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.

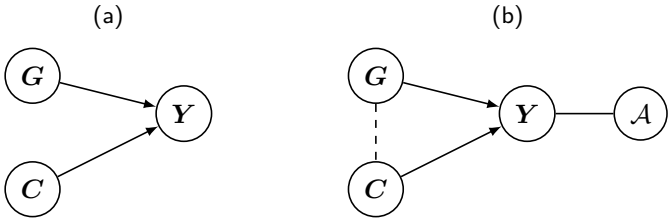

Figure 1.1: **Example to illustrate possible correlation structures among risk factors and a trait in (a) a random sample and (b) a case-control sample**. G: SNP, C: clinical or environmental covariate, Y: binary disease trait, A: ascertainment. Continuous arrows between two nodes connect variables that 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.