. Then (3.1) can be restated as follows

$$H_0 : \Theta_\delta^0 = \left\{(\boldsymbol{\delta}^{ca}, \boldsymbol{\delta}^{co}) \in \Delta^2 \mid \boldsymbol{\delta}^{ca} = \boldsymbol{\delta}^{co}\right\}, H_1 : \Theta_\delta^1 = \left\{(\boldsymbol{\delta}^{ca}, \boldsymbol{\delta}^{co}) \in \Delta^2\right\} \tag{3.6}$$

Again, $LR = 2\left(\log L\left(\mathbf{G}; \hat{\boldsymbol{\delta}}^{\mathbf{1}}\right) - \log L\left(\mathbf{G}; \hat{\boldsymbol{\delta}}^{\mathbf{0}}\right)\right) \sim \chi^2_{2^M-1}$ where $\hat{\boldsymbol{\delta}}^{\mathbf{0}}$, $\hat{\boldsymbol{\delta}}^{\mathbf{1}}$ are ML estimates under the null and alternative. We will refer to (3.6) as a *Full* test because we are testing all orders of LD parameters.

We now consider two families of tests with reduced d.f. The first family consists of tests that involve only lower order LD parameters. We will refer to them as *Bottom-Up* tests. Let P be the set containing the orders for which we would like to test for differences, *e.g.* $P = \{1, 2\}$ if we consider both allele frequencies and pairwise LD. The corresponding null and alternative hypotheses for any such set P is:

$$\begin{aligned} H_0 :& \Theta_{BU,P}^0 = \left\{(\boldsymbol{\delta}^{ca}, \boldsymbol{\delta}^{co}) \in \Delta^2 \middle| \quad \forall s \in S : |s| \in P \Rightarrow \delta_s^{ca} = \delta_s^{co}\right\} \\ H_1 :& \Theta_{BU,P}^1 = \Theta_\delta^1 \end{aligned} \tag{3.7}$$

Under H_0 we only constrain parameters of orders contained in P to be equal.

The second family consists of tests that involve only higher order LD parameters, *e.g.* for $M = 3$, $P = \{2, 3\}$ focuses only on second and third order LD parameters.

We will refer to them as *Top-Down* tests. The corresponding null and alternative hypotheses for any such set P is given by:

$$\begin{aligned} H_0 :&\Theta^0_{TD,P} = \Theta^0_\delta \\ H_1 :&\Theta^1_{TD,P} = \left\{(\boldsymbol{\delta}^{ca}, \boldsymbol{\delta}^{co}) \in \Delta^2 | \quad \forall s \in S : |s| \notin P \Rightarrow \delta^{ca}_s = \delta^{co}_s \right\} \end{aligned} \tag{3.8}$$

Here, parameters are constraint to be equal between cases and control both under H_0 and H_1 except for higher order parameters under the alternative. Both families of tests allow to employ direct optimization both under the null and the alternative. Since lower order parameters are estimated first, higher order parameters, which depend on the lower order parameters, will automatically be estimated to honor these constraints. On the other hand, had we constrained higher order parameters, lower order parameters would have to change once higher order constraints are considered. In these cases ML estimates would have to be found by joint optimization of parameters.

Top-Down tests can be interpreted as performing interaction tests without correcting for main effects. Uncorrected main effect can induce apparent interactions thereby allowing to reject some hypotheses where all differences come from main effects (or orders not included). For these reasons we will interpret these tests as global tests.

3.3 Simulation study

To evaluate the finite sample properties of the proposed re-parametrization and the association tests, we performed a simulation study. In the first part, we investigated type I error and power of the tests in data simulated based on real three-SNP haplotype frequencies from the WTCCC RA study. Here, we focus on the four most significant associations identified from the WTCCC data analysis. In the second part, we study the performance of the tests under several disease generating models, *e.g.* SNPs with main effects only, interacting pairs of SNPs and 'super-alleles'.

In each simulated data set, all tests described in the previous section were applied. For comparison purposes we also list results on the single SNP tests and score test performed using the `R` package `haplo.stats` [Sinnwell and Schaid, 2013]. For the scenarios under the null hypothesis, 10^3 data sets were simulated, each consisting of 2000 cases and 3000 controls. For the scenarios under the alternative hypothesis, 1000 data sets were simulated, also consisting of 2000 cases and 3000 controls.

3.3.1 Data simulation and results using real haplotype frequencies

For each of the four triplets identified as significant from the analysis of the WTCCC data, we estimated the haplotype frequencies in the sample of cases, the sample of controls and the pool of samples. We list these values in Table 3.1. The LD parameters to which these frequencies correspond are listed in Table A.1 of appendix A.4. In order to simulate data under the null hypothesis, we draw random samples from a multinomial distribution using the frequencies estimated from the pool of samples. In order to simulate data under the alternative hypothesis, we draw random

Table 3.1: Estimated haplotype frequencies in the cases (Ca), controls (Co) and pool (P) of cases and controls samples for each of the four triplets identified from the WTCCC data analysis.

	Triplet 1			Triplet 2		
	P	Ca	Co	P	Ca	Co
θ_{000}	.596	.569	.613	.499	.479	.512
θ_{001}	.059	.063	.056	.200	.189	.208
θ_{010}	.104	.098	.107	.015	.015	.015
θ_{011}	.003	.006	.002	.047	.054	.043
θ_{100}	.192	.211	.180	.147	.165	.135
θ_{101}	.028	.037	.022	.062	.059	.064
θ_{110}	.017	.014	.019	.025	.033	.020
θ_{111}	.002	.002	.001	.004	.006	.003

	Triplet 3			Triplet 4		
	P	Ca	Co	P	Ca	Co
θ_{000}	.477	.464	.486	.358	.340	.370
θ_{001}	.135	.172	.110	.088	.115	.071
θ_{010}	.029	.031	.028	.240	.228	.247
θ_{011}	.010	.008	.011	.101	.084	.112
θ_{100}	.115	.112	.115	.148	.166	.137
θ_{101}	.067	.060	.071	.004	.004	.004
θ_{110}	.132	.116	.142	.054	.058	.052
θ_{111}	.036	.035	.037	.006	.006	.006

samples separately for the group of cases and controls from a multinomial distribution using the frequencies estimated in the sample of case and controls, respectively.

Results on type I error rate for all tests and triplets are listed in Table 3.2. At the nominal level, type I error should lie in the interval (4.68, 5.31) for a test to properly maintain type I error. In general, the type I error rate is well maintained. For Triplet 1, the *Bottom-Up* test for allele frequencies and one of the single SNP tests is slightly deflated, while for Triplet 3, both tests are slightly inflated. For Triplet 2 and 4, the *Top-Down* test for third order LD is slightly inflated. Moreover, the *Full* test, is deflated for Triplet 2, type I error rate. All reject rates lie between 4.51 and 5.54.

The power for all tests and triplets is also listed in Table 3.2. For an association to be considered significant in the genome-wide associations study setting, the p-value of the test should be smaller than 5×10^{-8}. In all triplets the single SNP test and both *Top-Down* tests reach power below 80%. Regarding the other tests, different tests seem to be more powerful in each triplet with the *Bottom-Up* test for $P = \{1, 2\}$ being the one with the most consistent power across all triplets. In all triplets the score test from `haplo.stats` performs comparable to the *Full* test or the *Bottom-Up* test for $P = \{1, 2\}$.

Table 3.2: Result on type I error rate (%) and power (%) for the scenarios simulated based on parameters from significant findings from the WTCCC data. The parameter values for each scenario are listed in Table A.1. The *Bottom-Up* tests with $P = \{1\}$ and $P = \{1, 2\}$ test only for differences in allele frequencies and in allele frequencies and pairwise LD parameters; the *Full* test tests for differences in all parameters; the *Top-Down* tests with $P = \{3\}$ and $P = \{2, 3\}$ test only for differences in third order LD parameters and in second and third order LD parameters; the Single SNP tests are three separate one d.f. tests and `haplo.stats` is a score test from package `haplo.stats`. *The score tests from `haplo.stats` can have different d.f. in each data set because the package automatically groups rare haplotypes.

Test		d.f.	Triplet 1	Triplet 2	Triplet 3	Triplet 4
				Type I Error Rate (%)		
Bottom-Up	$P = \{1\}$	3	4.56	5.18	5.47	4.91
	$P = \{1, 2\}$	6	5.05	4.62	5.11	4.92
Full		7	5.13	4.51	5.18	5.16
Top-Down	$P = \{3\}$	1	4.93	5.54	5.36	5.97
	$P = \{2, 3\}$	4	4.98	4.87	5.03	5.35
Single SNP	SNP 1	1	5.1	5.14	5.54	5.26
	SNP 2	1	4.52	5.03	5.17	5.12
	SNP 3	1	4.99	4.71	5.28	4.78
`haplo.stats`		7*	5.29	4.93	5.17	4.86
				Power (%)		
Bottom-Up	$P = \{1\}$	3	66.90	74.80	89.30	71.30
	$P = \{1, 2\}$	6	71.43	70.00	94.80	97.90
Full		7	66.30	65.60	96.70	97.30
Top-Down	$P = \{3\}$	1	0.00	0.00	0.00	0.00
	$P = \{2, 3\}$	4	0.20	0.00	4.40	24.10
Single SNP	SNP 1	1	20.10	23.80	9.80	10.40
	SNP 2	1	0.00	15.70	1.60	9.80
	SNP 3	1	15.20	0.00	47.30	0.00
`haplo.stats`		7*	70.50	69.00	96.80	97.30

3.3.2 Data simulation and results under different disease generating models

In this section we further study the type I error rate and power properties of each test under different disease models and different LD structures. In all scenarios, we considered four SNPs with allele frequencies equal to .05, .18, .31 and .45, respectively. Two structures of LD among the SNPs are considered. In Scenario 1, the SNPs were in equilibrium, thus all second, third and fourth LD parameters were equal to zero. In Scenario 2, the second order standardized LD parameters were set to .4, the third order LD standardized parameters were set to .1 and the fourth order LD parameter was set to zero. In both cases, we mapped the LD parameters to haplotype frequencies, which are listed in Table A.2 of appendix A.4, and used those frequencies to generate haplotype data for a large population of individuals. The LD parameters in Scenario 1 correspond to frequencies in which 11 out of 16 haplotypes had frequencies below 5% and six had frequencies below 1%. On the other hand, in Scenario 2 only four haplotypes had frequencies below 5%.

Using different disease models, we generate the disease status Y of each individual and then sampled 2000 individual from the population of cases and 3000 individuals from the population of controls. For each disease model the following logistic model was used

$$\text{logit}(\text{P}(Y = 1 \mid \mathbf{D})) = \alpha_0 + \sum_{j=1}^{4} \alpha_j G_j + \sum_{j,k=1, j\neq k}^{4} \alpha_{ij} G_j \times G_k + \sum_{s \in S} \gamma_s SA_s \quad (3.9)$$

where α_0 is the intercept; $\alpha_j, j = 1, \ldots, 4$ are the main effect odds ratios of each SNP, $\alpha_{jk}, j, k = 1, \ldots, 4, j \neq k$ are the interaction effect for each pair of SNP; γ_s are the main effects of the 'super-allele' at loci $s \in S$, with $S = \{\{2, 3\}, \{1, 2, 3\}, \{1, 2, 3, 4\}\}$ and

$$SA_{23} = \begin{cases} 0 & \text{if both } h_1 \text{ and } h_2 \notin D_{23} \\ 1 & \text{if one of } h_1, h_2 \in D_{23} \\ 2 & \text{if both } h_1 \text{ and } h_2 \in D_{23} \end{cases}, SA_{123} = \begin{cases} 0 & \text{if both } h_1 \text{ and } h_2 \notin D_{123} \\ 1 & \text{if one of } h_1, h_2 \in D_{123} \\ 2 & \text{if both } h_1 \text{ and } h_2 \in D_{123} \end{cases},$$

$$\text{and } SA_{1234} = \begin{cases} 0 & \text{if } h_1 \neq \text{`1111'} \text{ and } h_2 \neq \text{`1111'} \\ 1 & \text{if } h_1 = \text{`1111'}, h_2 \neq \text{`1111'} \text{ or } h_1 \neq \text{`1111'}, h_2 = \text{`1111'} \\ 2 & \text{if } h_1 = \text{`1111'} \text{ and } h_2 = \text{`1111'} \end{cases},$$

where $D_{23} = \{\text{`0110'}, \text{`1110'}, \text{`0111'}, \text{`1111'}\}$, i.e. all haplotypes that contain the '1' allele at loci 2 and 3 and $D_{123} = \{\text{`1110'}, \text{`1111'}\}$ the haplotypes that contain the '1' allele at loci 1, 2 and 3.

Under the null hypothesis, all parameters in (3.9), besides the intercept, were zero. Results on type I error rate for all tests and scenarios are listed in Table 3.3. For Scenario 2, in which the four SNPs were in LD, all tests properly control the type I error rate. For Scenario 1, however, some tests are deflated, *Bottom-Up* tests with $P = \{1, 2, 3\}$, the *Full* test and all three *Top-Down* tests, while `haplo.stats` is inflated.

For scenarios under the alternative hypothesis, six different disease models were considered. In Model 1, the four SNPs had only main effects on disease risk. In Model 2, SNP 2 and 3 had main and interaction effects on disease risk. In model 3, SNPs 1, 2 and 3 had only interaction effects. We also studied the power of our approach in the presence of 'super-alleles'. In this case we assumed that the combination of alleles over two, three and

Table 3.3: Result on type I error rate for each test and each scenario listed in Table A.2. The *Bottom-Up* tests with $P = \{1\}$ and $P = \{1, 2\}$ test only for differences in allele frequencies and in allele frequencies and pairwise LD parameters; the *Full* test tests for differences in all parameters; the *Top-Down* tests with $P = \{3\}$ and $P = \{2, 3\}$ test only for differences in third order LD parameters and in second and third order LD parameters; the Single SNP tests are three separate one d.f. tests and `haplo.stats` is a score test from package `haplo.stats`. *The score tests from `haplo.stats` can have different d.f. in each data set because the package automatically groups rare haplotypes.

Test		d.f.	Type I Error Rate	
			Scenario 1	Scenario 2
Bottom-Up	$P = \{1\}$	4	4.67	5.24
	$P = \{1, 2\}$	10	4.86	5.00
	$P = \{1, 2, 3\}$	14	4.43	4.76
Full		15	4.31	4.91
Top-Down Tests	$P = \{4\}$	1	4.22	5.04
	$P = \{3, 4\}$	5	3.93	5.00
	$P = \{2, 3, 4\}$	11	4.36	4.79
Single SNP	SNP 1	1	4.97	5.45
	SNP 2	1	4.59	5.03
	SNP 3	1	4.83	5.13
	SNP 4	1	5.20	5.20
`haplo.stats`		15*	6.09	5.03

four SNPs also had an effect of disease risk. In model 4, SNP 2 and 3 and the haplotype '11' over these two loci had a main effect; in model 5, SNP 1, 2 and 3 and the haplotype '111' had a main effect and in model 6, all four SNPs and the haplotype '1111' had a main effect. Results on power for all tests and models, as well as the exact parameter values for each model, are listed in Table 3.4 for Scenario 1 and in Table 3.5 for Scenario 2.

Based on these results we make the following observations. First, as expected, in the presence only of main effects, i.e. Model 1, for both Scenarios, the most powerful test is the *Bottom-Up* test with $P = \{1\}$. Second, although the *Bottom-Up* test with $P = \{1\}$ does not include second order parameters, its power is comparable to the power of *Bottom-Up* test with $P = \{1, 2\}$ in the presence of both main and interaction effects, i.e. Model 2, or in the presence only of interacting effects, i.e. Model 3. In the presence of 'super-alleles', the power to detect association when the LD among the involved loci is zero and the effect is spread accross three or four loci, i.e. Model 5 and 6 in Scenario 1, is much lower compared to the power in the presence of LD, i.e. Model 5 and 6 in Scenario 2. For both scenarios and all models, except Model 6 for Scenario 1, at least one of the *Bottom-Up* tests is more powerful than `haplo.stats`. The *Bottom-Up* test with $P = \{1, 2\}$ was the one with the most consistent power across all models and scenarios.

Table 3.4: Result on power of each test on Scenario 1. Non-zero parameters for Model 1: $\alpha_i = log(1.2)$, $i = 1, 2, 3, 4$; Model 2: $\alpha_2 = log(1.2), \alpha_3 = log(1.1)$, $\alpha_{12} = log(1.2)$; Model 3: $\alpha_{jk} = log(1.3)$, $j, k = 1, 2, 3$, $j \neq k$; Model 4: $\alpha_1 = \alpha_2 = log(1.1)$, $\gamma_{23} = log(1.5)$; Model 5: $\alpha_i = log(1.1)$,$k = 1, 2, 3$, $\gamma_{123} = log(1.5)$; Model 6: $\alpha_i = log(1.1)$, $i = 1, 2, 3, 4$, $\gamma_{1234} = log(5)$.

Test		d.f.	Power with 95 % CI					
			Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
	$P = \{1\}$	4	90.32	90.72	68.67	96.50	32.04	13.18
Bottom-Up	$P = \{1, 2\}$	10	74.82	86.26	89.41	94.52	19.75	10.16
	$P = \{1, 2, 3\}$	14	63.37	79.20	82.89	91.26	15.67	8.52
Full		15	59.57	77.45	81.47	90.24	14.21	8.11
	$P = \{4\}$	1	.00	.00	.00	.00	.00	.00
Top-Down	$P = \{3, 4\}$	5	.00	.00	.00	.00	.00	.00
	$P = \{2, 3, 4\}$	11	.00	.00	1.01	.00	.00	.00
	SNP 1	1	.00	.00	4.10	.00	.00	.00
Single SNP	SNP 2	1	2.70	79.80	19.10	89.40	1.40	.10
	SNP 3	1	10.50	11.00	3.70	16.90	1.20	.40
	SNP 4	1	16.00	.00	.00	.00	1.50	.10
`haplo.stats`		15	65.60	80.10	83.70	90.50	18.32	22.10

Table 3.5: Result on power of each test on Scenario 2. Non-zero parameters for Model 1: $\alpha_1 = \alpha_2 = log(1.2)$, $\alpha_3 = \alpha_4 = log(1.1)$; Model 2: $\alpha_2 = \alpha_3 = log(1.1)$, $\alpha_{12} = log(1.2)$; Model 3: $\alpha_{jk} = log(1.2)$, $j, k = 1, 2, 3$, $j \neq k$; Model 4: $\alpha_1 = \alpha_2 = log(1.1)$, $\gamma_{23} = log(1.3)$; Model 5: $\alpha_i = log(1.1)$,$k = 1, 2, 3$, $\gamma_{123} = log(1.3)$; Model 6: $\alpha_i = log(1.1)$, $i = 1, 2, 3, 4$, $\gamma_{1234} = log(2)$.

Test		d.f.	Power with 95 % CI					
			Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Bottom-Up	$P = \{1\}$	4	75.92	91.67	83.24	77.42	87.37	80.47
	$P = \{1, 2\}$	10	48.37	88.42	79.07	65.59	75.79	71.59
	$P = \{1, 2, 3\}$	14	36.12	81.25	69.48	62.37	64.21	62.60
Full		15	32.96	78.12	68.07	61.29	60.00	60.93
Top-Down	$P = \{4\}$	1	.00	.00	.00	.00	.00	.00
	$P = \{3, 4\}$	5	.00	.00	.00	.00	.00	.00
	$P = \{2, 3, 4\}$	11	.00	.00	.00	.00	.00	.00
Single SNP	SNP 1	1	2.40	.00	32.40	.00	.00	18.30
	SNP 2	1	40.60	85.00	50.40	70.00	50.00	15.20
	SNP 3	1	10.40	65.00	10.40	34.00	31.00	15.60
	SNP 4	1	6.70	.00	.00	1.00	17.00	10.90
`haplo.stats`		15*	37.10	80.00	71.70	61.00	63.00	67.10

3.4 Data example

To illustrate an application of the proposed association tests, we performed an analysis of a data set from the WTCCC, consisting of 1860 cases of RA and 2938 controls. In the initial analysis, single SNP tests were performed and several SNPs, strongly associated with RA, were identified [Burton et al., 2007]. In addition, a list of 59 SNPs, showing 'moderate' association with RA, with nominal significance in the range of 10^{-3} to 10^{-6}, was provided in the initial article. Some of these SNPs map to genes with plausible biological relevance however the single SNP analyses failed to pass the significance threshold.

Here, we investigate possible increase in the significance level of the 59 SNPs when a three SNP haplotype based analysis is used. For each of these SNPs we choose 40 neighboring SNPs that had passed quality control, 20 to the left and 20 to the right side of the SNP and construct all possible triplets between the SNPs that contain the moderately associated SNP. For each of the 59 SNP, 780 triplets were constructed. To avoid problems caused by high LD, we excluded from the analysis all triplets in which at least one of the standardized pairwise LD parameters was above 0.8. For the remaining triplets, the tests mentioned in the previous section were applied. For comparison purposes, we also show results from single-SNP analysis. A triplet of SNPs was considered to be associated with RA if the p-value exceeded the threshold $5 \times 10^{-8} / (N_{tests} \times N_{triplets})$, where $N_{tests} = 5$ is the total number of tests performed on each triplet and $N_{triplets}$ the total number of triplets tested for each 'moderately' associated SNP.

Several triplets containing the SNPs rs12723859 and rs12205634 showed a strong association with RA. Specifically, for rs12723859 we identified 40 triplets with 20 unique SNPs, and for rs12205634 we identified 5 triplets with 4 unique SNPs. For rs6920220, 3 triplets consisting of 4 unique SNPs, had p-values smaller than the genome-wide significance threshold 5×10^{-8} but they were no longer significant when adjusting for the multiple number of tests and triplets. For the other 56 SNPs no strong association with RA was identified from the haplotype analysis. In Table 3.6 we list for each of rs6920220, rs12723859, and rs12205634, the p-values of all tests for the two triplets that show the strongest association with RA. For rs6920220 we tested a total of 21 triplets. Only the *Bottom-Up* test for allele frequencies yields a p-value below 5×10^{-8}. If we correct for the number of tests and triplets tested no test yields a significant p-value. For rs12723859 and rs12205634 we tested a total of 144 and 38 triplets respectively. The *Full* test and the *Bottom-Up* tests for $P = \{1\}$ and $P = \{1, 2\}$ yield p-values below 5×10^{-8}. After correcting for the number of tests performed the *Bottom-Up* tests for $P = \{1\}$ no longer gives a significant association, the *Bottom-Up* tests for $P = \{1, 2\}$ is still significant.

Table 3.6: Results on real data

SNPs in the triplet		*Bottom-Up* Test		*Full* Test	*Top-Down* Test		Single SNP tests		
		$P = \{1\}$	$P = \{1, 2\}$		$P = \{3\}$	$P = \{2, 3\}$			
				SNP rs6920220					
rs11961920	rs11970411	2.6e-08	7.9e-08	2.3e-07	.89	.21	5e-06	.16	1.2e-05
rs11970411	rs674451	8.5e-09	9.9e-08	2.8e-07	.81	.56	5e-06	1.2e-05	.25
				SNP rs12723859					
rs12739961	rs1113523	1.8e-10	4.4e-11	5.6e-12	7.78e-03	8.50e-04	3e-05	.0013	2.2e-07
rs12739961	rs17013326	2.4e-10	7.3e-11	7.9e-12	6.40e-03	9.29e-04	3e-05	.0013	3.1e-07
				SNP rs12205634					
rs411136	rs210137	1.9e-08	4.8e-12	1.2e-11	.41	2.26e-05	5.2e-05	4.3e-05	6.9e-02
rs411136	rs210138	2.1e-08	1.1e-11	2.1e-11	3.7e-05	5.1e-05	5.2e-05	4.3e-05	6.9e-02

3.5 Discussion

In this article, we propose a re-parametrization of the multinomial haplotype distribution into allele frequencies, standard pairwise LD parameters, and higher-order LD parameters. Our re-parametrization enables us to employ flexible testing strategies over a range of parameter sets. For example, joint tests of single-SNPs and joint tests of single-SNPs and their pairwise LD showed in both simulated and real data that such tests can often have increased power as compared to the full global haplotype or single-SNP based tests.

In this study, we use rather simplistic multiple testing strategies, namely using a Bonferroni correction for multiple tests performed on the same genotype data. This is certainly not optimal as the performed tests are usually highly correlated. Among our future interests is to develop iterative or sequential testing procedures, *e.g.* [Meinshausen, 2008], which better exhaust the α level. Moreover, we have not focused on the choice of haplotype size or region covered as an optimal strategy. It is likely that the optimal number of SNPs used for haplotype-based approaches will depend on the population history and the genomic region, which is beyond the scope of this report. We are currently working on implementation of the hierarchical LD model in the context of equivalence testing for reconstruction of independent haplotype blocks.

For a case-control sample, population substructure and cryptic relatedness among subjects leads to over-dispersion of the chi-square test statistic for association and causes spurious rejections of the null hypothesis. The data set we are using is known to be fairly homogeneous [Burton et al., 2007] and we do not expect population stratification artifacts. As presented, our method does not allow incorporation of additional covariates and can only handle a binary trait in the present form. One way to deal with covariates at the moment is to perform stratified analyses in a Mantel-Haenszel framework.

To avoid diminished power from the large number of haplotype configurations [Schaid et al., 2002] proposed to either pool rare haplotypes into a single baseline group or to scan a large chromosomal region for sub-segments that may be associated with the trait, starting with single-locus associations, followed by 'sliding' tests for two-locus haplotypes, followed by 'sliding' tests for three-locus haplotypes, and so forth. We saw from our simulation study that, as the number of haplotype configuration increases, pooling rare haplotypes does not avoid the diminished power problem. In addition, analyses involving a series of adjacent markers assume that the most informative markers are the physically closest. However, this is not always the case and tests based on such associations will not always be optimal. Consider for example the case when relatively recent mutations have introduced correlation among two SNPs in a low LD region, with for example 5 SNPs separating them. In order to include the pairwise correlation of the two SNPs of interest, we would have to use a sliding window of size 7 and perform a test with $2^7 - 1 = 127$ d.f.. Given the large number of haplotype configurations, most haplotype frequencies will be very low and pooling most haplotypes would be unavoidable. On the other hand, one could repeat the same procedure, using again a sliding window of 7, but testing only for allele frequencies and pairwise LD parameters. In this case, one would need to perform a test with $7 + \binom{7}{2} = 28$ d.f.. In this study, we followed a similar, heuristic strategy that lead to the identification of novel associations.

In a given population, the mutations that are causal in disease etiology will have arisen on one or more ancestral haplotypes [Degli-Esposti et al., 1992] and thereafter will have spread to other haplotypes by recombination. Early on in this process, very-high-order association will exist, and the most powerful test for association will be a very-high-order association test, since the strength of the high-order effect more than outweighs the large number of d.f. However, this advantage will not survive in perpetuity, since, as shown in Clayton and Jones [1999] high-order effects will be rapidly diluted by recombination, at

progressively more rapid rates than first order association between a single marker or a pair of markers and disease. As a result, tests based on lower order effects will in general be more powerful than the full haplotype tests. This result is also supported by our simulation study, since in the scenarios we considered, *Bottom Up* tests are the most powerful accross all different disease models. Our proposed method allows to flexibly accomodate both higher- and lower-order LD scenarios.

Appendix

A.1. Proof of lemma 1

Consider again the case of genotype measurements of M bi-allelic loci. Let $h \in H$ be a haplotype at these loci, with $H = \{0,1\}^M$ enumerating the 2^M possible haplotypes. Assume that $h \sim Mult(1, \boldsymbol{\theta})$ with $\boldsymbol{\theta} = (\theta_h)_{h \in H}$ the parameter vector of the haplotype frequencies of each haplotype, $\boldsymbol{\theta} \in \Theta$ and $\Theta = \{\boldsymbol{\theta} \mid \boldsymbol{\theta} \in (0,1)^{2^M}, \sum_{h \in H} \theta_h = 1\}$. Let $A = \{A_1, A_2, \ldots, A_M\}$ a set of M random variables with $A_j \in \{0,1\}$ the indicator random variable for either one of the two alleles at locus $j, j = 1, \ldots, M$. Let $S = 2^A \setminus \emptyset$ be the power set of A, in lexicographical order, without the empty set, that is, the set of all singletons, pairs, triplets, etc of allele indicator random variables.

In order to prove that $\boldsymbol{\delta}$ is a reparametrization of $\boldsymbol{\theta}$, we introduce an intermediate parameterization, denoted as $\boldsymbol{\eta}$. Then the proof goes as follows. First, we show that $\boldsymbol{\eta}$ is a reparametrization of $\boldsymbol{\theta}$. To prove this, we introduce the function $f(\boldsymbol{\theta}, S)$ and prove that f is bijective. Then, we show that $\boldsymbol{\delta}$ is a reparametrization of $\boldsymbol{\eta}$, which implies that $\boldsymbol{\delta}$ is also a reparametrization of $\boldsymbol{\theta}$. Similarly, to prove this, we introduce the function $q(\boldsymbol{\eta}, S)$ and prove that q is bijective.

. *Mapping function g.* *For a set of random variables $s \in S$, let $\tau_s = \{v \in \{0,1\}^M \mid A_j \in s \Leftrightarrow v[j] = 1\}$ a tuple of all haplotypes whose j-th element is 1 if $A_j \in s$. We define g to be a function which takes as an input the parameter vector $\boldsymbol{\theta} = (\theta_h)_{h \in H}$ and outputs the joint expectation of random variables in s, which we denote with η_s. That is,*

$$g(\boldsymbol{\theta}, s) = E\left(\prod_{A \in s} A\right) = \sum_{h \in \tau_s} \theta_h = \eta_s.$$

We illustrate g with the following example. Let $M = 3$ and $s = \{A_1, A_2\}$. Then τ_s will contain two haplotypes, i.e. $\tau_{\{A_1,A_2\}} = \{(1,1,0),(1,1,1)\}$, and $g(\boldsymbol{\theta}, \{A_1, A_2\}) = E(A_1 A_2) = \theta_{(1,1,0)} + \theta_{(1,1,1)}$. We are thus computing the joint expectation of A_1 and A_2 or the haplotype frequency for loci 1 and 2 with allele 1 chosen at each locus.

For a haplotype $(1,1,1)$, we will write θ_{111} as a shorthand of $\theta_{(1,1,1)}$. Similarly, for a set $\{A_1, A_2, A_3\}$ of random variables, we will write η_{123} as a shorthand of $\eta_{\{A_1,A_2,A_3\}}$.

. *Mapping function f.* *We define f to be a function which takes as input the parameter vector $\boldsymbol{\theta} = (\theta_h)_{h \in H}$ and outputs the joint expectation of random variables in s for all $s \in S$. That is,*

$$f(\boldsymbol{\theta}, S) = \{g(\boldsymbol{\theta}_{\tau_s}, s)\}_{s \in S} = \left(\sum_{h \in \tau_s} \theta_h\right)_{s \in S} = (\eta_s)_{s \in S}.$$

We illustrate f with the following example. For $M = 3$ markers,

$$S = \{\{A_1\}, \{A_2\}, \{A_3\}, \{A_1, A_2\}, \{A_1, A_3\}, \{A_2, A_3\}, \{A_1, A_2, A_3\}\}.$$

Moreover

$$\begin{aligned}
\tau_{A_1} &= \{(1,1,1),(1,1,0),(1,0,1),(1,0,0)\},\\
\tau_{A_2} &= \{(1,1,1),(1,1,0),(0,1,1),(0,1,0)\},\\
\tau_{A_3} &= \{(1,1,1),(0,1,1),(1,0,1),(0,0,1)\},\\
\tau_{\{A_1,A_2\}} &= \{(1,1,1),(1,1,0)\},\\
\tau_{\{A_1,A_3\}} &= \{(1,1,1),(1,0,1)\},\\
\tau_{\{A_2,A_3\}} &= \{(1,1,1),(0,1,1)\}, \text{ and}\\
\tau_{\{A_1,A_2,A_3\}} &= (1,1,1).
\end{aligned}$$

Hence

$$f(\boldsymbol{\theta},S) = \begin{pmatrix} g(\boldsymbol{\theta}_{\tau_{A_1}},\{A_1\})\\ g(\boldsymbol{\theta}_{\tau_{A_2}},\{A_2\})\\ g(\boldsymbol{\theta}_{\tau_{A_3}},\{A_3\})\\ g(\boldsymbol{\theta}_{\tau_{\{A_1,A_2\}}},\{A_1,A_2\})\\ g(\boldsymbol{\theta}_{\tau_{\{A_1,A_3\}}},\{A_1,A_3\})\\ g(\boldsymbol{\theta}_{\tau_{\{A_2,A_3\}}},\{A_2,A_3\})\\ g(\boldsymbol{\theta}_{\tau_{\{A_1,A_2,A_3\}}},\{A_1,A_2,A_3\}) \end{pmatrix} = \begin{pmatrix} \theta_{111}+\theta_{110}+\theta_{101}+\theta_{100}\\ \theta_{111}+\theta_{110}+\theta_{011}+\theta_{010}\\ \theta_{111}+\theta_{011}+\theta_{101}+\theta_{001}\\ \theta_{111}+\theta_{110}\\ \theta_{111}+\theta_{101}\\ \theta_{111}+\theta_{011}\\ \theta_{111} \end{pmatrix} = \begin{pmatrix} \eta_1\\ \eta_2\\ \eta_3\\ \eta_{12}\\ \eta_{13}\\ \eta_{23}\\ \eta_{123} \end{pmatrix}$$

We are thus computing the frequency of the 'marginal' haplotypes over sets of singletons, pairs and triplets of markers.

Lemma 2. *: **Reparametrization** $\boldsymbol{\eta}$. Let $\boldsymbol{\eta} = (\eta_s)_{s\in S} = f(\boldsymbol{\theta},S) = \{g(\boldsymbol{\theta}_{\tau_s},s)\}_{s\in S}$, with $\boldsymbol{\eta}\in\Lambda$, and $\Lambda = \{\boldsymbol{\eta} \mid \boldsymbol{\eta} = f(\boldsymbol{\theta},S), \boldsymbol{\theta}\in\Theta\}$. Then, $\boldsymbol{\eta}$ is a re-parametrization of $\boldsymbol{\theta}$. That is, $f:\Theta\to\Lambda$ is bijective.*

Notice here that we limit $\boldsymbol{\eta}$ to take values in the image of function f. This guarantees that when a bijective function is used to map $\boldsymbol{\eta}$ back to $\boldsymbol{\theta}$'s, those haplotype frequencies will be properly defined, i.e. $\boldsymbol{\theta}\in(0,1)^M$ and $\sum_{h\in H}\theta_h$=1. Before we proceed to prove Lemma 2, we introduce the inverse functions of g and f.

. *Mapping function g^{-1}.* *Let $H^* = \left\{v\in\{0,1\}^M \mid \langle v,v\rangle \neq 0\right\}$ the set of all possible haplotypes over M loci except the haplotype containing only '0' alleles. For a haplotype $h\in H^*$, let $s_h = \{s\in S \mid h[j]=1 \Leftrightarrow A_j\in s\}$ and $\tau_h = \{s\in S \mid s_h\subseteq s\}$. We define g^{-1} to be a function which takes as an input the parameter vector $\boldsymbol{\eta} = \{\eta_s\}_{s\in S}$ and outputs θ_h, the frequency of haplotype h, for $h\in H^*$. That is,*

$$g^{-1}(\boldsymbol{\eta},h) = \sum_{s\in\tau_h}(-1)^{|s_h|+|s|}\eta_s = \theta_h.$$

We illustrate g^{-1} with the following example. Let $M=3$ markers and $h=(1,0,0)$. Then $s_h=\{A_1\}$ and $\tau_h = \{\{A_1,A_2\},\{A_1,A_3\},\{A_1,A_2,A_3\}\}$. Thus

$$\begin{aligned}
\theta_{100} &= (-1)^1\{(-1)^1\eta_1 + (-1)^2\eta_{12} + (-1)^2\eta_{13} + (-1)^3\eta_{123}\}\\
&= \eta_1 - \eta_{12} - \eta_{13} + \eta_{123} = \eta_1 - (\eta_{12}-\eta_{123}) - (\eta_{13}-\eta_{123}) - \eta_{123}\\
&= \eta_1 - \theta_{110} - \theta_{101} - \theta_{111}.
\end{aligned}$$

. *Mapping function $\boldsymbol{f}^{-1}$.* *We define f^{-1} to be a function which takes as input the parameter vector $\boldsymbol{\eta}$ and outputs the haplotype frequencies $\boldsymbol{\theta}$. That is,*

$$f^{-1}(\boldsymbol{\eta},H) = \left[\left\{g^{-1}(\boldsymbol{\eta}_{s_h},h)\right\}_{h\in H^*}, 1-\sum_{h\in H^*} g^{-1}(\boldsymbol{\eta}_{s_h},h)\right].$$

Notice here that the frequency of the haplotype which contains the '0' allele at all the markers is cumputed as one minus the sum of the frequencies of all other haplotypes, i.e. all $h \in H^*$. This guarantees that $\sum_{h \in H} \theta_h = 1$.

We illustrate f^{-1} with the following example. For $M = 3$ markers

$$H^* = \{(1,0,0),(0,1,0),(0,0,1),(1,1,0),(1,0,1),(0,1,1),(1,1,1)\},$$

and

$$\begin{aligned}
&s_{(1,0,0)} = \{A_1\} && \text{and} && \tau_{(1,0,0)} = \{\{A_1\},\{A_1,A_2\},\{A_1,A_3\},\{A_1,A_2,A_3\}\}.\\
&s_{(0,1,0)} = \{A_2\} && \text{and} && \tau_{(0,1,0)} = \{\{A_2\},\{A_1,A_2\},\{A_2,A_3\},\{A_1,A_2,A_3\}\},\\
&s_{(0,0,1)} = \{A_3\} && \text{and} && \tau_{(0,0,1)} = \{\{A_3\},\{A_1,A_3\},\{A_2,A_3\},\{A_1,A_2,A_3\}\},\\
&s_{(1,1,0)} = \{A_1,A_2\} && \text{and} && \tau_{(1,1,0)} = \{\{A_1,A_2\},\{A_1,A_2,A_3\}\},\\
&s_{(1,0,1)} = \{A_1,A_3\} && \text{and} && \tau_{(1,0,1)} = \{\{A_1,A_3\},\{A_1,A_2,A_3\}\},\\
&s_{(0,1,1)} = \{A_2,A_3\} && \text{and} && \tau_{(0,1,1)} = \{\{A_2,A_3\},\{A_1,A_2,A_3\}\},\\
&s_{(1,1,1)} = \{A_1,A_2,A_3\} && \text{and} && \tau_{(1,1,1)} = \{A_1,A_2,A_3\}.
\end{aligned}$$

Hence

$$f^{-1}(\boldsymbol{\eta}, H) = \begin{pmatrix} \theta_{111}\\ \theta_{011}\\ \theta_{101}\\ \theta_{110}\\ \theta_{001}\\ \theta_{010}\\ \theta_{100}\\ \theta_{000} \end{pmatrix} = \begin{pmatrix} \eta_{123}\\ \eta_{23}-\eta_{123}\\ \eta_{13}-\eta_{123}\\ \eta_{12}-\eta_{123}\\ \eta_{3}-\eta_{13}-\eta_{23}+\eta_{123}\\ \eta_{2}-\eta_{12}-\eta_{23}+\eta_{123}\\ \eta_{1}-\eta_{12}-\eta_{13}+\eta_{123}\\ 1-\left\{\sum_{h\in H^*} g^{-1}(\boldsymbol{\eta}_{s_h}, h)\right\} \end{pmatrix}.$$

We now proceed with the proof of Lemma 2. To prove that f is bijective we need to show that f is both injective, i.e. $\forall \boldsymbol{\theta}, \boldsymbol{\theta}^* \in \Theta, f(\boldsymbol{\theta}, H) = f(\boldsymbol{\theta}^*, H) \Rightarrow \boldsymbol{\theta} = \boldsymbol{\theta}^*$, and surjective, i.e. $\forall \boldsymbol{\eta} \in \Lambda, \quad \exists \boldsymbol{\theta} \in \Theta : f(\boldsymbol{\theta}, H) = \boldsymbol{\eta}$. Now that we have defined the inverse function of f, it is easy to show that,

$$f(\boldsymbol{\theta}, H) = f(\boldsymbol{\theta}^*, H) \Rightarrow f^{-1}\{f(\boldsymbol{\theta}, H)\} = f^{-1}\{f(\boldsymbol{\theta}^*, H)\} \Rightarrow \boldsymbol{\theta} = \boldsymbol{\theta}^*$$

and $\forall$ arbitrary parameter vectors $\boldsymbol{\eta} \in \Lambda$ we can choose $\boldsymbol{\theta} = f^{-1}(\boldsymbol{\eta}, H)$ such that $f(\boldsymbol{\theta}, S) = f\left\{f^{-1}(\boldsymbol{\eta}, H), S\right\} = \boldsymbol{\eta}$. This concludes the proof of the bijectivness of f, which concludes also the proof of Lemma 2.

We now proceed to prove that $\boldsymbol{\delta}$ is a reparametrization of $\boldsymbol{\eta}$ and hence a reparametrization of $\boldsymbol{\theta}$. First, we introduce functions $c(\boldsymbol{\eta}, s)$ and $q(\boldsymbol{\eta}, S)$.

. *Mapping function $\boldsymbol{\kappa}$.* *Let P_s refer to the family of sets of all possible partitions of a set of random variables $s, s \in S$, into nonempty subsets (blocks). So, for $p \in P_s$, each $b \in p$ is a block. Moreover, let $\tau'_s = 2^s \setminus \emptyset$ the power set of s minus the empty set. We define κ to be a function which takes as an input the parameter vector $\boldsymbol{\eta} = (\eta_s)_{s \in S}$ and outputs the joint cumulant of the set of random variables in s, which we denote by δ_s. That is,*

$$\kappa(\boldsymbol{\eta}, s) = \sum_{p \in P_s} (-1)^{|p|-1}(|p|-1)! \prod_{b \in p} E\left(\prod_{A \in b} A\right) = \sum_{p \in P_s} (-1)^{|p|-1}(|p|-1)! \prod_{b \in p} \eta_b = \delta_s.$$

We illustrate function c with the following example. Let $M = 3$ and $s = \{A_1, A_2, A_3\}$. Then,

$$P_s = \Big\{ \{A_1, A_2, A_3\}, \{\{A_1, A_2\}, \{A_3\}\}, \{\{A_1, A_3\}, \{A_2\}\}, \{\{A_2, A_3\}, \{A_1\}\}, \{\{A_1\}, \{A_2\}, \{A_3\}\}\}$$

and $\boldsymbol{\eta}_{\tau_s'} = (\eta_1, \eta_2, \eta_3, \eta_{12}, \eta_{13}, \eta_{23}, \eta_{123})$.

Thus,

$$\begin{aligned}\kappa\left(\boldsymbol{\eta}_{\tau_s'}, s\right) &= (-1)^{1-1}(1-1)!\ \eta_{\{A_1,A_2,A_3\}} + (-1)^{2-1}(2-1)!\ \eta_{\{A_1,A_2\}}\eta_{\{A_3\}} \\ &+ (-1)^{2-1}(2-1)!\ \eta_{\{A_1,A_3\}}\eta_{\{A_2\}} + (-1)^{2-1}(2-1)!\ \eta_{\{A_2,A_3\}}\eta_{\{A_1\}} \\ &+ (-1)^{3-1}(3-1)!\ \eta_{\{A_1\}}\eta_{\{A_2\}}\eta_{\{A_3\}} \\ &= \eta_{123} + \eta_{12}\eta_3 + \eta_{13}\eta_2 + \eta_{23}\eta_1 + 2\,\eta_1\eta_2\eta_3\end{aligned}$$

. *Mapping function q.* *We define q to be a function which takes as input the parameter vector $\boldsymbol{\eta} = (\eta_s)_{s\in S}$ and outputs the joint cumulant of random variables in s for all $s \in S$. That is,*

$$q(\boldsymbol{\eta}, S) = \left\{\kappa\left(\boldsymbol{\eta}_{\tau_s'}, s\right)\right\}_{s\in S} = \{\delta_s\}_{s\in S}.$$

We illustrate q with the following example. For $M = 3$,

$$\begin{aligned}P_{\{A_1\}} &= \{A_1\}, \\ P_{\{A_2\}} &= \{A_2\}, \\ P_{\{A_3\}} &= \{A_3\}, \\ P_{\{A_1,A_2\}} &= \{\{A_1, A_2\}, \{\{A_1\}, \{A_2\}\}, \\ P_{\{A_1,A_3\}} &= \{\{A_1, A_3\}, \{\{A_1\}, \{A_3\}\}, \\ P_{\{A_2,A_3\}} &= \{\{A_2, A_3\}, \{\{A_2\}, \{A_3\}\}, \text{ and} \\ P_{\{A_1,A_2,A_3\}} &= \{\{A_1, A_2, A_3\}, \{\{A_1, A_2\}, \{A_3\}\}, \{\{A_1, A_3\}, \{A_2\}\}, \{\{A_2, A_3\}, \{A_1\}\}, \\ &\quad \{\{A_1\}, \{A_2\}, \{A_3\}\}\}.\end{aligned}$$

Hence $q(\boldsymbol{\eta}, S) = \begin{pmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ \eta_{12} - \eta_1\eta_2 \\ \eta_{13} - \eta_1\eta_3 \\ \eta_{23} - \eta_2\eta_3 \\ \eta_{123} - \eta_{12}\eta_3 - \eta_{13}\eta_2 - \eta_{23}\eta_1 + 2\eta_1\eta_2\eta_3 \end{pmatrix}$.

We are thus computing the joint cumulant of all possible sets of singletons, pairs and triplets of markers.

Lemma 3. ***Reparametrization $\boldsymbol{\delta}$*** *Let $\boldsymbol{\delta} = (\delta_s)_{s\in S} = q(\boldsymbol{\eta}, S) = \left(\kappa\left(\boldsymbol{\eta}_{\tau_s'}, s\right)\right)_{s\in S'}$ with $\boldsymbol{\delta} \in \Delta$ and $\Delta = \{\boldsymbol{\delta} \mid \boldsymbol{\delta} = q \circ f(\boldsymbol{\theta}, S), \boldsymbol{\theta} \in \Theta\}$. Then $\boldsymbol{\delta}$ is a re-parametrization of $\boldsymbol{\eta}$. That is, $q : \Lambda \to \Delta$ is a bijective mapping function.*

Notice here that we limit $\boldsymbol{\delta}$ to take values in the image of function q. This guarantees that when a bijective function is used to map $\boldsymbol{\delta}$ back to $\boldsymbol{\eta}$ and then back to $\boldsymbol{\theta}$'s, those haplotype frequencies will be properly defined. Before we proceed to prove Lemma 3, we introduce the inverse functions of c and q.

. Mapping function κ^{-1} *. We define κ^{-1} to be a function which takes as an input the parameter vector $\boldsymbol{\delta} = (\delta_s)_{s\in S}$ and outputs η_s, the joint expectation of the set of random variables in s. That is,*

$$\kappa^{-1}(\boldsymbol{\delta}, s) = \delta_s - \sum_{p\in P_s\setminus s} (-1)^{|p|}(|p|-1)! \prod_{b\in p}\left\{\delta_b - \sum_{p'\in P_b\setminus b}(-1)^{|p'|}(|p'|-1)! \prod_{b'\in p'}\delta_{b'}\right\}.$$

We illustrate function κ^{-1} with the following example. Let $M=3$ and $s=\{A_1,A_2,A_3\}$. Then, $\boldsymbol{\delta}_{\tau_s'} = (\delta_1,\delta_2,\delta_3,\delta_{12},\delta_{13},\delta_{23},\delta_{123})$. Hence,

$$\begin{aligned}
\kappa^{-1}\left(\boldsymbol{\delta}_{\tau_s'}, s\right) &= \delta_{\{A_1,A_2,A_3\}} + (-1)^2(2-1)!\ \left(\delta_{\{A_1,A_2\}} + (-1)^2(2-1)!\,\delta_{A_1}\delta_{A_2}\right)\delta_{A_3} \\
&\quad + (-1)^2(2-1)!\ \left(\delta_{\{A_1,A_3\}} + (-1)^2(2-1)!\,\delta_{A_1}\delta_{A_3}\right)\delta_{A_2} \\
&\quad + (-1)^2(2-1)!\ \left(\delta_{\{A_2,A_3\}} + (-1)^2(2-1)!\,\delta_{A_2}\delta_{A_3}\right)\delta_{A_1} \\
&\quad + (-1)^3(3-1)!\,\delta_{A_1}\delta_{A_2}\delta_{A_3} \\
&= \delta_{\{A_1,A_2,A_3\}} + \left(\delta_{\{A_1,A_2\}} + \delta_{A_1}\delta_{A_2}\right)\delta_{A_3} + \left(\delta_{\{A_1,A_3\}} + \delta_{A_1}\delta_{A_3}\right)\delta_{A_2} \\
&\quad + \left(\delta_{\{A_2,A_3\}} + \delta_{A_2}\delta_{A_3}\right)\delta_{A_1} - 2\delta_{A_1}\delta_{A_2}\delta_{A_3}
\end{aligned}$$

Which is the same expression we would get if we used the definition of $\delta_{\{A_1,A_2,A_3\}}$, and solved for η_{123}, that is

$$\begin{aligned}
\delta_{\{A_1,A_2,A_3\}} &= \eta_{234} - \eta_{12}\eta_3 - \eta_{13}\eta_2 - \eta_{23}\eta_1 + 2\eta_1\eta_2\eta_3 \\
\Rightarrow \eta_{123} &= \delta_{\{A_1,A_2,A_3\}} + \eta_{12}\eta_3 + \eta_{13}\eta_2 + \eta_{23}\eta_1 - 2\eta_1\eta_2\eta_3 \\
&= \delta_{\{A_1,A_2,A_3\}} + \left(\delta_{\{A_1,A_2\}} + \delta_{A_1}\delta_{A_2}\right)\delta_{A_3} + \left(\delta_{\{A_1,A_3\}} + \delta_{A_1}\delta_{A_3}\right)\delta_{A_2} \\
&\quad + \left(\delta_{\{A_2,A_3\}} + \delta_{A_2}\delta_{A_3}\right)\delta_{A_1} - 2\delta_{A_1}\delta_{A_2}\delta_{A_3}
\end{aligned}$$

For a set $\{A_1,A_2,A_3\}$ of random variables, we will write δ_{123} as a shorthand of $\delta_{\{A_1,A_2,A_3\}}$.

. Mapping function q^{-1}. *We define q^{-1} to be a function which takes as input the parameter vector $\boldsymbol{\delta} = (\delta_s)_{s\in S}$ and outputs $\boldsymbol{\eta} = (\eta_s)_{s\in S}$, the joint expectation of the set of random variables in s for all $s\in S$. That is,*

$$q^{-1}(\boldsymbol{\delta}, S) = \left\{\kappa^{-1}\left(\boldsymbol{\delta}_{\tau_s'}, s\right)\right\}_{s\in S} = (\eta_s)_{s\in S}.$$

We illustrate q^{-1} with the following example. For $M=3$,

$$q^{-1}(\boldsymbol{\delta}, S) = \begin{pmatrix}
\delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_{12}+\delta_1\delta_2 \\ \delta_{13}+\delta_1\delta_3 \\ \delta_{23}+\delta_2\delta_3 \\ \delta_{123}+\delta_{12}\delta_3+\delta_{13}\delta_2+\delta_{23}\delta_1-2\delta_1\delta_2\delta_3
\end{pmatrix} = \begin{pmatrix}
\eta_1 \\ \eta_2 \\ \eta_3 \\ \eta_{12} \\ \eta_{13} \\ \eta_{23} \\ \eta_{123}
\end{pmatrix}.$$

We now proceed with the proof of Lemma 3. To prove that q is bijective we need to show that q is both injective, i.e. $\forall\boldsymbol{\eta},\boldsymbol{\eta}^*\in\Lambda, q(\boldsymbol{\eta},S)=q(\boldsymbol{\eta}^*,S)\Rightarrow\boldsymbol{\eta}=\boldsymbol{\eta}^*$, and surjective, i.e. $\forall\boldsymbol{\delta}\in\Delta,\quad \exists\boldsymbol{\eta}\in\Lambda : q(\boldsymbol{\eta},S)=\boldsymbol{\delta}$. Now that we have defined the inverse function of q, it is easy to show that, $q(\boldsymbol{\eta},S)=q(\boldsymbol{\eta}^*,S)\Rightarrow q^{-1}\{q(\boldsymbol{\eta},S)\}=q^{-1}\{q(\boldsymbol{\eta}^*,S)\}\Rightarrow\boldsymbol{\eta}=\boldsymbol{\eta}^*$ and for all arbitrary parameter vectors $\boldsymbol{\delta}\in\Delta$, we can choose $\boldsymbol{\eta}=q^{-1}(\boldsymbol{\delta},S)$ such that $q(\boldsymbol{\eta},S)=q\left\{q^{-1}(\boldsymbol{\delta},S),S\right\}=\boldsymbol{\delta}$. This concludes the proof of the bijectivness of q, which concludes also the proof of Lemma 3 and thus of Lemma 1.

A.2. Standardized parameters

Recall that δ_A can be expressed as

$$\delta_A = \eta_A - \sum_{p \in P_A \setminus A} (-1)^{|p|}(|p|-1)! \prod_{b \in p} \eta_b = \eta_A - \sum_{p \in P_A \setminus A} R_\delta(p) = \eta_A - R_\delta$$

where $R_\delta(p)$'s depend on loci $b \in p$ with $|b| < M$. These rest terms $R_\delta(p)$ are considered fixed and bounds for δ_A are to be determined completely analogous to the two locus case based on η_A.

First, η_A is upper bound by all lower-order η_s and lower bound by 0. That is

$$\eta_A \leq U_1(A) := \min\{\eta_s | s \in S \setminus A\}.$$
$$\eta_A \geq L_1(A) = 0$$

Second, further constraints are imposed by the relationship between η_A and lower order haplotype frequencies η_s. It is straightforward to see that η_s can be restated as:

$$\eta_s = g(\theta, s) = \theta_s + \sum_{t \in S, t \supset s} (-1)^{|t|-|s|-1} \eta_t$$

Here, θ_{h_s} is the frequency for haplotype $h_s = \left\{v \in \{0,1\}^M \mid v[j] \Leftrightarrow A_j \in s\right\}$, the haplotype with M loci with 1-alleles at loci s and 0-alleles elsewhere. Note, that all the sums above include η_A Solving for η_A gives us:

$$\begin{aligned}
\eta_A &= (-1)^{|A|-|s|-1} \left\{ \eta_s - \theta_{h_s} - \sum_{t \in S \setminus A, t \supset s} (-1)^{|t|-|s|-1} \eta_t \right\} \\
&= (-1)^{|A|-|s|} \left\{ \theta_{h_s} - \eta_s + \sum_{t \in S \setminus A, t \supset s} (-1)^{|t|-|s|-1} \eta_t \right\} \\
&= (-1)^{|A|-|s|} \left\{ \theta_{h_s} - \eta_s - \sum_{t \in S \setminus A, t \supset s} (-1)^{|t|-|s|} \eta_t \right\} \\
&= (-1)^{|A|-|s|} \left[\theta_{h_s} - \left\{ \eta_s + \sum_{t \in S \setminus A, t \supset s} (-1)^{|t|-|s|} \eta_t \right\} \right] \\
&= (-1)^{|A|-|s|} \left\{ \theta_{h_s} - \sum_{t \in S \setminus A, t \supseteq s} (-1)^{|t|-|s|} \eta_t \right\} \\
&= \sigma_s \left(\theta_{h_s} - R_s \right),
\end{aligned}$$

where $\sigma = (-1)^{|A|-|s|}$ and $R_s = \sum_{t \in S \setminus A, t \supseteq s} (-1)^{|t|-|s|} \eta_t$. Each η_s therefore contributes an upper and lower bound to η_A by choosing $\theta_{h_s} = 0$ or $\theta_{h_s} = 1$:

$$\eta_A \leq U_s := \begin{cases} \max(1 - R_s, 0) & \text{if } \sigma \geq 0, \\ \min(-R_s, 0) & \text{if } \sigma < 0 \end{cases},$$

$$\eta_A \geq L_s := \begin{cases} \max(-R_s, 0) & \text{if } \sigma \geq 0, \\ \min(1 - R_s, 0) & \text{if } \sigma < 0 \end{cases}.$$

With $U_2(A) := \min\{U_s | s \in S \setminus \{A\}\}$ and $L_2(A) := \max\{L_s | s \in S \setminus \{A\}\}$, we get

$$\eta_A^{\max} := \min\{U_1(A), U_2(A)\},$$
$$\eta_A^{\min} := \max\{L_1(A), L_2(A)\}.$$

Then $\eta_A^{\max}$ and $\eta_A^{\min}$ can be used as above to standardize δ_A.

A.4. Additional Tables

Table A.1: Linkage disequilibrium parameters in the cases (Ca), controls (Co) and pool (P) of cases and controls samples for each of the four triplets identified from the WTCCC data analysis.

	Triplet 1			Triplet 2		
	P	Ca	Co	P	Ca	Co
δ_1	.239	.264	.223	.239	.264	.223
δ_2	.126	.120	.130	.091	.108	.081
δ_3	.091	.108	.081	.314	.308	.319
δ_{12}	-.374	-.502	-.279	.111	.135	.081
δ_{13}	.111	.135	.081	-.112	-.199	-.046
δ_{23}	-.544	-.382	-.663	.363	.351	.377
δ_{123}	.172	.084	.229	-.697	-.665	-.716

	Triplet 3			Triplet 4		
	P	Ca	Co	P	Ca	Co
δ_1	.350	.324	.366	.213	.234	.199
δ_2	.207	.191	.218	.401	.376	.417
δ_3	.247	.276	.229	.199	.209	.193
δ_{12}	.712	.696	.720	-.297	-.268	-.306
δ_{13}	.103	.033	.170	-.759	-.790	-.739
δ_{23}	-.105	-.174	-.040	.227	.088	.335
δ_{123}	-.160	-.119	-.190	.203	.521	-.126

Table A.2: Corresponding haplotype frequencies for SNPs in linkage equilibrium (Scenario 1), and SNPs in LD (Scenario 2).

Haplotype	Scenario 1	Scenario 2
θ_{0000}	.292	.358
θ_{0001}	.239	.088
θ_{0010}	.135	.240
θ_{0011}	.111	.101
θ_{0100}	.065	.148
θ_{0101}	.054	.004
θ_{0110}	.030	.053
θ_{0111}	.025	.007
θ_{1000}	.015	.358
θ_{1001}	.013	.088
θ_{1010}	.007	.240
θ_{1011}	.006	.101
θ_{1100}	.003	.148
θ_{1101}	.003	.004
θ_{1110}	.002	.053
θ_{1111}	.001	.007

4

Combining Information from Linkage and Association Mapping [1]

Summary

In this analysis, we investigate the contributions that linkage-based methods, such as identical-by-descent mapping, can make to association mapping to identify rare variants in next-generation sequencing data. First, we identify regions in which cases share more segments identical-by-descent around a putative causal variant than do controls. Second, we use a two-stage mixed-effect model approach to summarize the single-nucleotide polymorphism data within each region and include them as covariates in the model for the phenotype. We assess the impact of linkage disequilibrium in determining identical-by-descent states between individuals by using markers with and without linkage disequilibrium for the first part and the impact of imputation in testing for association by using imputed genome-wide association studies or raw sequence markers for the second part. We apply the method to next-generation sequencing longitudinal family data from Genetic Association Workshop 18 and identify a significant region at chromosome 3: 40249244-41025167 (p-value $= 2.3 \times 10^{-3}$).

4.1 Introduction

In genetic association studies, joint analysis of multiple single-nucleotide polymorphisms (SNPs) can be more powerful than separate SNP analysis because single markers typically either have small effect sizes (common variants) or minor allele frequencies that are too small to reliably fit models (rare variants) [Cantor et al., 2010]. If the rare variant effects were large, and the disease was not heterogeneous, they would have been found through previous family-based linkage studies. There may be a middle ground in which multiple rare variants of moderate effect size play a key role in the etiology of some diseases. Such situations might

[1] Published in *BMC Proceedings*.

Table 4.1: Description of genotypic data sets used in each part of the analysis. M: a million, K: a thousand. IBD: identical-by-descent. AllMark: data set containing approximately 50K GWAS markers. NoLD: data set containing only 784 LD-pruned GWAS markers. DOS: imputed dosage GWAS data. WGS: whole genome sequence data.

	IBD mapping		Association mapping	
	AllMark	NoLD	DOS	WGS
Type of data	GWAS (65K, Illumina chips)		Imputed WGS based on existing GWAS	WGS
No. markers	$\sim$ 50K	784	$\sim$ 1.2M	$\sim$1.7M
No. individ	939		939	464

be ideal for identity-by-descent (IBD) mapping [Browning and Thompson, 2012]. Moreover, with the availability of genome-wide SNP data, the density of SNP markers has increased dramatically, making it possible to detect segments of IBD as small as 2 centimorgans (cM) [Browning and Browning, 2011].

In this article, we investigate the contribution that linkage-based methods, such as IBD mapping, can make to association mapping to identify rare variants in next-generation sequencing data. In the first part of our analysis, we use the methods of Browning and Thompson [2012] to identify regions in which cases share more segments of IBD around a putative causal variant than do controls. After selecting these regions, we use a two-stage mixed-effects model approach, which was recently proposed by Tsonaka et al. [2012], to summarize the SNP data within each region and include them as covariates in the model for the phenotype. To increase our power to identify rare variants, we also include the number of rare variants per region as a covariate in the model.

To assess the impact of linkage disequilibrium (LD) on our analysis, we present results from estimating IBD probabilities using markers with and without LD. We assess the impact of imputation by analyzing both imputed dosage genome-wide association studies (DOS) and whole genome sequence (WGS) data. Table 4.1 provides a description of the data sets used for IBD and association mapping.

4.2 Material and Methods

4.2.1 Study sample

We consider data from 939 individuals from 20 families; 464 are directly sequenced individuals and imputed WGS data, based on existing genome-wide association studies (GWAS) data, are available for their family members. We restrict our work to real genotypic data from chromosome 3. For each individual, we have information on age at examination and current tobacco smoking for up to 4 time points. We use the binary trait hypertension diagnosis at the first time point for selection of regions with excess IBD sharing and the quantitative trait diastolic blood pressure (DBP) for the phenotype model.

4.2.2 Selection of regions with excess IBD sharing

We construct all possible case-case (CaCa) and case-control (CaCo) pairs, such that individuals within pairs are unrelated. This results in 9229 CaCa pairs and 10080 CaCo pairs. We estimate the IBD state using 2 data sets: one containing approximately 50,000 GWAS markers, which we refer to as the AllMark data set, and 1 containing only 784 LD-pruned GWAS markers, the NoLD data set. From both data sets we eliminate SNPs with minor allele frequencies (MAFs) $<5\%$ because shared alleles that are assumed to be rare represent strong evidence for IBD and can distort results if this assumption is violated [Brown et al., 2012]. In brief, the NoLD markers are selected using a sliding window 1 cM in size, removing markers based on linkage information content and excluding markers with the lowest MAF. At each marker we calculate the rate of IBD for each of the 2 groups and subtract their genomic average over all markers and pairs. If the ratio between CaCa pairs is larger than the maximum CaCo ratio, exceeding a certain threshold, we consider this region for association analysis.

To compute the IBD states between pairs of individuals, we use the method of Thompson [2008] implemented in their `ibd_haplo` software. This method uses a continuous - time Markov rate matrix to model and estimate IBD states among pairs of individuals, using data at dense SNP loci, ignoring the LD structure. However, LD remains a major confounding factor because LD is itself a reflection of co-ancestry at the population level. To assess the impact of LD on IBD estimation, we present results for both AllMark and NoLD data sets. In `ibd_haplo`, one needs to specify a value for parameters of the latent IBD process β, the pointwise pairwise probability of IBD, and α, the overall rate of change of IBD state along a chromosome. The choice of these parameters defines the time-depth of the IBD that is sought [Brown et al., 2012]. For the results shown in this paper, $\alpha = 0.05$ and $\beta = 0.01$. We use a calling threshold of 0.9 as the probability that each of the IBD states must reach for the state to be called.

4.2.3 Two-stage approach

After the regions have been selected, we use the two-stage approach of [Tsonaka et al., 2012] to test for their association with the longitudinal phenotype. In the first stage, a random-effects model is used to summarise the regions via their empirical Bayes (EB) estimates. Next, the EB estimates of a specific region r, obtained from the first stage, are added as covariates into the model for the phenotype to test for region effects. Below, we describe in brief the phenotypic model used in the second stage. Let DBP_{ijt} be the diastolic blood pressure for individual j from family i at time point t, where $i = 1, \ldots, N$, $j = 1, \ldots, n_i$, $t = 1, 2, 3, 4$, and n_i is the number of individuals in family i. We use the following linear mixed model for each region r:

$$DBP_{ijt} = \beta_0 + \beta_1 \mathbf{x}_{ijt} + \beta_2 eb_{ijr} + \beta_3 s_{ijr} + u_{ij} + e_{ijt} \tag{4.1}$$

where $\mathbf{x}_{ijt}$ is the vector with covariate values for age and smoking status, eb_{ijr} is the EB estimates of the region r, obtained from the first stage, and s_{ijr} is the number of rare variants, here variants with a MAF of less than 5%, within region r; u_{ij} is the random family effect and $\mathbf{u}_i = (u_{i1}, \ldots, u_{in_i})^\intercal$ follows a multivariate normal distribution with mean 0 and variance-covariance matrix $\sigma^2_{u_i} \times R$, where R is the coefficient of relationships matrix; e_{ijt} is a normally distributed residual with a 4×4 covariance matrix to model the correlation among 6 repeated measurements. We use a multivariate Wald statistic with 2 degrees of freedom to test the null hypothesis of no region effect; that is, $H_0 : \beta_2 = \beta_3 = 0$.

Table 4.2: Description of IBD between case-case (CaCa) and case-control (CaCo) pairs. IBD: identical-by-descent. AllMark: data set containing approximately 50K GWAS markers. NoLD: data set containing only 784 LD-pruned GWAS markers. DOS: imputed dosage GWAS data. WGS: whole genome sequence data.

		Mean proportions			Mean length of segments		
Data	Pairs	Any IBD	Not IBD	No call	Any IBD	Not IBD	No call
AllMark	CaCa	.295	.499	.206	58.27	144.48	25.98
	CaCo	.292	.503	.205	58.01	145.58	25.91
NoLD	CaCa	.006	.950	.044	44.81	316.00	21.27
	CaCo	.004	.951	.045	39.52	315.09	21.59

4.3 Results

Table 4.2 presents the mean proportions and lengths of IBD segments shared for both groups. Averages were taken over all markers and all pairs. For both AllMark and NoLD, we observed a small difference in both mean proportion and length. However, in AllMark, where LD is ignored, the mean proportion of IBD is increased, as compared to NoLD. We compared the rates between the 2 groups and found 8 and 7 regions with an excess of IBD between CaCa pairs for AllMark and NoLD, respectively. Table 4.3 lists the starting and ending physical positions of these regions, as well as the number of SNPs and rare variants they contain. Interestingly, we observed no overlap between regions when using markers with and without LD.

After selecting the regions, we tested their association with the longitudinal phenotype by fitting a linear mixed model to DBP with the EB estimates per region, smoking status, and age as covariates. To further increase our power, we considered a second model, where we adjusted also for the sum of rare variants. We used 2 different genotype data, DOS with imputed genotypes on 939 individuals and WGS with complete genomics on 464 individuals. To account for multiple testing, we used a Bonferroni correction, using a significance level of alpha divided by the maximum number of independent regions tested for each data set; that is, 7 for the NoLD and 8 for the AllMark. We used 6×10^{-3} as the significance level for AllMark and 7×10^{-3} for NoLD.

No significant results were found when the candidate regions were selected using the AllMark data (results not shown). Table 4.4 gives the results of the analysis based on NoLD. When NoLD and DOS were used, there was a significant result for the region 3:40249244-41025167 (p-value of the 2df Wald 2.3×10^{-3}). When WGS was used instead of DOS, the variance of the estimates increased and the signal was no longer significant. When the number of rare variants was removed from the model, the region again reached significance (p-value $= 2.1 \times 10^{-3}$).

4.4 Discussion

We have presented a method that combines linkage and association-based mapping to identify rare variants in next-generation sequencing data. Initially, we identify regions with an excess of IBD between case-case as compared to case-control pairs. Subsequently, we use a two-stage approach to summarize the regions via an EB estimate of the genetic variation and test for region effects. The two-stage approach captures the correlation between SNPs

Table 4.3: Descriptions of regions. N, number of SNPs per region; n, number of rare variants (MAF $< 5\%$) per region. AllMark: data set containing approximately 50K GWAS markers. NoLD: data set containing only 784 LD-pruned GWAS markers. DOS: imputed dosage GWAS data. WGS: whole genome sequence data.

Physical position	DOS		WGS	
Start-end	N	n	N	n
		AllMark		
27279401-27292557	77	38	100	61
52618319-52637439	105	46	168	111
52759860-52771468	77	44	117	82
52830547-52866115	291	156	379	244
86269515-86282586	60	24	96	58
99537305-99580268	211	120	322	260
99621002-99676384	270	144	386	299
99927237-100004117	396	185	575	427
		NoLD		
29239664-29531222	2153	919	2984	1659
34834899-35282759	2730	1284	4267	2715
35718847-36018767	1618	927	2446	1755
36815704-37526013	3738	2151	5669	4038
40249244-41025167	4247	2530	6168	4214
167635899-168125439	2665	1349	3926	2552
168621773-168859006	1508	708	2018	1207

Table 4.4: P-values for testing, marginally or jointly, region effects using the NoLD data set. Two different models are fitted; a: with and b: without including the number of rare variants as covariates. The regions are in the same order as in Table 4.3. NoLD: data set containing only 784 LD-pruned GWAS markers. DOS: imputed dosage GWAS data. WGS: whole genome sequence data.

DOS				WGS			
β_2^a	β_3^a	β_2, β_3^a	β_2^b	β_2^a	β_3^a	β_2, β_3^a	β_2^b
.03	.25	.04	.02	.04	.76	.12	.04
.93	.91	.99	.92	.81	.27	.54	.93
.99	.11	.27	.77	.35	.41	.50	.51
.18	.24	.25	.20	.32	.13	.15	.23
1.3×10^{-3}	.05	2.3×10^{-3}	3.6×10^{-3}	9.3×10^{-3}	.33	.01	2.1×10^{-3}
.29	1.00	.55	.22	.27	.75	.54	.28
.09	.66	.22	.09	.25	.26	.31	.33

within regions by using random effects. These types of approaches can be more powerful than methods that ignore the dependency structure between the SNPs [Chen et al., 2010]. The approach can be directly applied to family and longitudinal data and can deal with missing genotypes.

One main advantage of this method, as compared to an association-only approach [Houwing-Duistermaat et al., 2014], is that by using the IBD mapping in the first step, we reduce the number of candidate regions to areas more enriched for putative causal loci. This considerably reduces the number of tests that need to be performed, and testing for interactions becomes feasible. This method can also be used for non-gene regions, although cautiously, because possibly important regions might already have been excluded in the first part, if the parameters for the IBD are misspecified. Moreover, if the resulting regions contain too many markers, the effect of rare variants might be diluted. The regions are selected using the binary hypertension diagnosis phenotype at the first measurement and not the quantitative DBP phenotype analyzed in the association study. This may be a problem if the 2 phenotypes are different. In our case, the binary phenotype was created using a threshold for the quantitative phenotype or information on medications. If the effect of a variant changes over time, we might lose power by determining the IBD states only on the first measurement. For individuals receiving treatment, the recorded DBP could be considered as a right-censored value, because we know that it is less than what the untreated value would be. Our approach ignores this information, which again may result in power loss. One way to address this issue could be to use a nonparametric algorithm to adjust blood pressure for treatment effect [Soler and Blangero, 2003].

In this article, we do not present results for type I error or power. However, Tsonaka et al. [2012] and Houwing-Duistermaat et al. [2014] report results for both regarding the two-stage approach. Using extensive simulations, Tsonaka et al. [2012] showed that the test statistics preserve the type I error at nominal level for scenarios comparable to ours. Houwing-Duistermaat et al. [2014] analyzed the simulated phenotypes from this Genetic Analysis Workshop (GAW) and found that the power was as high as 96.5% and 72.5% using the imputed GWAS and WGS data, respectively.

We found significant results only when the candidate regions were selected using the NoLD and DOS data. One reason for the better performance of the NoLD data, as compared to the AllMark data, could be the presence of LD in the latter. LD leads to increased rates of false positive IBD results [Brown et al., 2012], which could erroneously indicate these regions as interesting. The absence of overlap between regions when using these 2 data sets also indicates the sensitivity of the method to the amount of LD in the data. Another reason for the better performance of the NoLD data set could be the region selection process. In the NoLD data, the markers are further apart from each other, as compared to the AllMark data set. Hence, when selecting a region (at least 2 markers), we automatically include more SNPs and rare variants.

When the NoLD and WGS data were used, the signal of the region found using DOS was no longer significant. This power loss could be a result of the smaller sample size in the WGS data, which leads to increased variances of the parameter estimates (results not shown). The same happens for the estimates of the genetic variance. On one hand, using the DOS data we estimate σ_u^2 =10.622 with a variance of 1.4366 (p-value 6.9×10^{-11}). On the other hand, when using WGS, the estimate becomes much smaller, σ_u^2= 1.153, and its variance increases to 27.23 (p-value = 0.99). Removing the number of rare variants from the model led to a significant p-value for this region.

Using the NCBI database, we found that the gene *CADM2*, which is 146 kilobase (kb) on the right of the region we identified, is associated, among other phenotypes, with blood pressure and body mass index [Speliotes et al., 2010]. More specifically, 3 SNPs in this gene are associated with blood pressure: rs1370032 (p-value = 7.22×10^{-5}), rs13074417

(p-value $= 7.625 \times 10^{-5}$), and rs4859048 (p- value $= 7.872 \times 10^{-5}$).

5

A Retrospective Likelihood Approach for Efficient Integration of Multiple Omics Factors in Case-Control Association Studies [1]

Summary

Integrative omics, the joint analysis of outcome and multiple types of omics data, such as genomics, epigenomics and transcriptomics data, constitutes a promising approach for powerful and biologically relevant association studies. These studies often employ a case-control design, and often include non-omics covariates, such as age and gender, that may modify the underlying omics risk factors. An open question is how to best integrate multiple omics and non-omics information to maximize statistical power in case-control studies that ascertain individuals based on the phenotype. Recent work on integrative omics have used prospective approaches, modeling case-control status conditional on omics and non-omics risk factors. Compared to univariate approaches, jointly analyzing multiple risk factors with a prospective approach increases power in non-ascertained cohorts. However, these prospective approaches often lose power in case-control studies. In this article, we propose a novel statistical method for integrating multiple omics and non-omics factors in case-control association studies. Our method is based on a retrospective likelihood function that models the joint distribution of omics and non-omics factors conditional on case-control status. The new method provides accurate control of Type I error rate and has increased efficiency over prospective approaches in both simulated and real data. The method is publicly available at `https://github.com/BrunildaBalliu/IntegrativeOmics`.

[1] Published in *Genetic Epidemiology*.

5.1 Introduction

Recent advances in technology have made it possible to collect multiple types of omics data, such as genomics, transcriptomics, and epigenomics in the same individuals. Genome-, transcriptome-, and epigenome-wide association studies have led to the identification of genetic variants, transcripts, and methylation sites associated with many complex diseases [Edgar et al., 2002; Hindorff et al., 2009; Lv et al., 2012]. However, due to the lack of integrative statistical approaches, these associations were mainly identified through their marginal effects on disease risk. As a result, underlying disease mechanisms through which omics factors affect phenotypes, e.g. joint effects or mediation effects, remain unknown for most complex diseases. Integrative omics studies, the joint analysis of outcome and multiple omics data, have emerged as a promising alternative to more powerful and biologically informative association studies [Chen et al., 2008; Li, 2013; Zhao et al., 2014; Huang et al., 2014].

Here, we are interested in leveraging integrative omics approaches to identify associations between a genetic variant G, a transcript E or a methylation site M and a binary outcome Y, accounting for environmental or clinical factors X. In randomly ascertained studies, when E, M, G and X have independent effects on Y, modeling them jointly can increase power to detect associations between Y and any of E, M, and G [Robinson and Jewell, 1991; Neuhaus and Jewell, 1993; Neuhaus, 1998]. However, E, M, and G can be correlated, e.g. genetic variants can alter gene expression and DNA methylation [Schadt et al., 2003; Zhang et al., 2010] and DNA methylation can regulate gene expression [Gutierrez-Arcelus et al., 2013]. In such scenarios, E and M can act as mediators of G, and testing for their joint effect on Y can be more powerful than testing only for genetic associations [Huang et al., 2014; Zhao et al., 2014]. Moreover clinical covariates X, such as age and gender, can be associated with M and/or E [Richardson, 2003; Horvath et al., 2012; Liu et al., 2010; Dimas et al., 2012; Glass et al., 2013]. Consistent with previous approaches, we make the assumption of independence between X and G [Umbach and Weinberg, 1997; Chatterjee and Carroll, 2005]. In addition to increasing power for G, clinical covariates can confound the effect of E and M on Y, thus including them in the analysis is necessary in order to control bias and prevent false discoveries. Figure 5.1.a illustrates the relationships between E, M, G, X, and Y in a randomly ascertained population cohort.

Integrative omics studies typically employ a case-control design. Since cases are enriched for all risk factors, ascertainment will induce additional correlation between E, M, G and X (Figure 5.1.b). Existing methods for integrative omics analysis use prospective approaches to model the distribution of the case-control status conditional on the risk factors, in our case $P(Y|E,M,G,X)$ [Huang et al., 2014; Zhao et al., 2014]. In these ascertained studies, prospective approaches will not account for the sampling scheme, potentially resulting in severe power loss relative to univariate analyses of each risk factor [Chatterjee and Carroll, 2005; Xing and Xing, 2010; Zaitlen et al., 2012a,b; Pirinen et al., 2012; Mefford and Witte, 2012]. The main reason for this power loss is that, in studies for which ascertainment is based on outcome, the distribution of the risk factors $P(E,M,G,X)$ contains information about the parameters in $P(Y|E,M,G,X)$ [Scott and Wild, 2001]. In a prospective approach $P(E,M,G,X)$ is ignored when making inference and as a result such methods will be less efficient than methods that explicitly make use of $P(E,M,G,X)$.

In this article, we propose a novel integrative omics approach that addresses these issues of power loss in case-control association studies. Our approach is based on a retrospective likelihood function that models the joint distribution of the omics and non-omics factors conditional on the case-control status, $P(E,M,G,X|Y) = P(Y|E,M,G,X) \times P(E,M,G,X)/P(Y)$. In order to model $P(E,M,G,X)$ as efficiently as possible, we exploit knowledge about the correlation structure between risk factors and the distribution of

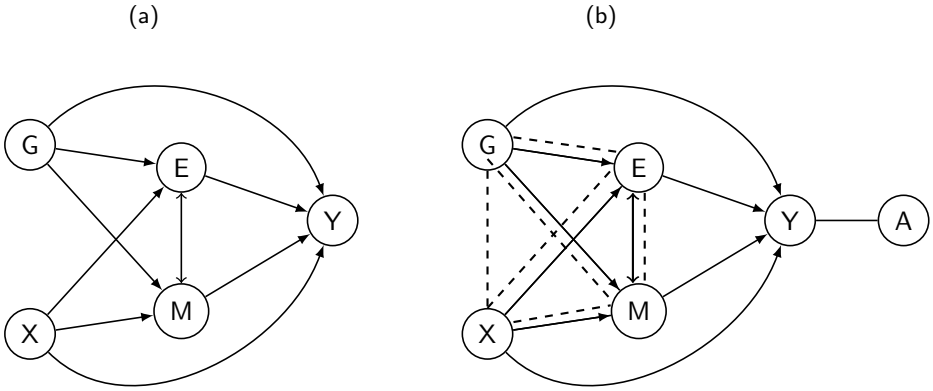


Figure 5.1: Example to illustrate possible correlation structures among risk factors and a trait in (a) a random sample and (b) a case-control sample. G: genetic variant, E: gene expression, M: DNA methylation, X: non-omics covariate, Y: trait/disease, A: ascertainment of cases and controls. Continuous arrows between two nodes connect variables that could be correlated in the population while dashed lines represent induced correlations due to ascertainment.

the risk factors in the population by making parametric assumptions about $P(E, M, G, X)$. When these distributional assumptions hold, the corresponding maximum likelihood estimates are unbiased and statistically efficient, in that they have the smallest variances among all valid estimators, and the corresponding association tests are the most powerful among all valid tests [Chatterjee and Carroll, 2005; Lin and Zeng, 2009].

The use of a retrospective approach to exploit the gene-environment independence assumption in case-control genetic associations studies was originally proposed by Chatterjee and Carroll [2005]. The method accommodates genetic and environmental covariates that are independent in the underlying population or that are independent conditional on some other factors. Our work is an extension of this method to accommodate situations in which independence or conditional independence assumptions for genetic and additional omics risk factors do not hold. Moreover, our approach can accommodate continuous risk factors by using parametric distributions.

The rest of the paper is organized as follows. In Section 5.2, we introduce the method, the assumptions about the distribution of omics and non-omics risk factors and describe the statistical testing. In Section 5.3, we evaluate the finite sample performance of the proposed method using an extensive simulation study. We compare our method with a prospective likelihood approach and show that our method has increased efficiency and power under many realistic disease models while maintaining a properly controlled type I error. In Section 5.4 we demonstrate that our approach is more efficient than the prospective approach when analyzing omics data from a multiple sclerosis study [Huynh et al., 2014]. We also describe how the models can be modified when not all types of data are available. We close with a discussion in Section 5.5.

5.2 Material and Methods

5.2.1 The Statistical Model

Consider a case-control study of N subjects, N_1 cases and N_0 controls, where for each subject, information on genetic variation, DNA methylation, expression, and one or more clinical or environmental covariates is available. If one or more of the data sources is not available, as is the case in our real data example, the following models can be modified accordingly. Here we focus on a single genetic, epigenetic and transcriptional measurement per subject. In the Discussion section, we consider extensions for the high dimensional setting. Let $\mathbf{Y} = (Y_1, \ldots, Y_N)$ be the vector of phenotypes for the subjects in the study, with Y_i a binary indicator of disease status, i.e. $Y_i = 1$ if i is affected and 0 if i is unaffected. Similarly, let $\mathbf{G}$, $\mathbf{M}$, and $\mathbf{E}$ be vectors of a genetic, an epigenetic and a transcriptional factor of the N subjects, respectively. Last, let $\mathbf{X} = (\mathbf{X_1}, \ldots, \mathbf{X_J})$ denote the matrix of J clinical covariates for the N subjects.

The prospective likelihood models the distribution of the disease status conditional on the potential risk factors and is given as follows,

$$\mathcal{PL}(\boldsymbol{\alpha}) = P\left(\mathbf{Y}|\mathbf{G}, \mathbf{M}, \mathbf{E}, \mathbf{X}\right), \tag{5.1}$$

where $\boldsymbol{\alpha}$ is the parameter vector of the effect of risk factors on disease risk. Moreover, the prospective risk model for subject i, given its risk factors, is given by the logistic regression model

$$P\left(Y_i = 1|G_i, M_i, E_i, \mathbf{X}_i\right) = logit^{-1}\{\alpha_0 + \alpha_G G_i + \alpha_M M_i + \alpha_E E_i + \mathbf{X}_i\boldsymbol{\alpha}_X\}, \tag{5.2}$$

where $logit^{-1}\left(x\right) = \frac{e^x}{1+e^x}$, α_0 the intercept and α_G, α_M, α_E and $\boldsymbol{\alpha}_{\mathbf{X}}$ the effect of G, E, M and X on disease risk. In this work we consider only main effects of the risk factors. However, more general models, with interaction of different orders between the risk factors, could also be used.

On the other hand, the retrospective likelihood models the distribution of risk factors conditional on the disease status and is given as follows,

$$\mathcal{RL}(\boldsymbol{\theta}) = P\left(\mathbf{G}, \mathbf{M}, \mathbf{E}, \mathbf{X}|\mathbf{Y}\right) = \frac{P\left(\mathbf{Y}|\mathbf{G}, \mathbf{M}, \mathbf{E}, \mathbf{X}\right) \times P\left(\mathbf{G}, \mathbf{M}, \mathbf{E}, \mathbf{X}\right)}{P\left(\mathbf{Y}\right)}, \tag{5.3}$$

where $\boldsymbol{\theta}$ is the parameter vector containing the effect of risk factors on disease risk and parameters for characterizing the distribution of risk factors. The numerator in (5.3) is a product of the prospective risk model and the joint distribution of the risk factors. The denominator represents the marginal disease probability in the population.

The challenge in maximizing the retrospective likelihood (5.3) with respect to $\boldsymbol{\theta}$ is due to the unknown covariate distribution $P\left(\mathbf{G}, \mathbf{M}, \mathbf{E}, \mathbf{X}\right)$. It is well known that if no assumption is made about the form of the covariate distribution, $P\left(\mathbf{G}, \mathbf{M}, \mathbf{E}, \mathbf{X}\right)$ is not identifiable from case-control data [Prentice and Pyke, 1979]. Furthermore, Rabinowitz [1997] and Breslow et al. [2000] showed that, if $P\left(\mathbf{G}, \mathbf{M}, \mathbf{E}, \mathbf{X}\right)$ is treated fully non-parametrically, the efficiencies of (5.1) and (5.3) are equivalent. To optimally model $P\left(\mathbf{G}, \mathbf{M}, \mathbf{E}, \mathbf{X}\right)$, and increase efficiency for estimating the parameters of interest, we exploit knowledge about the correlation structure between risk factors. Specifically, we assume that E and M can be correlated, G and X can be associated with E and M, and that G and X are independent of each other in the population.

Thus $\mathcal{RL}$ is further factorized as follows

$$\mathcal{RL}(\boldsymbol{\theta}) = \frac{P\left(\mathbf{Y}|\mathbf{G}, \mathbf{M}, \mathbf{E}, \mathbf{X}\right) \times P\left(\mathbf{M}, \mathbf{E}|\mathbf{G}, \mathbf{X}\right) \times P\left(\mathbf{G}\right) \times P\left(\mathbf{X}\right)}{P\left(\mathbf{Y}\right)}. \tag{5.4}$$

Explicitly imposing independence between G and X will result in efficiency gain for estimating α_G and $\boldsymbol{\alpha}_X$, compared to approaches that ignore their distribution or do not exploit this assumption [Chatterjee and Carroll, 2005].

To further increase efficiency, we exploit knowledge about the distribution of the omics factors. Specifically, we make the HWE assumption to model $P(G)$. Under this assumption, $G \sim Binomial(2, p)$ with p the minor allele frequency so that $P(G)$ is characterized by a single parameter. This will increase efficiency to estimate α_G. Moreover, we assume that, after proper transformations and normalization procedures, the epigenetic and transcriptional factors are normally distributed and therefore use a multivariate normal for their joint distribution [Calza and Pawitan, 2010; Yousefi et al., 2013]. This parametric model will result in efficiency gain for estimating α_M and α_E, compared to methods that treat the distribution of E and M non-parametrically. We model the conditional distribution of M and E using a multivariate linear regression model:

$$\begin{aligned} M_i &= \beta_{GoM} G_i + \mathbf{X}_i^T \boldsymbol{\beta}_{\mathbf{XoM}} + \epsilon_{i1} \\ E_i &= \beta_{GoE} G_i + \mathbf{X}_i^T \boldsymbol{\beta}_{\mathbf{XoE}} + \epsilon_{i2} \end{aligned} \tag{5.5}$$

where β_{GoM}, $\boldsymbol{\beta}_{\mathbf{XoM}}$, β_{GoE} and $\boldsymbol{\beta}_{\mathbf{XoE}}$ are the effects of the genetic and clinical factors on the epigenetic and transcriptional factor. We assume that the errors $(\epsilon_{i1}, \epsilon_{i2})^T$ follow a bi-variate normal distribution, $\mathbf{MVN}_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{bmatrix}\sigma^2_{\epsilon_1} & \sigma_{\epsilon_1\epsilon_2}\\ \sigma_{\epsilon_1\epsilon_2} & \sigma^2_{\epsilon_2}\end{bmatrix}\right)$, with $\sigma^2_{\epsilon_1}, \sigma^2_{\epsilon_2}$, and $\sigma_{\epsilon_1\epsilon_2}$ the variances and co-variance of ϵ_1 and ϵ_2. Notice that in the model we center M, E, G and $\mathbf{X}$ around zero such that there is no need for intercepts in (5.5).

For parametrization of the distribution of $\mathbf{X}$, we assume that $\mathbf{X_1}, \ldots, \mathbf{X_J}$ are mutually independent in the population and factorize their distribution as $P(\mathbf{X}) = \prod_{j=1}^{J} P(\mathbf{X_j})$. This assumption will increase efficiency to estimate each α_X and can be relaxed when independence is not plausible. For simplicity of exposition, we focus on binary $\mathbf{X}$ and model them using binomial distributions. In the Discussion section, we consider problems arising from deviations from the assumed correlation structure and distributions of the risk factors, and propose solutions to address them.

Last, we specify the marginal distribution of Y. Since $P(Y)$ is usually not known, we compute it by marginalizing the joint distribution $P(Y, G, M, E, \mathbf{X})$ over all possible values of G, M, E and $\mathbf{X}$, denoted by G^*, M^*, E^* and $\mathbf{X}^*$. Therefore, we need to compute a two dimensional integral over M and E for all possible values of G and $\mathbf{X}$:

$$P(\mathbf{Y}) = \sum_{G^*, \mathbf{X}^*} \int_{E^*, M^*} P(\mathbf{Y}|G^*, E^*, M^*, \mathbf{X}^*) P(E^*, M^*|G^*, \mathbf{X}^*) P(G^*) P(\mathbf{X}^*) d_{E^*, M^*}$$

There exists no closed form solutions for this integral, thus numerical methods need to be employed to compute the integral and maximize the likelihood. Here, we use the Gauss-Hermite Quadrature for numerical integration and the R package `optim` for numerical optimization.

5.2.2 Statistical Testing

We wish to test the null hypothesis of no omics effect on disease risk. The null hypothesis can be written using the regression coefficients in (5.2) as:

$$\begin{aligned} H_0 &: \alpha_G = \alpha_M = \alpha_E = 0 \text{ versus} \\ H_1 &: \text{at least one of } \alpha_G, \alpha_M, \alpha_E \neq 0. \end{aligned} \tag{5.6}$$

Likelihood-based statistics can be used to make inference about the parameters of main interest. Here, a likelihood ratio test (LRT) is used to test the null hypothesis. Following standard likelihood theory, the LRT statistic under the null hypothesis asymptotically follows a χ^2_3 distribution for a correctly specified model.

In the simulation study below, we examine the impact of model misspecification on the distribution of the test statistic under the null and alternative hypothesis.

5.3 Simulation Study

We wish to compare the relative performance of our proposed $\mathcal{RL}$ approach with the $\mathcal{PL}$. We present results on type I error rate, bias, efficiency and power. We also study the performance of the methods under the null hypothesis when the joint distribution of M and E deviates from normality. In each scenario described below, 1000 replication data sets were simulated. In each replication, we generated data for 500 cases and 500 controls by sampling the cases and controls from a larger random sample of subjects.

Since in our real data set we have information on age (A) and gender (S) of the subjects, we also consider these two clinical covariates. Thus, in all the formulas above, $\mathbf{X} = (\mathbf{S}, \mathbf{A})$, $\boldsymbol{\alpha}_X = (\alpha_S, \alpha_A)^T$, $\boldsymbol{\beta}_{XoM} = (\beta_{SoM}, \beta_{AoM})$, and $\boldsymbol{\beta}_{XoE} = (\beta_{SoE}, \beta_{AoE})$. Age was treated as binary, with 0 indicating an individual with age younger than or equal to the median age in the population, i.e. 35, and 1 otherwise. In all scenarios below, binary age and gender were considered to be mutually independent. Thus $P(\mathbf{S}, \mathbf{A}) = P(\mathbf{S}) \times P(\mathbf{A})$. Age was generated from a $Binomial(1, .5)$, gender was generated from a $Binomial(1, .5)$, with zero indicating a male and one indicating a female, and the genetic variant G was generated from a $Binomial(2, .20)$. In all scenarios presented in the main text, in order to speed up the computation time, $\sigma^2_{\epsilon_1}$, $\sigma^2_{\epsilon_2}$ and $\sigma_{\epsilon_1,\epsilon_2}$ were fixed to their sample estimates, $\hat{\sigma}^2_{\epsilon_1}$, $\hat{\sigma}^2_{\epsilon_2}$ and $\hat{\sigma}_{\epsilon_1,\epsilon_2}$, and they no longer were part of the optimization procedure. In the simulation scenarios presented in the Appendix, $\sigma^2_{\epsilon_1}$, $\sigma^2_{\epsilon_2}$ and $\sigma_{\epsilon_1,\epsilon_2}$ were part of the optimization procedure.

5.3.1 Type I Error

First we studied the performance of the two methods in terms of type I error rate, when the distribution of the errors in (5.5) was properly specified. M and E were generated from (5.5) with no genetic, age or gender effect and normally distributed errors with $\sigma^2_{\epsilon_1} = \sigma^2_{\epsilon_2} = 1$ and $\sigma_{\epsilon_1\epsilon_2} = 0$. Other values were also tested but results are similar and are not shown. The binary disease outcome was generated from (5.2) with no genetic, epigenetic or transcriptional effect. The effect of age and gender was also set to zero, although this is not necessary and different values can be chosen. We set the intercept $\alpha_0 = logit(1e-03)$, such that the marginal disease probability in the population would be approximately $P(Y = 1) = .1\%$, reflecting a common disease with relatively low prevalence, such as multiple sclerosis. The type I error rate for $\mathcal{PL}$ and $\mathcal{RL}$ was 5.4% and 5.6%, respectively.

Next, we studied the performance of the methods when the normality assumption for the distribution of the errors in (5.5) was violated. To mimic situations in which outliers are present, we simulated the errors from a bi-variate t-distribution with 10 degrees of freedom and same location and scale parameters as the normal case. All other parameters remain the same. All methods properly control for the type I error rate; type I error rate for $\mathcal{PL}$ and $\mathcal{RL}$ was 4.7% and 5.1%, respectively.

5.3.2 Bias and Efficiency

Two different scenarios were considered. In the first scenario, the risk factors had moderate effect on disease risk. Specifically, parameters in (5.2) were set to $\alpha_E = \alpha_M = .18$, corresponding to an OR of 1.2, $\alpha_G = \alpha_S = \alpha_A = .26$, corresponding to an OR of 1.3. Moreover, parameters in (5.5) were set to values similar to our real data example, that is $\beta_{GoE} = \beta_{SoE} = \beta_{AoE} = \beta_{GoM} = \beta_{bSoM} = .1$, $\beta_{bAoM} = .3$, $\sigma^2_{\epsilon_1} = \sigma^2_{\epsilon_2} = 1$ and $\sigma_{\epsilon_1\epsilon_2} = .3$. In the second scenario, we considered effect sizes for E, M, S and A on disease risk that were closer to our real data example. Specifically, parameters in (5.2) were set to $\alpha_E = \alpha_M = 1$, corresponding to an OR of 3, $\alpha_G = .26$, corresponding to an OR of 1.3, and $\alpha_S = \alpha_A = -.69$, corresponding to an OR of .5. α_0 was set to $logit(9e-04)$, such that the marginal disease probability in the population would again be $.1\%$. Results on bias and efficiency of parameter estimates, for both scenario and likelihood approaches, are listed in Table 5.1. To study the impact of fixing $\sigma^2_{\epsilon_1}, \sigma^2_{\epsilon_2}$ and $\sigma_{\epsilon_1\epsilon_2}$ to their sample estimates, we repeat the analysis in both the scenarios described above, but this time we estimate also $\sigma^2_{\epsilon_1}, \sigma^2_{\epsilon_2}$ and $\sigma_{\epsilon_1\epsilon_2}$. Results on bias and efficiency for this case are listed in Table 5.1 of the Appendix.

Based on these simulation results we make the following key observations. First, as expected from theory, both $\mathcal{PL}$ and the proposed $\mathcal{RL}$ estimators provide essentially unbiased estimators of all regression parameters. For scenario 2, the bias for both likelihood is slightly larger than the bias for Scenario 1. This small increase in bias stems from the fact that in scenario 2 the effect sizes are larger, as compared to scenario 1. As a consequence, the impact of the ascertainment is stronger and thus the information available to estimate the parameters of interest is more limited. As expected, part of the bias of the $\mathcal{RL}$ also comes from fixing $\sigma^2_{\epsilon_1}, \sigma^2_{\epsilon_2}$ and $\sigma_{\epsilon_1\epsilon_2}$ to their sample estimates. This can be seen by comparing the bias of the $\mathcal{RL}$ in Table 5.1 with the bias in Table 5.1 of the Appendix; bias for α_E decreases from 4.6% to 3% and bias for α_A decreases from 4.5% to 2.1%.

Secondly, ratios of variance estimates of the parameter estimates from $\mathcal{RL}$ and $\mathcal{PL}$ estimators show that, when the information on the distribution of covariates is exploited correcting for ascertainment in case-control data, there is a major efficiency gain for the estimation of the regression coefficients. The gain is larger for the scenario with larger effect sizes, as compared to smaller effect sizes; and for continuous, as compared to discrete covariates. Results for the LRT for testing (5.6) also agree with the efficiency results; the test based on $\mathcal{RL}$ offers a mean increase of 9.5 $\%$ in χ^2 test statistic for the first scenario and 22.2% for the second scenario (results not shown). Moreover, the gain is larger when $\sigma^2_{\epsilon_1}, \sigma^2_{\epsilon_2}$ and $\sigma_{\epsilon_1\epsilon_2}$ are estimated rather than fixed to their sample estimates. Third, comparison of the empirical standard errors (SE) with the estimated SE of the $\mathcal{RL}$ shows that the numerical approximation of the integral using Gauss-Hermite Quadrature and the numerical optimization algorithm perform well for realistic parameter values and modest sample sizes. Finally, the estimated SE when $\sigma^2_{\epsilon_1}, \sigma^2_{\epsilon_2}$ and $\sigma_{\epsilon_1\epsilon_2}$ are estimated are smaller, compared to the estimated SE when $\sigma^2_{\epsilon_1}, \sigma^2_{\epsilon_2}$ and $\sigma_{\epsilon_1\epsilon_2}$ are fixed to their sample estimates.

5.4 Data Example

In this section, we re-analyze data from a case-control study of 28 patients of multiple sclerosis and 19 controls. In the initial study, quantile-normalized DNA methylation, log_2 normalized gene expression data, as well as information on age and gender, was available for each subject. In the initial study, Huynh et al. [2014] analyzed DNA methylation and gene expression data sets separately, correcting for age and gender. Significant results from each analysis were compared and several genes showed overlapping signals in both DNA methylation and gene expression analysis.

Table 5.1: Simulation study for studying bias and efficiency of the prospective likelihood ($\mathcal{PL}$) and retrospective likelihood ($\mathcal{RL}$). The frequency of the genetic variant was .20, the frequency of category 0 for gender and age was .5. The disease prevalence in the population was .1%, corresponding to a common disease with low prevalence. VR, variance ratio; SE, standard error; Emp: Empirical, Est: Estimated. Results are based on the average over 1000 simulated data set. $\sigma^2_{\epsilon_1}$, $\sigma^2_{\epsilon_2}$ and $\sigma_{\epsilon_1,\epsilon_2}$ in (5.5) were fixed to their sample estimates and were no longer part of the optimization procedure.

True	Bias		Emp SE		Est SE		MSE		VR
Values	$\mathcal{PL}$	$\mathcal{RL}$	$\mathcal{PL}$	$\mathcal{RL}$	$\mathcal{PL}$	$\mathcal{RL}$	$\mathcal{PL}$	$\mathcal{RL}$	$\frac{\mathcal{RL}}{\mathcal{PL}}$
				Scenario 1: Moderate effect sizes.					
$\alpha_E = .18$	.000	-.004	.068	.062	.069	.062	.005	.004	.910
$\alpha_M = .18$	.000	-.004	.068	.062	.070	.063	.005	.004	.911
$\alpha_G = .26$	.009	.009	.111	.109	.112	.111	.012	.012	.983
$\alpha_S = .26$	.004	.004	.130	.129	.136	.133	.017	.017	.986
$\alpha_A = .26$	.006	.007	.132	.129	.137	.133	.017	.017	.983
				Scenario 2: Large effect sizes.					
$\alpha_E = 1$	.013	-.046	.103	.091	.102	.088	.011	.010	.885
$\alpha_M = 1$	.011	-.045	.103	.091	.104	.090	.011	.010	.885
$\alpha_G = .26$	.011	-.030	.148	.137	.150	.138	.022	.020	.927
$\alpha_S = -.69$	-.008	-.027	.181	.169	.185	.178	.033	.029	.935
$\alpha_A = -.69$	-.005	-.025	.181	.169	.184	.173	.033	.029	.935

Here, we study one of the significant genes identified from the original analysis, SLC47A22, and apply both the prospective and the proposed retrospective likelihood approach. Age was treated as a binary variable, with 0 indicating an individual younger than or equal to 60 years old, which was the median age in our sample. The binary age and gender were considered to be independent in the population. For a binary age, this assumption is realistic, since, in 2010 in the United States, where our sample comes from, 83% of males were younger than 60 as opposed to 81 % of females [Howden and Meyer, 2011].

Since we do not have information for the genetic covariates, the two likelihoods are modified as follows:

$$\mathcal{PL}(\boldsymbol{\alpha}) = P\left(\mathbf{Y}|\mathbf{M},\mathbf{E},\mathbf{A},\mathbf{S}\right) \tag{5.7}$$

$$\mathcal{RL}(\boldsymbol{\theta}) = \frac{P\left(\mathbf{Y}|\mathbf{M},\mathbf{E},\mathbf{A},\mathbf{S}\right) P\left(\mathbf{M},\mathbf{E}|,\mathbf{A},\mathbf{S}\right) P\left(\mathbf{A}\right) P\left(\mathbf{S}\right)}{P\left(\mathbf{Y}\right)} \tag{5.8}$$

and the null hypothesis for the parameters of interest is now the following

$$H_0 : \alpha_M = \alpha_E = 0.$$

We assume that under the null hypothesis the LRT statistic is asymptotically χ^2_2 distributed.

For this gene, DNA methylation is available for 15 sites. Given the small size of our sample, we applied (5.7) and (5.8) 15 times, keeping the same E, A and S and adding a different methylation site in the model each time. Parameter estimates, standard errors and p-value for the LRT test for each model and method used, are listed in Tables A.5.3 - A.5.5

of the Appendix. In Figure 5.2, we plot the parameter estimates with their 95% confidence intervals (CI) for both methods.

Based on these results, we make the following observations. Our approach had smaller standard errors than $\mathcal{PL}$ approach in 11 out of 15 estimates for α_M, with an increase in efficiency of $5-20\%$, comparable standard errors in 3 out of 15 sites, with a $< 5\%$ increase or decrease in efficiency, and larger standard errors in 1 out of 15 sites, with a $4-7\%$ decrease in efficiency. The largest reduction in standard errors, 20%, was for the fourth methylation site. Moreover, site 9 was significant at nominal level when the $\mathcal{RL}$ was used and not when $\mathcal{PL}$ was used. Standard errors of α_E, α_S and α_A for $\mathcal{RL}$ were $3-8\%$ smaller than for $\mathcal{PL}$ when averaging across the 15 models.

5.5 Conclusions and Discussion

In this paper, we have proposed a statistical framework for efficient integration of omics and non-omics factors in case-control association studies. We used a retrospective likelihood approach to model the distribution of the risk factors conditional on the case controls status and performed a LRT for the joint effect of omics factors on disease risk. We demonstrated via simulation studies and real data analysis that the retrospective likelihood approach can be more efficient than the prospective likelihood when integrating data from case-control studies.

In order to compute the retrospective likelihood, we made certain assumptions about the correlation structure between the risk factors in the population. If evidence about additional independences exists, e.g. independence between E and M or their independence from G and X, our method can be modified accordingly. If G and X are not independent, e.g. due to population stratification, estimates of α_G and α_X could be biased. To address this issue, Chatterjee and Carroll [2005] proposed to model the distribution of G and X conditional on other common measured factors, such as principal components. Alternatively, Mukherjee and Chatterjee [2008] proposed to use an empirical Bayes-type shrinkage estimator that corrects for falsely attributed independence of covariates. For X binary, a multinomial distribution can be used for the joint distribution $P(\mathbf{X}, \mathbf{G})$. In addition, if X is a discrete variable with many levels or a continuous variable, the joint distribution could be factorized as $P(X, G) = P(X|G) \times P(X)$ and a Poisson or linear regression could be used. Last, we assumed the non-omics X's to be mutually independent. For age and gender this assumption can be verified using population registries. If this assumption is violated, methods proposed above to address the violation of G-X independence assumption can be used.

In addition to assumptions about the correlation structure between the risk factors, our method makes assumptions about the distribution of the risk factors. We assume that after proper transformations and normalization procedures, E and M are normally distributed [Calza and Pawitan, 2010; Yousefi et al., 2013]. When this assumption is violated, e.g. heavy tails or skewed distributions, our method could give biased parameter estimates (see Table A.5.2 of the Appendix). To avoid this issue, more flexible or discrete distributions can be considered for the error distributions of E and M, e.g. Laplace or negative binomial distribution [Purdom and Holmes, 2005; Sun, 2012]. Alternatively, quantile normalization techniques can be used to align the quantiles of E and M to a normal distribution. Such techniques can result in the dilution of the effects of the risk factors on disease risk and should therefore be used with caution. Moreover, the interpretation of parameters after the quantile normalization is no longer possible, thus we advice the use of different distributions rather than normalization. In our real data example, E and M were normalized prior to analysis [Huynh et al., 2014]. However, in such a small sample, it is difficult to verify normality and we did not formally test the fit. Among other possible reasons, deviations from normality could explain why the $\mathcal{PL}$ had in some cases similar or smaller standard errors

for α_M and α_E than the $\mathcal{RL}$. In this work, we treated age as binary, which might have decreased the gain in efficiency from exploiting the independence assumption for estimating α_S and α_A. Last, it is known that in small samples the logistic regression can give biased estimates of the OR's [Nemes et al., 2009] thus results in the real data for both methods should be interpreted with caution.

Efficiency of parameter estimation can be further increased using external knowledge about the disease prevalence or distribution of the covariates in the population. This information could be incorporated in several ways. Chatterjee and Carroll [2005] and Tsonaka et al. [2013] show how to incorporate external information about disease prevalence. Huijts et al. [2014] show how to increase efficiency for estimating genetic effects using existing genotype data from controls. Zaitlen et al. [2012a] show how to incorporate external information about the distribution of the risk factors based on the liability threshold model with parameters informed by external epidemiological information. The latter approach provides increased efficiency not only for phenotype based ascertainment but also for phenotype and covariate based ascertainment. Our approach could be modified in a similar way to include such information and this is among our future extensions.

In this article, we have considered a simple association approach by focusing on the joint association of a single genetic, epigenetic and transcriptional factor per gene with the phenotype. One possible way to accommodate settings with several genetic or epigenetic factors per gene is to use a mixed logistic regression in which the regression coefficient of the main genetic or epigenetic effects are assumed to follow an arbitrary distribution, e.g. the normal distribution [Huang et al., 2014]. Alternatively, penalized likelihood approaches, which put a separate penalty for the genetic and epigenetic factors, could be used. Furthermore, concepts from mediation analysis framework can be used to construct more powerful testing procedures. For example, here we test for the joint effect of the omics factors. This test can be considered as a test for the total effect of G on Y, directly on Y or indirectly via E and M. In the case of complete mediation of the genetic effect via E and M, i.e. α_G=0, a more powerful approach for assessing genetic associations would be to test only for the indirect genetic effect [Kenny and Judd, 2013; Zhao et al., 2014]. These possibilities are among our future research interests.

In summary, the retrospective likelihood based inference can be more efficient than prospective based inference for joint analysis of multiple omics and non-omics risk factors in case-controls association studies. Efficiency gain is a function of the number of parameters used to model the distribution of the risk factors and the effect sizes of risk factors, with increased efficiency gain for continuous factors and for risk factors with large effect sizes.

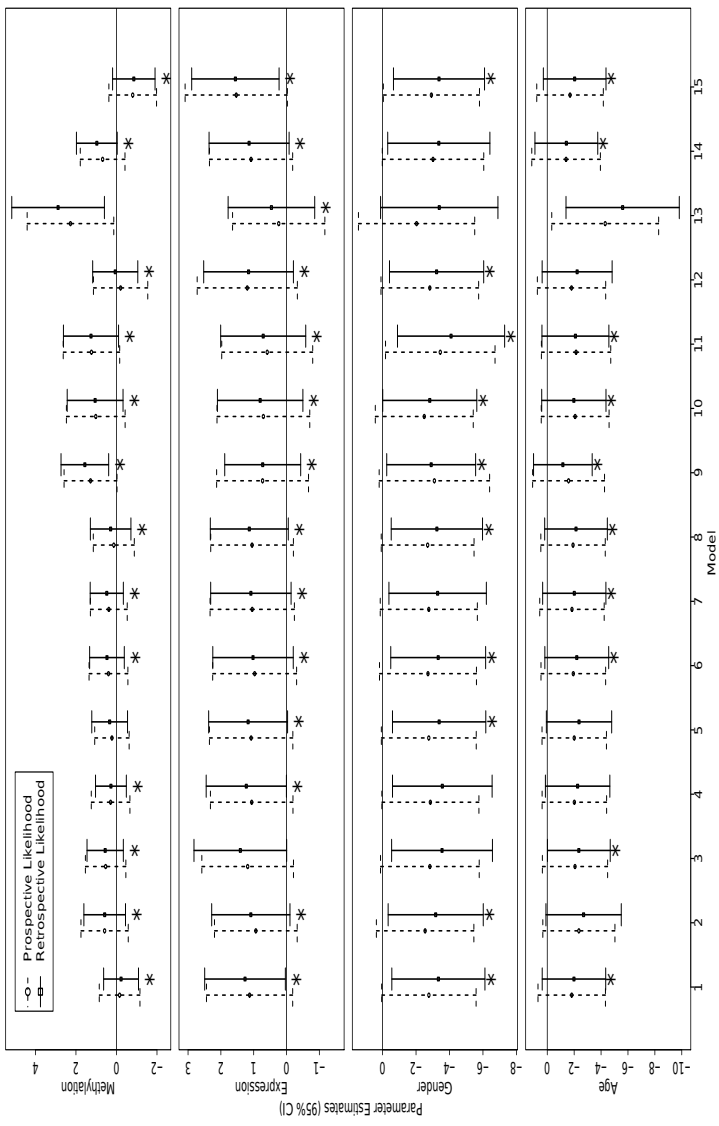


Figure 5.2: Results from applying prospective (5.1) and retrospective (5.4) approaches in the multiple sclerosis data. Each plot (top to bottom) shows estimates of the OR parameters, with their 95% confidence intervals, for the effect of methylation, gene expression, gender and age on multiple sclerosis in each of the 15 models fitted. The asterisk (*) denotes cases in which the confidence intervals of a parameter estimated using the retrospective likelihood are narrower compared to the confidence intervals estimated using the prospective approach.

Appendix

Table A.5.1: Simulation study for studying bias and efficiency of the prospective likelihood ($\mathcal{PL}$) and retrospective likelihood ($\mathcal{RL}$). The frequency of the genetic variant was .20, the frequency of category 0 for gender and age was .5. The disease prevalence in the population was .1%, corresponding to a common disease with low prevalence. VR, variance ratio ; SE, standard error; Emp: Empirical, Est: Estimated. Results are based on the average over 1000 simulated data set. $\sigma^2_{\epsilon_1}$, $\sigma^2_{\epsilon_2}$ and $\sigma_{\epsilon_1,\epsilon_2}$ in (5.5) were part of the optimization procedure.

True	Bias		Emp SE		Est SE		MSE		VR
Values	$\mathcal{PL}$	$\mathcal{RL}$	$\mathcal{PL}$	$\mathcal{RL}$	$\mathcal{PL}$	$\mathcal{RL}$	$\mathcal{PL}$	$\mathcal{RL}$	$\frac{\mathcal{RL}}{\mathcal{PL}}$
Scenario 1: Moderate effect sizes.									
$\alpha_E = .18$	.003	.006	.068	.065	.068	.051	.005	.004	.948
$\alpha_M = .18$	.003	.000	.068	.065	.070	.055	.005	.004	.952
$\alpha_G = .26$	-.001	.015	.111	.109	.116	.086	.012	.012	.986
$\alpha_S = .26$	.003	.026	.130	.130	.130	.103	.017	.017	.994
$\alpha_A = .26$	-.004	.024	.132	.131	.126	.098	.017	.018	.992
Scenario 2: Large effect sizes.									
$\alpha_E = 1$	.019	.030	.103	.091	.106	.061	.011	.009	.882
$\alpha_M = 1$	.011	.021	.103	.090	.104	.059	.011	.009	.879
$\alpha_G = .26$	.006	.028	.148	.135	.144	.084	.022	.019	.914
$\alpha_S = -.69$	-.009	.032	.181	.162	.190	.098	.033	.027	.893
$\alpha_A = -.69$	-.008	.029	.181	.162	.180	.097	.033	.027	.896

Table A.5.2: Simulation study for studying bias and efficiency of the $\mathcal{PL}$ and $\mathcal{RL}$ when the true error distribution for E and M was a bi-variate (a) normal, (b) t_{10}, (b) t_{50}, (d) $SN^{1,2}$ with slant $a = c(1,1)$ and (e) $SN^{1,2}$ with slant $a = c(2,2)$. Only E, M and G were considered as covariates. The frequency of the genetic variant was .20; disease prevalence was .1%; $\sigma^2_{\epsilon_1} = \sigma^2_{\epsilon_2} = 1$ and $\sigma_{\epsilon_1\epsilon_1} = .3$. The location and scale parameters of the t and SN distributions are the same as the normal. SN: skew-normal, VR, variance ratio ; SE, standard error; Emp: Empirical, MSE: mean squared error. Results are based on the average over 1000 simulated data set. $\sigma^2_{\epsilon_1}$, $\sigma^2_{\epsilon_2}$ and $\sigma_{\epsilon_1,\epsilon_2}$ in (5.5) were part of the optimization procedure.

True	Bias		Emp SE		Est SE		MSE		VR
Values	$\mathcal{PL}$	$\mathcal{RL}$	$\mathcal{PL}$	$\mathcal{RL}$	$\mathcal{PL}$	$\mathcal{RL}$	$\mathcal{PL}$	$\mathcal{RL}$	$\frac{\mathcal{RL}}{\mathcal{PL}}$
				(a) Bi-variate Normal					
$\alpha_E = 1$	.010	.012	.101	.087	.102	.080	.010	.008	.856
$\alpha_M = 1$	.010	.012	.101	.087	.103	.082	.010	.008	.856
$\alpha_G = .26$	.001	.008	.146	.132	.146	.124	.021	.018	.906
				(b) Bi-variate t_{10}					
$\alpha_E = 1$	.012	-.164	.101	.071	.099	.106	.010	.032	.701
$\alpha_M = 1$	.010	-.164	.101	.071	.103	.104	.010	.032	.702
$\alpha_G = .26$	.009	.032	.165	.141	.167	.204	.027	.021	.850
				(c) Bi-variate t_{50}					
$\alpha_E = 1$	.003	-.022	.101	.085	.104	.086	.010	.008	.840
$\alpha_M = 1$	.009	-.023	.101	.084	.099	.079	.010	.008	.836
$\alpha_G = .26$	.003	.011	.149	.133	.152	.139	.022	.018	.897
				(d) Bi-variate Skew-Normal with $a = (1,1)$					
$\alpha_E = 1$	.010	-.012	.099	.087	.100	.093	.010	.008	.873
$\alpha_M = 1$	.007	-.013	.099	.087	.102	.094	.010	.008	.875
$\alpha_G = .26$	-.003	-.006	.133	.124	.136	.127	.018	.015	.931
				(e) Bi-variate Skew-Normal with $a = (2,2)$					
$\alpha_E = 1$	.009	-.014	.099	.087	.098	.101	.010	.008	.876
$\alpha_M = 1$	.004	-.019	.099	.087	.099	.098	.010	.008	.878
$\alpha_G = .26$	-.005	-.007	.131	.122	.134	.131	.017	.015	.933

[1] Notation as Azzalini, A. with the collaboration of Capitanio, A. (2014). The Skew-Normal and Related Families. Cambridge University Press, IMS Monographs series.

[2] To generate data from a bi-variate SN, the R package `SN` was used.

Table A.5.3: Results from analysis of multiple sclerosis data using the prospective likelihood ($\mathcal{PL}$) and retrospective liklelihood ($\mathcal{RL}$). Estimates (Est) and standard errors (SE) of the effect of methylation and gene expression on multiple sclerosis α_M, in each of the 15 models. VR, variance ratio. If VR < 1, $\mathcal{RL}$ gives smaller standard errors for the parameter estimates.

Model	$\mathcal{PL}$		$\mathcal{RL}$		VR
	Est	SE	Est	SE	$\frac{\mathcal{RL}}{\mathcal{PL}}$
			Methylation		
1	$-.16$	.51	$-.23$	.44	.86
2	.59	.60	.58	.52	.88
3	.53	.51	.56	.46	.90
4	.29	.49	.27	.39	.80
5	.22	.43	.34	.45	1.04
6	.39	.49	.47	.44	.89
7	.38	.46	.48	.42	.90
8	.13	.52	.29	.51	.99
9	1.28	.67	1.56	.60	.90
10	1.02	.74	1.05	.70	.95
11	1.23	.71	1.26	.69	.98
12	$-.20$	.68	.06	.57	.83
13	2.27	1.09	2.88	1.17	1.07
14	.68	.56	.97	.51	.91
15	$-.80$	.60	$-.86$	.53	.89
			Gene Expression		
1	1.13	.67	1.26	.63	.94
2	.93	.64	1.09	.61	.95
3	1.18	.71	1.41	.72	1.01
4	1.06	.64	1.23	.62	.97
5	1.08	.65	1.17	.61	.95
6	.97	.65	1.02	.63	.96
7	1.04	.66	1.09	.62	.95
8	1.05	.64	1.13	.61	.94
9	.73	.71	.73	.59	.83
10	.70	.72	.80	.66	.92
11	.59	.71	.71	.66	.93
12	1.20	.78	1.16	.70	.90
13	.24	.72	.46	.67	.94
14	1.08	.65	1.14	.62	.96
15	1.53	.79	1.56	.68	.85

Table A.5.4: Results from analysis of multiple sclerosis data using the prospective likelihood ($\mathcal{PL}$) and retrospective liklelihood ($\mathcal{RL}$). Estimates (Est) and standard errors (SE) of the effect of gender and age on multiple sclerosis α_S, in each of the 15 models. VR, variance ratio. If VR < 1, $\mathcal{RL}$ gives smaller standard errors for the parameter estimates.

Model	$\mathcal{PL}$		$\mathcal{RL}$		VR
	Est	SE	Est	SE	$\frac{\mathcal{RL}}{\mathcal{PL}}$
			Gender		
1	-2.77	1.44	-3.34	1.42	.99
2	-2.54	1.49	-3.17	1.45	.98
3	-2.83	1.51	-3.55	1.54	1.02
4	-2.86	1.48	-3.57	1.52	1.03
5	-2.76	1.45	-3.38	1.42	.98
6	-2.71	1.48	-3.32	1.45	.98
7	-2.76	1.48	-3.29	1.49	1.00
8	-2.69	1.42	-3.25	1.39	.98
9	-3.10	1.68	-2.91	1.36	.80
10	-2.49	1.49	-2.82	1.43	.96
11	-3.45	1.67	-4.09	1.63	.98
12	-2.82	1.49	-3.23	1.44	.96
13	-2.03	1.77	-3.39	1.79	1.01
14	-3.02	1.55	-3.36	1.56	1.01
15	-2.92	1.47	-3.37	1.39	.95
			Age		
1	-1.82	1.28	-1.99	1.20	.94
2	-2.35	1.37	-2.70	1.43	1.04
3	-2.06	1.24	-2.35	1.19	.96
4	-2.02	1.22	-2.26	1.22	1.00
5	-2.01	1.22	-2.36	1.24	1.01
6	-1.94	1.23	-2.19	1.21	.99
7	-1.84	1.22	-2.01	1.20	.98
8	-1.92	1.22	-2.14	1.19	.97
9	-1.58	1.36	-1.16	1.11	.82
10	-2.07	1.29	-1.98	1.22	.95
11	-2.14	1.32	-2.09	1.27	.96
12	-1.80	1.30	-2.22	1.33	1.02
13	-4.31	2.03	-5.60	2.15	1.06
14	-1.40	1.30	-1.42	1.19	.92
15	-1.69	1.26	-2.04	1.19	.94

Table A.5.5: Results from analysis of multiple sclerosis data using the prospective likelihood ($\mathcal{PL}$) and retrospective liklelihood ($\mathcal{RL}$). Pvalues from the two degrees of freedom likelihood ratio test for testing the null hypothesis of no methylation and expression effect on multiple sclerosis, in each of the 15 models.

Model	$\mathcal{PL}$	$\mathcal{RL}$
1	1.8e-01	9.4e-02
2	1.1e-01	5.4e-02
3	1.0e-01	4.5e-02
4	1.6e-01	8.2e-02
5	1.7e-01	8.1e-02
6	1.4e-01	5.7e-02
7	1.4e-01	5.5e-02
8	1.9e-01	9.1e-02
9	2.1e-02	2.3e-03
10	5.9e-02	2.5e-02
11	2.8e-02	1.1e-02
12	1.8e-01	1.1e-01
13	5.0e-03	3.0e-04
14	8.6e-02	1.8e-02
15	6.7e-02	2.4e-02

6

Classification and Visualization Based on Derived Image Features: Application to Genetic Syndromes [1]

Summary

Data transformations prior to analysis may be beneficial in classification tasks. In this article we investigate a set of such transformations on 2D graph-data derived from facial images and their effect on classification accuracy in a high-dimensional setting. These transformations are low-variance in the sense that each involves only a fixed small number of input features. We show that classification accuracy can be improved when penalized regression techniques are employed, as compared to a principal component analysis (PCA) pre-processing step. In our data example classification accuracy improves from 47% to 62% when switching from PCA to penalized regression. A second goal is to visualize the resulting classifiers. We develop importance plots highlighting the influence of coordinates in the original 2D space. Features used for classification are mapped to coordinates in the original images and combined into an importance measure for each pixel. These plots assist in assessing plausibility of classifiers, interpretation of classifiers, and determination of the relative importance of different features.

6.1 Introduction

In clinical genetics, syndrome diagnosis presents a classification problem, namely whether and if so which syndrome is to be diagnosed for the presenting patient. We here focus on facial image data in order to facilitate this diagnosis. Facial features play an important role in syndrome diagnosis [Winter, 1996]. We have previously demonstrated that information from 2D images can help in this classification problem [Boehringer et al., 2006; Vollmar et al.,

[1] Published in *PLoS One*.

2008; Boehringer et al., 2011]. Similar work in 3D confirms this assessment [Hammond et al., 2005; Hennessy et al., 2007; Hammond et al., 2012].

This classification problem tends to be high-dimensional, i.e. the number of covariates is bigger than the number of observations. Previously, we employed classical dimension reduction by principal component analysis (PCA) and showed that PCA has a large contribution to classification errors [Boehringer et al., 2011]. This can be seen by comparing cross-validation (CV) runs used to estimate error once including a PCA within each fold and once performing PCA prior to CV. It is well-known that feature selection must occur within CV to accurately estimate prediction error [Molinaro et al., 2005] and indicates that this step plays a crucial role in our application. Principal components (PCs) can exhibit high variation in small data sets [Jolliffe, 2005] which is a possible explanation for our results. To test this assumption, PCA is compared to low-variance transformation and their classification performance is evaluated.

We here pursue penalized regression techniques that are applicable in the high-dimensional setting and can be applied to data directly without preceding dimension reduction [Tibshirani, 1996]. The process of fitting the regression model itself ensures that the final model is low dimensional and asymptotically only contains true predictors. Furthermore, in the low-dimensional setting, a trade-off between variance of predictors and their unbiasedness leads to improved accuracy (such as measured by classification accuracy or the mean-squared-error) as compared to least-squares regression [Hastie et al., 2001]. One advantage of being able to directly work with high-dimensional data is that the dimensionality of data can be even increased further prior to performing classification. We combine these ideas with geometric properties of our data set by applying low-variance transformations on coordinates that represent features in 2D images. For example, distances are computed between graph vertices depending on only two of them. By contrast, PCs in general depend on all vertices derived from a given 2D image. We evaluate the performance of classifiers resulting from such a strategy.

A second goal is to visualize resulting classifiers. If PCA is used together with a linear classification technique such as linear discriminant analysis (LDA) all transformations leading from one group to another in a two-class classification problem can be represented by a single direction in the original feature space. This can be used to create caricatures by moving data points or means away from each other along this direction [Boehringer et al., 2006]. If non-linear transformations are involved visualization becomes more challenging. We develop a general framework that allows to create visualizations that indicate importance of neighborhoods in the original 2D space. We apply this methodology to the original syndrome data.

6.2 Materials and Methods

6.2.1 Ethics statement

Written informed consent was received from all patients or their wardens and the study was approved by the medical ethical committee of the Universitätsklinikum Essen, Germany. Consent was documented on forms which were reviewed and approved by the medical ethical committee of the Universitätsklinikum Essen, Germany.

6.2.2 Data

Frontal 2D images of 205 individuals each diagnosed with one of 14 syndromes were included in the study. This data set was used in a previous study and is described in detail elsewhere

Table 6.1: Description of data set with numbers per class.

Syndrome	Number of Individuals
Microdeletion 22q11.2 [22q]	25
Wolf-Hirschhorn syndrome [4p]	12
Cri-du-chat syndrome [5p]	16
Cornelia de Lange syndrome [CDL]	17
Fragile X syndrome [fraX]	9
Mucopolysaccharidosis Type II [MPS2]	6
Mucopolysaccharidosis Type III [MPS3]	7
Noonan syndrome [Noon]	13
Progeria [Pro]	5
Prader-Willi syndrome [PWS]	13
Smith-Lemli-Opitz syndrome [SLO]	15
Sotos syndrome [Sot]	15
Treacher Collins syndrome [TCS]	10
Williams-Beuren syndrome [WBS]	42

[Boehringer et al., 2006]. Table 6.1 summarizes the number of individuals available per syndrome. In this study, we used coordinate from 48 manually placed landmarks (vertices) that were registered on 2D greyscale images (Figure 6.1). These landmarks represent anatomical features in the face. The process of picture pre-processing and landmark registration is described elsewhere [Boehringer et al., 2006].

6.2.3 Data pre-processing

Vertices were standardized according to translation, rotation and size analogously to a Procrustes analysis [Gower, 1975] (graphs were rotated so that the average angle of symmetric points was 0, the center of the graph was 0 (as defined by the sum of x and y coordinates, respectively) and the size of the graph was scaled to unit size; as defined by the bounding

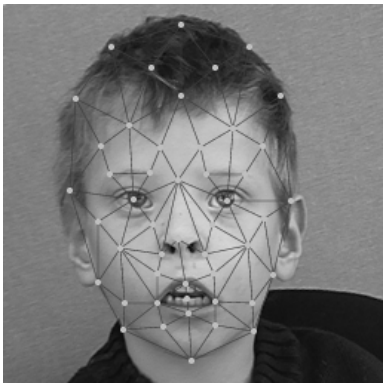


Figure 6.1: Illustration of data set with example of registered nodes.

rectangle). On this data, all possible pairwise distances between vertices were computed (D = 1128). To avoid multicollinearity problems, pairs of symmetric distances were averaged (Figure 6.2.a) reducing the number to 778 distances. Using a Delaunay triangulation of the set of averaged vertex positions, we constructed 41 triangles for which 41 areas and 123 angles were computed. Again, symmetric features were averaged. To assess the role of symmetry in syndrome discrimination, asymmetry scores for coordinate pairs, triangle areas and distances were calculated as the sum of squared residuals resulting from the averaging procedure between symmetric information. In order to be able to estimate possible non-linear effects, the square of each feature was also computed. In total, 2x1044=2088 covariates were derived per individual from the initial 96 values.

6.2.4 Statistical Analysis

We performed both simultaneous classification and pairwise classification of syndromes. Simultaneous classification serves to evaluate the problem of assigning a syndrome to a given face, that is, the problem of diagnosis. Pairwise comparisons of syndromes can be used to evaluate similarity of syndromes and to compare the performance achieved with the current data set to other data sets published thus far.

Due to the high dimensionality of the data set (number of individuals $= 205 <<$ number of covariates $= 2088$), dimension reduction techniques need to be employed. For simultaneous classification we trained classifiers using regularized multinomial regression with an elastic net penalty [Friedman et al., 2010]. Multinomial regression is a generalization of linear logistic regression model to a multi-logit model, when the categorical response variable has more than 2 levels. For pairwise classification we used regularized logistic regression with an elastic net penalty. Elastic net penalty is a penalized least squares method using a convex combination of the lasso and ridge penalty (with mixing parameter α). In contrast to the LASSO component, which as a general rule selects only one covariate from a group of correlated covariates, the ridge penalty has the effect of distributing effects over covariates that are highly correlated, entering them together into the model. Parameter α can therefore be chosen to control the sparsity of the final model.

We do not consider α to be a tuning parameter but instead consider twenty values of α between 0 and 1 as alternative models. To evaluate model performance, leave-one-out CV was performed. For each of the twenty elastic net models and the PCA analysis, four different covariate sets were used: coordinates of points only, points and their squares, all features and all features and their squared values. Comparisons between these covariate sets allow determining the trade-off between introducing more variation into the data by additional transformations and being able to potentially use more accurate features for the purpose of classification. Fitting an elastic-net model involves choosing a tuning parameter λ for the L_1-penalty, which was chosen by a nested loop of leave-one-out CV. Likewise, PCA uses an inner CV-loop to estimate principal components (PCs) and train a regression model based on these PCs. In the outer loop, data was mapped to these PCs onto which the prediction model was applied.To directly compare classification performance with a classical PCA approach, the outer CV loop was identical for the elastic net and PCA models, i.e. outer CV-folds were computed and identically used for all models.

To compute simultaneous accuracy for the PCA, we trained classifiers using multinomial logistic regression. 70 PCs were extracted from the whole data set. Subsequently, stepwise forward selection was performed to select PCs relevant for the classification decision based on the Akaike information criterion (AIC). The selected models were used to predict the samples in the test set of each CV-fold.

All statistical analyses were performed using the software package R (version 3.0.1) [R Core Team, 2014]. We used the package `geometry` for the Delaunay triangulation

and package glmnet to perform model selection and regularized multinomial and logistic regression with an elastic net penalty.

6.2.5 Visualization

The aim of our visualization strategy is to assign an importance value to each point in an average image of a class that represents how important features in that location are to discriminate the given class. While this strategy does not directly represent changes in, for example, distances, it allows to combine all features relevant for a classification decision in a single image. Figure 6.2.b illustrates the process of computing the color coefficient for a point δ based on the following significant features: a point p_1, a distance d_1, an area of triangle t_1 and an angle of a traingle a_1. We assume that a weight is assigned to each feature, in our case regression coefficients denoted with β_{p_1}, β_{d_1} , β_{t_1} and β_{a_1}. To calculate the importance of point δ we define the distances of this point to the significant features. For p_1 we compute the Euclidean distance of δ to p_1 , for d_1 we compute the Euclidean distance of δ to m_1, the midpoint of d_1 , for t_1 we compute the Euclidean distance of δ to c_1, the centroid of t_1 and for a_1 we compute the Euclidean distance to c_1, the vertex of a_1, respectively. The importance of each point is then defined as the sum of the weights, in our case regression coefficients, inversely weighted by the distances. This definition assumes that all weights are measured on the same scale, which can be assured by standardizing covariates in the regression setting. Finally, we normalize these importance values to (0, 1) by using the logistic function and we map resulting values to a color palette. As we symmetrized our data set, we also create symmetrized plots, i.e., one half is computed and mirrored to the other part. We overlay these maps on average facial images for the class corresponding to the respective classifier. The procedure of producing average images is described elsewhere [Günther, 2012].

For glmnet we used the regression coefficient of each feature as weights. To obtain the coefficients of each feature when PCA was performed, regression coefficients of PCs are back-calculated to the original feature space using the loadings matrix. The weight for each feature is the sum of contributions over all PCs.

6.3 Results

6.3.1 Model Selection

Average misclassification error (AME) rate for each choice of the mixing parameter α and feature set are reported in Table 6.2. In the last row of the table we list the results for the PCA. In Figure 6.3 we illustrate these results together with the 95% confidence intervals. The best model for glmnet is obtained for $\alpha = .105$ when the set of all features was used with an AME = 0.38 (95% CI: 0.31 - 0.44). PCA performed best when only points were used with AME = 0.53 (95% CI: 0.46 - 0.60). The AME of glmnet decreased with increasing number of features. In contrast, the AME of PCA increases. Results from the inner leave-one-out CV for glmnet models for $\alpha = .105$ to choose tuning parameter λ that gives the lowest AME rate are plotted in Figure 6.4. The lowest AME rate was obtained for λ=0.047.The difference between the best glmnet model for all features and best PCA model (points) is significant (Z-test for 2 population proportions, p-value=.0015).

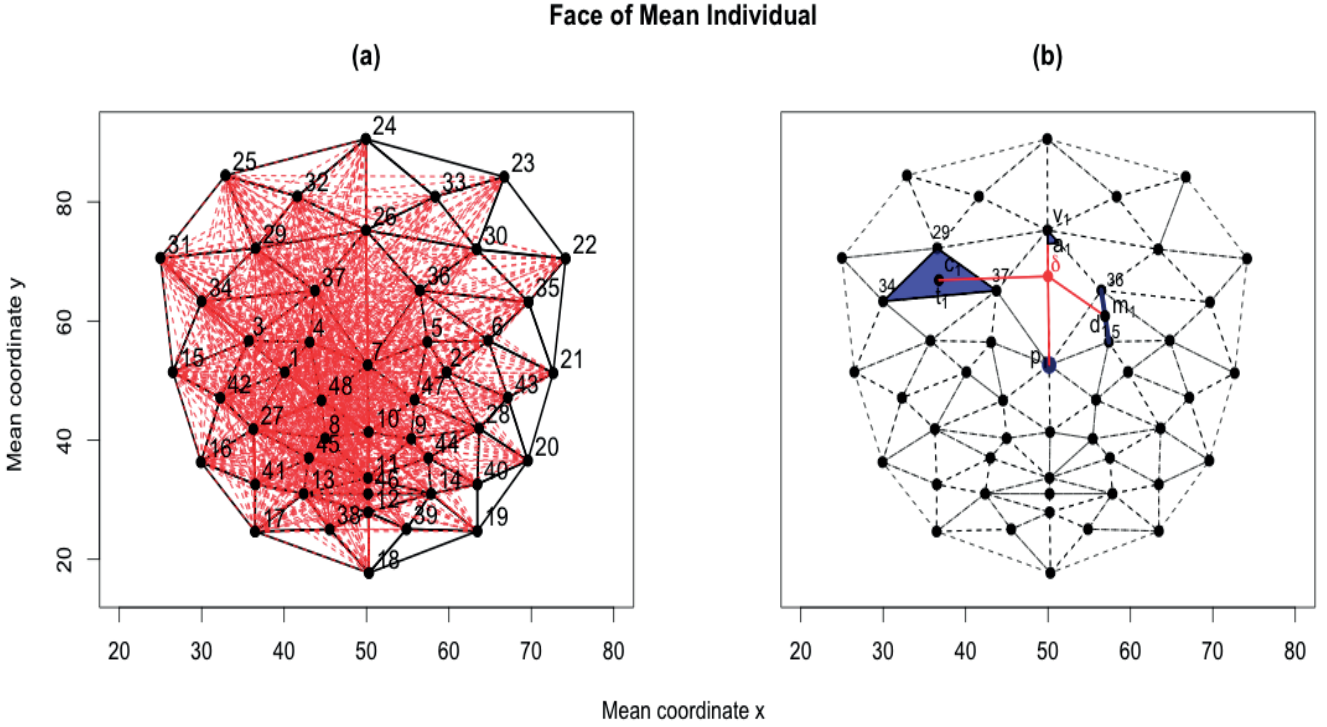


Figure 6.2: Illustration of data set and importance weighting. (a) Distances between coordinate pairs excluding symmetries. Numbers 1 to 48 correspond to landmarks; red: pairwise connections, excluding symmetries; black : Delaunay triangulation. (b) Illustration of the procedure to compute importance for a point δ. Significant point p_1, triangle t_1, distance d_1, and angle a_1 used to compute importance of point δ are highlighted in dark blue. c_1:centroid of t_1; m_1: midpoint of d_1; v_1: vertex of a_1. Red : distance of δ from c_1, m_1, p_1, and v_1.

Table 6.2: Average misclassification error (AME) with 95% confidence interval for leave-one-out cross validation for `glmnet`, 20 different values of α (see text), and PCA using only points (p), all features (a), only points and their squares (p+p^2) and all features and their squares (a+a^2).

	p	a	p+p^2	a+a^2
α=0	.400 (.333 - .467)	.444 (.376 - .512)	.498 (.429 - .566)	.493 (.424 - .561)
α=.053	.415 (.347 - .482)	.390 (.323 - .457)	.488 (.419 - .556)	.454 (.385 - .522)
α=.105	.410 (.342 - .477)	**.376 (.309 - .442)**	.502 (.434 - .571)	.468 (.400 - .537)
α=.158	.415 (.347 - .482)	.380 (.314 - .447)	.488 (.419 - .556)	.478 (.410 - .547)
α=.211	.415 (.347 - .482)	.385 (.319 - .452)	.483 (.414 - .552)	.493 (.424 - .561)
α=.263	.405 (.338 - .472)	.405 (.338 - .472)	.498 (.429 - .566)	.502 (.434 - .571)
α=.316	.395 (.328 - .462)	.410 (.342 - .477)	.498 (.429 - .566)	.493 (.424 - .561)
α=.368	.415 (.347 - .482)	.405 (.338 - .472)	.493 (.424 - .561)	.498 (.429 - .566)
α=.421	.415 (.347 - .482)	.415 (.347 - .482)	.488 (.419 - .556)	.507 (.439 - .576)
α=.474	.429 (.361 - .497)	.405 (.338 - .472)	.483 (.414 - .552)	.512 (.444 - .581)
α=.526	.434 (.366 - .502)	.415 (.347 - .482)	.498 (.429 - .566)	.522 (.453 - .590)
α=.579	.439 (.371 - .507)	.420 (.352 - .487)	.502 (.434 - .571)	.517 (.448 - .586)
α=.632	.434 (.366 - .502)	.420 (.352 - .487)	.512 (.444 - .581)	.537 (.468 - .605)
α=.684	.434 (.366 - .502)	.434 (.366 - .502)	.517 (.448 - .586)	.527 (.458 - .595)
α=.737	.444 (.376 - .512)	.434 (.366 - .502)	.512 (.444 - .581)	.532 (.463 - .600)
α=.789	.439 (.371 - .507)	.424 (.357 - .492)	.512 (.444 - .581)	.541 (.473 - .610)
α=.842	.463 (.395 - .532)	.424 (.357 - .492)	.507 (.439 - .576)	.541 (.473 - .610)
α=.895	.493 (.424 - .561)	.424 (.357 - .492)	.512 (.444 - .581)	.541 (.473 - .610)
α=.947	.493 (.424 - .561)	.439 (.371 - .507)	.507 (.439 - .576)	.541 (.473 - .610)
α=1	.493 (.424 - .561)	.439 (.371 - .507)	.507 (.439 - .576)	.546 (.478 - .615)
PCA	.532 (.463 - .600)	.810 (.756 - .864)	.527 (.458 - .595)	.727 (.666 - .788)

6.3.2 Simultaneous classification

Results for simultaneous classification using the best `glmnet` model are reported in Table 6.3 and 6.4. Specifically, Table 6.3 shows breakup of AME per syndrome. The best performance was achieved for WBS (AME=9.5%) and 22q (AME=20%). The lowest performance was achieved for the syndromes with the smallest sample sizes, MPS2 (AME=100%) and MPS3 (AME=70%). Table 6.4 shows the corresponding confusion matrix, i.e. what were the classification decisions per syndrome? For example, 22q was confused with 5p, Sot and WBS, whereas MPS2 was confused with MPS3, 22q, SLO and WBS.

We summarize the number of components used for the classification decision in Table 6.5. Approximately 200 features were selected per syndrome. Distances seemed to be more important (ca. 150 distances per syndrome) as compared to the other features (points between 10 and 25, angles between 20 and 40, < 20 for areas and coordinates).

Table 6.3: Simultaneous average misclassification error (AME) per syndrome

Syndromes	AME
22q	.200
4p	.583
5p	.500
CDL	.529
fraX	.333
MPS2	1.000
MPS3	.714
Noon	.462
Pro	.400
PWS	.615
Slo	.333
Sot	.333
TCS	.400
WBS	.095

6.3.3 Pairwise classification

Results for pairwise comparisons of syndromic conditions are reported in Table 6.6, which lists AME. For many pairs, such as FraX/22q or FraX/4p, we achieve an AME of 0% . The highest AME was observed when discriminating between MPS2/MPS3, two syndromes with similar facial appearance (38%).

6.3.4 Visualization

Results from the visualization process are depicted in Figure 6.5 and 6.6, for best `glmnet` and PCA model, respectively. For these figures, importance below a threshold is ignored to better show the underlying average image. The same color mapping scheme and scale is used for all sub-figures, making colors comparable. As a comparison, features were also visualized by drawing line segments, points, areas, and small triangles to visualize the importance of distances, coordinates, areas, and angles, respectively. In supplementary images we provide importance plots for the different data components.

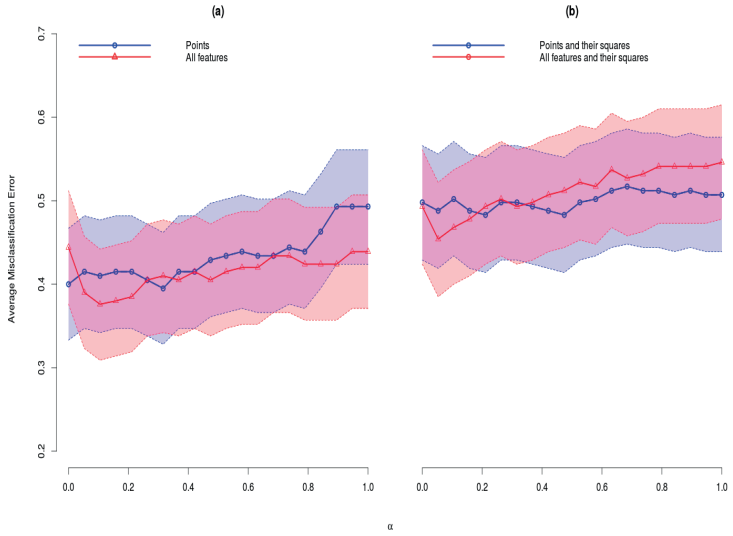


Figure 6.3: Average misclassification error for `glmnet` with 95% confidence intervals across leave-one-out cross-validation for models with different values of mixing parameter α. (a) all features (red) and only points (blue) were used and (b) all features and their squares (red) and only points and their squares (blue) were used.

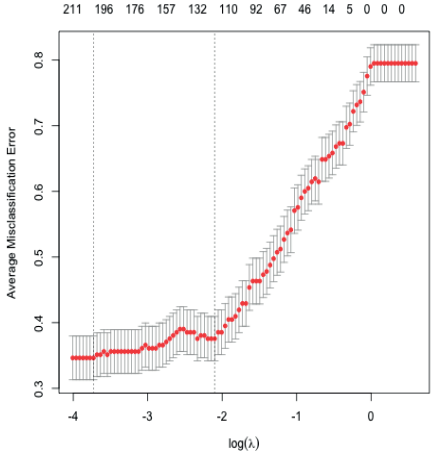


Figure 6.4: Average miclassifcation errors for tuning parameter λ for the L_1-elastic net penalty when $\alpha = .105$.

Table 6.4: Confusion matrix for the best glment model, $\alpha = .105$, using all features. Rows indicate the percentages of predicted syndromes for each of the syndromes in the study.

		Predicted Class												
	22q	4p	5p	CDL	fraX	MPS2	MPS3	Noon	Pro	PWS	Slo	Sot	TCS	WBS	
True Class	22q	.800	.000	.120	.000	.000	0	.000	.000	.0	.000	.000	.040	.000	.040
	4p	.000	.417	.000	.167	.000	0	.000	.167	.0	.000	.000	.167	.000	.083
	5p	.188	.062	.500	.000	.000	0	.000	.000	.0	.000	.000	.062	.000	.188
	CDL	.000	.000	.000	.471	.176	0	.000	.059	.0	.059	.059	.000	.000	.176
	fraX	.000	.000	.000	.111	.667	0	.000	.000	.0	.000	.000	.000	.000	.222
	MPS2	.333	.000	.000	.000	.000	0	.167	.000	.0	.000	.333	.000	.000	.167
	MPS3	.000	.000	.143	.000	.000	0	.286	.000	.0	.000	.000	.143	.000	.429
	Noon	.077	.077	.000	.000	.000	0	.000	.538	.0	.000	.000	.154	.077	.077
	Pro	.200	.000	.000	.000	.000	0	.000	.000	.6	.000	.000	.200	.000	.000
	PWS	.154	.000	.077	.154	.000	0	.000	.000	.0	.385	.000	.000	.000	.231
	Slo	.000	.067	.067	.000	.000	0	.067	.000	.0	.000	.667	.000	.000	.133
	Sot	.133	.067	.000	.000	.000	0	.000	.000	.0	.067	.067	.667	.000	.000
	TCS	.100	.100	.000	.000	.000	0	.000	.100	.0	.000	.000	.000	.600	.100
	WBS	.024	.000	.000	.024	.000	0	.000	.000	.0	.000	.048	.000	.000	.905

All visualizations show distinct patterns of important regions in the face. In general, the central part of the face is included for all syndromes. As an example, progeria is described to exhibit midface hypoplasia and micrognathia (MIM # 17667016) thus featuring a relatively enlarged forehead. Overall importance is focused around the nose whereas the coordinate component shows importance in forehead regions as well as the nose (supplementary Figures S1, S2, and S3), a finding that is discussed below.

6.4 Discussion

Dimension reduction can pose a formidable problem in classification problems if data sets are small. It is well known that methods like PCA can induce big additional variation in data sets thereby reducing classification accuracy. Partly in response to problems like this, penalized regression techniques were developed to estimate classifiers that trade unbiasedness (i.e., parameter estimates that are correct on average) for more stable estimation of classifiers (as measured by the variance of parameter estimates) [Tibshirani, 1996; Hastie et al., 2001]. We have used these ideas in the current study and demonstrate that additional data transformations can even improve classification accuracy. We chose data transformations with low variance as compared to variation of PCs. If these derived features better describe differences between groups, the tradeoff (more variation, more accurate features) can result in a net benefit in terms of classification accuracy, as was the case in this study. As a conclusion, carefully chosen data transformations that increase dimensionality of data sets can improve classification accuracy even if a problem is already high-dimensional. Which transformations to choose is data set specific. As a general rule, each transformation should only depend on few original features (e.g., distances, angles, areas in our case depend on maximally 6 coordinates) in contrast to many (PCA at the other extreme).

Table 6.5: Number of non zero coefficients for each syndrome for the best `glmnet` model ($\alpha = .105$ using all features). t: total , p: points, d: distances, ar: areas and an: angles.

	t	p	d	ar	an
22q	244	27	157	12	46
4p	204	28	138	9	28
5p	243	26	173	15	28
CDL	200	22	120	13	43
fraX	170	14	106	8	40
MPS2	150	12	99	10	28
MPS3	187	17	118	11	40
Noon	197	17	118	15	46
Pro	150	10	105	6	28
PWS	203	20	144	9	28
SLO	235	20	183	8	21
Sot	220	25	153	9	31
TCS	171	16	111	10	33
WBS	257	19	181	17	38
Total	1045	96	778	41	123

Table 6.6: Pairwise average misclassification error rate for the best `glmnet` model.

	22q	4p	5p	CDL	fraX	MPS2	MPS3	Noon	Pro	PWS	SLO	Sot	TCS
4p	.05												
5p	.20	.14											
CDL	.05	.00	.09										
fraX	.03	.00	.04	.15									
MPS2	.10	.11	.18	.04	.00								
MPS3	.09	.11	.22	.00	.06	.38							
Noon	.11	.28	.14	.07	.00	.11	.05						
Pro	.03	.12	.05	.00	.00	.00	.00	.00					
PWS	.16	.04	.24	.27	.14	.11	.10	.04	.00				
SLO	.05	.11	.16	.06	.00	.10	.18	.04	.05	.11			
Sot	.02	.19	.19	.00	.00	.10	.05	.14	.00	.04	.07		
TCS	.06	.18	.12	.04	.00	.12	.00	.13	.00	.04	.04	.04	
WBS	.06	.06	.09	.08	.04	.08	.08	.02	.00	.09	.12	.00	.02

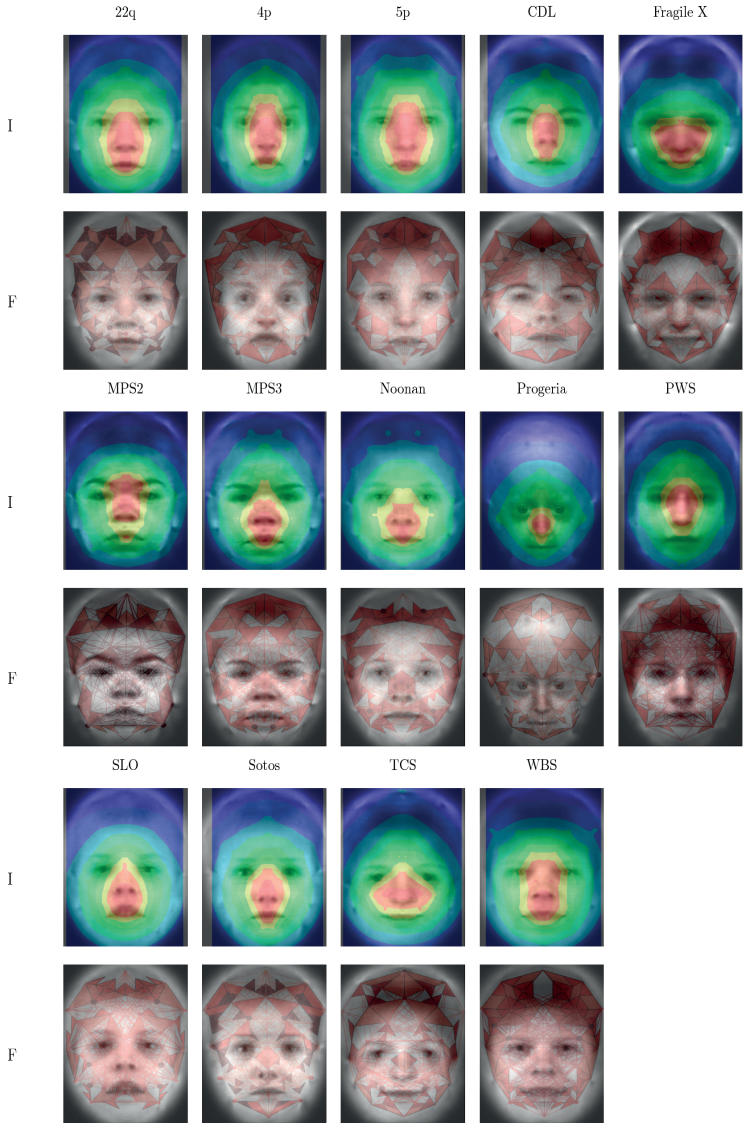


Figure 6.5: Importance plots for `glmnet`. Visualization of simultaneous classification for syndromes. For each syndrome an importance plot (row I) and a plot visualizing classification features (row F) is provided. Importance plot assign an importance with respect to classification to each point as described in the text. Feature plots visualize absolute regression coefficients by thickness of line segments (distances), size of points (coordinates), color of areas (areas; dark red more important than light red) and small triangles (angles; dark red more important than light red).

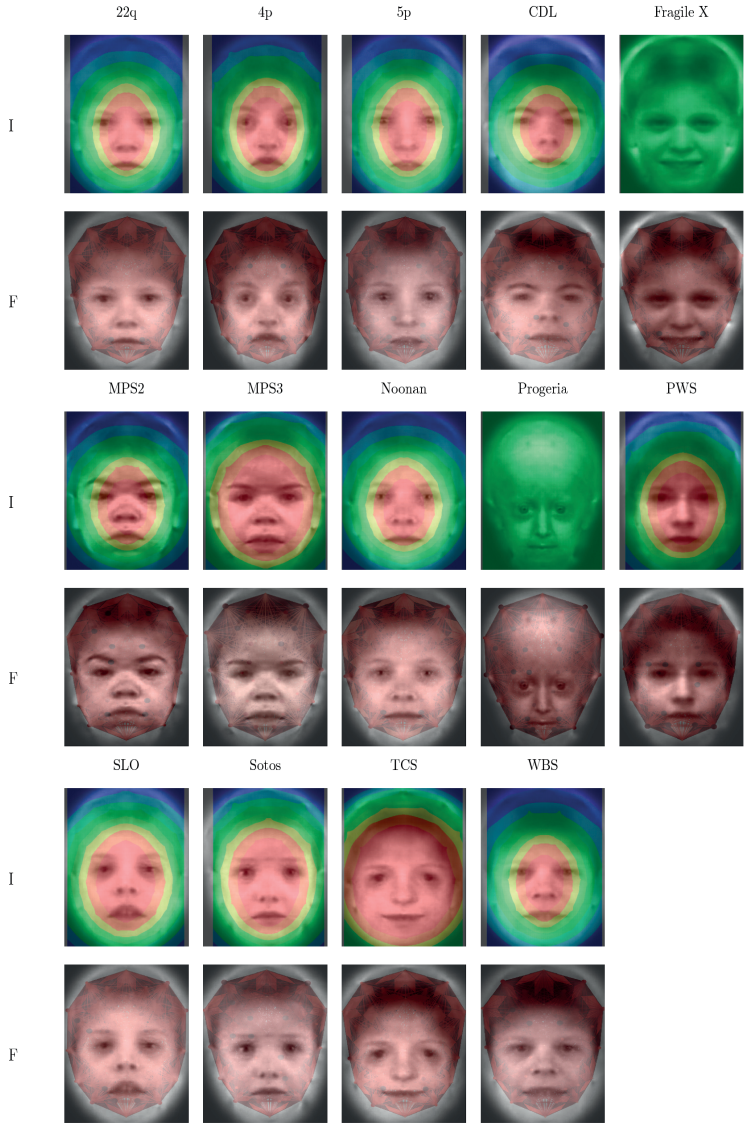


Figure 6.6: Importance plots PCA. Visualizations analogous to Figure 6.5 for PCA based classification.

Pair-wise classification results can be used to get exploratory insights. For example, the pair MPS2/MPS3 has an AME close to 40% implying that the features used in this study do not allow to distinguish this pair of syndromes. In the genetic context, pair-wise classification accuracies can be used as a descriptive measure of phenotypic distinctness.

Our attempt at visualization has the advantage of being generic. As long as a distance of a feature with a point can be defined, we can apply this approach and produce images representing importance of image neighborhoods for the classification decision. At the same time this is a disadvantage as no distinction is made between different types of features and it is impossible to derive such information from our images in general. This shortcoming can be partly addressed by visualizing different data components, which might give important additional information. For example, in the progeria example mentioned above, the nose was visualized as the most important feature in this data set. A narrow nose bridge is a distinguishing feature for progeria in our data set, however, visualizing coordinates alone also indicates that the size of the forehead is a selected feature for this syndrome and would be a more expected feature from the genetic perspective. It is therefore possible to get a better understanding of classifiers by means of such stratified importance plots.

A related problem is that in high-dimensional problems penalized methods have to be selective and choose few features for the final model from the set of all input features. This can well lead to the omission of features that are more easily recognized by human raters. We tried to mitigate this problem by two approaches. First, by using elastic net regression we tried to create less sparse models, thereby retaining more features as compared to a pure LASSO. As a striking example, had we not symmetrized our data, the LASSO would have ignored one of the highly correlated symmetric features whereas elastic net (for an appropriate value of α) would have split the effect almost equally between the two. Second, our means of creating importance plots takes into account the locality of features. If two distances share one vertex, and their vectors are not linearly independent, they are likely to be correlated. Even if one of the distances would be omitted from the model its importance would still be mapped through the correlated distance that shares close proximity.

It follows that the best performing classifier is not necessarily the most intuitive to visualize and we accept that our approach has limitations in overcoming all possible difficulties. Yet, we believe that the visualizations presented here have several merits. First, plausibility of classifiers can be checked. In our case the more variable positions in the hair should be less likely to be important as is the case. Second, these visualizations could be used to refine data pre-processing. In our case we could decide to omit coordinates from the upper rim of the graph altogether, as they do not appear to be important. Third, these visualizations can make it more easy to interpret the actual regression models and can potentially lead to deeper insights for the data expert, in our case the clinical geneticist.

Finally, it is challenging but possible to produce actual caricatures, which would overemphasize images features relevant for the classification decisions. Such caricatures would have to account for the potentially selective nature of the model selection discussed above and presents a computational problem due to the high dimensionality of the feature space ($D = 2088$ in our case). We intend to pursue such an approach.

In conclusion, we have demonstrated the importance of small variance transformations in classification problems of facial data to improve accuracy. Visualization and interpretation remains challenging and can be guided by importance plots that can summarize highly complex classifiers in a single figure or few figures.

Supporting Information

The supplementary material can be found online at `http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0109033#s6`

Bibliography

Ahlbach C, Usatine J, and Pippenger N 2012. A combinatorial interpretation of the joint cumulant. arXiv:12110652 [math] arXiv: 1211.0652.

Ainsworth HF, Unwin J, Jamison DL, and Cordell HJ 2011. Investigation of maternal effects, maternal-fetal interactions and parent-of-origin effects (imprinting), using mothers and their offspring. Genet Epidemiol 35, no. 1:19–45.

Akey J, Jin L, and Xiong M 2001. Haplotypes vs single marker linkage disequilibrium tests: what do we gain? Eur J Hum Genet 9, no. 4:291–300.

Azzouz D, Martin M, Roudier J, and Lambert NC 2011. Could microchimerism be a source of disease-associated HLA alleles in patients with scleroderma? Ann Rheum Dis 70:A34.

Bennett JH 1952. On the theory of random mating. Annals of Eugenics 17, no. 1:311–317.

Boehringer S, Guenther M, Sinigerova S, Wurtz RP, Horsthemke B, and Wieczorek D 2011. Automated syndrome detection in a set of clinical facial photographs. American Journal of Medical Genetics Part A 155, no. 9:2161–2169.

Boehringer S and Pfeiffer RM 2009. A model for fine mapping in family based association studies. Hum Hered 67, no. 4:226–236.

Boehringer S, Vollmar T, Tasse C, Wurtz RP, Gillessen-Kaesbach G, Horsthemke B, and Wieczorek D 2006. Syndrome identification based on 2d analysis software. European Journal of Human Genetics: EJHG 14, no. 10:1082–9.

Breslow NE and Day NE 1980. Statistical methods in cancer research , 1: The analysis of case-control studies. Lyon: International Agency for Research on Cancer .

Breslow NE, Robins JM, and Wellner JA 2000. On the semi-parametric efficiency of logistic regression under case-control sampling. Bernoulli 6:447–455.

Brillinger DR 1991. Some history on the study of higher order moments and spectra. Statistica Sinica 1 1:465–476.

Brown MD, Glazner CG, Zheng C, and Thompson EA 2012. Inferring coancestry in population samples in the presence of linkage disequilibrium. Genetics 190, no. 4:1447–1460.

Browning BL and Browning SR 2011. A fast, powerful method for detecting identity by descent. Am J Hum Genet 88, no. 2:173–182.

Browning SR and Thompson EA 2012. Detecting rare variant associations by identity-by-descent mapping in case-control studies. Genetics 190, no. 4:1521–1531.

Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, Kwiatkowski DP, . . ., and Compston A 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447, no. 7145:661–678.

Calza S and Pawitan Y 2010. Normalization of gene-expression microarray data. Methods Mol Biol pp. 37–52.

Cantor RM, Lange K, and Sinsheimer JS 2010. Prioritizing GWAS results: A review of statistical methods and recommendations for their application. Am J Hum Genet 86, no. 1:6–22.

Cardon LR and Palmer LJ 2003. Population stratification and spurious allelic association. Lancet 361, no. 9357:598–604.

Chatterjee N and Carroll RJ 2005. Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. Biometrika 92:399–418.

Chen J, Lin D, and Hochner H 2012. Semiparametric maximum likelihood methods for analyzing genetic and environmental effects with case-control mother-child pair data. Biometrics pp. 10.1111/j.1541–0420.2011.01728.x.

Chen LS, Hutter CM, Potter JD, Liu Y, Prentice RL, Peters U, and Hsu L 2010. Insights into colon cancer etiology via a regularized approach to gene set analysis of GWAS data. Am J Hum Genet 86, no. 6:860–871.

Chen Y, Zhu J, Lum PY, Yang X, Pinto S, MacNeil DJ, Zhang C, Lamb J, Edwards S, Sieberts SK, Leonardson A, Castellini LW, Wang S, Champy MF, Zhang B, Emilsson V, Doss S, Ghazalpour A, Horvath S, Drake TA, Lusis AJ, and Schadt EE 2008. Variations in DNA elucidate molecular networks that cause disease. Nature 452:429–435.

Clark AG, Weiss KM, Nickerson DA, Taylor SL, Buchanan A, Stengård J, Salomaa V, Vartiainen E, Perola M, Boerwinkle E, and Sing CF 1998. Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. Am J Hum Genet 63, no. 2:595–612.

Clayton D and Jones H 1999. Transmission/disequilibrium tests for extended marker haplotypes. The American Journal of Human Genetics 65, no. 4:1161–1169.

Degli-Esposti MA, Leelayuwat C, and Dawkins RL 1992. Ancestral haplotypes carry haplotypic and haplospecific polymorphisms of BAT1: possible relevance to autoimmune disease. Eur J Immunogenet 19, no. 3:121–127.

Dempster AP, Laird NM, and Rubin DB 1977. Maximum likelihood from incomplete data via the EM algorithm. J R Statist Soc B 39, no. 1:1–38.

Diggle Pj, Liang KY, and Zeger SL 1994. Analysis of longitudinal data. Oxford Science Publications.

Dimas AS, Nica AC, Montgomery SB, Stranger BE, Raj T, Buil A, Giger T, Lappalainen T, Gutierrez-Arcelus M, McCarthy MI, and Dermitzakis ET 2012. Sex-biased genetic effects on gene regulation in humans. Genome Res 22:2368–2375.

Drysdale CM, McGraw DW, Stack CB, Stephens JC, Judson RS, Nandabalan K, Arnold K, Ruano G, and Liggett SB 2000. Complex promoter and coding region beta 2-adrenergic receptor haplotypes alter receptor expression and predict in vivo responsiveness. Proc Natl Acad Sci USA 97, no. 19:10483–10488.

Edgar R, Domrachev M, and Lash AE 2002. Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucl Acids Res 30:207–210.

Epstein MP and Satten GA 2003. Inference on haplotype effects in case-control studies using unphased genotype data. Am J Hum Genet 73, no. 6:1316–1329.

Evangelou E, Trikalinos TA, Salanti G, and Ioannidis JPA 2006. Family-based versus unrelated case-control designs for genetic associations. PLoS Genet 2, no. 8:e123.

Excoffier L and Slatkin M 1995. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 12, no. 5:921–927.

Feitsma AL, Worthington J, van der Helm-van Mil AHM, Plant D, Thomson W, Ursum J, van Schaardenburg D, van der Horst-Bruinsma IE, van Rood J, Huizinga TWJ, Toes REM, and de Vries RRP 2007. Protective effect of noninherited maternal HLA-DR antigens on rheumatoid arthritis development. Proc Natl Acad Sci USA 104:19966–70.

Fisher R 1930. The Genetical Theory of Natural Selection. Oxford University Press.

Fitzmaurice BY, Garrett M, and Laird N 1993. A likelihood-based method for analysing longitudinal binary responses. Biometrika 80:141–51.

Friedman J, Hastie T, and Tibshirani R 2010. Regularization paths for generalized linear models via coordinate descent. Journal of statistical software 33, no. 1:1–22.

Glass D, Viñuela A, Davies MN, Ramasamy A, Parts L, Knowles D, Brown AA, Hedman AK, Small KS, Buil A, Grundberg E, Nica AC, Meglio PD, Nestle FO, Ryten M, the UK Brain Expression consortium, the MuTHER consortium, Durbin R, McCarthy MI, Deloukas P, Dermitzakis ET, Weale ME, Bataille V, and Spector TD 2013. Gene expression changes with age in skin, adipose tissue, blood and brain. Genome Biology 14:R75.

Gower JC 1975. Generalized procrustes analysis. Psychometrika 40, no. 1:33–51.

Günther M 2012. Statistical Gabor graph based techniques for the detection, recognition, classification, and visualization of human faces. Ph.D. thesis, Shaker, Aachen.

Gutierrez-Arcelus M, Lappalainen T, Montgomery SB, Buil A, Ongen H, Yurovsky A, Bryois J, Giger T, Romano L, Planchon A, Falconnet E, Bielser D, Gagnebin M, Padioleau I, Borel C, Letourneau A, Makrythanasis P, Guipponi M, Gehrig C, Antonarakis SE, and Dermitzakis ET 2013. Passive and active DNA methylation and the interplay with genetic variation in gene regulation. eLife Sciences 2:e00523.

Hammond P, Hannes F, Suttie M, Devriendt K, Vermeesch JR, Faravelli F, Forzano F, Parekh S, Williams S, McMullan D, South ST, Carey JC, and Quarrell O 2012. Fine-grained facial phenotype-genotype analysis in wolf-hirschhorn syndrome. European Journal of Human Genetics 20, no. 1:33–40.

Hammond P, Hutton TJ, Allanson JE, Buxton B, Campbell LE, Clayton-Smith J, Donnai D, Karmiloff-Smith A, Metcalfe K, Murphy KC, Patton M, Pober B, Prescott K, Scambler P, Shaw A, Smith ACM, Stevens AF, Temple IK, Hennekam R, and Tassabehji M 2005. Discriminating power of localized three-dimensional facial morphology. The American Journal of Human Genetics 77, no. 6:999–1010.

Hastie T, Tibshirani R, and Friedman J 2001. The elements of statistical learning: data mining, inference, and prediction. Springer, New York, USA.

Hattersley AT and McCarthy MI 2005. What makes a good genetic association study? Lancet 366, no. 9493:1315–1323.

Hay EM, Olliver WFR, and Silman AJ 1993. The arthritis and rheumatism council's national family material repository. Br J Rheumatol 32:443–444.

Heagerty PJ and Kurland BF 2001. Misspecified maximum likelihood estimates and generalised linear mixed models. Biometrika 88:973–985.

Hedrick PW 1987. Gametic disequilibrium measures: Proceed with caution. Genetics 117, no. 2:331–341.

Hennessy RJ, Baldwin PA, Browne DJ, Kinsella A, and Waddington JL 2007. Three-dimensional laser surface imaging and geometric morphometrics resolve frontonasal dysmorphology in schizophrenia. Biological Psychiatry 61, no. 10:1187–1194.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, and Manolio TA 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 106:9362–9367.

Hollox EJ, Poulter M, Zvarik M, Ferak V, Krause A, Jenkins T, Saha N, Kozlov AI, and Swallow DM 2001. Lactase haplotype diversity in the old world. Am J Hum Genet 68, no. 1:160–172.

Horvath S, Zhang Y, Langfelder P, Kahn RS, Boks MP, van Eijk K, van den Berg LH, and Ophoff RA 2012. Aging effects on DNA methylation modules in human brain and blood tissue. Genome Biol 13:R97.

Houwing-Duistermaat JJ, Helmer Q, Balliu B, Akker Evd, Tsonaka R, and Uh HW 2014. Gene analysis for longitudinal family data using random-effects models. BMC Proceedings 8, no. Suppl 1:S88.

Houwing-Duistermaat JJ, van Houwelingen HC, and de Winter J P 2000. Estimation of individual genetic effects from binary observations on relatives applied to a family history of respiratory illnesses and chronic lung disease of newborns. Biometrics 56, no. 3:808–14.

Howden L and Meyer J 2011. Age and sex composition: 2010. US Department of Commerce, Economics and Statistics Administration, US Census Bureau .

Hsieh H, Palmer CGS, Harney S, Chen H, Bauman L, Brown MA, and Sinsheimer JS 2007. Using the maternal-fetal genotype incompatibility test to assess non-inherited maternal HLA-DRB1 antigen coding alleles as rheumatoid arthritis risk factors. BMC Proc I(Suppl I):S124.

Hsieh H, Palmer CGS, Harney S, Newton JL, Wordsworth P, Brown MA, and Sinsheimer JS 2006. The v-MFG test: Investigating maternal, offspring and maternal-fetal genetic incompatibility effects on disease and viability. Genet Epidemiol 30:333–347.

Huang YT, VanderWeele TJ, and Lin X 2014. Joint analysis of SNP and gene expression data in genetic association studies of complex diseases. The Annals of Applied Statistics 8:352–376.

Huijts PEA, Hollestelle A, Balliu B, Houwing-Duistermaat JJ, Meijers CM, Blom JC, Ozturk B, Krol-Warmerdam EMM, Wijnen J, Berns EMJJ, Martens JWM, Seynaeve C, Kiemeney LA, van der Heijden HF, Tollenaar RAEM, Devilee P, and van Asperen CJ 2014. CHEK2*1100delC homozygosity in the netherlands–prevalence and risk of breast and lung cancer. Eur J Hum Genet 22:46–51.

Huynh JL, Garg P, Thin TH, Yoo S, Dutta R, Trapp BD, Haroutunian V, Zhu J, Donovan MJ, Sharp AJ, and Casaccia P 2014. Epigenome-wide differences in pathology-free regions of multiple sclerosis-affected brains. Nat Neurosci 17:121–130.

Ioannidis JPA 2003. Genetic associations: false or true? Trends Mol Med 9, no. 4:135–138.

Jolliffe I 2005. Principal component analysis. Wiley Online Library.

Joosten PH, Toepoel M, Mariman EC, and Van Zoelen EJ 2001. Promoter haplotype combinations of the platelet-derived growth factor alpha-receptor gene predispose to human neural tube defects. Nat Genet 27, no. 2:215–217.

Kenny DA and Judd CM 2013. Power anomalies in testing mediation. Psychological Science 25:334–339.

Koch L 2014. Epigenetics: An epigenetic twist on the missing heritability of complex traits. Nat Rev Genet 15, no. 4:218–218.

Kraft P, Hsieh Hj, Cordell HJ, and Sinsheimer J 2005. A conditional-on-exchangeable-parental-genotypes likelihood that remains unbiased at the causal locus under multiple-affected-sibling ascertainment. Genet Epidemiol 29, no. 1:87–90.

Kraft P and Thomas DC 2000. Bias and efficiency in family-based gene-characterization studies: conditional, prospective, retrospective, and joint likelihoods. Am J Hum Genet 66:1119–31.

Lewontin RC 1964. The interaction of selection and linkage. i. general considerations; heterotic models. Genetics 49, no. 1:49–67.

Li H 2013. Systems biology approaches to epidemiological studies of complex diseases. WIREs Syst Biol Med 5:677–686.

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

Lin DY and Zeng D 2006. Likelihood-based inference on haplotype effects in genetic association studies. Journal of the American Statistical Association 101, no. 473:89–104.

Lin DY and Zeng D 2009. Proper analysis of secondary phenotype data in case-control association studies. Genet Epidemiol 33:256–265.

Liu J, Morgan M, Hutchison K, and Calhoun VD 2010. A study of the influence of sex on genome wide methylation. PLoS ONE 5:e10028.

Liu N, Zhang K, and Zhao H 2008. Haplotype-association analysis. In: Advances in Genetics, Academic Press, volume 60 of *Genetic Dissection of Complex Traits*, pp. 335–405.

Lv J, Liu H, Su J, Wu X, Liu H, Li B, Xiao X, Wang F, Wu Q, and Zhang Y 2012. DiseaseMeth: a human disease methylation database. Nucleic Acids Res 40 (Database issue):D1030–1035.

Mefford J and Witte JS 2012. The covariate's dilemma. PLoS genetics 8:e1003096.

Meinshausen N 2008. Hierarchical testing of variable importance. Biometrika 95, no. 2:265–278.

Molinaro AM, Simon R, and Pfeiffer RM 2005. Prediction error estimation: a comparison of resampling methods. Bioinformatics 21, no. 15:3301–3307.

Morris RW and Kaplan NL 2002. On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. Genet Epidemiol 23, no. 3:221–233.

Mukherjee B and Chatterjee N 2008. Exploiting gene-environment independence for analysis of case-control studies: an empirical bayes-type shrinkage estimator to trade-off between bias and efficiency. Biometrics 64:685–694.

Nelson JL, Furst DE, Maloney S, Gooley T, Evans PC, Smith A, Bean MA, Ober C, and Bianchi D 1998. Microchimerism and HLA-compatible relationships of pregnancy in scleroderma. Lancet 351:559–62.

Nemes S, Jonasson JM, Genell A, and Steineck G 2009. Bias in odds ratios by logistic regression modelling and sample size. BMC Med Res Methodol 9:56.

Neuhaus JM 1998. Estimation efficiency with omitted covariates in generalized linear models. Journal of the American Statistical Association 93:1124–1129.

Neuhaus JM and Jewell NP 1993. A geometric approach to assess bias due to omitted covariates in generalized linear models. Biometrika 80:807–815.

Pfeiffer RM, Hildesheim A, Gail MH, Pee D, Chen CJ, Goldstein AM, and Diehl SR 2003. Robustness of inference on measured covariates to misspecification of genetic random effects in family studies. Genet Epidemiol 24:14–23.

Pfeiffer RM, Pee D, and Landi MT 2008. On combining family and case-control studies. Genet Epidemiol 32:638–46.

Pirinen M, Donnelly P, and Spencer CCA 2012. Including known covariates can reduce power to detect genetic effects in case-control studies. Nature Genetics 44:848–851.

Prentice RL and Pyke R 1979. Logistic disease incidence models and case-control studies. Biometrika 66:403–411.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, and Reich D 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38, no. 8:904–909.

Purdom E and Holmes SP 2005. Error distribution for gene expression data. Statistical applications in genetics and molecular biology 4:1544–6115.

R Core Team 2014. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rabinowitz D 1997. A note on efficient estimation from case-control data. Biometrika 84:486–488.

Richardson B 2003. Impact of aging on DNA methylation. Ageing Research Reviews 2:245–261.

Risch N and Merikangas K 1996. The future of genetic studies of complex human diseases. Science 273, no. 5281:1516–1517.

Robinson LD and Jewell NP 1991. Some surprising results about covariate adjustment in logistic regression models. International Statistical Review / Revue Internationale de Statistique 59:227–240.

Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, Ruff TG, Milligan SB, Lamb JR, Cavet G, Linsley PS, Mao M, Stoughton RB, and Friend SH 2003. Genetics of gene expression surveyed in maize, mouse and man. Nature 422:297–302.

Schaid DJ 2004. Evaluating associations of haplotypes with traits. Genet Epidemiol 27, no. 4:348–364.

Schaid DJ, Rowland CM, Tines DE, Jacobson RM, and Poland GA 2002. Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am J Hum Genet 70, no. 2:425–434.

Scott AJ and Wild CJ 2001. Maximum likelihood for generalised case-control studies. Journal of Statistical Planning and Inference 96:3–27.

Silman AJ and Hochberg MC, editors 2001. Epidemiology of the Rheumatic Diseases. Oxford ; New York: Oxford University Press, 2 edition edition.

Silman AJ, MacGregor AJ, Thomson W, Holligan S, Carthy D, Farhan A, and Ollier WER 1993. Twin concordance rates for rheumatoid arthritis: results from a nationwide study. Br J Rheumatol 32:903–7.

Sinnwell J and Schaid D 2013. haplo.stats: Statistical Analysis of Haplotypes with Traits and Covariates when Linkage Phase is Ambiguous. R package version 1.6.8.

Sinsheimer JS, Palmer CGS, and Woodward JA 2003. Detecting genotype combinations that increase risk for disease: The maternal-fetal genotype incompatibility test. Genet Epidemiol 24:1–13.

Smits JMA, Claas FHJ, van Houwelingen HC, and Persijn GG 1998. Do noninherited maternal antigens (NIMA) enhance renal graft survival? Transpl Int 11:82–88.

Soler JM and Blangero J 2003. Longitudinal familial analysis of blood pressure involving parametric (co)variance functions. BMC Genetics 4, no. Suppl 1:S87.

Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, Jackson AU, Allen HL, ..., and Loos RJF 2010. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nat Genet 42, no. 11:937–948.

Spinka C, Carroll RJ, and Chatterjee N 2005. Analysis of case-control studies of genetic and environmental factors with missing genetic information and haplotype-phase ambiguity. Genet Epidemiol 29, no. 2:108–127.

Sun W 2012. A statistical framework for eQTL mapping using RNA-seq data. Biometrics 68:1–11.

Tavtigian SV, Simard J, Teng DH, Abtin V, Baumgard M, Beck A, Camp NJ, Carillo AR, Chen Y, Dayananth P, Desrochers M, Dumont M, Farnham JM, Frank D, Frye C, Ghaffari S, Gupte JS, Hu R, Iliev D, Janecki T, Kort EN, Laity KE, Leavitt A, Leblanc G, McArthur-Morrison J, Pederson A, Penn B, Peterson KT, Reid JE, Richards S, Schroeder M, Smith R, Snyder SC, Swedlund B, Swensen J, Thomas A, Tranchant M, Woodland AM, Labrie F, Skolnick MH, Neuhausen S, Rommens J, and Cannon-Albright LA 2001. A candidate prostate cancer susceptibility gene at chromosome 17p. Nat Genet 27, no. 2:172–180.

Thiele TN 1899. Om iagttagelseslaere halvinvarianter. Overs Vid Sels Forh pp. 135–141.

Thompson EA 2008. The IBD process along four chromosomes. Theor Popul Biol 73, no. 3:369–373.

Tibshirani R 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B (Methodological) 58, no. 1:267–288.

Tsonaka R, De Visser MCH, and Houwing-Duistermaat J 2013. Estimation of genetic effects in multiple cases family studies using penalized maximum likelihood methodology. Biostatistics 14:220–231.

Tsonaka R, van der Helm-van Mil AHM, and Houwing-Duistermaat JJ 2012. A two-stage mixed-effects model approach for gene-set analyses in candidate gene studies. Statist Med 31, no. 11-12:1190–1202.

Umbach DM and Weinberg CR 1997. Designing and analysing case-control studies to exploit independence of genotype and exposure. Stat Med 16:1731–1743.

van der Woude D, Houwing-Duistermaat JJ, Toes REM, Huizinga TWJ, Thomson W, Worthington J, van der Helm-van Mil AHM, and de Vries RRP 2009. Quantitative heritability of anti-citrullinated protein antibody-positive and anti-citrullinated protein antibody-negative rheumatoid arthritis. Arthritis Rheum 60:916–23.

van der Woude D, Lie BA, Lundström E, Balsa A, Feitsma AL, Houwing-Duistermaat JJ, Verduijn W, Nordang GBN, Alfredsson L, Klareskog L, Pascual-Salcedo D, Gonzalez-Gay MA, Lopez-Nevot MA, Valero F, Roep BO, Huizinga TWJ, Kvien TK, Martín J, Padyukov L, de Vries RRP, and Toes REM 2010. Protection against anti-citrullinated protein antibody-positive rheumatoid arthritis is predominantly associated with HLA-DRB1*1301: a meta-analysis of HLA-DRB1 associations with anti-citrullinated protein antibody-positive and anti-citrullinated protein . Arthritis and rheumatism 62:1236–45.

Vollmar T, Maus B, Wurtz RP, Gillessen-Kaesbach G, Horsthemke B, Wieczorek D, and Boehringer S 2008. Impact of geometry and viewing angle on classification accuracy of 2d based analysis of dysmorphic faces. European Journal of Medical Genetics 51, no. 1:44–53.

Wang WYS, Barratt BJ, Clayton DG, and Todd JA 2005. Genome-wide association studies: theoretical and practical concerns. Nat Rev Genet 6, no. 2:109–118.

Weinberg, C R 1999. Methods for detection of parent-of-origin effects in genetic studies of case-parents triads. Am J Hum Genet 65:229–35.

Weir BS 1990. Genetic Data Analysis: Methods for Discrete Population Genetic Data. Sinauer Associates Incorporated.

Winter RM 1996. What's in a face? Nature Genetics 12, no. 2:124–129.

Witte JS 2010. Genome-wide association studies and beyond. Annu Rev Public Health 31:9–20.

World Health Organization 2008. Atlas: Multiple sclerosis resources in the world. WHO - Atlas Project Neurology .

Worthington J, Olliver WFR, Leach MK, Smith I, Hay EM, Thomson W, Pepper L, Carthy D, Farhan A, Martin S, Dyer P, Davison J, Bamber S, and Silman AJ 1994. Research practice the arthritis and rheumatism council's national repository of family material: Pedigrees from the first 100 rheumatoid arthritis families containing affected sibling pairs. Br J Rheumatol 33:970–976.

Xing G and Xing C 2010. Adjusting for covariates in logistic regression models. Genet Epidemiol 34:769–771.

Yousefi P, Huen K, Schall RA, Decker A, Elboudwarej E, Quach H, Barcellos L, and Holland N 2013. Considerations for normalization of DNA methylation data by illumina 450K BeadChip assay in population studies. Epigenetics 8:1141–1152.

Zaitlen N, Lindström S, Pasaniuc B, Cornelis M, Genovese G, Pollack S, Barton A, Bickeböller H, Bowden DW, Eyre S, Freedman BI, Friedman DJ, Field JK, Groop L, Haugen A, Heinrich J, Henderson BE, Hicks PJ, Hocking LJ, Kolonel LN, Landi MT, Langefeld CD, Le Marchand L, Meister M, Morgan AW, Raji OY, Risch A, Rosenberger A, Scherf D, Steer S, Walshaw M, Waters KM, Wilson AG, Wordsworth P, Zienolddiny S, Tchetgen ET, Haiman C, Hunter DJ, Plenge RM, Worthington J, Christiani DC, Schaumberg DA, Chasman DI, Altshuler D, Voight B, Kraft P, Patterson N, and Price AL 2012a. Informed conditioning on clinical covariates increases power in case-control association studies. PLoS Genet 8:e1003032.

Zaitlen N, Pasaniuc B, Patterson N, Pollack S, Voight B, Groop L, Altshuler D, Henderson BE, Kolonel LN, Marchand LL, Waters K, Haiman CA, Stranger BE, Dermitzakis ET, Kraft P, and Price AL 2012b. Analysis of case-control association studies with known risk variants. Bioinformatics 28:1729–1737.

Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ, and Ehm MG 2002. Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. Hum Hered 53, no. 2:79–91.

Zhang D, Cheng L, Badner JA, Chen C, Chen Q, Luo W, Craig DW, Redman M, Gershon ES, and Liu C 2010. Genetic control of individual differences in gene-specific methylation in human brain. The American Journal of Human Genetics 86:411–419.

Zhao S, Cai T, and Li H 2014. More powerful genetic association testing via a new statistical framework for integretive genomics. To appear in Biometrics -.

Zheng Y, Heagerty PJ, Hsu L, and Newcomb P 2010. On combining family-based and population-based case-control data in association studies. Biometrics 66:1024–33.

English Summary

This dissertation describes new statistical methods designed to improve the power of genetic association studies. Of particular interest are studies with a response-selective sampling design, i.e. case-control studies of unrelated individuals and case-control studies of family members. In the pages that follow, we detail novel statistical methods that (a) take advantage of information available in the distribution of the covariates in case-control studies by modeling the ascertainment process; (b) incorporate information from both family-based studies and case-control studies of unrelated individuals; (c) use "richer" models of the relationship between genetic variants and phenotypes, compared to models used in standard genetic association studies; and (d) integrate different types of data, such as genomic, epigenomic, transcriptomic and environmental information. Together, these methods will improve the ability of the genetics community to identify the genetic basis of complex human phenotypes.

Chapter 1 provides a general introduction to existing methods for the statistical analysis of genetic association studies with response-selective sampling designs. We start by introducing the relevant terminology and the key concepts of genetic association studies. Next, we present and compare the two most popular response-selective sampling designs in genetic studies: case-control studies of unrelated individuals and case-control studies of family members. We proceed to explain the two main advantages of accounting for ascertainment in such studies: the potential increase in power to detect associations and proper secondary phenotype analysis.

The rest of the introduction is split in two parts. In the first part, we present three different likelihood approaches for modelling the ascertainment in family-based case-control studies. The first approach is based on the prospective likelihood, which models the distribution of phenotypes conditional on covariates and ascertainment. The second approach is the ascertainment-corrected joint likelihood, which models the joint distribution of phenotypes and covariates conditional on ascertainment. The last approach is the retrospective likelihood, which models the distribution of covariates conditional on phenotypes. The latter is also appropriate for the analysis of case-control data of unrelated individuals. The likelihoods are compared in terms of efficiency of parameter estimates and computational efficiency.

In the second part of the introduction we describe different models for the relation between the genetic variants and the phenotype. The current standard analysis protocol for genome wide association studies is to individually evaluate the relationship between each SNP and disease. However, most common complex diseases do not arise from a single genetic cause, but rather a combination of multiple genetic and environmental factors (i.e., they are polygenic). Here, we present alternative approaches, which more closely model the underlying biological mechanisms, such as jointly modeling multiple genetic variants, or jointly modeling genetic variants and intermediate cellular phenotypes.

Chapter 2 describes a novel method to improve the power of genetic association studies by combining data from multi-case family studies and twin studies and modeling the ascertainment process of such studies. In order to maximize efficiency in parameter estimation, inference about the parameters of interest is based on an ascertainment-corrected joint like-

lihood. To take into account the correlation of disease risks among family members, due to shared but unmeasured genetic or environmental factors, a family-specific random term is used.

Simulations and real data analysis show that this ascertainment-corrected joint likelihood combining family and twin data is more efficient for estimating the parameters of interest, as compared to a families-only ascertainment-corrected joint likelihood approach or a prospective likelihood approach which ignores the ascertainment. The combined approach, not only enhances the statistical power to detect direct offspring allelic effects, but also effects depending on maternal-offspring genotype combinations, such as non-inherited maternal antigen effects. The efficiency improvement of the joint likelihood over the prospective likelihood is higher when information is limited, i.e. when the families are small (three offspring per family) and ascertained such that at least two out of three offspring are affected. The efficiency improvement of the combined families-and-twins approach against the families-only approach is noticeably high when the sample size is small, i.e. the number of families in the study is 100 or less.

Chapter 3 considers an alternative haplotype based strategy to the current gold standard of marginal testing. Marginal tests based on individual SNPs have dominated association analyses in the past decade. Although single SNP analyses have led to the identification of hundreds of genetic variants associated with many complex diseases, greater power might be gained by using haplotype-based approaches to analyze multiple markers simultaneously. Haplotype-based association methods incorporate linkage disequilibrium (LD) information from multiple markers and can be more powerful for gene mapping than methods based on single SNPs. A limitation of haplotype-based methods is that the number of parameters increases exponentially with the number of SNPs, inducing a commensurate increase in the degrees of freedom and weakening the power to detect associations.

Here we consider a hierarchical linkage disequilibrium model for trait mapping that enables flexible testing strategies over a range of hypotheses: from single SNP analyses through the haplotype distribution tests. Many such models reduce d.f. and increase the power to detect associations. These models are based on a re-parametrization of the multinomial haplotype distribution, where every parameter corresponds to the joint cumulant of each possible subset of a set of loci. Extensive simulations and a real data analysis show that such tests, which make plausible restrictions on the parameter space, have often increased power against the unrestricted global haplotype test for association or the single-SNP tests.

Genetic studies aim to assess the association between genetic variants and common complex traits. For the analysis of such traits, two different methods can be used: linkage mapping and association mapping. In Chapter 4, we consider the trade-offs between these two methods. Linkage mapping methods are more powerful for identifying rare variants with large effect on disease susceptibility while association-mapping methods are more suitable for identifying more common variants with moderate effect sizes. However, SNPs typically have small effect sizes (common variants) or minor allele frequencies that are too small to reliably fit models (rare variants). If the rare variant effects were large, and the disease was not heterogeneous, they would have been found through previous family-based linkage studies. Thus, there may be a middle ground in which multiple rare variants of moderate to low effect size play a key role in the etiology of some diseases. Such situations might be ideal for combining linkage- and association-mapping.

We develop a two-part analysis in order to investigate the contribution that linkage-based methods, such as IBD mapping, can make to association mapping to identify rare variants in next-generation sequencing data. In the first part we use identity-by-descent (IBD) mapping to identify regions in which cases share more segments of IBD around a putative causal variant than do controls. In the second part we perform association-mapping

by using a two-stage mixed-effects model approach to summarize the SNP data within the regions identified in the first part and including them as covariates in the model for the phenotype. To increase our power to identify rare variants, we also include the number of rare variants per region as a covariate in the model. The method was applied to next-generation sequencing longitudinal family data from Genetic Association Workshop 18 and a significant association was identified.

Chapter 5 examines integrative omics, the joint analysis of outcome and multiple types of omics data, such as genomics, epigenomics and transcriptomics data. Integrative omics has emerged as a promising approach for powerful and biologically relevant association studies. These studies often employ a case-control design, and often include non-omics covariates, such as age and gender, that may modify the underlying omics risk factors. An open question is how to best integrate multiple omics and non-omics information to maximize statistical power in case-control studies that ascertain individuals based on the phenotype. Recent works on integrative omics have used prospective approaches, modeling case-control status conditional on omics and non-omics risk factors. Compared to univariate approaches, jointly analyzing multiple risk factors with a prospective approach increases power in non-ascertained cohorts. However, in case-control studies this is no longer the case and these prospective approaches often lose power compared to univariate approaches.

We present a novel statistical method for integrating multiple omics and non-omics factors that addresses these issues of power loss in case-control association studies. This method is based on a retrospective likelihood function that models the joint distribution of omics and non-omics factors conditional on case-control status. In order to model the distribution of the risk factors as efficiently as possible, knowledge about the correlation structure between risk factors in the population is exploited and parametric assumptions about the distribution of the risk factors are made. The new method provides accurate control of Type I error rate and has increased efficiency over prospective approaches in both simulated and real data. Efficiency gain is a function of the number of parameters used to model the distribution of the risk factors and the effect sizes of risk factors, with increased efficiency gain for continuous factors and for risk factors with large effect sizes.

Chapter 6 considers the problem of phenotype description. Sometimes an outcome is based on rating of multiple underlying features and might thereby be prone to inter-rater variability. In these cases the outcome definition can be made objective by learning a predictor for the outcome based on the underlying multivariate data. Potentially this can improve power of ensuing studies and improve the understanding of the outcome variable.

Here, we consider genetic syndromes as such a phenotype and 2D graph-data derived from facial images as features. We present a method for automated syndrome classification and visualization of the classifier. In order to optimize the classifier, we investigate a set of data transformations prior to analysis and their effect on classification accuracy in a high-dimensional setting. These transformations are low-variance in the sense that each involves only a fixed small number of input features. It is shown that classification accuracy can be improved when penalized regression techniques are employed, as compared to a principal component analysis pre-processing step.

A second goal is to visualize the resulting classifiers. We develop importance plots highlighting the influence of coordinates in the original 2D space. Features used for classification are mapped to coordinates in the original images and combined into an importance measure for each pixel. These plots assist in assessing plausibility of classifiers, interpretation of classifiers, and determination of the relative importance of different features.

Nederlandse Samenvatting

Dit proefschrift behandelt nieuwe statistische methoden, die ontwikkeld zijn om de statistische power in genetische associatiestudies te verbeteren. De focus ligt op epidemiologische studies met een *response-selective sampling design*, zoals *case-control* studies met niet-verwante individuen en *case-control* studies met families. In deze samenvatting beschrijven we in detail nieuwe statistische methoden die (a) profiteren van de beschikbare informatie in de verdeling van de covariabelen in *case-control* studies door het *ascertainment* proces te modelleren; (b) informatie van familie-gebaseerde en *case-control* studies met niet-verwante individuen combineren; (c) gebruik maken van uitgebreidere modellen voor het beschrijven van de relatie tussen genetische varianten en fenotypen in standaard genetische associatiestudies; en (d) verschillende soorten data, zoals genomische, epigenomische, transcriptomische informatie integreren. Deze viertal punten kunnen samen de power verbeteren om de genetische basis van complexe menselijke eigenschappen te achterhalen.

Hoofdstuk 1 geeft een algemene inleiding op de bestaande methoden voor de statistische analyse van genetische associatiestudies met *response-selective sampling designs*. We introduceren de relevante terminologie en de belangrijkste concepten binnen genetische associatiestudies. Daarna vergelijken we de twee meest populaire *response-selective sampling designs* in genetische studies, namelijk *case-control* studies met niet-verwante individuen en die met families. De voordelen van methoden die rekening houden met *ascertainment* zijn de potentiële toename in statistische power voor het detecteren van associaties en het uitvoeren van een secundaire fenotype analyse.

De rest van de introductie is opgesplitst in twee delen. Het eerste deel laat drie verschillende likelihoods zien voor het modelleren van *ascertainment* in *case-control* studies met familiedata. De eerste is de *prospective likelihood*, waarin de verdeling van de uitkomst conditioneel op de covariabelen en de *ascertainment* wordt gemodelleerd. De tweede is de *ascertainment* gecorrigeerde *joint likelihood* die de gezamenlijke (joint) verdeling van de uitkomst en de covariabelen modelleert conditioneel op de *ascertainment*. De laatste is de *retrospective likelihood*. Deze modelleert de verdeling van de covariabelen gegeven de uitkomst. De *retrospective likelihood* is ook geschikt voor het analyseren van *case-control* data met niet-verwante individuen. We vergelijken de drie likelihoods met betrekking tot efficiëntie van de parameterschatters en de computationele kosten.

Het tweede gedeelte van de introductie beschrijft verschillende modellen voor de relatie tussen genetische varianten en de uitkomst. De huidige standaard voor genoombrede analyse is om voor elke *SNP* apart de relatie met de uitkomst te evalueren. Dit terwijl complexe ziekten meestal niet één enkele genetische oorzaak hebben, maar het gevolg zijn van een combinatie van meerdere genetische en omgevingsfactoren (bijv. polygenetisch). In dit gedeelte van de introductie presenteren we alternatieve methoden, die beter de onderliggende biologische mechanismen modelleren door het effect van meerdere genetische varianten of het effect van genetische en intermediare cellulaire fenotypen mee te nemen.

Hoofdstuk 2 beschrijft een nieuwe methode die de power van genetische associatie studies verbetert door data uit *multi-case* familie- en tweelingen studies met elkaar te combineren. Hierbij wordt ook het proces van *ascertainment* gemodelleerd. Om de efficiëntie van de parameterschatters te verhogen, gebruiken we de *ascertainment* gecorrigeerde *joint*

likelihood. Door gebruik te maken van een familie-specifiek random effect houden we rekening met de correlatiestructuur binnen families die veroorzaakt wordt door ongemeten genetische of omgevingsfactoren.

Met behulp van simulaties en echte data-analyse laten we zien dat belangrijke parameters efficiënter worden geschat door gebruik te maken van de *ascertainment* gecorrigeerde *joint likelihood*, waarin de familie en tweelingen data worden gecombineerd. Deze methode is efficienter dan de *ascertainment* gecorrigeerde *joint likelihood* met alleen familie data en de *prospective likelihood* waarin het *ascertainment* proces niet meegenomen wordt. De gecombineerde aanpak heeft niet alleen meer statistische power voor het vinden van effecten van individuele genotypen, maar ook het effect van het genotype van de moeder op de uitkomst, bijv. niet-overerfbare maternale antigen effecten. Deze verbetering in efficiëntie van de *joint likelihood* ten opzichte van de *prospective likelihood* is groter wanneer er minder informatie is. Bijvoorbeeld voor datasets met kleine families (3 kinderen per gezin) met tenminste twee aangedane kinderen. We zien vooral een verbetering van de efficiëntie bij een kleine steekproef van 100 of minder families voor de *joint likelihood* waarin families en tweelingen gecombineerd worden ten opzichte van de *joint likelihood* met alleen de familiedata.

Hoofdstuk 3 beschouwt een alternatieve strategie voor het testen van *haplotypes* in vergelijking met de huidige gouden standaard, namelijk marginale testen. Deze *single SNPs* testen zijn het afgelopen decennium het meest gebruikt. Alhoewel deze *single SNPs analyses* voor vele ziekten geleid hebben tot het identificeren van honderden geassocieerde genetische varianten, zou meer statistische power verkregen kunnen worden wanneer er gebruik gemaakt wordt van op haplotype gebaseerde statistische methoden. Deze methodoen analyseren namelijk meerdere genetische markers tegelijkertijd door gebruik te maken van *linkage disequilibrium* (LD) informatie. Hierdoor kan de power verbeteren voor het vinden van genetische varianten voor een bepaalde eigenschap (ziekte). Een nadeel van deze haplotype-gebaseerde statistische methoden is dat het aantal parameters exponentieel toeneemt met het aantal *SNPs*. Dit gaat samen met een overeenkomstige toename van het aantal vrijheidsgraden wat tot een afname in power om associaties te detecteren kan leiden.

Wij introduceren een hiërarchisch *linkage disequilibrium model* dat flexibele teststrategieën geeft voor het vinden van genetische varianten van eigenschappen over een serie van statistische hypothesen: van standaard *single SNP* analyses tot en met testen van associatie met volledige haplotypeverdelingen. Voor veel van deze hiërarchisch *linkage disequilibrium modellen* blijft het aantal vrijheidsgraden relatief laag, en daarmee is de power voor het detecteren van associaties dan ook beter. Het model is gebaseerd op een *reparametrisering* van de multinomiale haplotype verdeling waarin iedere parameter overeenkomt met een *joint cumulant* van elke mogelijke deelverzamelingen van *loci*. Een uitgebreide simulatiestudie en echte data-analyses laten zien dat testen binnen het hiërarchisch *linkage disequilibrium model* vaak een hogere statistische power hebben dan de *global haplotype test* en de *single SNP* associetietesten.

Genetische associatie studies hebben tot doel de associatie tussen genetische varianten en complexe genetische eigenschappen te detecteren. Voor de analyse van deze eigenschappen, kunnen twee verschillende methoden gebruikt worden: *linkage mapping* en *association mapping*. In hoofdstuk 4 bestuderen we de eigenschappen van deze twee methoden. *Linkage mapping* methoden zijn krachtiger voor het identificeren van zeldzame varianten die een effect hebben op vatbaarheid voor ziekte terwijl *association mapping* meer geschikt is voor het detecteren van algemeen voorkomende varianten met matige effect groottes. Echter, genetische varianten komen of vaak voor en hebben een klein effect op de uitkomts, of zijn te zeldzaam om hen effect op een betrouwbare wijze te schatten. Wanneer de effecten van de zeldzame variant groot waren geweest, en het fenotype niet heterogeen, dan kunnen deze varianten gedetecteerd worden met bijvoorbeeld *linkage* studies gebaseerd op familiedata.

Oftewel, er zou een methode moeten zijn waarbij meerdere zeldzame varianten met matige tot kleine effecten of de uitkomst moeten zijn. Een dergelijk uitgangspunt zou ideaal zijn voor het combineren van *linkage*- en *association mapping*.

We hebben een twee staps methode ontwikkeld om te onderzoeken of linkage gebaseerde methoden, zoals *identity by descent (IBD) mapping*, een bijdrage kunnen leveren aan het detecteren van zeldzame varianten via associatie in (*next-generation sequencing data*). In de eerste stap passen we IBD mapping toe om regio's te vinden waar *cases* meer segmenten IBD met elkaar delen dan *controls* rondom een vermeende causale variant. In de tweede stap doen we een associatie analyse met behulp van een *two stage mixed-effect model*. Met dit model kunnen we een overzicht creëren van de *SNP* data binnen de gevonden regio's en vervolgens nemen we deze *SNPs* mee als covariabelen in het model voor de uitkomst. Om de power te verbeteren, nemen we ook een variabele op die het aantal zeldzame varianten per regio telt. Met deze methode hebben we een significante associatie gevonden in de *next generation sequencing* longitudinale familiedata van de *Genetic Association Workshop 18*.

Hoofdstuk 5 bestudeert de analyse van de uitkomst variabele met meerdere soorten *omics* data, zoals genomische, epigenomische, en transcriptomische data (*integrative omics*). De *integrative omics* methode heeft zich ontpopt tot een krachtige en biologisch relevante richting van onderzoek voor associatiestudies. In *integrative omics* methoden maakt men vaak gebruik van het *case-control* design. Ook is het van belang om het effect van andere risicofactoren en covariabelen op de *omics* data te modeleren. Een open vraag is hoe je het beste meerdere *omics* datasets en deze risicofactoren en covariabelen het best kan integreren. Recente studies van *integrative omics* data maken gebruik van een *prospective model* waarin de *case-control* status conditioneel op de *omics* en de risciofactoren gemodelleerd wordt. In vergelijking met de univariate modellen heeft het analyseren van de meerdere risciofactoren in een prospectief model meer statistische power wanneer de individuen niet geselecteerd zijn. Echter, in *case-control studies* is dit niet het geval, daarom is de *power* vaak minder in vergelijking met de univariate aanpak.

Wij presenteren een nieuwe statistische methode voor *case-control* associatiestudies die het verlies in power kunnen opvangen en ook de meerdere *omics* en niet-*omics* factoren kunnen modelleren. Deze methode is gebaseerd op een *retrospective likelihood* functie waarin, conditioneel op de *case-control* status, de gezamenlijke verdeling van *omics* en risicofactoren gemodelleerd wordt. Om de verdeling van de risciofactoren efficiënt te kunnen modelleren, benutten we kennis over de correlatiestructuur tussen de risciofactoren in de populatie, en maken we gebruik van parametrische aannamen over de verdeling van de risciofactoren.

Uit simulatiestudies blijkt dat deze nieuwe statistische methode voldoet met betrekking tot de Type I fout en meer efficiëntie heeft, ten opzichte van de prospectieve benadering. De winst in efficiëntie hangt af van het aantal parameters in het model en de effectgrootten van de risciofactoren. Deze winst is groter wanneer de risciofactoren continu zijn en wanneer ze elk groot effect hebben.

Hoofdstuk 6 beschouwt het probleem van het correct beschrijven van een fenotype. Voor bepaalde genetische aandoeningen is het fenotype slecht gedefinieerd. Heterogeniteit in fenotypes leidt tot een lage power voor het detecteren van genetische associaties. Daarnaast zijn gevonden significante associaties vaak moeilijk te interpreteren. Het doel van de fenotypische classificatiemethode is het verfijnen van de classificatie van het fenotype met behulp van een, vaak hoogdimensionele, verzameling aan *features*. We bestuderen genetische syndromen als fenotypes en de pixels van tweedimensionale afbeeldingen van gezichten als *features*. We presenteren een methode voor geautomatiseerde classificatie en visualisatie van dit soort data.

Wanneer de data eerst getransformeerd wordt kan een betere classificatie verkregen worden. We onderzoeken het effect van verschillende transformaties op de nauwkeurigheid

van de classificatie in een hoog-dimensionele ruimte van *features*. Deze transformaties hebben betrekking tot een klein aantal *features*. Wanneer geregulariseerde regressietechnieken toegepast wordt op deze features, kan de classificatie naukeuriger zijn dan een principale componenten analyse.

Een tweede doel is het visualiseren van de classificatiefactoren die we gevonden hebben. We hebben *importance* plots ontwikkeld die de invloed van coördinaten in het originele tweedimensionale afbeelding weergeven. *Features* die gebruikt worden in de classificatie worden toegewezen aan coördinaten in het oorspronkelijke beeld en samengebracht tot een maat van *importance* voor elke pixel. Deze plots dienen als hulp bij het beoordelen van de plausibiliteit en de interpretatie van de classificaties en het bepalen van de relevante *features*.

List of Publications

Balliu B, Tsonaka R, Boehringer S and Houwing-Duistermaat JJ (2015). " A Retrospective Likelihood Approach for Efficient Integration of Multiple Omics Factors in Case-Control Association Studies". *Genetic Epidemiology. Advance online publication. DOI: 10.1002/gepi.21884.*

Tissier R, **Balliu B**, Tsonaka R, Uh HW, Houwing-Duistermaat JJ (2015). "Impact of the Family Structure in Weighted Correlation Network Analysis for Gene-Expression Studies". *BMC Proceedings. In Press.*

Balliu B and Zaitlen N (2015). "Leveraging Family Trios To Remove Biases And Increase Power In Tests Of Epistatic Interaction". *Manuscript submitted for publication.*

Balliu B and Boehringer S (2015). "Powerful Testing via Hierarchical Linkage Disequilibrium in Haplotype Association Studies". *Manuscript submitted for publication.*

Balliu B Wortz D, Horsthemke B, Wieczorekand D and Boehringer S (2014). "Classification and visualization based on derived image features: application to genetic syndromes". *PLoS ONE 9(11): e109033. doi:10.1371/journal.pone.0109033.*

Huijts PE, Hollestelle A, **Balliu B**, Houwing-Duistermaat JJ et al. (2014). CHEK2 * 1100delC homozygosity in the Netherlands-prevalence and risk of breast and lung cancer. *European Journal of Human Genetics 22: p.46-51.*

Balliu B, Uh H, Tsonaka R, Boehringer S, Helmer Q and Houwing-Duistermaat JJ (2014). "Combining Information from Linkage and Association Mapping for Next Generation Sequencing Longitudinal Family Data". *BMC Proceedings 8(Suppl 1):S34.*

Houwing-Duistermaat JJ, Helmer Q, **Balliu B**, van den Akker E, Tsonaka R, Uh, H.W. (2014). Gene Analysis for Longitudinal Family Data using Random-Effects Models. *BMC Proceedings 8(Suppl 1):S88.*

Chen H, Malzahn D , **Balliu B**, Li C, Bailey JN (2014) Testing Genetic Association with Rare and Common Variants in Family Data. *Genetic Epidemiology 38 (Suppl 1):S37-43.*

B Balliu, Tsonaka R, van der Woude D, Boehringer S and Houwing-Duistermaat J (2012). "Combining Family and Twin Data in Association Studies to Estimate the Noninherited Maternal Antigens Effect". *Genet Epidemiol. 36(8):811-819.*

Curriculum Vitae

Brunilda was born on the 30th of September 1987, in Vlorë, Albania. She finished her secondary education in 2005 at the 19ο Γενικό Λύκειο Αθηνών in Athens, Greece, where she graduated first in her class. She studied Statistics at the Athens University of Economics and Business, where she graduated as B.Sc. in 2010, again first in her class.

In 2010 she started her PhD at the Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, under the supervision of Prof.dr. Jeanine Houwing Duistermaat and dr. Stefan Boehringer. Her work focused on the development of novel statistical methodology and computational tools for the analysis of studies with response-selective sampling designs, such as case-control or family-based studies, with applications to genetic association studies. The results of this research are presented in this thesis. Chapter 2 and 5 of this thesis have been awarded with the Best Student Presentation Award at the 6th Eastern Mediterranean Region of the International Biometric Society Conference (2011) and 27th International Biometric Society Conference (2014).

Since 2015, she is working as a post-doctoral fellow at the Departments of Pathology and Genetics at the Stanford University School of Medicine developing statistical methods for understanding the effects of genome variation on gene expression and disease, identification of causal regulatory variation using both family and large population cohorts, and methods development for understanding the causal basis of rare diseases.