Statistical methods for analyzing genetic sequencing association studies

Statistical methods for analyzing genetic sequencing association studies

A dissertation presented

by

Godwin Yuen Han Yung

to

The Department of Biostatistics

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Biostatistics

Harvard University

Cambridge, Massachusetts

May 2016

Dissertation Advisor: Professor Xihong Lin Godwin Yuen Han Yung

Statistical methods for analyzing genetic sequencing association studies

Abstract

Case-control genetic sequencing studies are increasingly being conducted to identify rare variants associated with complex diseases. Oftentimes, these studies collect a variety of secondary traits–quantitative and qualitative traits besides the case-control disease status. Reusing the data and studying the association between rare variants and secondary phenotypes provide an attractive and cost effective approach that can lead to discovery of new genetic associations.

In Chapter 1, we carry out an extensive investigation of the validity of ad hoc methods, which are simple, computationally efficient methods frequently applied in practice to study the association between secondary phenotypes and single common genetic variants. Though other researchers have investigated the same problem, we make two key contributions to existing literature. First, we show that in taking an ad hoc approach, it may be desirable to adjust for covariates that affect the primary disease in the secondary phenotype model, even though these covariates are not necessarily associated with the secondary phenotype in the population. Second, we show that when the disease is rare, ad hoc methods can lead to severely biased estimation and inference if the true disease model follows a non-logistic model such as the probit model. Spurious associations can be avoided by including interaction terms in the fitted regression model. Our results are justified theoretically and via simulations, and illustrated by a genome-wide association study of smoking using a lung cancer case-control study.

In Chapter 2, we consider the problem of testing associations between secondary phenotypes and sets of rare genetic variants. We show that popular region-based methods such as the burden test and the sequence kernel association test (SKAT) can only be applied under the same conditions as those applicable to ad hoc methods (Chapter 1). For a more robust alternative, we propose an inverse-probability-weighted version of the optimal SKAT (SKAT-O) to account for unequal sampling of cases and controls. As an extension of SKAT-O, our approach is data adaptive and includes the weighted burden test and weighted SKAT as special cases.

In addition to weighting individuals to account for the biased sampling, we can also consider weighting the variants in SKAT-O. Decreasing the weight of non-causal variants and increasing the weight of causal variants can improve power. However, since researchers do not know which variants are actually causal, it is common practice to weight genetic variants as a function of their minor allele frequencies. This is motivated by the belief that rarer variants are more likely to have larger effects. In Chapter 3, we propose a new unsupervised statistical framework for predicting the functional status of genetic variants. Compared to existing methods, the proposed algorithm integrates a diverse set of annotations—which are partitioned beforehand into multiple groups by the user—and predicts the functional status for each group, taking into account within- and between-group correlations. We demonstrate the advantages of the algorithm through application to real annotation data and conclude with future directions.

Contents

Acknowledgments

First and foremost, I thank my Lord and savior Jesus Christ for His blessings. In Him, I am loved, redeemed, forgiven, freed, and sanctified. These spiritual blessings far outweigh my fortunate circumstances, including my family, education, and graduating from Harvard University. They remind me that whatever is entrusted to me is not for myself, but for others.

I am grateful to my advisor Xihong Lin for her unwavering patience and guidance. Without her support, none of this would have been possible. I would like to thank the other members of my committee, Peter Kraft and Eric Tchetgen Tchetgen, for generously sharing their time and insights with me. I also send my appreciation to the staff at the Harvard T.H. Chan School of Public Health, especially Jelena Follweiler and Kerri Noonan who provided me and my wife with support during a tough stretch in our second year in Boston.

I owe a great debt to my various mentors over the years. To Eric Kriekhaus, who encouraged me to attend Reed College as an undergraduate. To Alybn Jones, for our coffee conversations at JSM. To Parag Mallick and Robert Rice, who kindled my interest in statistics. To Martha Bulyk and Leila Shokri, for giving me the opportunity to work in a genetics lab and entrusting a rookie with expensive equipment. And to Iuliana Ionita-Laza and Richa Saxena, for making science and collaborating fun.

I would also like to thank all the friends that I've made in Boston, particularly those from Highrock Church and the Department of Biostatistics. There are too many people to name here, but they've made my time in Boston full of joy and discovery.

Finally, I express my deepest love and appreciation for my family. I am forever grateful to my parents, who worked hard every day of their lives. They provided for me and my sister while giving us the space to discover ourselves and our interests. And to my wife Cathy, my best friend and my greatest gift, thank you for joining me on this five-year journey. You are the Penny to my Leonard and you encourage me to be a better person.

To

MY PARENTS
in admiration of their spirit,

and

MY WIFE
who didn't say yes,
but at least said,
"Oh! You got the ring I wanted!"

Chapter 1

Validity of using ad hoc methods to analyze secondary traits in case-control association studies

Godwin Yuen Han Yung[1], Xihong Lin[1]

[1]Department of Biostatistics, Harvard T.H. Chan School of Public Health

1.1 Introduction

Genome-wide association studies (GWAS) examine associations between genetic variants and disease status, often by employing a case-control design. Many of these studies also collect a variety of secondary traits—quantitative and qualitative traits besides the case-control status. In view of high genotyping costs, the resulting data provide a cost-effective way to identify genetic associations with secondary traits. For example, in a lung cancer GWAS conducted at the Massachusetts General Hospital (MGH), detailed smoking histories were collected from each study participant. It is of interest to reuse the data to identify SNPs associated with smoking behavior (Schifano et al., 2013).

A number of methods have been proposed for the analysis of a binary or continuous secondary trait. They include: (a) the *naïve method* which analyzes the combined sample of cases and controls, ignoring case-control ascertainment (Nagelkerke et al., 1995); (b) the *case-only or control-only analysis* (Nagelkerke et al., 1995) (c) the *"adjusted" analysis* where the case-control status is included as a covariate in the fitted model (Jiang et al., 2006); (d) *meta-analytic methods* (Li et al., 2010); (e) the *inverse probability weighted (IPW) method*

(Richardson et al., 2007); and (f) the *semiparametric likelihood method* that explicitly accounts for the case-control sampling scheme (Jiang et al., 2006; Lin and Zeng, 2009; He et al., 2012; Tchetgen Tchetgen, 2014).

We focus here on studying the validity of using the simple and computationally efficient methods (a)-(c), commonly referred to as the "ad hoc" or "standard" methods. Though these methods are widely popular, a deeper understanding of when they are valid is required for proper analysis of secondary traits. It has been argued previously that ad hoc methods can lead to biased estimates of marker-secondary trait associations, except under special conditions (Nagelkerke et al., 1995; Lin and Zeng, 2009; Monsees et al., 2009):

(i) If the disease is not associated with the secondary trait given the genotype, then ad hoc methods are valid.

(ii) For a binary secondary trait, if the disease is not associated with the genotype given the secondary trait, then ad hoc methods are valid. For a continuous secondary trait, the same is true if, in addition, the null hypothesis of no marker-secondary trait association holds.

(iii) If the disease is rare, then methods (b)-(c) are approximately valid.

Consequently, other methods such as (e) and (f) have been proposed as general solutions to secondary trait analysis.

In spite of their limitations and the emergence of other approaches, ad hoc methods have remained popular. Recent years have seen a steady stream of publications on genetic variants influencing human quantitative traits such as body mass index (Speliotes et al., 2010; Wen et al., 2012; Monda et al., 2013). It is common practice to obtain data from multiple case-control association studies of complex diseases (e.g., diabetes, cancer, and hypertension), analyze the data from each study separately using an ad hoc approach, and combine the study-specific results via meta-analysis.

There are several reasons why ad hoc methods have remained popular. First, considering the majority of tested markers in a GWAS are unlikely to be associated with disease risk, and diseases of interest are usually rare, conditions (ii) and (iii) are often met in practice, making ad hoc methods a seemingly valid option. Second, ad hoc methods are straightforward to apply. They require little model building and can be easily performed using linear or logistic regression. In contrast, methods (e)-(f) are more complex. (e) requires that the disease prevalence in a population is known and weighting the sampled subjects in such a way that the weighted subjects approximate the underlying population, which itself might not be well defined. (f) accounts for the case-control sampling by modeling certain nuisance terms in the retrospective likelihood, such as the distribution of the disease given the genotype and secondary trait in the underlying population, which might not be known in practice and requires the knowledge of the population prevalence. In addition, methods (e)-(f), despite their added complexity, may not necessarily be more efficient or robust than ad hoc methods when the assumptions under which the ad hoc methods are valid are met. It has been shown that the weighted approach generally has less power than ad hoc methods that use the entire case-control sample when the ad hoc methods are valid (Monsees et al., 2009). If any of the assumed nuisance models in a semiparametric likelihood are misspecified, then inference may be invalid (Jiang et al., 2006).

Here, we revisit the problem of when ad hoc methods can and cannot be used. This problem is of practical interest because previous discussions by Nagelkerke et al. (1995), Lin and Zeng (2009), and Monsees et al. (2009) leading to (ii) and (iii) make two limiting assumptions: that there are no covariates in the regression models for the disease and secondary trait, and that the disease follows a correctly specified logistic regression model. These assumptions may not be true in practice. Indeed, there may be confounders that need to be adjusted for in order to protect against spurious associations in GWAS. A familiar example of such confounders in GWAS is the presence of population structure, which can be correlated with both the disease and the tested genetic markers (Rosenberg et al., 2002; Price et al., 2006). On the other hand, researchers often assume a logistic model for the disease model in case-control studies. In some cases, the logistic model that is used for analysis might be misspecified, e.g., the probit model for the disease status instead of the logistic model might be true.

Therefore, the purpose of this chapter is to study the performance of ad hoc methods on estimation and inference for the genetic effect on a secondary trait in the presence of covariates and possible disease model misspecification. Our first key contribution is that we show theoretically and with simulations that the presence of covariates confounding the effect of a genetic marker on the disease can lead to spurious genetic associations even when condition (ii) is met. We identify conditions under which the ad hoc methods are valid in the presence of confounders. We show that the spurious associations can be easily and effectively avoided by including the covariates in the fitted regression model for secondary phenotypes. Our second key contribution is that when the disease is rare, we show that the case-only and adjusted analyses can lead to severely biased estimation and incorrect inference if the true disease model is a probit model instead of a logistic model. In this case, spurious associations can be avoided by including interaction terms between the disease status, genetic marker, and covariates in the secondary regression model.

The remainder of this chapter is organized as follows. In Section 1.2, we describe in more detail the study setting, notation, and ad hoc methods. In Section 1.3, we derive the conditions for valid ad hoc analysis in the presence of covariates. Some details are relegated to the Appendix. We present simulation results to examine the conditions in finite samples and to compare existing methods. We also illustrate various methods by applying them to a GWAS of smoking behavior in a sample of lung cancer cases and controls. Finally, in Section 1.4, we discuss the implications of our results for the design and analysis of GWAS of secondary traits using samples ascertained on the basis of another trait.

1.2 Methods

1.2.1 Study setting and notation

Consider a case-control study with n_1 cases and n_0 controls. Let D denote the disease status (1=case, 0=control), Y a binary or continuous secondary trait, $\mathbf{G}$ the genotypes, and $\mathbf{Z}$ and $\mathbf{X}$ the covariates associated with D and Y, respectively. We assume that in the population, disease and secondary trait are distributed with conditional means $\mu_D(Y) = E(D|\mathbf{Z}, \mathbf{G}, Y)$ and $\mu_Y = E(Y|\mathbf{X}, \mathbf{G})$, which follow the generalized linear models:

$$\begin{aligned} g_D\{\mu_D(Y)\} &= \beta_0 + \mathbf{Z}'\boldsymbol{\beta}_Z + \mathbf{G}'\boldsymbol{\beta}_G + Y\beta_Y & (1.1) \\ g_Y(\mu_Y) &= \alpha_0 + \mathbf{X}'\boldsymbol{\alpha}_X + \mathbf{G}'\boldsymbol{\alpha}_G, & (1.2) \end{aligned}$$

where $g_D(\cdot)$ is the link function for the primary phenotype (disease) D model; $g_Y(\cdot)$ is the link function for the secondary phenotype Y model; $(\beta_0, \boldsymbol{\beta}_Z, \boldsymbol{\beta}_G, \beta_Y)$ are the regression coefficients in the D model; and $(\alpha_0, \boldsymbol{\alpha}_X, \boldsymbol{\alpha}_G)$ are the regression coefficients in the Y model.

For binary Y, we assume $g_Y(\cdot) = \text{logit}$. For continuous Y, we assume $g_Y(\cdot)$ is the identity link function and Y follows a normal distribution with the conditional population mean $\mu_Y = E(Y|\mathbf{X}, \mathbf{G})$ and variance σ^2. Our main interest is in estimating and making inference on $\boldsymbol{\alpha}_G$, the population parameter capturing the genetic marker-secondary trait association.

As discussed in the Introduction, existing literature regarding the validity of ad hoc methods often assume a logistic disease model and no covariates. Here, we allow $g_D(\cdot)$ to be any smooth link function. For a rare disease, we consider more closely the choice between the logistic model and the probit model in order to show that misspecification of the disease model by using a misspecified link function can be consequential for the secondary phenotype analysis, which is of primary interest. It is natural to compare the logistic disease model to the probit disease model, because the latter is arguably the most popular alternative parametric model for analyzing binary response data. Also, there is increasing interest to use the probit model (also known as the liability threshold model) in studies of genetic association, heritability, and risk prediction (Wray et al., 2010; So and Sham, 2010; Lee et al., 2011; Zaitlen et al., 2012).

1.2.2 Ad hoc methods

The typical ad hoc approach in the presence of covariates is to regress Y on $\mathbf{X}$, $\mathbf{G}$, and perhaps D, using only the n_1 cases, the n_0 controls, or all $n = n_1 + n_0$ subjects. However, we have found that such a simple ad hoc approach may be invalid in the presence of confounders under the previously established conditions where the ad hoc methods are valid in the absence of covariates. We will show in the next section that including a linear effect of disease-related confounders $\mathbf{Z}$ in the regression model for Y can correct for bias under suitable conditions similar to the existing conditions. Therefore, in the presence of covariates, there are two types of ad hoc methods that one can consider applying. The first type, which we shall refer to as the ad hoc methods with Y-related covariates, takes the typical approach by regressing

Y on $\mathbf{X}$, $\mathbf{G}$, and perhaps D. The second type regresses Y on $\mathbf{X}$, $\mathbf{G}$, perhaps D, *and* $\mathbf{Z}$. Since this type includes both $\mathbf{X}$ and $\mathbf{Z}$ as covariates in the model for Y, we shall refer to them as the ad hoc methods with pooled covariates. Note that if $\mathbf{Z} \subseteq \mathbf{X}$, then the two types of ad hoc methods are equivalent. Furthermore, if an ad hoc method with Y-related covariates (e.g., control-only analysis with Y-related covariates) is valid, then its pooled counterpart (e.g., control-only analysis with pooled covariates) is also valid.

1.3 Results

Let $P(\cdot)$ denote the population-based probability, $\kappa = P(D=1)$ denote the disease prevalence, S indicate with the values 1 versus 0 whether or not an individual from the population is sampled in the case-control study, and $\pi(D) = P(S=1|D)$ be the probability of being sampled in the case-control study for an individual with disease status D. Also, let $\widetilde{P}(\cdot) = P(\cdot|S=1)$, $\widetilde{\mu}_Y = E(Y|\mathbf{X}, \mathbf{G}, \mathbf{Z}, S=1)$, $\widetilde{\mu}_{Y|D} = E(Y|\mathbf{X}, \mathbf{G}, \mathbf{Z}, D, S=1)$, $\widetilde{\sigma}^2 = Var(Y|\mathbf{X}, \mathbf{G}, \mathbf{Z}, S=1)$, and $\widetilde{\sigma}^2_D = Var(Y|\mathbf{X}, \mathbf{G}, \mathbf{Z}, D, S=1)$ denote the case-control probability, conditional means of Y, and conditional variances of Y, all observed under the case-control design.

1.3.1 Common disease, binary secondary trait

When the secondary phenotype Y is binary, we can show that the conditional means of Y in case-control samples satisfy

$$\begin{aligned} \text{logit}(\widetilde{\mu}_Y) &= \alpha_0 + \mathbf{X}'\boldsymbol{\alpha}_X + \mathbf{G}'\boldsymbol{\alpha}_G + r(\mathbf{Z}, \mathbf{G}) \quad (1.3) \\ \text{logit}(\widetilde{\mu}_{Y|d}) &= \alpha_0 + \mathbf{X}'\boldsymbol{\alpha}_X + \mathbf{G}'\boldsymbol{\alpha}_G + r_d(\mathbf{Z}, \mathbf{G}) \quad (1.4) \end{aligned}$$

where

$$\begin{aligned} r(\mathbf{Z}, \mathbf{G}) &= \log\left\{ \frac{\sum_{d=0}^{1} \pi(d)[\mu_D(1)]^d[1-\mu_D(1)]^{1-d}}{\sum_{d=0}^{1} \pi(d)[\mu_D(0)]^d[1-\mu_D(0)]^{1-d}} \right\} \\ r_d(\mathbf{Z}, \mathbf{G}) &= \log\left\{ \left(\frac{\mu_D(1)}{\mu_D(0)}\right)^d \left(\frac{1-\mu_D(1)}{1-\mu_D(0)}\right)^{1-d} \right\} \end{aligned}$$

and $d = 0, 1$. Equivalent expressions for (1.3) and (1.4) were derived by Lin and Zeng (2009) and Tchetgen Tchetgen (2014). From (1.3) and (1.4), it is easy to see that differences between the mean models for the secondary phenotype in case-control studies and in the population (1.2) are given by $r(\mathbf{Z}, \mathbf{G})$ and $r_d(\mathbf{Z}, \mathbf{G})$. Therefore, validity of ad hoc methods depends on the value of these extra terms and whether the methods properly adjust for them. It should be noted that the true means of the secondary phenotypes Y in case-control studies not only depend on the Y-related covariates $\mathbf{X}$ but also the D-related covariates $\mathbf{Z}$.

If $\beta_Y = 0$, i.e., the secondary phenotype Y is not associated with the disease D, then $\mu_D(1) = \mu_D(0)$ and $r(\mathbf{Z}, \mathbf{G}) = r_d(\mathbf{Z}, \mathbf{G}) = 0$. It follows that (1.3) and (1.4) reduce to (1.2), and ad hoc methods with only Y-related covariates $\mathbf{X}$ can be used as a valid tool to estimate and perform inference on all the population parameters α_0, $\boldsymbol{\alpha}_X$, and $\boldsymbol{\alpha}_G$.

Alternatively, if $\boldsymbol{\beta}_G = \mathbf{0}$, i.e., when a SNP is not associated with disease, then $r(\mathbf{Z}, \mathbf{G}) = r(\mathbf{Z})$ and $r_d(\mathbf{Z}, \mathbf{G}) = r_d(\mathbf{Z})$ are functions of $\mathbf{Z}$ but not of $\mathbf{G}$. In this situation, validity of ad hoc methods depends on whether $\mathbf{Z}$ and $\mathbf{G}$ are associated, whether $r(\cdot)$ and $r_d(\cdot)$ are linear in $\mathbf{Z}$, and whether $r_1(\cdot)$ and $r_0(\cdot)$ differ by a constant. When $\mathbf{Z}$ and $\mathbf{G}$ are independent, i.e., when $\mathbf{Z}$ is not a confounder for the genetic association with disease, it is not necessary to adjust for $r(\cdot)$ and $r_d(\cdot)$ in the secondary phenotype regression in order to obtain valid estimation and inference of $\boldsymbol{\alpha}_G$. Hence the ad hoc methods with only Y-related covariates $\mathbf{X}$ can be used. When $\mathbf{Z}$ and $\mathbf{G}$ are correlated, i.e., $\mathbf{Z}$ is a confounder for the genetic association with disease, failure to adjust for $r(\cdot)$ and $r_d(\cdot)$ can lead to spurious associations between G and Y, because an estimate of the association between $\mathbf{G}$ and Y may also capture the association between $\mathbf{Z}$ and Y induced by $r(\cdot)$ and $r_d(\cdot)$. This leads us to consider ad hoc methods with pooled covariates $(\mathbf{X}, \mathbf{Z})$.

Suppose, in addition to $\boldsymbol{\beta}_G = \mathbf{0}$, that $r(\cdot)$ and $r_d(\cdot)$ are linear in $\mathbf{Z}$, and $r_0(\cdot)$ and $r_1(\cdot)$ differ by a constant. Then we can write $\text{logit}(\widetilde{\mu}_Y) = \alpha_0^* + \mathbf{X}'\boldsymbol{\alpha}_X + \mathbf{G}'\boldsymbol{\alpha}_G + \mathbf{Z}'\boldsymbol{\alpha}_Z^*$ and $\text{logit}(\widetilde{\mu}_{Y|d}) = \alpha_{0d}^{**} + \mathbf{X}'\boldsymbol{\alpha}_X + \mathbf{G}'\boldsymbol{\alpha}_G + \mathbf{Z}'\boldsymbol{\alpha}_Z^{**}$, from which it is easy to see that ad hoc methods with pooled covariates are valid. In Appendix A.1.1, we generalize this result by first showing theoretically that for any smooth link function $g_D(\cdot)$, $r(\cdot)$ and $r_d(\cdot)$ are approximately linear in $\mathbf{Z}$ as long as $|\beta_Y|$ and $|\boldsymbol{\beta}_Z|$ are not exceedingly large. We then show for several choices of link function (logit, probit, complementary log-log) that $r_0(\cdot)$ and $r_1(\cdot)$ differ by approximately, if not exactly, a constant. These theoretical results, confirmed by our simulation studies (not provided), show that for typical values of β_Y and $\boldsymbol{\beta}_Z$, ad hoc methods with pooled covariates lead to approximately unbiased estimates of $\boldsymbol{\alpha}_G$ and nominal type I error rates. We conclude that for practical purposes, if $\boldsymbol{\beta}_G = \mathbf{0}$, then ad hoc methods with pooled covariates can be used and provide approximately correct inference.

1.3.2 Common disease, continuous secondary trait

In the case that Y is continuous, we have for the case-control conditional distributions,

$$\begin{aligned} \widetilde{P}(Y|\mathbf{X},\mathbf{G},\mathbf{Z}) &= \frac{P(S=1|\mathbf{Z},\mathbf{G},Y)P(Y|\mathbf{X},\mathbf{G})}{P(S=1|\mathbf{Z},\mathbf{G})} && (1.5) \\ \widetilde{P}(Y|\mathbf{X},\mathbf{G},\mathbf{Z},D) &= \frac{P(D|\mathbf{Z},\mathbf{G},Y)P(Y|\mathbf{X},\mathbf{G})}{P(D|\mathbf{Z},\mathbf{G})}. && (1.6) \end{aligned}$$

If $\beta_Y = 0$, then factors cancel in the numerators and denominators so that $\widetilde{P}(Y|\mathbf{X},\mathbf{G},\mathbf{Z}) = \widetilde{P}(Y|\mathbf{X},\mathbf{G},\mathbf{Z},D) = P(Y|\mathbf{X},\mathbf{G})$ and ad hoc methods with only Y-related covariates $\mathbf{X}$ can be used to estimate and perform inference on all the population parameters α_0, $\boldsymbol{\alpha}_X$, and $\boldsymbol{\alpha}_G$. On the other hand, if $\beta_Y \neq 0$, then calculations of the case-control conditional means and variances of Y, such as $\widetilde{\mu}_{Y|D} = \int y\widetilde{P}(y|\mathbf{X},\mathbf{G},\mathbf{Z},D)dy$, are generally intractable. There is however one exception. When $g_D(\cdot) = \Phi^{-1}$, it can be shown that

$$\begin{aligned} \widetilde{\mu}_{Y|d} &= \mu_Y + r_d(\mathbf{Z},\mathbf{G},\mathbf{X}) && (1.7) \\ \widetilde{\sigma}^2_d &= \sigma^2 + s_d(\mathbf{Z},\mathbf{G},\mathbf{X}) && (1.8) \end{aligned}$$

where

$$\begin{aligned} r_d(\mathbf{Z},\mathbf{G},\mathbf{X}) &= \frac{(-1)^{1-d} \times c \times \phi(\eta)}{[\Phi(\eta)]^d[1-\Phi(\eta)]^{1-d}} \\ s_d(\mathbf{Z},\mathbf{G},\mathbf{X}) &= \frac{(-1)^{1-d} \times c^2 \times \phi'(\eta)}{[\Phi(\eta)]^d[1-\Phi(\eta)]^{1-d}} - [r_d(\mathbf{Z},\mathbf{G},\mathbf{X})]^2 \\ c &= \frac{\sigma^2\beta_Y}{\sqrt{\sigma^2\beta_Y^2+1}} \\ \eta &= \frac{g_D(\mu_D(\mu_Y))}{\sqrt{\sigma^2\beta_Y^2+1}}. \end{aligned}$$

Derivations for (1.7) and (1.8) as well as closed form expressions for $\widetilde{\mu}_Y$ and $\widetilde{\sigma}^2$ can be found in Appendix A.1.2. Given that the logit and probit functions are very close in the mid-range (Amemiya, 1981), we can also find approximate expressions for $g_D(\cdot) = \text{logit}$. Together, these expressions can be useful for investigating what happens when $\beta_Y \neq 0$ and ad hoc methods are applied.

If $\boldsymbol{\alpha}_G = \boldsymbol{\beta}_G = \mathbf{0}$, then $r(\cdot)$ and $r_d(\cdot)$ are functions of $\mathbf{Z}$ and $\mathbf{X}$ but not of $\mathbf{G}$. In this situation, validity of ad hoc methods depends on whether $\mathbf{Z}$ is associated with $\mathbf{G}$, whether $r(\cdot)$ and $r_d(\cdot)$ are linear functions of $(\mathbf{Z}', \mathbf{X}')'$, whether $r_0(\cdot)$ and $r_1(\cdot)$ differ by a constant, and whether $s(\cdot)$ and $s_d(\cdot)$ are constants. For example, if $r_0(\cdot)$ and $r_1(\cdot)$ are linear functions of $(\mathbf{Z}', \mathbf{X}')'$ that differ by a constant, and $s_d(\cdot)$ are constants, then we can write $\widetilde{\mu}_{Y|d} = \alpha_{0d}^{**} + \mathbf{X}'\boldsymbol{\alpha}_X^{**} + \mathbf{G}'\boldsymbol{\alpha}_G + \mathbf{Z}'\boldsymbol{\alpha}_Z^{**}$ and $\widetilde{\sigma}_d^2 = \sigma_d^2$. It follows that for large samples, ad hoc method (b) with pooled covariates $(\mathbf{X}, \mathbf{Z})$ provides valid estimation and inference of $\boldsymbol{\alpha}_G$. The adjusted analysis (c) with pooled covariates can be used too if $\widetilde{\sigma}_0^2 = \widetilde{\sigma}_1^2$.

In Appendix A.1.2, we show theoretically that $r(\cdot)$ and $r_d(\cdot)$ are approximately linear in $(\mathbf{Z}', \mathbf{X}')'$ and $s(\cdot)$ and $s_d(\cdot)$ are approximately constants as long as $|\beta_Y|$ and $|(\boldsymbol{\beta}'_Z, \boldsymbol{\alpha}'_X)'|$ are not exceedingly large. In Section 1.3.5, we show with simulations that for typical values of β_Y and $(\boldsymbol{\beta}'_Z, \boldsymbol{\alpha}'_X)'$, ad hoc methods (a) and (b) with pooled covariates lead to approximately unbiased estimates of $\boldsymbol{\alpha}_G$ and nominal type I error rates. Therefore, we conclude that for practical purposes, if $\boldsymbol{\alpha}_G = \boldsymbol{\beta}_G = \mathbf{0}$, then ad hoc methods (a) and (b) with pooled covariates are approximately valid.

As mentioned, the adjusted analysis with pooled covariates is valid if, in addition, $r_1(\cdot) - r_0(\cdot)$ is a constant and $\widetilde{\sigma}_0^2 = \widetilde{\sigma}_1^2$. While it is easy to show that the first condition is approximately true for common disease (Appendix A.1.2), $\widetilde{\sigma}_0^2$ is generally not equal to $\widetilde{\sigma}_1^2$. Nevertheless, the difference between the sample variance of the case-only and control-only analyses with pooled covariates seemed to be small enough for inference to be approximately correct in our simulations.

1.3.3 Rare disease

For rare disease, $P(D = 0|\mathbf{Z}, \mathbf{G}, Y)$ and $P(D = 0|\mathbf{Z}, \mathbf{G})$ in (1.6) are approximately equal to 1, so $\widetilde{P}(Y|\mathbf{X}, \mathbf{G}, \mathbf{Z}, D = 0) \approx P(Y|\mathbf{X}, \mathbf{G})$. It follows that a control-only analysis is approximately valid for binary and continuous secondary traits. Intuitively, when the disease is rare, the controls closely resemble the general population. Therefore, any conclusion about the population based on the controls will be approximately correct.

As for ad hoc methods that use cases, these methods may or may not be valid depending on the underlying disease model. If $g_D(\cdot) = \text{logit}$, then we have for binary Y that $r_1(\mathbf{Z}, \mathbf{G}) \approx \beta_Y$, and for continuous Y that $r_1(\mathbf{Z}, \mathbf{G}, \mathbf{X}) \approx \beta_Y \sigma^2$ and $s_1(\mathbf{Z}, \mathbf{G}, \mathbf{X}) \approx 0$. In fact, for continuous Y, $\widetilde{P}(Y|\mathbf{X}, \mathbf{G}, \mathbf{Z}, D = 1)$ is approximately proportional to $\exp\{-[Y - \mu_Y - \beta_Y \sigma^2]/2\sigma^2\}$ (Lin and Zeng, 2009). Thus, for both binary and continuous secondary traits, ad hoc methods (b)-(c) with only Y-related covariates $\mathbf{X}$ yield approximately valid estimation and inference for $\boldsymbol{\alpha}_X$ and $\boldsymbol{\alpha}_G$.

If instead, $g_D(\cdot) = \Phi^{-1}$, then we have for binary Y that $r_1(\mathbf{Z}, \mathbf{G}) \approx \text{constant} - \beta_Y \cdot \Phi^{-1}(\mu_D(0))$, and for continuous Y, $r_1(\mathbf{Z}, \mathbf{G}, \mathbf{X}) \approx -\frac{\sigma^2 \beta_Y}{\sigma^2 \beta_Y^2 + 1} \Phi^{-1}(\mu_D(\mu_Y))$. Derivations are available in Appendix A.1.3. Note that the first remainder is a linear function of $\mathbf{Z}$ and $\mathbf{G}$ and the latter is a linear function of $\mathbf{Z}$, $\mathbf{G}$, and $\mathbf{X}$. These results are substantially different from those obtained under $g_D(\cdot) = \text{logit}$, where r_1 for both types of secondary traits were constants. Results assuming $g_D(\cdot) = \Phi^{-1}$ imply that an estimate of $\boldsymbol{\alpha}_G$ from the case-only analysis with pooled covariates is generally biased:

$$E(\widehat{\boldsymbol{\alpha}}_G - \boldsymbol{\alpha}_G) \approx \begin{cases} -\beta_Y \boldsymbol{\beta}_G & \text{binary } Y \\ -\frac{\sigma^2 \beta_Y}{\sigma^2 \beta_Y^2 + 1} (\boldsymbol{\beta}_G + \beta_Y \boldsymbol{\alpha}_G) & \text{continuous } Y \end{cases}$$

By extension, the adjusted analysis is also invalid. Finally, one might consider extending the adjusted analysis with pooled variates to include D-Z, D-G, and D-X interactions. In doing so, the main effect of $\mathbf{G}$ will encode the marginal association of interest $\boldsymbol{\alpha}_G$. However, if $\mathbf{Z}$ and $\mathbf{X}$ include large numbers of possibly confounding covariates for population stratification, it is unlikely that adding a large number of interactions will lead to an increase in power compared to the control-only analysis.

1.3.4 Conditions for ad hoc analysis in the presence of covariates

We have conducted a thorough investigation into the properties of the ad hoc methods. We state here the main conclusions. Ad hoc methods can lead to invalid estimation and inference of $\boldsymbol{\alpha}_G$, except under special conditions:

(i) If the disease is not associated with the secondary trait ($\beta_Y = 0$), then ad hoc methods are valid.

(ii*) For binary Y, if the disease is not associated with the genotype ($\boldsymbol{\beta}_G = \mathbf{0}$), then ad hoc methods with pooled covariates $(\mathbf{X}, \mathbf{Z})$ are approximately valid. Ad hoc methods with only Y-related covariates $\mathbf{X}$ are also approximately valid if, in addition, the D-related covariates $\mathbf{Z}$ is not associated with $\mathbf{G}$, i.e., are not confounders for gene-disease association. Similarly, for continuous Y, if neither the disease nor secondary trait are associated with the genotype ($\boldsymbol{\alpha}_G = \boldsymbol{\beta}_G = \mathbf{0}$), then ad hoc methods with pooled covariates $(\mathbf{X}, \mathbf{Z})$ are approximately valid. Ad hoc methods with Y-related covariates are also approximately valid if, in addition, $\mathbf{Z}$ is not associated with $\mathbf{G}$.

(iii*) If the disease is rare, then the control-only analysis is approximately valid. The case-only and adjusted analyses are also approximately valid if, in addition, $g_D(\cdot) = \text{logit}$; however, if $g_D(\cdot) = \Phi^{-1}$, then these two analyses can lead to biased estimation and incorrect inference.

1.3.5 Simulation study

To quantify the type I error rate, bias, and power of ad hoc methods for secondary trait analysis, we simulated case-control association studies drawn from an underlying cohort of size N. Our simulation procedure extends that of Monsees et al. (2009) by allowing for covariates and a non-logistic disease model.

First, covariates Z_{1i} and X_{1i} for subjects $i = 1, ..., N$ were drawn from a standard normal distribution, and $Z_{2i} = X_{2i}$ was sampled as a Bernoulli random variable with probability of success 0.5. Diallelic genotype G_i was sampled conditional on Z_{1i} as a binomial random variable of size 2 with probability of success $\text{expit}(\gamma_0 + \gamma_1 Z_{1i})$. Continuous secondary trait Y_i was drawn from a normal distribution with mean $\alpha_0 + X_{1i}\alpha_{X1} + X_{2i}\alpha_{X2} + G_i\alpha_G$ and variance 1. (In the original journal article, we consider a binary secondary trait.) Disease D_i was sampled conditional on $\mathbf{Z}_i = (Z_{1i}, Z_{2i})'$, G_i, and Y_i as a Bernoulli random variable with $g_D(P(D_i = 1|\mathbf{Z}_i, Y_i, G_i)) = \beta_0 + \lambda(\mathbf{Z}_i'\boldsymbol{\beta}_Z + \beta_Y Y_i + \beta_G G_i)$. Finally, case-control samples were selected by randomly sampling n_1 cases and n_0 controls from the simulated cohort. Note that, depending on the values of γ_1 and β_{Z1}, Z_1 was or was not a confounder of the effect of G on D.

We simulated a wide variety of scenarios, varying seven parameters: disease prevalence $\kappa \in \{0.01, 0.10\}$; link function $g_D(\cdot) \in \{\text{logit}, \Phi^{-1}\}$; the increase in log odds of inheriting a minor allele from a specific parent per unit change in $Z_1 = \gamma_1 \in \{0, \ln(1.7)/2, \ln 1.7\}$; the percent of variance in Y explained by $G = r^2_{YG} \in \{0, 0.005, 0.01\}$; the association between Z_1 and $D = \beta_{Z1} \in \{0, \ln(1.7)/2, \ln 1.7\}$; the association between Y and $D = \beta_Y \in \{0, \ln(2)/2, \ln 2\}$; and the association between G and $D = \beta_G \in \{0, \ln(1.7)/2, \ln 1.7\}$.

We fixed $(\alpha_0, \alpha_{Xj}, \beta_{Z2}) = (0, 0.2, \log(1.7)/2)$. For $g_D(\cdot) = \text{logit}$, we set $\lambda = 1$ so that a non-intercept coefficient in the disease model could be interpreted as the increase in log odds of disease per unit change in the corresponding explanatory variable. For $g_D(\cdot) = \Phi^{-1}$, we set $\lambda = \sqrt{3}/\pi$ so that the association between D and $(\mathbf{Z}, Y, G)$ were comparable between the logistic and probit disease model (Amemiya, 1981). γ_0 was chosen so that the genotype had a minor allele frequency of approximately 0.13. The mean change in Y per copy of the minor allele (α_G) and the baseline odds parameter β_0 were chosen to be consistent with r^2_{YG} and κ. We generated large cohorts and sampled from each $n_1 = 1,000$ cases and $n_0 = 1,000$ controls. In order to estimate type I error rate (power) accurately, a total of 10^8 (10^4) replicate data sets were simulated for each scenario.

For an example of a scenario with different confounders for the disease models and the secondary phenotype models, consider Crohn's disease (D) and lactase persistence (Y). Genetic lactase persistence has been linked to risk of Crohn's disease, lactase persistence has been shown to vary from northeast to southeast Europe (X_1), and Jews of European descent (Z_1) are at significantly higher risk of Crohn's disease (Nolan et al., 2010; Price et al., 2006; Kenny et al., 2012). Another example is lung cancer (D) and smoking behavior (Y). It is well known that first and second hand smoking (Z_1) causes lung cancer (U.S. Department of Health and Human Services, 2006). While there is no data to suggest that the two are themselves associated, the practice of smoking differs from culture to culture, so it is possible that first and second hand smoking are associated with certain genetic markers.

We conducted the following nine analyses for each simulated dataset:

1. Naïve analysis with Y-related covariates: regress Y on $(\mathbf{X}, G)$ in the case-control sample.

2. Control-only analysis with Y-related covariates: regress Y on $(\mathbf{X}, G)$ among controls.

3. Case-only analysis with Y-related covariates: regress Y on $(\mathbf{X}, G)$ among cases.

4. Adjusted analysis with Y-related covariates: regress Y on $(\mathbf{X}, G, D)$ in the case-control sample.

5. Naïve analysis with pooled covariates: regress Y on $(\mathbf{X}, \mathbf{Z}, G)$ in the case-control sample.

6. Control-only analysis with pooled covariates: regress Y on $(\mathbf{X}, \mathbf{Z}, G)$ among controls.

7. Case-only analysis with pooled covariates: regress Y on $(\mathbf{X}, \mathbf{Z}, G)$ among cases.

8. Adjusted analysis with pooled covariates: regress Y on $(\mathbf{X}, \mathbf{Z}, G, D)$ in the case-control sample.

9. IPW regression: regress Y on $(\mathbf{X}, G)$ using weights $w_1 = \kappa$ for cases and $w_0 = 1 - \kappa$ for controls.

We included Analysis 9 for the purpose of generalizing previous results by Monsees et al. (2009) comparing the performance of ad hoc methods to IPW regression. For each method and scenario, the probability of rejecting the null hypothesis H_0: $\alpha_G = 0$ was estimated by applying a nominal significance threshold of $\alpha \in \{10^{-4}, 10^{-5}, 10^{-6}\}$. Bias was obtained by taking the average of $\widehat{\alpha}_G - \alpha_G$.

Figures 1.1–1.3 summarize the type I error rates and bias for the control-only, adjusted, and IPW regression analyses across the null scenarios ($\alpha_G = 0$) that were considered. Results for the naïve and case-only analyses can be found in the original journal article. Results for $\alpha \in \{10^{-4}, 10^{-5}\}$ are omitted but similar. As expected, IPW regression (Analysis 9) was unbiased for all of the scenarios considered. However, interestingly, its type I error rates were consistently slightly inflated due to the instability of the sandwich estimator. Increasing the sample size (n_1,n_0) improved type I error control (not shown). Ad hoc methods with pooled covariates (Analyses 5–8) had appropriate type I error rates and no perceptible bias whenever $\beta_Y = 0$ or $\beta_G = 0$. Likewise, ad hoc methods with Y-related covariates (Analyses 1–4) were

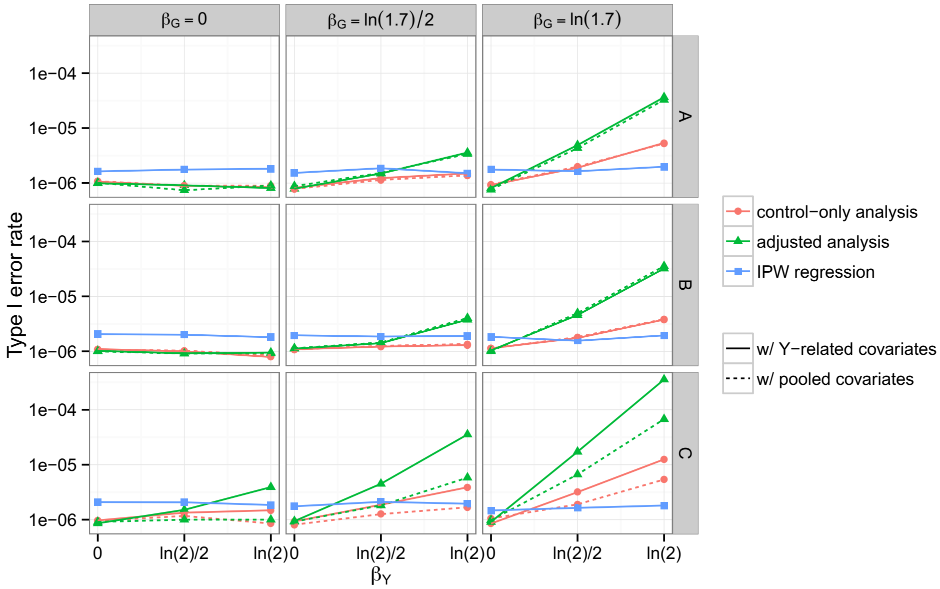

Figure 1.1: Empirical type I error rates for testing genetic associations with a continuous secondary trait, at genome-wide $\alpha = 10^{-6}$ level and across scenarios with different combinations of β_Y, β_G, γ_1 and β_{Z1}. The disease is assumed to be common (10% prevalence) and to follow a logistic model. In row A, covariate Z_1 is assumed to be associated with G but not with D ($\gamma_1 = \ln 1.7$, $\beta_{Z1} = 0$). In row B, Z_1 is associated with D but not with G ($\gamma_1 = 0$, $\beta_{Z1} = \ln 1.7$). In row C, Z_1 is a confounder of the association between G and D ($\gamma_1 = \beta_{Z1} = \ln 1.7$).

valid whenever $\beta_Y = 0$, or $\beta_G = 0$ and Z is not a confounder for the effect of G on D ($\gamma_1 = 0$ or $\beta_{Z1} = 0$).

For common disease ($\kappa = 0.10$; Figures 1.1 and 1.2), we detected an inflation in type I error rates and bias for all eight ad hoc methods when $\beta_Y \neq 0$ and $\beta_G \neq 0$. We also detected an inflation in type I error rates and bias for Analyses 1–4 when $\beta_Y \neq 0$, $\beta_G = 0$, and Z_1 confounded the association between G and D ($|\gamma_1| > 0, |\beta_{Z1}| > 0$).

For rare disease ($\kappa = 0.01$; Figure 1.3) with a logistic link function, all ad hoc methods that condition on case-control status (Analyses 2–4, 6–8) had little to no inflation in type I error rates and bias regardless of whether $\beta_Y = 0$ or $\beta_G = 0$. However, for rare disease with a probit link function, only the control-only analysis (Analyses 2 and 6) and IPW regression were approximately valid in general. All other ad hoc methods had highly inflated type I error rates and severe bias when $\beta_Y \neq 0$ and $\beta_G \neq 0$.

We compared the power of Analyses 1–9 whenever the analyses were approximately valid

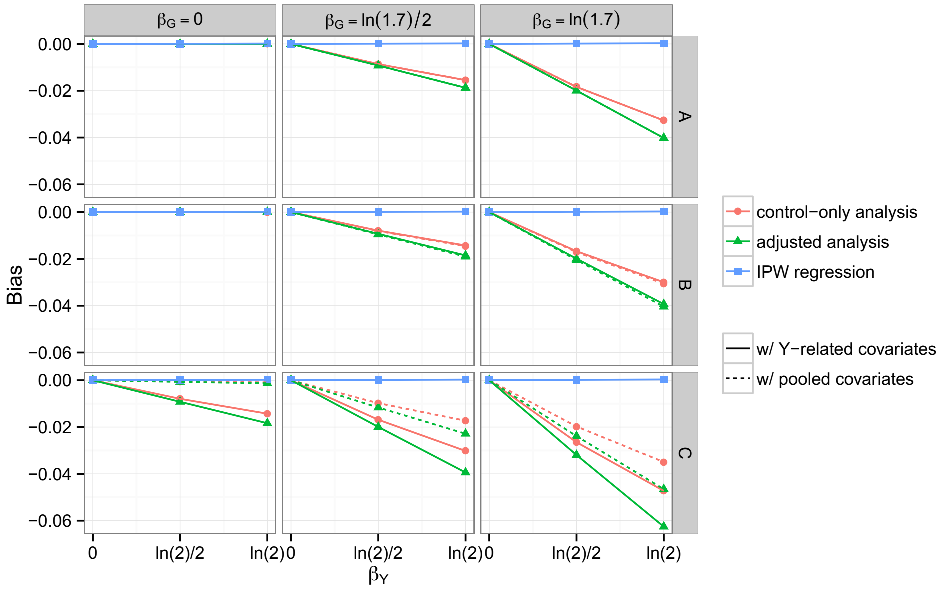

Figure 1.2: Empirical bias for the estimated genetic effect $\widehat{\alpha}_G$ on a continuous secondary trait, across null scenarios ($\alpha_G = 0$) with different combinations of β_Y, β_G, γ_1 and β_{Z1}. The disease is assumed to be common (10% prevalence) and to follow a logistic model ($g_D(\cdot) = \text{logit}$). In row **A**, covariate Z_1 is assumed to be associated with G, but not with D ($\gamma_1 = \ln 1.7$, $\beta_{Z1} = 0$). In row **B**, Z_1 is associated with D, but not with G ($\gamma_1 = 0$, $\beta_{Z1} = \ln 1.7$). In row **C**, Z_1 is a confounder of the association between G and D ($\gamma_1 = \beta_{Z1} = \ln 1.7$).

by varying $\alpha_G \in \{0, \ln(1.7)/2, \ln(1.7)\}$. The naïve analyses (Analyses 1 and 5) tended to be the most powerful, followed by the adjusted analyses (Analyses 4 and 8), IPW regression (Analysis 9), and finally the ad hoc analyses restricted to cases or controls (Analyses 2, 3, 6, and 7) . In addition, ad hoc methods with Y-related covariates were slightly more powerful than their corresponding ad hoc methods with pooled covariates.

1.3.6 Data example: GWAS of smoking behavior

To demonstrate the application of ad hoc methods, we performed a genome-wide association analysis of smoking behavior using a set of 696 lung cancer cases and 730 controls.

Study population

Our study population was derived from a large ongoing case-control study of the molecular epidemiology of lung cancer at MGH, and has been described in detail elsewhere (Schifano et al., 2013). Briefly, the controls were recruited from the friends or spouses of cancer

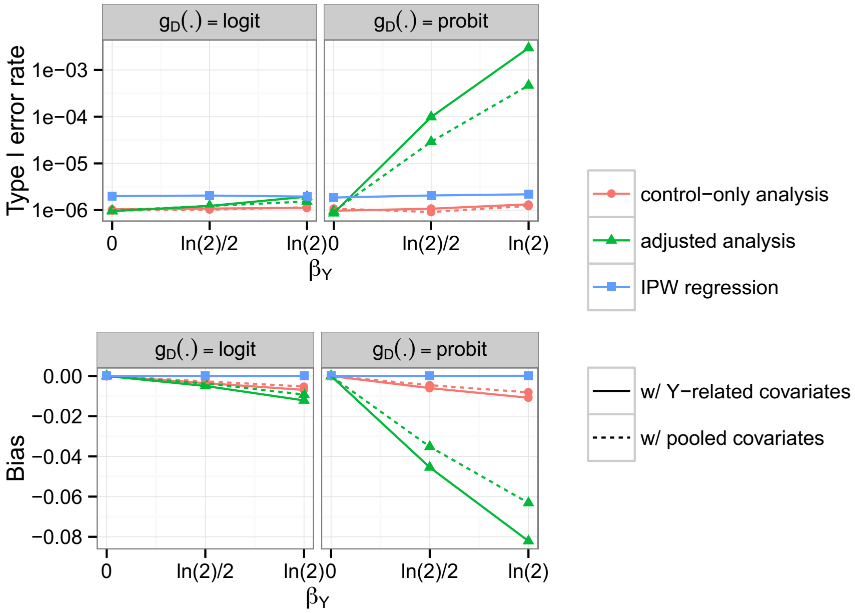

Figure 1.3: Empirical type I error rates and bias for testing and estimating genetic associations with a continuous secondary trait, at genome-wide $\alpha = 10^{-6}$ level and across null scenarios ($\alpha_G = 0$) with different combinations of β_Y and link function $g_D(\cdot)$ for the disease model. The disease is assumed to be rare (1% prevalence) and to follow either a logistic or probit model ($g_D(\cdot)$ = logit or Φ^{-1}). G is assumed to be associated with D ($\beta_G = \ln 1.7$). Z_1 is assumed to be a confounder of the association between G and D ($\gamma_1 = \beta_{Z1} = \ln 1.7$). The scenarios with a logistic disease model (left column) are the same as the scenarios in the bottom right plots of Figures 1.1 and 1.2, except here the disease is not common but rather rare.

patients or the friends or spouses of other surgery patients in the same hospital. To reduce confounding due to population structure, the study was limited to individuals of self-reported European descent.

Genotyping

Peripheral blood samples were obtained from all study participants at the time of enrollment. DNA was extracted from samples using the Puregene DNA Isolation Kit (Gentra Systems), and genoyping was performed with the Illumina Human610-Quad BeadChip. For quality control, SNPs that had call rate less than 95%, that failed the Hardy-Weinberg equilibrium test at 10^{-6}, or that had minor allele frequency less than 5%, were excluded. Blood samples with genotyping call rates less than 95% were also excluded. There were 513,271 SNPs remaining after frequency and quality control. To further control for population structure, EIGENSTRAT was used to perform a principal components (PCs) analysis (Price et al., 2006). We included the first four PCs, on the basis of significant Tracy-Widom tests ($p < 0.05$) and genomic control inflation factor, as covariates for all analyses. Of the remaining six out of ten top PCs, we decided to also include the ninth PC as a covariate in our secondary linear regression models because we found this PC to be significantly associated with lifetime smoking exposure ($p < 0.05$).

Covariate and phenotypic data collection

Interviewer-administered questionnaires collected information on sociodemographic variables from each subject, including age (years; continuous), gender, education history (college degree or more; yes/no), and smoking intensity (cigarettes/day and number of years smoked). Subjects were classified as either never smokers (less than 100 cigarettes in their lifetime), former smokers (quit smoking at least 1 year prior to interview date), or current smokers (at time of interview). Only ever-smokers (former and current) were used in our data analysis, as we were interested in studying the genetic effects on smoking intensity measured by pack-years.

We used square root pack-years (number of packs of cigarettes smoked daily times the number of years smoked) as our secondary outcome measure of smoking behavior. The square root transformation was applied to better satisfy assumptions of normality.

We performed the naïve, control-only, case-only, adjusted, and IPW analyses for each SNP by regressing square root of pack-years on genotype (number of minor alleles), age, gender, college education, and PCs 1–4 and 9. For the adjusted analysis, lung cancer status was included in the regression model. For IPW regression, we estimated the prevalence of lung cancer amongst ever-smokers in Massachusetts to be 0.00148, and used this prevalence to calculate the inverse probability weight for each study individual.

On conditions (i)-(iii*)

Since conditions (i)-(iii*) play an important role in determining which results from a genome-wide ad hoc analysis of a secondary trait are credible, we sought to verify these conditions

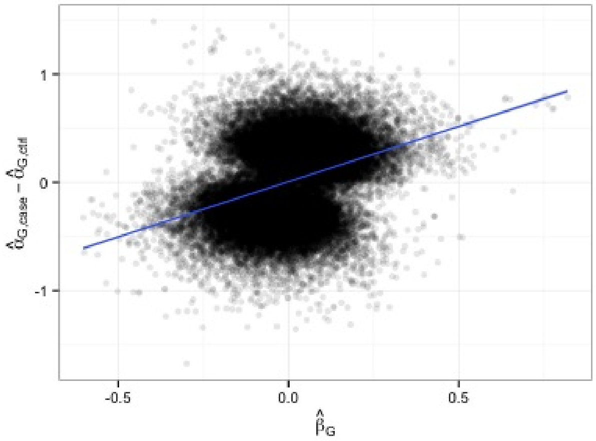

Figure 1.4: Top 50k SNPs from IPW regression. Observed difference between case-only and control-only estimates has a significant tendency to increase as the log odds-ratio of a genetic marker and lung cancer increases (slope of best fit line = 1.02, $p < 10^{-15}$). Under the assumption of a rare disease with a logistic model, one would expect the best fit line to be $y = 0$.

in our dataset.

For condition (i), we found that smoking intensity is significantly associated with lung cancer risk ($OR = 1.45$, $p < 10^{-15}$). For condition (ii*), we fitted for each SNP a logistic regression model to test for genetic associations with lung cancer, adjusting for square root pack-years, age, gender, college education, and the first four PCs. For condition (iii*), given an estimated prevalence of 0.00148, lung cancer can be considered a rare disease within the at-risk population of ever-smokers in Massachusetts. We looked at diagnostic plots to investigate whether a logistic model for (1) is a reasonable fit for lung cancer risk (Figure 1.4). Under such a model, one would expect case- and control-only estimates to be unbiased and uninfluenced by marker-disease associations. However, we see from Figure 1.4 that for our dataset the case- and control-only analyses were generally estimating different quantities, and that the difference between their estimates ($\widehat{\alpha}_{G,case} - \widehat{\alpha}_{G,ctrl}$) tended to increase as the log odds ratio of SNPs and lung cancer ($\widehat{\beta}_G$) increased. It was only when a SNP was weakly associated with lung cancer ($\widehat{\beta}_G \approx 0$) that the expected difference between case- and control-only estimates equalled 0.

These observations led us to conclude that for our purpose of analyzing genome-wide associations with smoking behavior, condition (i) does not hold, the disease is rare, and

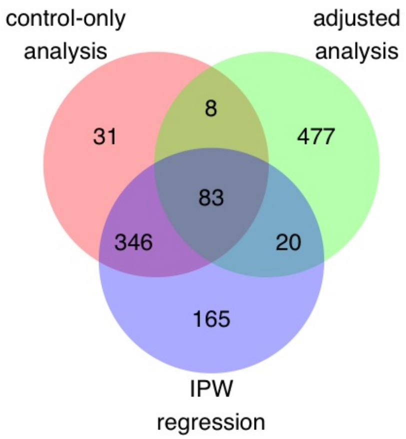

Figure 1.5: Number of nominally significant SNPs ($p < 10^{-3}$) from the control-only, adjusted, and IPW analysis of $\sqrt{\text{pack-years}}$. p values from a 1-DF Wald test assuming an additive genetic model.

the disease model (1.1) with $g_D(\cdot) = \text{logit}$ is somehow misspecified. Consequently, we may prefer results from IPW regression, the control-only analysis (because the disease is rare), and the adjusted analysis of SNPs with weak evidence of an association with lung cancer risk (because the adjusted analysis is one of the more powerful valid ad hoc approaches under condition (ii*). Yet, when condition (ii*) is not satisfied, it is not as severely biased as the naïve analysis.)

Results

Manhattan plots for the naïve, control-only, case-only, adjusted, and IPW analyses can be found in the original journal article. In total, 1130 SNPs were identified as nominally significant at $p < 10^{-3}$ by the control-only, adjusted, or IPW analysis (see Figure 1.5). Comparing the control-only analysis to IPW regression, SNPs identified as nominally significant by the control-only analysis were roughly a subset of the SNPs identified by IPW regression. Indeed, 429 of the 468 (91.7%) SNPs identified by the control-only analysis were also identified by IPW regression. Meanwhile, IPW regression identified 185 other SNPs. Of the 429 SNPs identified by both analyses, the majority (328, 76.5%) were more significant when analyzed by IPW regression than by the control-only analysis.

The adjusted analysis generally identified different SNPs as nominally significant than the control-only and adjusted analyses. Specifically, the adjusted analysis identified 477 novel

SNPs, novel in the sense that they were nominally significant ($p < 10^{-3}$) when analyzed by the adjusted analysis, but nominally insignificant ($p \geq 10^{-3}$) when analyzed by the control-only and IPW analyses. Likewise, the control-only and IPW analyses identified 31 and 165 novel SNPs. However, taken together, the control-only and IPW analyses collectively identified 542 SNPs that were nominally insignificant when analyzed by the adjusted analysis (Figure 5).

A large number of the novel SNPs identified by adjusted analysis had weak evidence of an association with lung cancer risk; when tested for H_0: $\beta_G = 0$, 139 (29.1%), 220 (46.1%), and 118 (24.7%) SNPs had p value in the range $[0.0, 0.1)$, $[0.1, 0.5)$, and $[0.5, 1.0]$, and odds ratio in the range $[0.63, 1.62]$, $[0.74, 1.35]$, and $[0.92, 1.09]$, respectively. Therefore, applying condition (ii*), the adjusted analysis of many of these SNPs are likely to be valid.

Table 1.1 displays the top ten SNPs for the control-only analysis. Looking at the top SNPs and the top ten novel SNPs for the control-only, adjusted, and IPW analyses, we found SNPs from several genes identified in previous GWASs of smoking cessation: *ARHGAP24*, *C1orf95*, *CDH18*, *CDYL2*, *DOK6*, *FAM189A1*, *HSD17B2*, *KSR1*, *NBEA*, *PDE10A*, *SLC9A2* (a paralog of *SLC9A9*), and *TACR1* (Rose et al., 2010; Uhl et al., 2010; Tang et al., 2014).

In Figure 1.6, we see that the control-only analysis and IPW regression performed similarly for nominally significant SNPs from the previously known genes. Meanwhile, for some SNPs their association with smoking behavior were much more significant when analyzed by the adjusted analysis than by the control-only or IPW analysis (e.g., SNPs from *HSD17B2*, *NBEA*, *SLC9A2*), and vice versa (e.g., SNPs from *CDH18*). This is consistent with simulation results that the adjusted analysis is more powerful than the control-only analysis and the IPW analysis in the situations when they are valid. Only *TACR1* had similar results across the three methods. We note that SNPs which were nominally significant only when analyzed by the adjusted analysis had weak evidence of an association with lung cancer risk.

1.4 Discussion

In this paper, we have given new conditions for using ad hoc methods. Our findings extend previous work by demonstrating that if there are covariates confounding the effect of a genetic

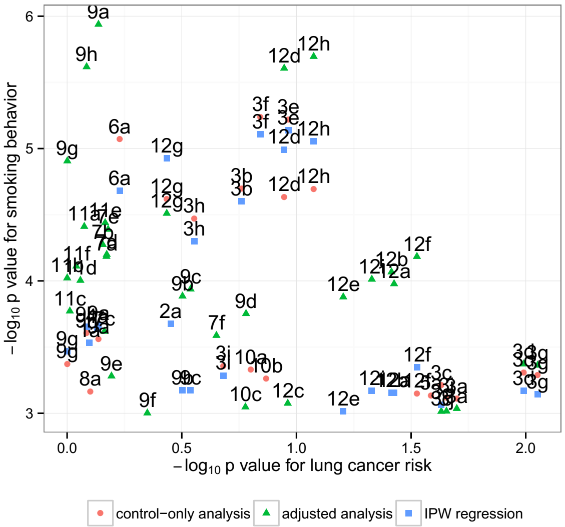

Figure 1.6: p values from the genome-wide association analysis of $\sqrt{\text{pack-years}}$ and lung cancer risk for nominally significant SNPs ($p < 10^{-3}$) from twelve selected genes: (1) *ARHGAP24*, (2) *C1orf95*, (3) *CDH18*, (4) *CDYL2*, (5) *DOK6*, (6) *FAM189A1*, (7) *HSD17B2*, (8) *KSR1*, (9) *NBEA*, (10) *PDE10A*, (11) *SLC9A2*, and (12) *TACR1*. All genes have been identified in previous studies of smoking cessation. Here, we compare the results from the control-only, adjusted, and IPW analyses of $\sqrt{\text{pack-years}}$. Results can be distinguished by gene (number), SNP (letter), and the secondary analysis applied (shape and color).

marker on the disease but that are not adjusted for in the secondary trait analysis, then ad hoc analysis can lead to spurious associations even when the genetic marker is not associated with the disease. Futhermore, for a rare disease, the case-only and adjusted analyses can lead to severely biased estimation and incorrect inference if the true disease model is not strictly logistic.

The conditions set forth in this paper apply to the setting where there are no gene-environment interactions in the disease. We now briefly discuss the validity of the ad hoc methods for G-E interaction models. It is easy to show that if there is an interaction between gene and covariates, but no interaction between gene and secondary trait, then conditions (i) and (iii*) hold, but not condition (ii*). On the other hand, if there is an interaction between gene and secondary trait on disease risk, then (i) and (ii*) do not hold, and the only valid analysis for a rare disease is the control-only analysis. The fact that the case-only and adjusted analyses lead to incorrect estimation and inference has been discussed previously by Li et al. (2010). As a solution, the authors proposed an adaptively weighted method that combines the case-only and control-only estimates, while reducing to the control-only analysis if there is strong evidence of a gene-secondary trait interaction.

We considered the possibility of interaction between SNPs and smoking behavior for lung cancer risk in our data analysis. We found that SNPs identified as nominally significant by the adjusted analysis tended not to modify the effect of smoking behavior on lung cancer risk, but SNPs identified by the control-only or IPW analysis had moderate to strong evidence of G-E interaction. This difference explains why we observed relatively little overlap in Figure 1.5, and why some previously known genes were identified by only the adjusted analysis, or by the control-only and IPW analyses but not the adjusted analysis (Figure 1.6).

The results in this paper have several important implications for secondary trait analysis. First, when applying ad hoc methods, one should consider including potential confounders of the association between the genetic marker and the disease, even if these covariates are not predictors of the secondary trait. For example, one might adjust for population structure associated with the secondary trait *and* population structure associated with the disease. However, one should be aware that when including additional covariates, power may be reduced if the secondary trait is binary and the covariates are not actually confounders

(Pirinen et al., 2012).

Second, for a rare disease, it is crucial to verify disease model assumptions or to perform sensitivity analysis. A potential pitfall is misspecifying the link function of the disease model (e.g., logit vs probit). Another is ignoring gene-environment interactions in the linear predictor. The importance of having a robust analysis applies not only to ad hoc methods, but also to complex approaches. For instance, Li et al.'s adaptively weighted method and Lin and Zeng's semi-parametric approach both assume that the disease follows a logistic model. Lin and Zeng's semi-parametric approach further assumes that there are no G-E interactions. It is important when applying either of these methods to verify their assumptions.

Finally, researchers may benefit from applying multiple methods rather than a one-size-fits-all solution. In our data analysis of smoking behavior, the adjusted analysis identified a large number of promising SNPs that were otherwise missed by the control-only analysis and IPW regression, and vice versa. Meanwhile, the control-only analysis and IPW regression performed similarly when analyzing SNPs from previously known genes. However, the control-only analysis was easier and computationally much faster to perform, while IPW regression was slightly more powerful because it used both the lung cancer cases and controls. Therefore, whether it is to save computational time or to improve the identification of promising genetic markers, researchers would do well to apply several ad hoc and complex methods.

Table 1.1: Top 10 SNPs from the genome-wide control-only analysis of $\sqrt{\text{pack-years}}$. Estimates of the additive genetic effect on smoking behavior ($\widehat{\alpha}_G$) and their p values from a 1-DF Wald test for the naïve, control-only, case-only, adjusted, and IPW analysis. Marker-lung cancer effect estimates ($\widehat{OR}_{DG} = \exp(\widehat{\beta}_G)$) and their p values from a 1-DF Wald test are also provided.

			Lung cancer	**Smoking behavior**				
SNP	Chr.	Gene	$\widehat{OR}_{DG}$	Naïve	Control-only	Case-only	Adjusted	IPW
rs7588326	2	TACR1	1.16	-0.40	-0.51	-0.30	-0.40	-0.51
			(8.41×10^{-2})	(9.82×10^{-6})	(2.02×10^{-5})	(9.97×10^{-3})	(2.02×10^{-6})	(8.82×10^{-6})
rs4461636	5	CDH18	1.25	-0.44	-0.98	0.01	-0.46	-0.98
			(1.44×10^{-1})	(5.57×10^{-3})	(5.78×10^{-6})	(9.49×10^{-1})	(1.65×10^{-3})	(7.78×10^{-6})
rs4242066	5	CDH18	1.28	-0.40	-0.99	0.06	-0.44	-0.99
			(1.08×10^{-1})	(1.13×10^{-2})	(6.02×10^{-6})	(7.65×10^{-1})	(2.93×10^{-3})	(7.26×10^{-6})
rs1391429	5	CDH18	1.22	-0.43	-0.90	-0.03	-0.45	-0.90
			(1.73×10^{-1})	(5.80×10^{-3})	(2.00×10^{-5})	(8.95×10^{-1})	(1.99×10^{-3})	(2.50×10^{-5})
rs7842063	8	N/A	0.91	0.44	0.64	0.17	0.41	0.64
			(3.79×10^{-1})	(8.69×10^{-5})	(1.29×10^{-5})	(2.52×10^{-1})	(7.77×10^{-5})	(4.41×10^{-5})
rs4404875	8	RP1L1	0.86	0.35	0.66	0.05	0.36	0.66
			(1.53×10^{-1})	(2.19×10^{-3})	(1.89×10^{-5})	(7.23×10^{-1})	(6.89×10^{-4})	(1.74×10^{-5})
rs1655645	15	FAM189A1	0.95	0.28	0.55	-0.03	0.26	0.55
			(5.89×10^{-1})	(1.99×10^{-3})	(8.48×10^{-6})	(8.05×10^{-1})	(2.17×10^{-3})	(2.080×10^{-5})
rs1893213	18	N/A	0.89	0.28	0.54	0.00	0.28	0.54
			(1.68×10^{-1})	(2.30×10^{-3})	(1.03×10^{-5})	(9.91×10^{-1})	(1.04×10^{-3})	(5.98×10^{-6})
rs4805573	19	ZNF536	1.36	-0.65	-1.06	-0.27	-0.67	-1.06
			(6.63×10^{-2})	(1.77×10^{-4})	(4.52×10^{-6})	(2.38×10^{-1})	(2.91×10^{-5})	(1.29×10^{-6})
rs4805574	19	ZNF536	1.34	-0.65	-1.02	-0.30	-0.67	-1.02
			(8.40×10^{-2})	(1.84×10^{-4})	(1.13×10^{-5})	(1.85×10^{-1})	(3.51×10^{-5})	(3.99×10^{-6})

Chapter 2

Optimal test for rare variant effects on secondary traits in case-control sequencing studies

Godwin Yuen Han Yung[1], Seunggeun Lee[2], Sara Lindstroem[3], Rulla Tamimi[4], Peter Kraft[4], Xihong Lin[1]

[1]Department of Biostatistics, Harvard T.H. Chan School of Public Health

[2]Department of Biostatistics, University of Michigan

[3]Department of Epidemiology, University of Washington

[4]Department of Epidemiology, Harvard T.H. Chan School of Public Health

2.1 Introduction

Despite the success of genome-wide association studies (GWASs) in identifying common single nucleotide polymorphisms (SNPs) that contribute to complex diseases, the majority of genetic variants identified so far confer relatively small increments in risk, leaving many more to be discovered (Manolio et al., 2009). The rapid evolution of massively parallel sequencing platforms makes it possible to sequence entire genomes and has the potential to identify rare genetic variants that contribute to disease susceptibility (Cirulli and Goldstein, 2010). In Chapter 1, we discussed the problem of using ad hoc methods to identify common genetic variants associated with secondary traits using available case-control GWAS. We now develop a procedure that re-uses data from case-control sequencing studies to identify rare variants associated with secondary traits.

There has been substantial work on developing powerful methods to identify rare variants associated with complex traits. Region-based analysis has become the standard approach, because individual variant tests, typically used to analyze single common variants, are underpowered to detect rare variant effects due to the low allele frequencies and the large number of rare variants in the genome (Bansal et al., 2010). Commonly used region-based tests include burden and non-burden tests. Burden tests assume all rare variants in a genomic region have effects on the phenotype in the same direction and of similar magnitude (Li and Leal, 2008; Madsen and Browning, 2009; Morris and Zeggini, 2010; Price et al., 2010). In contrast, the sequence kernel association test (SKAT) is particularly powerful in the presence of protective and deleterious variants and null variants, but is less powerful than burden tests when a large number of variants in a region are causal and in the same direction (Wu et al., 2011). In practice the underlying biological mechanisms are unknown and vary from one gene to another across the genome. To incorporate this uncertainty, Lee et al. (2012b) proposed a data-adaptive test (SKAT-O) that is optimal within a class of tests that include both burden tests and SKAT as special cases.

Burden tests, SKAT, and SKAT-O can be used to infer about genetic associations in a population when the study subjects are a random sample of the population or the outcome of interest is the disease status in a case-control study. However, when the outcome of interest is a secondary trait in a case-control study, applying these methods may be extremely misleading. This is because cases and controls are selected at different rates from their respective subpopulations. As a result, the study subjects do not constitute a random sample of the general population, and the population association between genetic variants and the secondary trait can be distorted in the case-control sample. This phenomenon was explored in the context of case-control GWAS in Chapter 1.

In this chapter, we propose a weighted version of SKAT-O to account for unequal sampling of cases and controls. Our approach uses inverse-probabilities-of-selection as weights and is applicable to binary and quantitative secondary traits. As an extension of SKAT-O, the inverse-probability-weighted test (IPW SKAT-O) is likewise data adaptive and includes inverse-probability-weighted burden tests and inverse-probability-weighted SKAT as special cases. We derive the asymptotic distribution of the IPW SKAT-O statistic, which allows us to calculate the p-value analytically with high accuracy and efficiency in the tail. Because IPW SKAT-O can be conservative for small to medium sample sizes, we further derive an adjustment procedure for IPW SKAT-O by precisely estimating the small-sample variance and kurtosis.

We compare, analytically and numerically, the performance of IPW SKAT-O with unweighted applications of SKAT-O that either ignore the sampling design, use only a subset of the samples, or include the case-control status as an additional covariate. We demonstrate that the finite-sample adjusted IPW SKAT-O has proper type I error rates, is more robust than unweighted tests, and has higher power than weighted burden tests and weighted SKAT.

2.2 Methods

2.2.1 Tests for rare variant effects

Before we describe the study setting and our method, we review existing rare variant testing methods. Suppose n subjects are sequenced in a region with p genotyped rare variants. For the ith subject, let Y_i denote the outcome variable of interest, $\mathbf{G}_i = (g_{i1}, ..., g_{ip})$ denote the genotypes for the p variants ($g_{ij} = 0, 1, 2$ for 0, 1, or 2 copies of the minor allele), and $\mathbf{X}_i = (x_{i1}, ..., x_{iq})$ the covariates for which we would like to adjust. Assume Y_i follows an exponential family distribution with first two moments $E(Y_i) = \mu_i$ and $Var(Y_i) = \phi v(\mu_i)$, and link function

$$g(\mu_i) = \mathbf{X}_i \boldsymbol{\alpha}_X + \mathbf{G}_i \boldsymbol{\alpha}_G \tag{2.1}$$

where $v(\cdot)$ is a variance function. $\boldsymbol{\alpha}_X$ and $\boldsymbol{\alpha}_G$ are the vectors of regression coefficients for the covariates and rare variants, respectively. Under this generalized linear model (GLM) framework, the association between the p rare variants and the phenotype Y_i can be tested by evaluating the null hypothesis that $H_0 : \boldsymbol{\alpha}_G = \mathbf{0}$. A p degree of freedom (df) test however may lose power when p is large. To reduce the d.f., additional assumptions need to be made.

Popular burden-based tests reduce the df by assuming that $\alpha_{Gj} = w_j \alpha_{G0}$ for all j, where each w_j is some known constant that may depend on MAF. Under this assumption, (2.1) becomes $g(\mu_i) = \mathbf{X}_i \boldsymbol{\alpha}_X + \alpha_{G0} \sum_{j=1}^{p} w_j g_{ij}$ and the association between the genetic variants and the phenotype can be tested by conducting a standard 1 d.f. test with $H_0 : \alpha_{G0} = 0$.

SKAT takes a different approach to reducing the df. It assumes that each α_{Gj} independently follows an arbitrary distribution with mean zero and variance $w_j\psi$, where w_j is again a fixed number that may depend on MAF. As a result, the null hypothesis $H_0 : \boldsymbol{\alpha}_G = \mathbf{0}$ is equivalent to $H_0 : \psi = 0$, i.e. the variance component test in generalized linear mixed models where α_{Gj}s are treated as random effects.

Burden tests assume that all variants are causal with the same direction of association and common α_{G0}. Violation of these assumptions can result in a loss of power. On the other hand, if a large percentage of variants in the target region are associated with the phenotype with the same direction of effect, burden tests can outperform SKAT because SKAT assumes that α_{Gj}'s are independent. To explicitly account for possible correlation among the variant effects, Lee et al. (2012b) proposed to allow $\boldsymbol{\alpha}_G$ to follow a multivariate distribution with exchangeable correlation structure. That is, they assumed the correlation matrix of $\boldsymbol{\alpha}_G$ to be $\mathbf{R}_\rho = (1-\rho)\mathbf{I}+\rho\mathbf{1}\mathbf{1}'$. Then for a fixed ρ, the score test statistic of the variance component ψ is:

$$Q_\rho = (\mathbf{Y} - \widehat{\boldsymbol{\mu}})'\mathbf{K}_\rho(\mathbf{Y} - \widehat{\boldsymbol{\mu}})/\widehat{\phi}^2 \qquad (2.2)$$

where $\widehat{\boldsymbol{\mu}}$ is an $n \times 1$ vector of estimates of $\boldsymbol{\mu}$, $\mathbf{W} = \text{diag}[w_1, ..., w_p]$ is a $p \times p$ diagonal matrix of weights for the p rare variants, and $\mathbf{K}_\rho = \mathbf{G}\mathbf{W}\mathbf{R}_\rho\mathbf{W}\mathbf{G}'$. When $\rho = 0$, Q_ρ reduces to SKAT. When $\rho = 1$, Q_ρ is equivalent to the burden score test statistic.

In practice, the optimal ρ is unknown and needs to be estimated from the data to maximize power. Therefore, Lee and others further proposed to select ρ by using the minimum of p-values as a test statistic. Specifically, their test statistic is $Q_{optimal} = \inf_{0\leq p \leq 1} p_\rho$, where p_ρ is the p-value computed based on Q_ρ. The resulting optimal test SKAT-O corresponds to a best linear combination of SKAT and burden tests that maximizes power.

2.2.2 Study setting and notation

Now consider a target population of N individuals. For the ith individual, let D_i denote the disease status (1=case, 0=control), Y_i the binary or continuous secondary trait, $\mathbf{G}_i = (g_{i1}, ..., g_{ip})$ the genotypes for the p variants, and $\mathbf{X}_i = (x_{i1}, ..., x_{iq})$ the covariates for which we would like to adjust (e.g. demographic or environmental variables). To related genotypes to the secondary trait Y_i such that, *in the target population*, Y_i follows an exponential family distribution with first two moments $E(Y_i) = \mu_i$ and $Var(Y_i) = \phi v(\mu_i)$, and link function (2.1). For binary Y_i, we assume $g(\cdot) = \text{logit}$. For continuous Y_i, we assume $g(\cdot)$ is the identity link function and Y_i follows a normal distribution.

For a case-control study with a total of $n < N$ subjects (n_1 cases, n_0 controls), the data consist of $(D_i, Y_i, \mathbf{X}_i, \mathbf{G}_i)$ $(i = 1, ..., n)$. Interest is still in evaluating the null hypothesis that $H_0 : \boldsymbol{\alpha}_G = \mathbf{0}$. However, unlike in the previous subsection where $\boldsymbol{\alpha}_G$ described the association between the variants and phenotype Y_i in the population as well as in the n sampled individuals, here $\boldsymbol{\alpha}_G$ describes the variant-phenotype association in the population but not necessarily in the n sampled cases and controls. Consequently, applying SKAT-O ad hoc by using (a) only the controls, (b) only the cases, (c) the combined sample of cases and controls, or (d) the combined sample of cases and controls and adjusting for the disease status D_i in the fitted model, may lead to misleading results. For example, if the disease is common (e.g., prevalence of 0.10) and both the secondary trait and rare variants are associated with the disease, then (a)-(d) can produce highly inflated type I error rates. In Appendix A.2.1, we provide detailed conditions under which ad hoc applications of SKAT-O are appropriate. Some of these conditions will also be demonstrated through simulations in Section.

2.2.3 IPW SKAT-O

If all of the individuals in the target population were selected, we could test $H_0 : \boldsymbol{\alpha}_G = \mathbf{0}$ by directly applying SKAT-O. However, in a case-control study, the individuals are selected with unequal probabilities. Let π_i denote the inclusion probability of the ith individual. In practice, if the disease prevalence κ is known, as it often is, then one can instead use $\pi = D_i \frac{n_1}{n} \frac{1}{\kappa} + (1 - D_i) \frac{n_0}{n} \frac{1}{1-\kappa}$. A Horvitz-Thompson (Horvitz and Thompson, 1952) type "estimator" of Q_ρ is

$$Q^*_\rho = (\mathbf{Y} - \widehat{\boldsymbol{\mu}})'\mathbf{W}^*\mathbf{K}_\rho\mathbf{W}^*(\mathbf{Y} - \widehat{\boldsymbol{\mu}})/\widehat{\phi}^2 \tag{2.3}$$

where $\widehat{\boldsymbol{\mu}} = g^{-1}(\mathbf{X}\widehat{\boldsymbol{\alpha}}_X)$, $w_i^* = 1/\pi_i$, and $\mathbf{W}^* = \text{diag}[w_1^*, ..., w_n^*]$. The estimators $\widehat{\boldsymbol{\alpha}}_X$ and $\widehat{\phi}$ of $\boldsymbol{\alpha}_X$ and ϕ can be obtained by using IPW linear or logistic regression to fit the null model $g(\mu_i) = \mathbf{X}_i \boldsymbol{\alpha}_X$. Equation (2.3) can also be written as

$$Q_\rho^* = \rho Q_1^* + (1-\rho) Q_0^*, \tag{2.4}$$

which is a weighted average of IPW SKAT and IPW burden score test statistics. One can easily see that the unified test statistic reduces to an IPW SKAT statistic when $\rho = 0$ and to an IPW burden score test statistic when $\rho = 1$.

We show in Appendix A.2.2 that for a fixed ρ, Q_ρ^* asymptotically follows a mixture of χ^2 distributions. Specifically, if $(\lambda_1, ..., \lambda_m)$ are the eigenvalues of $\widehat{\mathbf{V}}^{-1/2}\mathbf{K}_\rho\widehat{\mathbf{V}}^{-1/2}$ where $\widehat{\mathbf{V}} = \text{diag}[\widehat{\phi}^2/w_1^{*2}/(y_1 - \widehat{\mu}_1)^2, ..., \widehat{\phi}^2/w_n^{*2}/(y_n - \widehat{\mu}_n)^2]$,. then the null distribution of Q_ρ^* can be approximated by $\sum_{j=1}^m \lambda_j \chi_{1,j}^2$, where $\chi_{1,j}^2$ are independent χ_1^2 random variables. To reduce small sample bias, the restricted maximum likelihood (REML) estimator of the variance component can be used (Zhang and Lin, 2003). Define $\mathbf{P} = \widehat{\mathbf{V}}^{-1} - \widehat{\mathbf{V}}^{-1}\mathbf{X}(\mathbf{X}'\widehat{\mathbf{V}}\mathbf{X})^{-1}\mathbf{X}'\widehat{\mathbf{V}}^{-1}$. We use the eigenvalues of $\mathbf{P}^{1/2}\mathbf{K}_\rho\mathbf{P}^{1/2}$ to obtain the null distribution of Q_ρ^*. A p-value can be calculated by matching moments (Liu et al., 2009) or by inverting the characteristic function (Davies, 1980).

To select ρ, we follow Lee et al. (2012b) by using the minimum of p-values as a test statistic. Our final test statistic is

$$Q_{optimal}^* = \inf_{0 \le \rho \le 1} p_\rho^* \tag{2.5}$$

where p_ρ^* is the p-value computed based on Q_ρ^*. $Q_{optimal}^*$ can be obtained by simple grid search across a range of ρ: set a grid $0 = \rho_1 < \rho_2 < \cdots < \rho_b = 1$, then the test statistic $Q_{optimal}^* = \min\{\rho_1, \rho_2, ..., \rho_b\}$.

We show in Appendix A.2.3 that, for large samples each test statistic Q_ρ^* can be decomposed into a mixture of two random variables, one of which asymptotically follows a χ_1^2 distribution, the other of which can be asymptotically approximated by a mixture of chi-square distributions with a variance component. As a result, the p-value of $Q_{optimal}^*$ can be quickly obtained analytically by one-dimensional numeric integration.

2.2.4 Small-sample adjustment

One of the key strengths of IPW SKAT-O is its ability to efficiently compute asymptotic p-values without the need for resampling. It can also easily adjust for covariates. These features are advantageous in whole-genome and whole-exome sequencing studies wherein a large number of tests are performed and one needs to control for multiple comparisons and account for population stratification. However, even for sample sizes of 2000, the large-sampled based p-value calculations can produce conservative results, leading to incorrect type I error control and power loss (see simulation results). We propose in this section small-sample-adjusted p-value calculations for Q^*_ρ and $Q^*_{optimal}$. Our derivations follow closely to that of Lee et al. (2012a).

We first consider p-value calculations for Q^*_ρ. Our approach is to readjust the moments of the null distribution of Q^*_ρ. In Appendix A.2.2, we show that $\mu_{Q,\rho} = \sum \lambda_j$, where $(\lambda_1, .., \lambda_m)$ are eigenvalues of $\mathbf{P}^{1/2}\mathbf{K}_\rho\mathbf{P}^{1/2}$, is an unbiased estimate of $E(Q^*_\rho|H_0)$, the mean of Q^*_ρ under the null hypothesis. Unfortunately, because of the case-control sampling scheme, deriving the analytical formula of the variance and kurtosis of Q^*_ρ is infeasible. Hence, we propose to estimate the variance and kurtosis by generating resampled phenotypes from the parametric bootstrap (Davison and Hinkley, 1999).

Specifically, suppose $Q^{*(b)}_\rho$ $(b = 1, ..., B)$ is the test statistic Q^*_ρ from the bootstrap sample $\mathbf{Y}^{(b)}$. The sample variance and kurtosis are

$$\widehat{\sigma}^2 = \frac{1}{B}\sum_{b=1}^{B}(Q^{*(b)}_\rho - \mu_{q,\rho})^2 \quad \text{and} \quad \widehat{\gamma} = B\frac{\sum_{b=1}^{B}(Q^{*(b)}_\rho - \mu_{q,\rho})^4}{\left(\sum_{b=1}^{B}(Q^{*(b)}_\rho - \mu_{q,\rho})^2\right)^2} - 3.$$

Using the estimated moments, the p-value can be calculated as

$$1 - F\left(\frac{(Q^*_\rho - \mu_{Q,\rho})\sqrt{2df}}{\widehat{\sigma}} + df|\chi^2_{df}\right)$$

where $df = 12/\widehat{\gamma}$ and $F(\cdot|\chi^2_{df})$ is the cumulative distribution function of χ^2_{df}. We can apply the same approach to $Q^*_{optimal}$. Details are shown in Appendix A.2.3.

It should be noted that our method requires substantially less computation time than methods that compute p-values by calculating the proportion of permutation or bootstrap test statistics larger then the observed test statistic, .e.g., $P(Q^{*(b)}_\rho \geq Q^*_\rho)$. For whole-exome sequencing studies, one must be able to obtain p-values at the 10^{-3}-10^{-6} level to account for multiple comparisons when testing 20,000 genes. This requires more than 10^7-10^8 permutations or bootstraps for each gene. In contrast, our approach requires sampling phenotypes under the null model only 10,000 times to obtain stable estimates of the higher moments. And because the null model is the same across different genes, the same resampled bootstrap phenotypes can be used for all the genes across the genome. This saves a substantial amount of computation time.

2.3 Results

2.3.1 Simulation study

To compare the performance of the proposed methods to the performance of applying SKAT-O ad hoc, we simulated case-control sequencing studies drawn from an underlying cohort of size N. First, we generated sequence data of European ancestry from 10,000 chromosomes over 1 Mb regions using the calibrated coalescent model (Schaffner et al., 2005). We then randomly selected regions with lengths of 3 kb and chose from each region potential causal variants from the rare variants with true MAF < 0.03. Secondary and disease phenotypes were generated for each individual in the cohort using the linear and generalized linear regression models

$$Y_i = 0.5x_{i1} + 0.5x_{i2} + \sum_{j=1}^{s} \alpha_{gj} g_{ij} + \epsilon_i$$

$$\text{logit}(P(D_i = 1)) = \beta_0 + 0.5x_{i1} + 0.5x_{i2} + \beta_Y Y_i + \sum_{j=1}^{s} \beta_{Gj} g_{ij},$$

where X_1 and ϵ were standard normal random variables, X_2 was a Bernoulli random variable with probability of success 0.5, and $(g_1, ..., g_s)$ were the variants in a 3 kb region. Finally, case-control samples were selected by randomly sampling n_1 cases and n_0 controls from the simulation cohort.

We applied six different implementations of SKAT-O to each of the randomly selected 3 kb regions by adapting six approaches: (1) using only the controls (Ctrl); (2) using only the cases (Case); (3) using both cases and controls (Naïve); (4) joint analysis of cases and controls adjusting for disease status in the fitted null model (Joint); (5) IPW without small-sample-adjustment (IPW); and (6) small-sample adjusted IPW (IPW-S). For all the implementations, $Beta(1, 25)$ weights were used to upweight variants (Lee et al., 2012b). The p-values of the optimal tests were computed using the 11 values of ρ equally spaced between 0 and 1. For IPW-S, the sample variance and kurtosis were estimated from 10,000 bootstrapped phenotype sets. We also fixed ρ at 0 and 1 to obtain corresponding SKAT and burden test p-values for approaches (1)-(6).

We simulated a wide variety of scenarios by varying a number of parameters, including the disease prevalence $\kappa = P(D_i = 1) \in \{0.01, 0.10\}$, the increase in log-odds of D per unit increase in $Y = \beta_Y \in \{0, \ln(2)/2, \ln 2\}$, and sample size $n_1 = n_0 \in \{1000, 2000\}$. The baseline odds parameter β_0 was chosen to be consistent with κ.

To study the effects of varying proportions of variants being causal variants (causal with respect to Y or D), we followed Lee and others Lee et al. (2012a) by considering four different settings in which 0%, 10%, 20% or 50% of the rare variants were causal variants. For each setting, we considered three different sign configurations of the nonzero α_G's (β_G's): 50% of α_G's (β_G's) were positive, 80% of α_G's (β_G's) were positive, and all α_G's (β_G's) were positive. We used $|\alpha_{Gj}| = c_Y|\log_{10}(p_j)|/2$ and $|\beta_{Gj}| = c_D|\log_{10}(p_j)|/2$, where p_j was the MAF of the jth variant. When 10%, 20%, and 50% of the rare variants were causally associated with Y, we set $c_Y = 0.6$, $c_Y = 0.4$, and $c_Y = 0.2$, respectively. Similarly, when 10%, 20%, and 50% of the rare variants were causally associated with D, we set $c_D = \log 7$, $c_D = \ln 5$, and $c_D = \ln 2.5$. In doing so, we used decreasing effects to compensate for the increased number of causal variants.

For each scenario, we evaluated type I error rates at level $\alpha \in \{0.01, 0.05\}$ by simulating a total of 10,000 replicate data sets, one for each of the 10,000 randomly selected 3 kb regions. To investigate type I error rates at a level for exome-wide testing, we reduced the computational burden by generating 1000 phenotype sets for each 3 kb region, giving a total of 10^7 phenotypes. Type I error rates and power were estimated by the proposition of p-values smaller than the given α-level.

Tables 2.1 and 2.2 show that, in all scenarios, IPW-S SKAT-O accurately controls type I error with moderate α levels. Meanwhile, IPW SKAT-O produces slightly conservative

Table 2.1: Empirical type I error rates for six different implementations of SKAT-O aimed at testing an association between randomly selected 3 kb regions with a continuous secondary trait. From left to right, the table considers scenarios with different disease-secondary trait associations (β_Y) and % of variants in the region that are causally associated with the disease. The effects of causal variants decrease as the % of causal variants increases. The disease is assumed to be common (10% prevalence) and to follow a logistic model. Sample size is fixed at 1000 cases and 1000 controls.

	0% causal			20% causal			50% causal		
	$\beta_Y = 0$	$\beta_Y = \ln 1.4$	$\beta_Y = \ln 2$	$\beta_Y = 0$	$\beta_Y = \ln 1.4$	$\beta_Y = \ln 2$	$\beta_Y = 0$	$\beta_Y = \ln 1.4$	$\beta_Y = \ln 2$
Ctrl									
$\alpha = 0.05$	0.047	0.048	0.047	0.048	0.051	0.067	0.051	0.055	0.059
$\alpha = 0.01$	0.009	0.009	0.010	0.010	0.012	0.016	0.010	0.011	0.014
Case									
$\alpha = 0.05$	0.053	0.052	0.052	0.052	0.235	0.539	0.050	0.210	0.476
$\alpha = 0.01$	0.011	0.010	0.011	0.012	0.111	0.368	0.010	0.096	0.317
Naïve									
$\alpha = 0.05$	0.047	0.049	0.051	0.052	0.064	0.094	0.049	0.063	0.081
$\alpha = 0.01$	0.010	0.010	0.010	0.011	0.016	0.026	0.010	0.015	0.023
Joint									
$\alpha = 0.05$	0.047	0.048	0.050	0.053	0.204	0.457	0.049	0.183	0.402
$\alpha = 0.01$	0.010	0.011	0.010	0.011	0.095	0.300	0.010	0.084	0.261
IPW									
$\alpha = 0.05$	0.044	0.044	0.043	0.046	0.044	0.047	0.046	0.045	0.043
$\alpha = 0.01$	0.008	0.008	0.008	0.008	0.010	0.009	0.009	0.007	0.008
IPW-S									
$\alpha = 0.05$	0.049	0.048	0.047	0.050	0.048	0.048	0.051	0.048	0.046
$\alpha = 0.01$	0.009	0.009	0.009	0.010	0.011	0.010	0.010	0.009	0.010

type I error rates. The four ad hoc implementations of SKAT-O control type I error well if the secondary trait is not associated with the disease (i.e. $\beta_Y = 0$) or none of variants in the region are causally associated with the disease (i.e. $\beta_{Gj} = 0$ for all j); however, if both conditions are not satisfied, then these methods can result in highly inflated type I error rates. For rare diseases (Table 2.2), the control-only SKAT-O accurately controls type I error since the control population closely resembles a random sample of the target population.

Figure 2.1 shows the empirical power at $\alpha = 2.5 \times 10^{-6}$ under various considered configurations. When the percentage of SNPs causally associated with the secondary trait Y was low, SKATs and SKAT-Os had higher power than the burden tests. SKATs and SKAT-Os also outperformed the burden tests when 50% of the causal SNPs were deleterious, regardless of what percentage of variants in the genomic region were causal. The burden tests performed better than SKATs when both the percentage of variants that were causal and the percentage of causal variants that were deleterious were high. Even then, SKAT-Os performed better or similar to the burden tests. This suggests that the performance of SKAT-O is data adaptive. Comparing the six different implementations (Ctrl, Case, Naïve,

Table 2.2: Empirical type I error rates for six different implementations of SKAT-O aimed at testing an association between randomly selected 3 kb regions with a continuous secondary trait. From left to right, the table considers scenarios with different disease-secondary trait associations (β_Y) and % of variants in the region that are causally associated with the disease. The effects of causal variants decrease as the % of causal variants increases. The disease is assumed to be rare (1% prevalence) and to follow a logistic model. Sample size is fixed at 1000 cases and 1000 controls.

	0% causal			20% causal			50% causal		
	$\beta_Y = 0$	$\beta_Y = \ln 1.4$	$\beta_Y = \ln 2$	$\beta_Y = 0$	$\beta_Y = \ln 1.4$	$\beta_Y = \ln 2$	$\beta_Y = 0$	$\beta_Y = \ln 1.4$	$\beta_Y = \ln 2$
Ctrl									
$\alpha = 0.05$	0.049	0.049	0.048	0.050	0.051	0.051	0.050	0.050	0.051
$\alpha = 0.01$	0.011	0.011	0.009	0.010	0.010	0.012	0.010	0.011	0.010
Case									
$\alpha = 0.05$	0.050	0.051	0.051	0.051	0.211	0.486	0.049	0.178	0.394
$\alpha = 0.01$	0.009	0.011	0.012	0.011	0.108	0.342	0.011	0.091	0.269
Naïve									
$\alpha = 0.05$	0.049	0.052	0.050	0.051	0.160	0.356	0.051	0.156	0.328
$\alpha = 0.01$	0.011	0.012	0.010	0.010	0.065	0.195	0.012	0.061	0.167
Joint									
$\alpha = 0.05$	0.049	0.052	0.052	0.050	0.189	0.418	0.049	0.165	0.344
$\alpha = 0.01$	0.011	0.012	0.010	0.011	0.093	0.287	0.011	0.079	0.229
IPW									
$\alpha = 0.05$	0.046	0.045	0.044	0.046	0.046	0.045	0.048	0.044	0.044
$\alpha = 0.01$	0.008	0.007	0.007	0.008	0.009	0.009	0.009	0.009	0.007
IPW-S									
$\alpha = 0.05$	0.051	0.050	0.049	0.050	0.050	0.050	0.054	0.048	0.048
$\alpha = 0.01$	0.010	0.008	0.008	0.010	0.011	0.009	0.010	0.010	0.009

Joint, IPW, and IPW-S) for each rare variant test (SKAT, burden, SKAT-O), we see that the naïve and joint analyses were most powerful. Meanwhile, IPW-S, IPW, the control-only, and case-only analyses had similar power. At less stringent levels of α (0.001–0.05), we have found that IPW-S and IPW perform slightly better than the control-only and case-only analyses (Figure A.1 in Appendix A.2.4). On the other hand, the control-only and case-only analyses has slightly more power than IPW-S and IPW for studies with smaller sample sizes (Figure A.2 in Appendix A.2.4).

2.3.2 Data example: Sequencing analysis of mammographic density

We applied the proposed IPW SKAT-O and other competing methods to next-generation sequencing data from the Nurses' Health Studies I and II (Mensah-Ablorh et al., 2016) to test for association between mammographic density (MD) and rare variants. MD is regarded as an intermediate phenotype in breast cancer development. The identification of genes that regulate MD might enhance the ability to identify women at risk of developing breast cancer

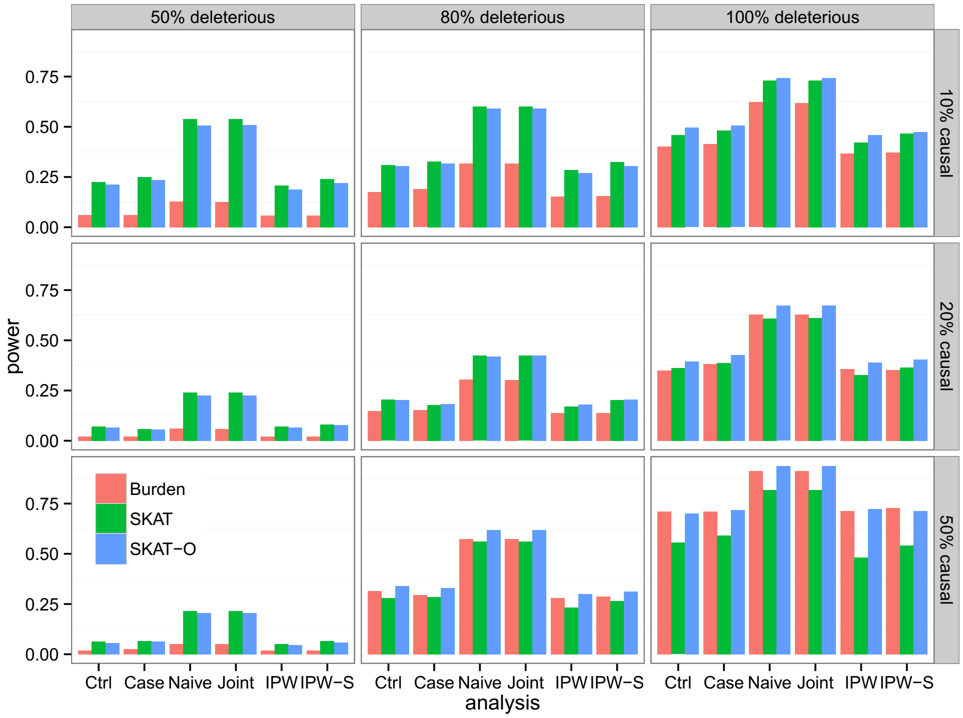

Figure 2.1: Empirical power at $\alpha = 2.5 \times 10^{-6}$ of methods for testing an association between randomly selected 3 kb regions with a continuous secondary trait. From top to bottom, the plots consider settings in which 10%, 20% and 50% of rare variants were causally associated with the secondary trait. From left to right, the plots consider settings in which 50%/50%, 80%/20%, and 100%/0% of the causal variants were deleterious/protective. The secondary trait and variants are assumed to be not associated with the disease, i.e. $\beta_Y = 0$ and $\beta_{Gj} = 0$ for all j. Sample size is fixed at 2000 cases and 2000 controls.

(Kelemen et al., 2008).

Boundaries of 12 target regions were defined by recombination hotspots flanking SNPs with published genome-wide significant associations to breast cancer risk. These regions contained 75 genes, ranging in length from 0.5 to 770 kb. A total of 27,102 variants across the 75 genes, including variants in exons, introns, and UTRs, were observed in 467 breast cancer cases and 591 controls. 22,364 of the 27,102 observed variants (83%) had MAF < 0.03. All subjects were female and of European ancestry.

We first applied a rank-based inverse normal transformation to MD (percentage of dense breast tissue). We then considered each gene separately and tested the association between variants in each gene and transformed MD. All 6 methods used in the simulation studies were applied, adjusting for age (year), body mass index, menopause (yes or no), and top 10 principal components of ancestry. To calculate subject-specific weights for IPW and AIPW SKAT-O, we used statistics from the National Cancer Institute (2012) and the 2010 US Consensus to estimate the prevalence of breast cancer at 2.67%.

To determine which ad hoc adaptations of SKAT-O might or might not in fact be valid, we fitted the null model for the primary trait association analysis by regressing breast cancer status on age, body mass index, menopause, top 10 principal components of ancestry, and transformed MD. The estimated positive increase in log-odds of breast cancer risk per unit increase in transformed MD (OR=1.62, p-value $= 4.02 \times 10^{-9}$) confirmed what is well known, that having dense breast tissue increases risk of getting breast cancer (Vachon et al., 2007). We also tested the association between variants in each gene and breast cancer status. Based on what we learned from simulation studies, we concluded that ad hoc analysis of a gene was valid only if there was little evidence of association between variants in the gene and breast cancer risk.

The top 5 genes identified by IPW-S SKAT-O are shown in Table 2.3. The results show that IPW-S SKAT-O was often the most powerful test. Four of the five genes (*AC008937.3*, *AC026462.2*, *ATE1* and *NXN1*) were mild to strongly associated with breast cancer risk. Therefore, ad hoc analyses for these genes are likely invalid. Although no gene was significant after Bonferroni correction (p-value $< 0.05/75$), there is reported evidence of association with MD for variant rs3803662 on *TOX3* (Fernandez-Navarro et al., 2013). This variant was included in our set of *TOX3* SNPs. However, since rs3803662 is a common variant (MAF=0.29), more than 96% of the other SNPs on *TOX3* are rare (MAF < 0.03), and $Beta(1, 25)$ weights were used in IPW-S SKAT-O to upweight rare variants and downweight common variants, the association here between MD and *TOX3* is driven by rare variants and independent of the previously identified rs3803662.

To explore the individual variants within the top 5 genes and their effects, we performed single-variant association analysis. Specifically, we applied IPW linear regression on transformed mammographic density, adjusting for a single variant, age, body mass index, menopause, and top 10 principal components. t-statistics based on the estimated variant effects were computed (Figures 2.2(a)-(e)). There is no clear evidence that *AC026462.2*, *ATE1*, and *TOX3* have variants with opposing or similar effects. However, in *AC008937.3*

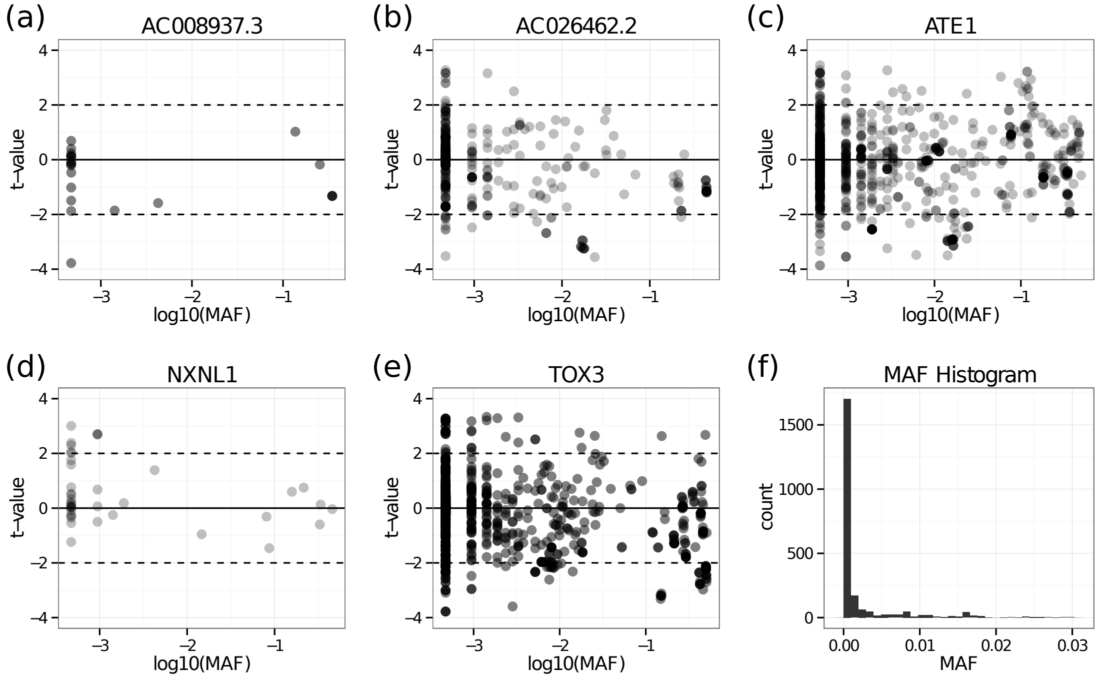

Figure 2.2: Single variant analysis results from IPW linear regression of transformed mammographic density. (a)-(e) Plots of log10(MAF) verses t-statistic values of each variant for the top 5 genes identified by the small-sample adjusted IPW SKAT-O. The dashed line represents the 95% confidence interval of no association. (f) Histogram of minor allele frequencies for 2,307 variants with MAF< 0.03.

and *NXNL1*, the majority of variants have noticeable effects in the same direction. 2,307 of the 2,670 variants (86.4%) observed in the top five genes have MAF < 0.03. The histogram of the estimated allele frequencies of the 2,307 variants with MAF < 0.03 is presented in Figure 2.2(f) and indicates that the majority of variants are extremely rare.

2.4 Discussion

In this chapter, we propose a weighted version of SKAT-O to account for the biased sampling when testing for rare variant effects on secondary traits using case-control sequencing data. As an extension of SKAT-O, our proposed IPW SKAT-O includes both IPW burden and IPW SKAT as special cases. IPW SKAT-O is computationally efficient and easily adjusts for covariates such as age, gender, and principal components for population stratification. We show in simulation studies that using SKAT-O ad hoc can result in highly inflated type I error rates. Meanwhile, IPW SKAT-O maintains good control of type I error. We also show that IPW SKAT and IPW burden tests can lose power when underlying assumptions are violated. In contrast, IPW SKAT-O is more robust in the wide range of circumstances we considered.

In simulation and real data analysis, we used a flexible beta weight to upweight the influence of rarer variants. In addition to using a function of the MAF of variants as weights, we can also consider choosing variants to be tested or constructing weights based on functional information. For example, evolutionary biologists use computational tools like PhastCons, PhyloP, and GERP to identify genetic regions that show preferential conservation across evolutionary time. A variant that shows strong selective constraint might be deemed likely functional, and therefore given a larger weight in SNP-set analyses. However, it is important to recognize that there is no universal definition of what constitutes function. As a result, there is currently a diverse set of functional annotations for every genetic variant. The challenge of integrating different pieces of functional information to obtain a comprehensive picture of the biological relevance of a variant is the topic of Chapter 3.

Table 2.3: Top 5 genes from small-sample adjusted IPW SKAT-O analysis of transformed mammographic density. p-values from all 6 implementations of SKAT-O for secondary trait analysis are presented, as well as p-values from applying SKAT-O to study associations between each gene and the primary trait breast cancer status. The number of observed variants in each gene and selected ρ values by SKAT-O are included in the second column and in parentheses, respectively.

		Breast cancer	**Mammographic density**					
Gene	No.	Primary	Naïve	Control-only	Case-only	Joint	IPW	IPW-S
AC008937.3	22	1.02×10^{-1}	1.83×10^{-2}	1.17×10^{-2}	6.33×10^{-1}	2.21×10^{-2}	2.67×10^{-2}	2.08×10^{-2}
		$(\rho = 0.0)$	$(\rho = 0.5)$	$(\rho = 1.0)$	$(\rho = 0.0)$	$(\rho = 1.0)$	$(\rho = 1.0)$	$(\rho = 1.0)$
AC026462.2	437	1.73×10^{-1}	2.21×10^{-2}	3.94×10^{-2}	8.03×10^{-1}	5.37×10^{-2}	1.41×10^{-2}	1.05×10^{-2}
		$(\rho = 1.0)$	$(\rho = 0.0)$	$(\rho = 0.0)$	$(\rho = 1.0)$	$(\rho = 0.0)$	$(\rho = 0.0)$	$(\rho = 0.0)$
ATE1	1147	1.62×10^{-1}	8.99×10^{-1}	3.14×10^{-2}	1.71×10^{-1}	7.43×10^{-1}	4.17×10^{-2}	3.94×10^{-2}
		$(\rho = 1.0)$	$(\rho = 1.0)$	$(\rho = 0.1)$	$(\rho = 0.0)$	$(\rho = 1.0)$	$(\rho = 0.1)$	$(\rho = 0.1)$
NXNL1	41	5.43×10^{-2}	1.83×10^{-1}	1.11×10^{-2}	4.65×10^{-1}	2.53×10^{-1}	4.14×10^{-2}	3.83×10^{-2}
		$(\rho = 0.0)$	$(\rho = 1.0)$	$(\rho = 1.0)$	$(\rho = 1.0)$	$(\rho = 0.1)$	$(\rho = 1.0)$	$(\rho = 1.0)$
TOX3	1023	6.51×10^{-1}	1.01×10^{-1}	1.63×10^{-2}	6.91×10^{-1}	9.53×10^{-2}	9.28×10^{-3}	7.34×10^{-3}
		$(\rho = 0.0)$	$(\rho = 1.0)$	$(\rho = 1.0)$	$(\rho = 0.0)$	$(\rho = 1.0)$	$(\rho = 1.0)$	$(\rho = 1.0)$

Chapter 3

Multivariate mixed models for predicting functional regions in the human genome

Godwin Yuen Han Yung[1], Iuliana Ionita-Laza[2], Xihong Lin[1]

[1]Department of Biostatistics, Harvard T.H. Chan School of Public Health

[2]Department of Biostatistics, Columbia University

3.1 Introduction

Since the completion of the human genome sequence, substantial effort has been put into identifying and annotating its functional DNA elements. With no universal definition of what constitutes function, we now have for any genetic variant, whether protein coding or noncoding, a diverse set of functional annotations. For example, the computational tool PolyPhen (Adzhubei et al., 2010) predicts damaging effects of missense mutations. Other tools such as phastCons (Siepel et al., 2005), PhyloP (Siepel et al., 2006), and GERP++ (Davydov et al., 2010) leverage comparative sequence information by looking for regions that show preferential conservation across evolutionary time. The Encyclopedia of DNA Elements (ENCODE) is a large-scale genomic project that has mapped regions of transcription, transcription factor association, chromatin structure and histone modification, effectively assigning biochemical functions for 80% of the genome (The ENCODE Project Consortium, 2012). Although the set of available functional annotations vary considerably with respect to the specific elements they predict and the extent of the human genome annotated by each, it is well understood that they provide complementary lines of evidence (Kellis et al., 2014). Therefore, in order to obtain a comprehensive picture of the biological relevance of genomic segments, all of the information acquired by the different annotations need to be taken into account.

Several unsupervised statistical learning algorithms have been proposed recently which integrate large, diverse sets of annotations into single measures of functional importance for a variant (Lu et al., 2015; Ionita-Laza et al., 2016). Compared to existing supervised algorithms (Kircher et al., 2014), these approaches do not rely on any labelled training data. This is particularly advantageous because with our current limited knowledge of non-coding regions, labelled training data are inevitably biased. Summarizing multiple annotations with a single measure also enables easy application. For instance, GenoCanyon (Lu et al., 2015) integrates a collection of 22 comparative genomic conservation scores and biochemical signals from the ENCODE project to calculate the posterior probability of a genomic position being functional. These posterior probabilities can help guide researchers in prioritizing variants for association analysis.

Existing unsupervised algorithms are not without limitations. GenoCanyon does not fully take into account correlations between the functional scores. Instead, it assumes that all annotations for a variant are conditionally independent given the variant's functional status. Another algorithm, EIGEN (Ionita-Laza et al., 2016), is arguably more robust. In estimating the predictive accuracy for each annotation, it assumes that the annotations are block-wise conditionally independent given, again, the variant's functional status. However, when EIGEN derives its final aggregate functional meta-score for each variant as a weighted linear combination of the individual annotations, the applied weights only take into account predictive accuracy of the annotations but not correlation between the annotations. In practice, certain sets of annotations measure similar elements (e.g., evolution conservation) and are therefore highly correlated.

In addition, GenoCanyon and EIGEN assume function as a binary outcome (functional or non-functional), which may be over-simplistic and unrealistic. The concept of functionality may be defined similarly between annotations measuring similar elements. However, what constitutes function may be quite different between annotations measuring different elements.

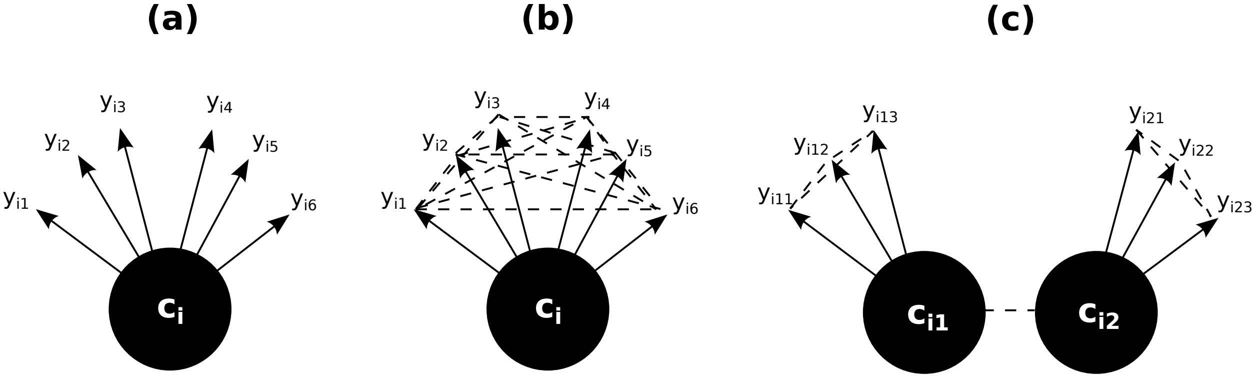

Figure 3.1: Models for the causal relations among variables. **(a)** All annotations y_{ij} (e.g., conservation measures, open chromatin data) are treated as consequences of a single latent dichotomous variable of function c_i. Annotations are assumed to be independent conditional on c_i. **(b)** All annotations y_{ij} are treated as consequences of c_i. Annotations may be correlated conditional on c_i. **(c)** There are multiple, possibly related, latent dichotomous variables of function $c_{i1}, ..., c_{iM}$. For each functional status c_{ij}, a subset of annotations $y_{ij1}, ..., y_{ijL_j}$ are observed as consequences. Annotations measuring the same c_{ij} may be correlated conditional on c_{ij}.

A single binary outcome cannot effectively summarize multiple, complementary concepts of functionality.

In this chapter, we propose to use a mixed model approach to integrate multiple annotations (Figure 3.1). Our model defines function as a vector of binary outcomes, each meant to capture functionality defined by a specific group of annotations. It also allows for correlations within and between the different groups of annotations. Using the EM algorithm, our approach calculates the posterior probability of a genomic position being functional.

We show that failure to take into account correlations between the functional scores can result in an algorithm that is biased in favor of larger groups of correlated annotations. We also show that if one assumes function is a binary outcome, then one also assumes that all annotations are unconditionally correlated. This is in stark contrast to the observed correlation structure between available functional annotations. Finally, we apply the proposed algorithm to real annotations of non-coding and synonymous variants.

3.2 Methods

3.2.1 Study setting and notation

Suppose that for SNP i and annotation group j, we have a set $\mathbf{y}_{ij} = (y_{ij1}, ..., y_{ijL_j})'$ of L_j annotations. Each SNP has $L = \sum_{j=1}^{M} L_j$ annotations in total. We wish to estimate the binary functional statuses $\mathbf{c}_i = (c_{i1}, ..., c_{iM})$ corresponding to each group. Conditionally on c_{ij} and the random effect variable b_{ijk}, assume that the elements of $\mathbf{y}_i$ are independent observations, each from a one-parameter exponential family with canonical parameterization and a mean that is a function of c_{ij} and b_{ijk}. That is, for $j = 1, ..., M$ and $k = 1, ..., L_j$ $m = 1, ..., M$,

$$f_{jk}(y_{ijk}|c_{ij}, b_{ijk}) = \exp[\{y_{ijk}\eta_{ijk} - d_{jk}(\eta_{ijk})\}/\phi_{jk} + h_{jk}(y_{ijk}, \phi_{jk})] \tag{3.1}$$

with

$$\mu_{ijk} = E(y_{ijk}) = d'_{jk}(\eta_{ijk}),$$
$$V_{ijk} = \text{var}(y_{ijk}) = d''_{jk}(\eta_{ijk})\phi_{jk},$$

where $\eta_{ijk} = g_{jk}(\mu_{ijk})$ is a linear function of the functional status c_{ij} and random effect variable b_{ijk} such that

$$\eta_{ijk} = \beta_{0jk} + c_{ij}\beta_{1jk} + b_{ijk} = \mathbf{x}_{ij}^T \boldsymbol{\beta}_{jk} + b_{ijk}$$

for $\mathbf{x}_{ij} = (1, c_{ij})^T$ and $\boldsymbol{\beta}_{jk} = (\beta_{0jk}, \beta_{1jk})^T$. To allow for additional correlations between elements of $\mathbf{y}_{ij}$, we assume that

$$\mathbf{b}_{ij} = \begin{pmatrix} b_{ij1} \\ b_{ij2} \\ \vdots \\ b_{ijL_j} \end{pmatrix} \overset{iid}{\sim} MVN(\mathbf{0}, \boldsymbol{\Sigma}_j(\boldsymbol{\theta}))$$

The marginal distribution of $\mathbf{y}_i$ can be obtained by integrating over the distribution of c_i and $\mathbf{b}_i$,

$$f(\mathbf{y}_i) = \sum_{c_{i1}=0,...,c_{iM}=0}^{1,...,1} \left(\prod_{j=1}^{M} \int f(\mathbf{y}_{ij}|c_{ij}, \mathbf{b}_{ij}) f(\mathbf{b}_{ij}, \boldsymbol{\theta}) \mathrm{d}\mathbf{b}_{ij} \right) p(c_{i1}, ..., c_{iM}) \tag{3.2}$$

The primary focus is on calculation of $p(\mathbf{c}_i|\mathbf{y}_i)$, the posterior probability of $\mathbf{c}_i$ conditional on the observed data. Because of the conditional independence of $\mathbf{y}_i$ given $\mathbf{c}_i$ and $\mathbf{b}_i$, an EM algorithm (Dempster et al., 1977) provides a natural approach. However, the integration necessary in equation (3.2) cannot be evaluated in closed form for many choices of $f(\mathbf{y}_{ij}|c_{ij},\mathbf{b}_{ij})$. An EM algorithm becomes complicated then because calculating $p(\mathbf{c}_i|\mathbf{y}_i) = f(\mathbf{y}_i|\mathbf{c}_i)p(\mathbf{c}_i)/f(\mathbf{y}_i)$ involves this integration. As described in more detail, we apply an EM algorithm using approximations where necessary for the required expectations with respect to the posterior distribution.

In the following, we let $\mathbf{1}_m$ be the vector of length m with each element being one and let $\mathbf{J}_m$ be the $m \times m$ matrix of ones, i.e. $\mathbf{J}_m = \mathbf{1}_m\mathbf{1}_m^T$. We let $\mathbf{I}_m$ be the $m \times m$ identity matrix. Subscripts are dropped whenever the dimensions of the vector or matrix is obvious. Our derivations follow closely that of Sammel et al. (1997), who considered a general class of latent variable models that allows for linear effects of covariates on multiple outcomes.

3.2.2 The EM algorithm

Maximization step

If c_i and b_{im} were directly observable we would simply maximize the complete data log-likelihood,

$$\log f(\mathbf{y},\mathbf{c},\mathbf{b}) = \sum_{i=1}^{N}\left(\sum_{j=1,k=1}^{M,L_j}\log f_{jk}(y_{ijk}|c_{ij},b_{ijk};\boldsymbol{\beta}_{jk},\phi_{jk}) + \sum_{j=1}^{M}\log f(\mathbf{b}_{ij};\boldsymbol{\theta}) + \log p(\mathbf{c}_i;\boldsymbol{\gamma})\right) \tag{3.3}$$

to estimate the unknown parameters $\boldsymbol{\zeta} = (\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\gamma},\boldsymbol{\theta})$. Since $\mathbf{c}_i$ and $\mathbf{b}_i$ are unobservable, the EM algorithm can be applied by solving instead the expected score functions, where the expectation is taken with respect to the posterior distribution

$$f(\mathbf{c}_i,\mathbf{b}_i|\mathbf{y}_i) = f(\mathbf{b}_i|\mathbf{y}_i,\mathbf{c}_i)p(\mathbf{c}_i|\mathbf{y}_i) = \prod_{j=1}^{M} f(\mathbf{b}_{ij}|\mathbf{y}_{ij},c_{ij})\cdot p(\mathbf{c}_i|\mathbf{y}_i)$$

of the missing data, conditionally on the observed data (Little and Rubin, 1987). In particular, if we let $S_i(\zeta)$ denote the complete data score function $\partial \log f(\mathbf{y}_i,\mathbf{c}_i,\mathbf{b}_i)/\partial\zeta$, then each

SNP's contribution to the expected score function for $\boldsymbol{\gamma}$ is given by

$$E_{c,b}S_i(\gamma_{z_1,...,z_M}) = \frac{p(c_{i1} = z_1, ..., c_{iM} = z_M|\mathbf{y}_i)}{\gamma_{z_1,...,z_M}} \tag{3.4}$$

for all $(z_1, ..., z_M) \in \{0, 1\}^M$. Therefore,

$$\widehat{\gamma}^{(k+1)}_{z_1,...,z_M} = \frac{\sum_{i=1}^N \widehat{p}^{(k)}(c_{i1} = z_1, ..., c_{iM} = z_M|\mathbf{y}_i)}{N} \tag{3.5}$$

For $\boldsymbol{\zeta}_{jk}$, the subset of parameters corresponding to the jkth outcome, the contribution to the expected score equation for SNP i is

$$E_{c,b}S_i(\boldsymbol{\zeta}_{jk}) = \sum_{\mathbf{c}_i \in \{0,1\}^M} \left(\int S_i(\boldsymbol{\zeta}_{jk}) f(\mathbf{b}_i|\mathbf{y}_i, \mathbf{c}_i) \; \mathrm{d}\mathbf{b}_i \right) p(\mathbf{c}_i|\mathbf{y}_i) \tag{3.6}$$

Depending on the form of the score function associated with the complete data log-likelihood $S(\boldsymbol{\zeta}_{jk}) = \sum_{i=1}^N S_i(\boldsymbol{\zeta}_{jk})$, the solution to $E_{c,b}S(\boldsymbol{\zeta}_{jk}) = \mathbf{0}$ may or may not be available in closed form. In the absence, of a closed form solution, we update the estimates $\boldsymbol{\zeta}_{jk}$ by using a one-step Fisher scoring algorithm. The usual method of estimation for this model is iteratively reweighed least squares (McCullagh and Nelder, 1989) where the weight function is updated at every iteration.

Expectation step

Given the current estimates of the parameters, $\boldsymbol{\zeta} = (\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\gamma}, \boldsymbol{\theta})$, the E-step is complicated by the need to compute expectations with respect to the posteriors distributions $f(\mathbf{b}_i|\mathbf{y}_i, \mathbf{c}_i)$ and $p(\mathbf{c}_i|\mathbf{y}_i)$ of the missing data, conditionally on the observed data. Only for normal outcomes will the posterior distributions have closed form solutions. First consider the Monte Carlo approximation. If we could generate a sample $(\mathbf{c}_1, \mathbf{b}_1), ..., (\mathbf{c}_T, \mathbf{b}_T)$ directly from $f(\mathbf{b}_i|\mathbf{y}_i, \mathbf{c}_i)$ and $p(\mathbf{c}_i|\mathbf{y}_i)$, we could estimate the expectation of functions of the data $g(\mathbf{c}_i, \mathbf{b}_i) = g(\mathbf{y}_i, \mathbf{c}_i, \mathbf{b}_i)$ by

$$E_{c,b}g(\mathbf{c}_i, \mathbf{b}_i) = \sum_{\mathbf{c}_i \in \{0,1\}^M} \left(\int g(\mathbf{y}_i, \mathbf{c}_i, \mathbf{b}_i) f(\mathbf{b}_i|\mathbf{y}_i, \mathbf{c}_i) \; \mathrm{d}\mathbf{b}_i \right) p(\mathbf{c}_i|\mathbf{y}_i) \approx \frac{1}{T} \sum_{t=1}^T g(\mathbf{y}_i, \mathbf{c}_t, \mathbf{b}_t) \tag{3.7}$$

However, in our setting, there are generally no closed form expression for $f(\mathbf{b}_i|\mathbf{y}_i, \mathbf{c}_i)$ and $p(\mathbf{c}_i|\mathbf{y}_i)$. Rewriting the posterior distributions as

$$f(\mathbf{b}_i|\mathbf{y}_i, \mathbf{c}_i) = \prod_{j=1}^M f(\mathbf{y}_{ij}|c_{ij}, \mathbf{b}_{ij}) f(\mathbf{b}_{ij}) \Big/ \int f(\mathbf{y}_{ij}|c_{ij}, \mathbf{b}_{ij}) f(\mathbf{b}_{ij}) \; \mathrm{d}\mathbf{b}_{ij},$$

$$p(\mathbf{c}_i|\mathbf{y}_i) = \frac{\prod_{j=1}^{M}\left[\int f(\mathbf{y}_{ij}|c_{ij},\mathbf{b}_{ij})f(\mathbf{b}_{ij})\ \mathrm{d}\mathbf{b}_{ij}\right]\cdot p(\mathbf{c}_i)}{\sum_{\mathbf{c}\in\{0,1\}^M}\prod_{j=1}^{M}\left[\int f(\mathbf{y}_{ij}|c_{j},\mathbf{b}_{ij})f(\mathbf{b}_{ij})\ \mathrm{d}\mathbf{b}_{ij}\right]\cdot p(\mathbf{c})}$$

and substituting into equation (3.7) we obtain

$$E_{c,b}g(\mathbf{c}_i,\mathbf{b}_i) = \frac{\sum_{\mathbf{c}_i\in\{0,1\}^M}\left[\int g(\mathbf{y}_i,\mathbf{c}_i,\mathbf{b}_i)\prod_{j=1}^{M} f(\mathbf{y}_{ij}|c_{ij},\mathbf{b}_{ij})f(\mathbf{b}_{ij})\ \mathrm{d}\mathbf{b}_i\right]\cdot p(\mathbf{c}_i)}{\sum_{\mathbf{c}_i\in\{0,1\}^M}\prod_{j=1}^{M}\left[\int f(\mathbf{y}_{ij}|c_{ij},\mathbf{b}_{ij})f(\mathbf{b}_{ij})\ \mathrm{d}\mathbf{b}_{ij}\right]\cdot p(\mathbf{c}_i)} \tag{3.8}$$

Now, given a sample of $\mathbf{b}_i$s generated from the distribution $f(\mathbf{b}_i)$, the Monte Carlo approximation of each integral, top and bottom, is

$$E_{c,b}g(\mathbf{c}_i,\mathbf{b}_i) = \frac{\sum_{\mathbf{c}_{t_1}\in\{0,1\}^M}\left[\sum_{t_2=1}^{T_2} g(\mathbf{y}_i,\mathbf{c}_{t_1},\mathbf{b}_{t_2})\prod_{j=1}^{M} f(\mathbf{y}_{ij}|c_{t_1 j},\mathbf{b}_{t_2 j})\right]\cdot p(\mathbf{c}_{t_1})}{\sum_{\mathbf{c}_{t_1}\in\{0,1\}^M}\left[\sum_{t_2=1}^{T_2}\prod_{j=1}^{M} f(\mathbf{y}_{ij}|c_{t_1 j},\mathbf{b}_{t_2 j})\right]\cdot p(\mathbf{c}_{t_1})} \tag{3.9}$$

Since this approximation is based on the weak law of large numbers, and our quantity is a ratio of two approximate integrals, a large value of T_2 is needed to yield precise estimates. In practice, this method may be quite slow.

If $g(\mathbf{y}_i,\mathbf{c}_i,\mathbf{b}_i) = g(\mathbf{y}_{ij'},c_{ij'},\mathbf{b}_{ij'})$ for some $j'\in\{1,...,M\}$, then the integral in the numerator of Equation (3.8) is equivalent to

$$\prod_{j=1}^{M}\left[\int g(\mathbf{y}_{ij'},c_{ij'},\mathbf{b}_{ij'})^{1(j=j')}f(\mathbf{y}_{ij}|c_{ij}\mathbf{b}_{ij})f(\mathbf{b}_{ij})\ \mathrm{d}\mathbf{b}_{ij}\right] \tag{3.10}$$

where $1(j=j')$ is equal to 1 if $j=j'$ and 0 otherwise. In this case, an alternative to Monte Carlo approximation is to use multivariate Gauss-Hermite quadratures. To approximate quantity (3.10), select T fixed abscissae $\{z_t\}_{t=1}^{T}$ and corresponding weights $\{w_t\}_{t=1}^{T}$ for a quadrature whose integration kernel is given by the density of a standard normal distribution (Abramowitz and Stegun, 1987). Given the spectral decomposition of $\mathbf{\Sigma}_j = \mathbf{S}_j\mathbf{\Lambda}_j\mathbf{S}_j^T$, let $\sigma_{jt} = \{\sigma_{jt}(1),...,\sigma_{jt}(L_j)\}$ be an ordered set of L_j integers obtained by sampling with replacement from $\{1,...,T\}$, $\mathbf{z}_{jt} = (z_{\sigma_{jt}(1)},...,z_{\sigma_{jt}(L_j)})^T$ the corresponding set of abscissae, and $\mathbf{b}_{jt} = \mathbf{S}_j\mathbf{\Lambda}_j^{1/2}\mathbf{z}_{jt}$. Then each term in the product (3.10)

$$\int g(\mathbf{y}_{ij'},c_{ij'},\mathbf{b}_{ij'})^{1(j=j')}f(\mathbf{y}_{ij}|c_{ij}\mathbf{b}_{ij})f(\mathbf{b}_{ij})\ \mathrm{d}\mathbf{b}_{ij}$$

can be approximated as

$$\sum_{\sigma_{jt}}\left(\prod_{k=1}^{L_j} w_{\sigma_{jt}(k)}\right) g(\mathbf{y}_{ij'},c_{ij'},\mathbf{b}_{j't})^{1(j=j')}f(\mathbf{y}_{ij}|c_{ij},\mathbf{b}_{jt})$$

where the sum is over all the possible ordered sets σ_{jt}. For some ordered sets σ_{jt}, the weights $\prod_{k=1}^{L_j} w_{\sigma_{jt}(k)}$ are very small, and thus contribute little to the sum. We may choose to remove these quantities by pruning a specified fraction of the smallest weights.

3.2.3 CAMM: EM algorithm for mixture of binary and normal annotations

The general formulation (3.1) allows different link functions $g_{jk}(\cdot)$ for different annotations and different covariance structures $\mathbf{\Sigma}_j(\boldsymbol{\theta})$ to accommodate for correlations between the annotations. In this section, we derive specific results for the EM algorithm when annotations are either conditionally bernoulli or normal random variables, i.e. all link functions $g_{jk}(\cdot)$ are either the identity or logistic link. We also introduce restrictions on the covariance matrices $\mathbf{\Sigma}_j(\boldsymbol{\theta})$ that allow for accurate approximations while greatly reducing the algorithm's computational speed. We call this algorithm CAMM for combined annotation mixed models.

Suppose that conditionally on c_{ij} and $\mathbf{b}_{ij}$, the first $L_j^{(1)}$ of the L_j outcomes $\mathbf{y}_{ij}$ follow a bernoulli distribution and the remaining $L_j^{(2)} = L_j - L_j^{(1)}$ outcomes follow a normal distribution. That is, for $k = 1, 2, ..., L_j^{(1)}$, y_{ijk} has distribution

$$f_{jk}(y_{ijk}|c_{ij}, b_{ijk}) = \exp[y_{ijk}\eta_{ijk} - \log\{1 + \exp(\eta_{ijk})\}]$$

where $\mu_{ijk} = \exp(\eta_{ijk})/\{1 + \exp(\eta_{ijk})\}$ and $V_{ijk} = \mu_{ijk}(1 - \mu_{ijk})$, and for $k = L_j^{(1)} + 1, L_j^{(1)} + 2, ..., L_j$, y_{ijk} has distribution

$$f_{jk}(y_{ijk}|c_{ij}, b_{ijk}) = \exp[\{y_{ijk}\mu_{ijk} - \mu_{ijk}^2/2\}/\phi_{jk} - \frac{1}{2}\{y_{ikl}^2/\phi_{jk} + \log(2\pi\phi_{jk})\}]$$

where $\mu_{ijk} = \eta_{ijk}$ and $V_{ijk} = \phi_{jk}$.

If $\mathbf{\Sigma}_j$ is left unstructured, then the EM algorithm will need to estimate $L_j(L_j + 1)/2$ parameters for the covariance matrix of group j. An even greater computational challenge is that the multivariate Gauss-Hermite quadrature will require T^{L_j} fixed abscissas. For $L_j = 22$ and $T = 10$, that amounts to 253 parameters and 10 sextillion abscissas. Thus, to reduce the number of model parameters and to make the algorithm computationally feasible, we assume that $\mathbf{b}_{ij} = \mathbf{\Lambda}_j \mathbf{f}_{ij}$ where $\mathbf{f}_{ij}$ is an unobserved vector of length $P_j < L_j$ that follows

$MVN(0, \mathbf{I})$. Then for the E-step,

$$\int g(\mathbf{y}_{ij'}, c_{ij'}, \mathbf{b}_{ij'})^{1(j=j')} f(\mathbf{y}_{ij}|c_{ij}\mathbf{b}_{ij}) f(\mathbf{b}_{ij}) \text{ d}\mathbf{b}_{ij}$$

$$\int g(\mathbf{y}_{ij'}, c_{ij}, \mathbf{b}_{ij'})^{1(j=j')} f(\mathbf{y}_{ij}|c_{ij}, \mathbf{b}_{ij}) f(\mathbf{b}_{ij}) \text{ d}\mathbf{b}_{ij} = \int g(\mathbf{y}_{ij}, c_{ij}, \mathbf{b}_{ij}) f(\mathbf{y}_{ij}|c_{ij}, \mathbf{b}_{ij}) f(\mathbf{f}_{ij}) \text{ d}\mathbf{f}_{ij}$$

so that integration is over a P_j-dimensional space as opposed to an L_j-dimensional space. The assumption $\mathbf{b}_{ij} = \mathbf{\Lambda}_j \mathbf{f}_{ij}$ forms the basis of factor analysis models (Lawley and Maxell, 1962) and is appropriate when the relationship between L_j manifest variables is thought to be primarily a result of the relationship between P_j latent variables. For functional annotations, the latent variables are likely to correspond to difference approaches measuring the same element. As in factor analysis, the larger the factor loading λ_{jkp}, the more the jkth annotation is said to "load" on the pth factor.

For the L_{j1} binary outcomes, substituting the appropriate quantities into equation (3.6) leads to the following expected score functions for variant i on outcome jk:

$$E_{c,b}S_i(\boldsymbol{\beta}_{jk}) = \sum_{c_{ij}=0}^{1} \left(\int \mathbf{x}_{ij}(y_{ijk} - \mu_{ijk}) \cdot f(\mathbf{b}_{ij}|\mathbf{y}_{ij}, c_{ij}) \text{ d}\mathbf{b}_{ij} \right) p(c_{ij}|\mathbf{y}_i) \tag{3.11}$$

$$E_{c,b}S_i(\mathbf{\Lambda}_{jk}) = \sum_{c_{ij}=0}^{1} \left(\int \mathbf{f}_{ij}(y_{ijk} - \mu_{ijk}) \cdot f(\mathbf{b}_{ij}|\mathbf{y}_{ij}, c_{ij}) \text{ d}\mathbf{b}_{ij} \right) p(c_{ij}|\mathbf{y}_i) \tag{3.12}$$

To update estimates for $\boldsymbol{\beta}_{jk}$ using a one-step Fisher scoring algorithm, consider a Taylor series expansion of the expected score function (3.11) about the true parameter $\boldsymbol{\beta}_{jk}$,

$$E_{c,b}S_i(\widehat{\boldsymbol{\beta}}_{jk}) \approx E_{c,b}S_i(\boldsymbol{\beta}_{jk}) + \left\{ \frac{\partial}{\partial \boldsymbol{\beta}_{jk}^T} E_{c,b}S_i(\boldsymbol{\beta}_{jk}) \right\} (\widehat{\boldsymbol{\beta}}_{jk} - \boldsymbol{\beta}_{jk}).$$

Since $E_{c,b}S(\widehat{\boldsymbol{\beta}}_{jk}) = \sum_{i=1}^N E_{c,b}S_i(\widehat{\boldsymbol{\beta}}_{jk}) = \mathbf{0}$ and assuming regularity conditions that allow the interchange of differentiation and integration, we have

$$E_{c,b}S(\boldsymbol{\beta}_{jk}) \approx \left\{ \sum_{i=1}^{N} I_i(\boldsymbol{\beta}_{jk}) \right\} (\widehat{\boldsymbol{\beta}}_{jk} - \boldsymbol{\beta}_{jk}),$$

where I_i is the ith SNP's contribution to the observed complete data Fisher information associated with the jkth outcome:

$$I_i(\boldsymbol{\beta}_{jk}) = -\frac{\partial}{\partial \boldsymbol{\beta}_{jk}^T} E_{c,b}S_i(\boldsymbol{\beta}_{jk}) = -\sum_{\mathbf{c}_i \in \{0,1\}^M} \left[\int \frac{\partial}{\partial \boldsymbol{\beta}_{jk}^T} S_i(\boldsymbol{\beta}_{jk}) f(\mathbf{b}_i|\mathbf{y}_i, \mathbf{c}_i) d\mathbf{b}_i \right] \cdot p(\mathbf{c}_i|\mathbf{y}_i).$$

The expected information is obtained by taking an additional expectation with respect to the observed outcomes $\mathbf{y}_i$:

$$J_i(\boldsymbol{\beta}_{jk}) = -E_{\mathbf{y}_i}\frac{\partial}{\partial \boldsymbol{\beta}_{jk}^T}E_{c,b}S_i(\boldsymbol{\beta}_{jk}).$$

Interchanging derivatives and expectations yield

$$J_i(\boldsymbol{\beta}_{jk}) = -\sum_{\mathbf{c}_i\in\{0,1\}^M}\left[\int E_{\mathbf{y}_i}\left\{\frac{\partial}{\partial \boldsymbol{\beta}_{jk}^T}S_i(\boldsymbol{\beta}_{jk})\right\}f(\mathbf{b}_i|\mathbf{y}_i,\mathbf{c}_i)\ \mathrm{d}\mathbf{b}_i\right]\cdot p(\mathbf{c}_i|\mathbf{y}_i)$$

For binary outcomes with logistic link, the expected information is

$$J_i(\boldsymbol{\beta}_{jk}) = \sum_{\mathbf{c}_i\in\{0,1\}^M}\left[\int \mathbf{x}_{ij}\mu_{ijk}(1-\mu_{ijk})\mathbf{x}_{ij}^T f(\mathbf{b}_i|\mathbf{y}_i,\mathbf{c}_i)\mathrm{d}\mathbf{b}_i\right]\cdot p(\mathbf{c}_i|\mathbf{y}_i) \tag{3.13}$$

Equations (3.11) and (3.13) yield the following scoring algorithm at iteration $r+1$:

$$\widehat{\boldsymbol{\beta}}_{jk}^{r+1} = \widehat{\boldsymbol{\beta}}_{jk}^{(r)} + \left(\sum_{i=1}^N E_{c,b}[\mathbf{x}_{ij}\widehat{\mu}_{ijk}^{(r)}(1-\widehat{\mu}_{ijk}^{(r)})\mathbf{x}_{ij}^T]\right)^{-1}\sum_{i=1}^N E_{c,b}[\mathbf{x}_{ij}(y_{ijk}-\widehat{\mu}_{ijk}^{(r)})] \tag{3.14}$$

Similarly,

$$\widehat{\boldsymbol{\Lambda}}_{jk}^{r+1} = \widehat{\boldsymbol{\Lambda}}_{jk}^{(r)} + \left(\sum_{i=1}^N E_{c,b}[\mathbf{f}_{ij}\widehat{\mu}_{ijk}^{(r)}(1-\widehat{\mu}_{ijk}^{(r)})\mathbf{f}_{ij}^T]\right)^{-1}\sum_{i=1}^N E_{c,b}[\mathbf{f}_{ij}(y_{ijk}-\widehat{\mu}_{ijk}^{(r)})] \tag{3.15}$$

For the $L_j^{(2)}$ normal outcomes, contributions to the complete data score functions for each SNP i are

$$\begin{aligned} S_i(\boldsymbol{\beta}_{jk}) &= \frac{1}{\phi_{jk}}\mathbf{x}_{ij}e_{ijk} \\ S_i(\boldsymbol{\Lambda}_{jk}) &= \frac{1}{\phi_{jk}}\mathbf{f}_{ij}e_{ijk} \\ S_i(\phi_{jk}) &= -\frac{1}{2\phi_{jk}} + \frac{1}{2\phi_{jk}^2}e_{ijk}^2 \end{aligned}$$

where $e_{ijk} = y_{ijk} - \mathbf{x}_{ij}^T\boldsymbol{\beta}_{jk} - \mathbf{f}_{ij}^T\boldsymbol{\Lambda}_{jk}$. It follows that

$$\widehat{\boldsymbol{\beta}}_{jk}^{(r+1)} = \left(\sum_{i=1}^N E_{c,b}[\mathbf{x}_{ij}\mathbf{x}_{ij}^T]\right)^{-1}\sum_{i=1}^N E_{c,b}[\mathbf{x}_{ij}(y_{ijk}-\mathbf{f}_{ij}^T\widehat{\boldsymbol{\Lambda}}_{jk}^{(r)})] \tag{3.16}$$

$$\widehat{\mathbf{\Lambda}}_{jk}^{(r+1)} = \left(\sum_{i=1}^{N} E_{c,b}[\mathbf{f}_{ij}\mathbf{f}_{ij}^T]\right)^{-1} \sum_{i=1}^{N} E_{c,b}[\mathbf{f}_{ij}(y_{ijk} - \mathbf{x}_{ij}^T \widehat{\boldsymbol{\beta}}_{jk}^{(r)})] \tag{3.17}$$

$$\widehat{\phi}_{jk} = \frac{1}{N}\sum_{i=1}^{N} E_{c,b}[\widehat{e}_{ijk}^2] \tag{3.18}$$

Beginning with reasonable initial estimates of the parameters, CAMM proceeds by first using the E-step to obtain the desired expectations relative to the posterior distribution. Given those estimates, we then solve the expected score equations to obtain new parameter estimates or one-step updates. The algorithm proceeds until the relative change in the estimated parameters is sufficiently small.

3.2.4 GenoCanyon

GenoCanyon is a special case of CAMM: it assumes that there is a single group of annotations and that the annotations are independent conditional on the univariate functional status c_{i1}, i.e. $M = 1$ and $\mathbf{\Sigma}_1 = \mathbf{0}$. However, assuming a single binary functional status implies that all annotations are unconditionally correlated. Also, failure to take into account correlations between annotations, conditional on the functional status, can result in an algorithm that is biased in favor of certain types of functional measures. These concepts are best demonstrated by a simple example.

Suppose we have two normally distributed annotations, y_{i11} and y_{i12}, equally informative in predicting c_{i1} but independent conditional on c_{i1}. For simplicity, we drop the jth index, e.g. $y_{i1k} = y_{ik}$. Then, since

$$cov(y_{i1}, y_{i2}) = \beta_1\beta_2 var(c_i),$$

the annotations are unconditionally correlated provided that $var(c_i) > 0$, i.e. there are functional and non-functional variants present. Meanwhile, the posterior probability of c_i given y_{i1} and y_{i2} is

$$p(c_i|y_{i1}, y_{i2}) = \frac{f(y_{i1}|c_i) \cdot f(y_{i2}|c_i) \cdot p(c_i)}{\sum_{c=0}^{1} f(y_{i1}|c) \cdot f(y_{i2}|c) \cdot p(c)}.$$

Now suppose we have a third normally distributed annotation, y_{i3}, that is independent of y_{i2} but highly correlated with y_{i1}, conditionally on c_i. Intuitively, y_{i3} provides little additional information beyond what is already available from y_{i1} and y_{i2}. Thus, a sensible algorithm would calculate

$$p(c_i|y_{i1}, y_{i2}, y_{i3}) \approx p(c_i|y_{i1}, y_{i2}).$$

On the other hand, an algorithm that incorrectly assumes conditional independence between y_{i2} and y_{i3} would calculate

$$p(c_i|y_{i1}, y_{i2}, y_{i3}) \approx \frac{f(y_{i1}|c_i)^2 \cdot f(y_{i2}|c_i) \cdot p(c_i)}{\sum_{c=0}^{1} f(y_{i1}|c)^2 \cdot f(y_{i2}|c) \cdot p(c)}.$$

This effectively says that y_{i2} is twice as informative as y_{i1}, even though in reality the two annotations are equally informative. A similar argument can be made for EIGEN.

GenoCanyon integrates 22 different annotations to calculate posterior probabilities, including 2 genomic conservation measures, 2 indicators of open chromatin, 8 histone modifications, and 10 transcription factor binding site (TFBS) peaks. While some groups may be correlated with others, some may be independent. Also, annotations within each of the four groups of measures are likely correlated. In Figure 3.2, we show the empirical matrix of correlations between 7 conservation measures and 3 open chromatin data across non-variant or synonymous variants of chromosome 1. It is easy to see that there is minimal correlation between the two groups, but measures within each group are highly correlated. The block-diagonal correlation structure contradicts the assumption of a single binary functional status, because under this assumption, all groups should be unconditionally correlated. Correlation between measures of the same group and the unequal number of measures within each group also lead us to conclude that GenoCanyon is biased in favor of histone modifications and TFBS peaks, which have more annotations than genomic conservation measures and open chromatin.

3.3 Results

We downloaded a set of 7 evolutionary conservation annotations (GERP_RS, PhyloP, PhyloM, PhyloV, PhastP, PhastM, PhastV) and 3 open chromatin data (OCPval, PolIIPval, ctcfPval) for noncoding and synonymous coding variants on chromosome 1 from the UCSC genome browser (January 2015). GERP_RS, PhyloP, PhyloM, and PhyloV were continuous measures. The rest were probabilities. In total, there were 38,402 variants with scores for

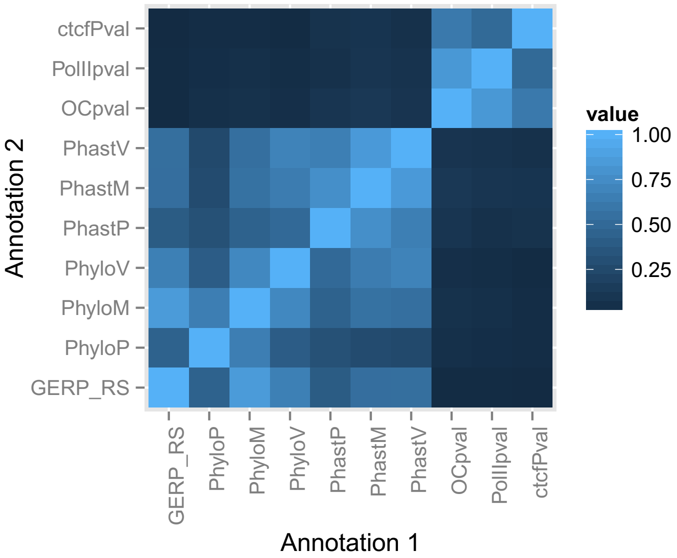

Figure 3.2: Matrix of correlations between 7 conservation scores and 3 open chromatin data.

all 10 annotations. The matrix of correlations between the annotations is provided in Figure 3.2. We chose from the list of functional annotations several subsets to integrate: (A) GERP_RS, PhyloP, PhyloM, PhyloV; (B) PhastP, PhastM, PhastV; (C) OCPval, PolIIPval, ctcfPval; (D) GERP_RS, PhyloP, PhyloM, PhyloV, PhastP, PhastM, PhastV; (E) PhastP, PhastM, PhastV, OCPval, PolIIPval, ctcfPval; and (F) all 10 annotations.

For subsets A-D, which included only annotations measuring similar elements, we defined function of a variant as a univariate bernoulli outcome. We considered defining function of a variant as a univariate as well as a bivariate bernoulli outcome for the remaining subsets E and F, which included both conservation measures and open chromatin data. To provide CAMM with initial parameter estimates, for each group j of annotations, we partitioned the variants into 2 clusters by either applying (1) 2-means clustering to the annotations of group j or (2) a threshold to one of the annotations in group j. We then fitted marginal linear or logistic regression models on each score, adjusting for the predicted cluster assignment. We experimented with different number of factors P_j for each conditional covariance matrix $\mathbf{\Sigma}_j$ of random effects.

The chosen number of factors as well as the estimated intercepts and slopes from CAMM are provided in Table 3.1. Many annotations provided information for predicting functional

Table 3.1: Parameter settings and estimates for Models A-F, including the method for calculating initial parameters (2-means clustering, thresholding, or both), dimension of functional status (univariate or bivariate), chosen number of factors for the conditional covariance matrices $\mathbf{\Sigma}_j$ of random effects, estimated intercepts β_{0jk}, and estimated slopes β_{1jk} for the effect of functional status c_{ij} on annotation y_{ijk}.

Model			A	B	C	D	E1	E2	F
Initial			both	both	both	both	2-means	thresh	both
Dim function			1	1	1	1	1	1	2
Num factors			2	1	1	2	1	1	2, 1
Conservation	GERP_RS	Intercept	-0.580			-0.494			-0.494
		Slope	4.948			4.913			4.916
	PhyloP	Intercept	-0.177			-0.166			-0.166
		Slope	0.059			0.570			0.570
	PhyloM	Intercept	-0.203			-0.170			-0.170
		Slope	2.110			2.084			2.085
	PhyloV	Intercept	-0.180			-0.152			-0.152
		Slope	6.903			6.850			6.845
	PhastP	Intercept		-2.268		-2.223	-2.122	-2.272	-2.223
		Slope		3.135		3.181	0.318	3.088	3.181
	PhastM	Intercept		-3.124		-5.191	-3.608	-3.150	-5.191
		Slope		5.536		11.748	0.658	5.380	11.758
	PhastV	Intercept		-2.775		-4.678	-3.191	-2.792	-4.677
		Slope		5.121		17.769	0.326	4.950	17.744
Open chromatin	OCPval	Intercept			-2.542		-2.619	-4.726	-2.542
		Slope			7.282		7.482	6.411	7.261
	PolIIPval	Intercept			-6.508		-6.391	-8.490	-6.559
		Slope			8.521		8.396	6.603	8.572
	ctcfPval	Intercept			-17.779		-13.604	-8.315	-17.928
		Slope			19.108		14.936	6.300	19.257

status, as indicated by the magnitude of the estimate slopes relative to the estimated conditional variances (not shown). In Figure A.3 (Appendix A.3.1), we show that CAMM accurately decomposed each of the four continuous scores in set A into a mixture of two normal distributions.

For models B, C, E1, and E2, which only used probability scores, the estimated loading factors converged to 0. Thus, it is reasonable to assume that the scores in these models are independent conditional on functional status. On the other hand, for models A, D, and F which used continuous scores, we found that annotations were correlated even after adjusting for functional status. In Table A.1 (Appendix A.3.1), we compare the empirical unconditional covariance matrix of scores in A to the estimated conditional covariance matrix $\widehat{\mathbf{\Sigma}}_1 + \text{diag}(\widehat{\phi}_{11}, ..., \widehat{\phi}_{14})$. Though the two matrices measure slightly different quantities, it is clear that CAMM accurately estimates the correlation structure between the different annotations and that the annotations are correlated conditional on the estimated functional status.

Table 3.2: Correlation between the posterior probabilities of functional status from applying CAMM to annotation sets A-E. Using different initial parameters and applying CAMM to annotation set E resulted in two converged models, E1 and E2.

	A	B	C	D	E1	E2
A	1.000	0.575	0.027	0.995	0.564	0.025
B	0.575	1.000	0.108	0.583	0.988	0.103
C	0.027	0.108	1.000	0.028	0.157	1.000
D	0.995	0.583	0.028	1.000	0.572	0.025
E1	0.564	0.988	0.157	0.572	1.000	0.152
E2	0.025	0.103	1.000	0.025	0.152	1.000

For models A-D, which defined function as a univariate bernoulli outcome and only included annotations measuring similar elements, using 2-means clustering or thresholding to propose initial parameters for the EM algorithm resulted in the same fitted model upon convergence. When applied to the set E of annotations, however, CAMM either had trouble converging or it converged to different models depending on the initial parameters. In Table 3.1, we see that, compared to models B and C, models E1 and E2 shrink the estimated slopes for certain annotations towards zero, but leave estimates for other annotations relatively unchanged. The posterior probabilities of function from models E1 and E2 were also near perfectly correlated with either the posteriors from B or the posteriors from C (Table 3.2). This indicates that, although set E consists of both conservation and open chromatin scores, assuming function as a single binary outcome resulted in models that, rather than integrating all the scores, used only one type of score, conservation or open chromatin, to predict functional status.

Model F used both conservation and open chromatin scores. However, unlike models E1 and E2, F assumed variant function to be a bivariate bernoulli outcome, the first outcome being functional status defined by conservation scores, the second being functional status defined by open chromatin data. The model was not sensitive to initial parameters. The estimated posterior probabilities of function were also highly discrete; the probability of function for each variant tended to be high for one of the four functional states (0,0), (0,1), (1,0), or (1,1), and low for the remaining states (Figure A.4 of Appendix A.3.1).

In Table 3.3, we estimate the percentage of variants in each state. A chi-squared test comparing the observed to the expected percentages under independence gives a significant

Table 3.3: Distribution of posterior probabilities, averaged across all ∼38,000 variants. The expected distribution assuming the two binary functional outcomes are independent is provided in parentheses.

		conservation		
		0	1	
open chromatin	0	68.71% (68.41%)	3.79% (4.09%)	72.50%
	1	25.65% (25.95%)	1.85% (1.55%)	27.50%
		94.36%	5.64%	

p-value of 1.5×10^{-8}. We estimate that $100 - 68.71 = 31.29\%$ of the variants are either (1,0), (0,1), or (1,1), that is, functional in terms of evolutionary conservation, chromatin structure, or both. Since the observed percentage of functional variants that are (1,1) is greater than that expected percentage under independence ($1.85\% > 1.55\%$), there is evidence of enrichment between conservation and open chromatin scores. Finally, we estimate that 5.64% and 27.50% of the variants are conserved and part of open chromatin, respectively. It is worth noting that comparative genomic studies estimate that 5% of mammalian genomes are under strong evolutionary constraint (Kellis et al., 2014).

3.4 Discussion

Recently proposed statistical frameworks including GenoCanyon and EIGEN assume that there exists a single latent dichotomous variable summarizing functional status. They also either ignore or do not fully take into account correlation between annotations conditional on function. In view of what biologists have observed, that the diverse set of available functional annotations provide complementary–rather than entirely overlapping–information, a more realistic model is one that assumes annotation scores are measured responses of multiple possibly related yet distinct latent variables. In this chapter, we proposed CAMM, a mixed model approach that allows for multiple, possibly correlated, binary functional statuses. Individually, each status captures functionality defined by a certain group of annotations. Annotations within each group can also be correlated conditional on function.

We demonstrated using real data that CAMM can integrate annotations of different types and predict a variant's posterior distribution of multivariate functional status. The posterior can be highly intuitive if it is clear how function is defined by each group of annotations. (In our data example, we grouped conservation measures into one group and open chromatin data into another. Therefore, function in the two groups corresponded to strong evolutionary constraint and open chromatin, respectively.) Also, if needed, the posterior can be conveniently summarized by single measures such as the "probability of function according to at least one group of annotations" or "probability of function according to all groups of annotations". This will be particularly useful with increasing number of groups.

CAMM is a flexible and informative approach for predicting functional regions in the genome. However, there are many ways in which we can apply CAMM, and more thought needs to be given on what annotations to include, how to group the annotations, and what assumptions can be reasonably made to increase computational speed. If we are interested in eventually identifying variants associated with some phenotype, then it may be desirable to use certain cell-type specific, species specific, or phenotype-related annotations to predict functionality. Also, certain groups of annotations may be correlated, for example, histone marker and open chromatin data. Should they constitute two different functions or just one? Finally, having a large number of groups may better reflect the true nature of the data. However, fitting an algorithm to estimate all the parameters may be computationally expensive. Therefore, although some groups of annotations may be correlated, it may be computationally advantageous to assume they are not and estimate the parameters of the two groups separately, especially if the correlation is weak.

Appendix

A.1 Chapter 1 Appendix

A.1.1 Derivations for common disease, binary secondary trait

Here, we determine the conditions under which r and r_d are approximately linear functions of $\mathbf{Z}$. First, note that r and r_d are functions of $\mathbf{Z}$ and $\mathbf{G}$ through the conditional means $\mu_D(1)$ and $\mu_D(0)$, which are themselves functions of $\eta = \Phi^{-1}(\mu_D(0)) = \beta_0 + \mathbf{Z}'\boldsymbol{\beta}_Z + \mathbf{G}'\boldsymbol{\beta}_G$. It follows that r and r_d are functions of η and we can write $r(\mathbf{Z},\mathbf{G}) = r(\eta)$ and $r_d(\mathbf{Z},\mathbf{G}) = r_d(\eta)$. We will use the different forms interchangeably. Now consider the second order Taylor series expansions of $r(\eta)$ and $r_d(\eta)$ centered at $\eta_0 = g_D(\kappa)$:

$$\begin{aligned} r(\eta) &= r(\eta_0) + r'(\eta_0)(\eta-\eta_0) + \frac{r''(\eta^*)}{2}(\eta-\eta_0)^2 \\ r_d(\eta) &= r_d(\eta_0) + r_d'(\eta_0)(\eta-\eta_0) + \frac{r_d''(\eta_d^*)}{2}(\eta-\eta_0)^2 \end{aligned}$$

where η^* and η_d^* are some real numbers between η and η_0. One can show that for $g_D(\cdot) = \text{logit}$ and $\kappa \in (0.1, 0.5)$,

$$\max_{\eta}\left|\frac{r''(\eta)}{2}\right| \approx \begin{cases} \frac{1}{4}\left(1 - \frac{\kappa}{\widetilde{P}(D=1)}\right)|\beta_Y| & \text{if } \kappa \le \widetilde{P}(D=1) \\ \frac{1}{2}(\kappa - \widetilde{P}(D=1))|\beta_Y| & \text{if } \kappa > \widetilde{P}(D=1) \end{cases}$$

and

$$\left|\frac{r_d''(\eta)}{2}\right| < \frac{1}{20}|\beta_Y| \quad \forall \eta.$$

Similar bounds can be found for $g_D(\cdot) = \Phi^{-1}$ and, in general, any smooth $g_D(\cdot)$. These bounds suggest that if $\boldsymbol{\beta}_G = \mathbf{0}$ and $|\beta_Y|$ and $|\eta - \eta_0|$ are not exceedingly large (i.e., Y and $\mathbf{Z}$ are not strongly associated with D), then the quadratic terms in the Taylor expansions will be small, and the remainders $r(\eta)$ and $r_d(\eta)$ will be approximately linear in $\eta = \beta_0 + \mathbf{Z}'\beta_Z$.

An interesting aside: $r(\eta)$ becomes increasingly linear in η as κ tends to $\widetilde{P}(D=1)$. In fact, if $\kappa = \widetilde{P}(D=1)$, then $\pi(0) = \pi(1)$ and it is easy to show from Equation (3) in the article that $r(\cdot)$ is exactly equal to 0. This result reflects the notion that a naïve analysis is valid when the study population is a random sample of the general population. Of course, this condition is not true in the setting of case-control studies.

When $r_0(\cdot)$ and $r_1(\cdot)$ are linear functions of $\mathbf{Z}$, the control-only and case-only analyses can be applied to estimate and make inference on $\boldsymbol{\alpha}_G$. An adjusted analysis is also valid if, in addition, $r_1(\cdot) - r_0(\cdot)$ is a constant. It is easy to show that this required condition is true for $g_D(\cdot) = \text{logit}$ and approximately true for $g_D(\cdot) = \Phi^{-1}$ when the disease is common. Specifically, if $g_D(\cdot) = \text{logit}$, then

$$r_1(\mathbf{Z}, \mathbf{G}) - r_0(\mathbf{Z}, \mathbf{G}) = \beta_Y.$$

Meanwhile, the probit and logit link functions are very close in the mid-range. For η such that $\Phi(\eta) \in (0.2, 0.8)$, the standard normal cumulative distribution can be approximated accurately by a transformed logistic distribution $\Phi(\eta) \approx \text{expit}(\eta/\lambda)$. Popular choices for λ include $\sqrt{3}/\pi$ and $5/8$ [Amemiya, 1981]. This approximation implies that for common disease and $g_D(\cdot) = \Phi^{-1}$,

$$r_1(\mathbf{Z}, \mathbf{G}) - r_0(\mathbf{Z}, \mathbf{G}) \approx \beta_Y/\lambda.$$

For $g_D(p) = \log(-\log(1-p))$, it can be shown, by taking a second order Taylor series expansion of

$$T(\eta + \beta_Y) = \log\left\{\frac{g_D^{-1}(\eta + \beta_Y)}{1 - g_D^{-1}(\eta + \beta_Y)}\right\}$$

centered at η, that

$$r_1(\mathbf{Z}, \mathbf{G}) - r_0(\mathbf{Z}, \mathbf{G}) = T(\eta + \beta_Y) - T(\eta) \approx T'(\eta^*)\beta_Y + 0.5T''(\eta^*)\beta_Y^2 \approx 1.3\beta_Y + 0.22\beta_Y^2.$$

A.1.2 Derivations for common disease, continuous secondary trait

Here, we derive Equations (1.7) and (1.8) from the article and provide the closed form expressions for $\widetilde{\mu}_Y$ and $\widetilde{\sigma}^2$. We then determine the conditions under which $r(\cdot)$ and $r_d(\cdot)$ are approximately linear in $\mathbf{Z}$ and $\mathbf{X}$. First, suppose that θ is a parameter in $\mathbb{R}$ and $Y \sim N(\mu_Y + \theta\sigma^2, \sigma^2)$. Our interest is in calculating $E(Y|\mathbf{X}, \mathbf{G}, \mathbf{Z}, D, S = 1, \theta = 0)$ and

$Var(Y|\mathbf{X}, \mathbf{G}, \mathbf{Z}, D, S = 1, \theta = 0)$, but since $\widetilde{P}(Y|\mathbf{X}, \mathbf{G}, \mathbf{Z}, D) = P(Y|\mathbf{X}, \mathbf{G}, \mathbf{Z}, D)$, it suffices to calculate the mean and variance of $Y|\mathbf{X}, \mathbf{G}, \mathbf{Z}, D, \theta = 0$. With that in mind, assume $g_D = \Phi^{-1}$ and define D^* to be the random variable such that $D^* = g_D(\mu_D(Y)) + \varepsilon$, $\varepsilon \sim N(0, 1)$. Then $P(D^* > 0|\mathbf{Z}, \mathbf{G}, Y) = P(D = 1|\mathbf{Z}, \mathbf{G}, Y)$. Furthermore,

$$\begin{aligned}
P(D = 1|\mathbf{Z}, \mathbf{G}, \mathbf{X}, \theta) &= \int P(D = 1|\mathbf{Z}, \mathbf{G}, \mathbf{X}, \theta, y)P(y|\mathbf{X}, \mathbf{G}, \mathbf{Z}, \theta)dy \\
&= \int P(D^* > 0|\mathbf{Z}, \mathbf{G}, \mathbf{X}, \theta, y)P(y|\mathbf{X}, \mathbf{G}, \mathbf{Z}, \theta)dy \quad D|\mathbf{Z}, \mathbf{G}, y \perp\!\!\!\perp \mathbf{X}, \theta \\
&= P(D^* > 0|\mathbf{Z}, \mathbf{G}, \mathbf{X}, \theta) \\
&= P(\varepsilon^* > -g_D(\mu_D(0))) \\
&= \Phi(f(\theta))
\end{aligned}$$

where $\varepsilon^* \sim N\left((\mu_Y + \theta\sigma^2)\beta_Y, \sigma^2\beta_Y^2 + 1\right)$ and

$$f(\theta) = \frac{g_D(\mu_D(\mu_Y + \theta\sigma^2))}{\sqrt{\sigma^2\beta_Y^2 + 1}}.$$

We can now calculate the first two moments of $Y|\mathbf{X}, \mathbf{G}, \mathbf{Z}, D, \theta = 0$ via its moment generating function:

$$\begin{aligned}
E(e^{tY}|\mathbf{X}, \mathbf{G}, \mathbf{Z}, D = d, \theta = 0) &= \exp\left(t\mu_Y + \frac{t^2\sigma^2}{2}\right)\left\{\frac{\Phi(f(t))}{\Phi(f(0))}\right\}^d \left\{\frac{1 - \Phi(f(t))}{1 - \Phi(f(0))}\right\}^{1-d} \\
E(Y|\mathbf{X}, \mathbf{G}, \mathbf{Z}, D = d, \theta = 0) &= \mu_Y + (-1)^{1-d} \cdot c \cdot \frac{\phi(f(0))}{\{\Phi(f(0))\}^d\{1 - \Phi(f(0))\}^{1-d}} \\
E(Y^2|\mathbf{X}, \mathbf{G}, \mathbf{Z}, D = d, \theta = 0) &= \sigma^2 + (-1)^{1-d} \cdot c^2 \cdot \frac{\phi'(f(0))}{\{\Phi(f(0))\}^d\{1 - \Phi(f(0))\}^{1-d}} \\
&\quad + \mu_Y^2 + 2\mu_Y(-1)^{1-d} \cdot c \cdot \frac{\phi(f(0))}{\{\Phi(f(0))\}^d\{1 - \Phi(f(0))\}^{1-d}}
\end{aligned}$$

The variance follows immediately:

$$\begin{aligned}
&Var(Y|\mathbf{X}, \mathbf{G}, \mathbf{Z}, D = d, \theta = 0) = \\
&\qquad \sigma^2 + c^2\left(\frac{(-1)^{1-d} \cdot \phi'(f(0))}{\{\Phi(f(0))\}^d\{1 - \Phi(f(0))\}^{1-d}} - \frac{\phi(f(0))^2}{\{\Phi(f(0))\}^{2d}\{1 - \Phi(f(0))\}^{2(1-d)}}\right)
\end{aligned}$$

Letting η denote $f(0)$ gives us Equations (1.7) and (1.8).

Next, to calculate $\widetilde{\mu}_Y$ and $\widetilde{\sigma}^2$, note that

$$\widetilde{P}(D|\mathbf{Z},\mathbf{G},\mathbf{X},\theta) = \frac{P(S=1|D)\cdot P(D|\mathbf{Z},\mathbf{G},\mathbf{X},\theta)}{\sum_{d=0}^{1} P(S=1|D=d)\cdot P(D=d|\mathbf{Z},\mathbf{G},\mathbf{X},\theta)}.$$

Therefore,

$$\begin{aligned} E(Y|\mathbf{X},\mathbf{G},\mathbf{Z},S=1,\theta=0) &= \mu_Y + c\cdot\phi(\eta)\cdot g(\mathbf{Z},\mathbf{G},\mathbf{X}) \\ &= \mu_Y + r(\mathbf{Z},\mathbf{G},\mathbf{X}) \\ Var(Y|\mathbf{X},\mathbf{G},\mathbf{Z},S=1,\theta=0) &= \sigma^2 + c^2\cdot\left\{\phi'(\eta)\cdot g(\mathbf{Z},\mathbf{G},\mathbf{X}) - \phi(\eta)^2\cdot g(\mathbf{Z},\mathbf{G},\mathbf{X})^2\right\} \\ &= \sigma^2 + s(\mathbf{Z},\mathbf{G},\mathbf{X}) \end{aligned}$$

where

$$g(\mathbf{Z},\mathbf{G},\mathbf{X}) = \frac{P(S=1|D=1) - P(S=1|D=0)}{\sum_{d=0}^{1} P(S=1|D=d)\cdot P(D=d|\mathbf{Z},\mathbf{G},\mathbf{X},\theta=0)}.$$

Finally, we determine the conditions under which $r(\cdot)$ and $r_d(\cdot)$ are approximately linear in $\mathbf{Z}$ and $\mathbf{X}$, and s and s_d are approximately constants. We begin by again noting that the all remainders are a function of η, which is itself a linear function of $\mathbf{Z}$, $\mathbf{G}$, and $\mathbf{X}$. Thus, we can write $r(\mathbf{Z},\mathbf{G},\mathbf{X}) = r(\eta)$, $r_d(\mathbf{Z},\mathbf{G},\mathbf{X}) = r_d(\eta)$, $s(\mathbf{Z},\mathbf{G},\mathbf{X}) = s(\eta)$, and $s_d(\mathbf{Z},\mathbf{G},\mathbf{X}) = s_d(\eta)$. If $\boldsymbol{\alpha}_G = \boldsymbol{\beta}_G = \mathbf{0}$, then the remainders are functions of $\mathbf{Z}$ and $\mathbf{X}$ alone. Meanwhile, consider the second and first order order Taylor series expansions of $r_d(\eta)$ and $s_d(\eta)$ centered at $\eta_0 = g_D(\kappa)/\sqrt{\sigma^2\beta_Y^2+1}$. One can show that in these expansions the quadratic and linear coefficients are bounded:

$$\left|\frac{r_d''(\eta)}{2}\right| < \frac{3}{20}|c|$$

and

$$|s_d'(\eta)| < \frac{3}{10}c^2$$

for all η and $d = 0, 1$. Similar bounds can be derived for $r(\eta)$ and $s(\eta)$. Therefore, if $\boldsymbol{\alpha}_G = \boldsymbol{\beta}_G = \mathbf{0}$ and $|\beta_Y|$ and $|\eta-\eta_0|$ are not exceedingly large (i.e., Y and $\mathbf{Z}$ are not strongly associated with D and $\mathbf{X}$ is not strongly associated with Y), then the quadratic and linear terms in the Taylor expansion of $r(\eta)$, $r_d(\eta)$, $s(\eta)$, and $s_d(\eta)$ will be small, $r(\eta)$ and $r_d(\eta)$ will be approximately linear in η—hence in $\mathbf{X}$ and $\mathbf{Z}$—and $s(\eta)$ and $s_d(\eta)$ we be approximately constant.

An adjusted analysis is unbiased if, in addition, $\boldsymbol{\alpha}^{**}_{Z0} = \boldsymbol{\alpha}^{**}_{Z1}$ and $\boldsymbol{\alpha}^{**}_{X0} = \boldsymbol{\alpha}^{**}_{X1}$ or, equivalently, $r_1(\cdot) - r_0(\cdot)$ is a constant. It is easy to show that this required condition is approximately true for common disease by using the logit approximation for the probit:

$$r_1(\mathbf{X}, \mathbf{G}, \mathbf{Z}) - r_0(\mathbf{X}, \mathbf{G}, \mathbf{Z}) \approx c/\lambda$$

While $s_0(\eta)$ is generally not equal to $s_1(\eta)$, in our simulations, the difference between the sample variance of the case-only and control-only analyses with pooled covariates seemed to be small enough for inference to be approximately correct.

A.1.3 Derivations for rare disease

Here, we derive the theoretical bias for the case-only analysis with pooled covariates when the disease is rare and $g_D = \Phi^{-1}$. If Y is binary, then

$$\begin{aligned}
\lim_{\eta\to-\infty} r_1'(\eta) &= \lim_{\eta\to-\infty} \frac{\phi(\eta+\beta_Y)}{\Phi(\eta+\beta_Y)} - \frac{\phi(\eta)}{\Phi(\eta)} \\
&= \lim_{\eta\to-\infty} \frac{\phi(\eta)}{\Phi(\eta)} + \frac{\phi'(\eta^*)\Phi(\eta^*) - \phi(\eta^*)^2}{\Phi(\eta^*)^2}\beta_Y - \frac{\phi(\eta)}{\Phi(\eta)} \quad \eta^* \text{ between } \eta \text{ and } \eta+\beta_Y \\
&= \beta_Y \cdot \lim_{\eta\to-\infty} \frac{\phi'(\eta)\Phi(\eta) - \phi(\eta)^2}{\Phi(\eta)^2} \\
&= \beta_Y \cdot \lim_{\eta\to-\infty} \frac{\eta\phi(\eta) + (\eta^2-1)\Phi(\eta)}{2\Phi(\eta)} \qquad \text{L'Hopital's rule} \\
&= \beta_Y \cdot \lim_{\eta\to-\infty} \frac{\eta\Phi(\eta)}{\phi(\eta)} \qquad \text{L'Hopital's rule} \\
&= -\beta_Y
\end{aligned}$$

and

$$\begin{aligned}
\lim_{\eta\to-\infty} r_1''(\eta) &= \lim_{\eta\to-\infty} \frac{d}{d(\eta+\beta_Y)} \frac{\phi(\eta+\beta_Y)}{\Phi(\eta+\beta_Y)} - \frac{d}{d(\eta)} \frac{\phi(\eta)}{\Phi(\eta)} \\
&= (-1) - (-1) \\
&= 0.
\end{aligned}$$

It follows that

$$\lim_{\eta\to-\infty} r_1(\eta) = r_1(\Phi^{-1}(\kappa)) - \beta_Y(\eta - \Phi^{-1}(\kappa)),$$

or equivalently,

$$\lim_{\kappa\to0} r_1(\mathbf{Z}, \mathbf{G}) = \left\{r_1(\Phi^{-1}(\kappa)) + \beta_Y(\Phi^{-1}(\kappa) - \beta_0)\right\} - \mathbf{Z}'\beta_Y\boldsymbol{\beta}_Z - \mathbf{G}'\beta_Y\boldsymbol{\beta}_G.$$

If instead Y is continuous, then because $\lim_{\eta \to -\infty} \frac{\phi(\eta)}{\eta\Phi(\eta)} = -1$,

$$r_1(\mathbf{Z}, \mathbf{G}, \mathbf{X}) = c\frac{\phi(\eta)}{\Phi(\eta)} \approx -c\eta.$$

Meanwhile,

$$\lim_{\kappa \to 0} s_1(\mathbf{Z}, \mathbf{G}, \mathbf{X}) = \lim_{\eta \to -\infty} c^2 \cdot \left\{ \frac{\phi'(\eta)\Phi(\eta) - \phi(\eta)^2}{\Phi(\eta)^2} \right\} = -c^2.$$

A.2 Chapter 2 Appendix

A.2.1 Using SKAT-O ad hoc

Define the working vector $\mathbf{Y}^* = \mathbf{X}\boldsymbol{\alpha}_X + \boldsymbol{\Delta}(\mathbf{Y} - \boldsymbol{\mu})$ where $\boldsymbol{\Delta} = \text{diag}\{g'(\mu_i)\}$, and the variance matrix $\mathbf{V} = \text{diag}\{\phi v(\mu_i)[g'(\mu_i)]^2\}$. Their estimates under the null hypothesis are $\widehat{\mathbf{Y}}^* = \mathbf{X}\widehat{\boldsymbol{\alpha}}_X + \widehat{\boldsymbol{\Delta}}(\mathbf{Y} - \widehat{\boldsymbol{\mu}})$, $\widehat{\boldsymbol{\Delta}} = \text{diag}\{g'(\widehat{\mu}_i)\}$, and $\widehat{\mathbf{V}} = \text{diag}\{\widehat{\phi} v(\widehat{\mu}_i)[g'(\widehat{\mu}_i)]^2\}$, where $\widehat{\boldsymbol{\alpha}}_X$ and $\widehat{\phi}$ are obtained by using generalized linear regression to fit the null model $g(\mu_i) = \mathbf{X}\boldsymbol{\alpha}_X$. Given that $g(\cdot)$ is a canonical link function,

$$Q_\rho = (\mathbf{Y} - \widehat{\boldsymbol{\mu}})'\mathbf{K}_\rho(\mathbf{Y} - \widehat{\boldsymbol{\mu}})/\widehat{\phi}^2 = (\mathbf{Y} - \widehat{\boldsymbol{\mu}})'\widehat{\boldsymbol{\Delta}}\widehat{\mathbf{V}}^{-1}\mathbf{K}_\rho\widehat{\mathbf{V}}^{-1}\widehat{\boldsymbol{\Delta}}(\mathbf{Y} - \widehat{\boldsymbol{\mu}})/\widehat{\phi}^2.$$

Now, if $\widehat{\mathbf{V}}^{-1/2}\widehat{\boldsymbol{\Delta}}(\mathbf{Y} - \widehat{\boldsymbol{\mu}}) \sim MVN(\mathbf{0}, \mathbf{I})$, then it can be easily shown that $Q_\rho \sim \sum \lambda_j \chi^2_{1j}$ where λ_j's are eigenvalues of $\widehat{\mathbf{V}}^{-1/2}\mathbf{K}_\rho\widehat{\mathbf{V}}^{-1/2}$ and χ^2_{1j}'s are independent χ^2_1 random variables. A p-value for Q_ρ can then be calculated. The requirement $\widehat{\mathbf{V}}^{-1/2}\widehat{\boldsymbol{\Delta}}(\mathbf{Y} - \widehat{\boldsymbol{\mu}}) \sim MVN(\mathbf{0}, \mathbf{I})$ is satisfied when the study subjects are a random sample of the population or Y is the disease status in a case-control study. However, it is not necessarily satisfied when Y is a secondary trait in a case-control study.

Indeed, $\widehat{\mathbf{V}}^{-1/2}\widehat{\boldsymbol{\Delta}}(\mathbf{Y} - \widehat{\boldsymbol{\mu}}) \sim MVN(\mathbf{0}, \mathbf{I})$ if and only if the score

$$\mathbf{X}'\widehat{\mathbf{V}}^{-1}\widehat{\boldsymbol{\Delta}}(\mathbf{Y} - \widehat{\boldsymbol{\mu}}) \sim MVN(\mathbf{0}, \mathbf{X}'\widehat{\mathbf{V}}^{-1}\mathbf{X})$$

if and only if generalized linear regression can be applied to properly estimate and make inference on $\boldsymbol{\alpha}_G$. Conditions under which generalized linear regression can be applied to estimate and make inference on $\boldsymbol{\alpha}_G$ have been studied previously [Lin and Zeng, 2009] and was the topic of Chapter 1. We summarize here without proof the conditions under which a standard p df test, and therefore Q_ρ and SKAT-O, can be applied ad hoc: if the secondary trait or rare variants is not associated with the disease (i.e. $\beta_Y = 0$ or $\boldsymbol{\beta}_G = \mathbf{0}$), then all ad hoc applications of Q_ρ and SKAT-O (e.g., methods (a)-(d)) can be used; if the disease is rare and is assume to follow a logistic model, then ad hoc applications of Q_ρ and SKAT-O that condition on disease status (e.g., (a), (b), and (d)) can be used; if, however, the disease is rare but is assumed to follow a non-logistic model such as the probit model, then only ad hoc applications of Q_ρ and SKAT-O based on only the controls (e.g., (a)) can be used.

A.2.2 Asymptotic distribution and small-sample mean of Q_ρ^* under H_0

Suppose $\boldsymbol{\alpha}_X$ and ϕ are known. Let s_i indicated with the values 1 versus 0 whether or not an individual from the target population is sampled in the case-control study. Use the subscript N to denote matrices or vectors for the target population, e.g., $\mathbf{Y}_N$, $\boldsymbol{\mu}_N$, $\mathbf{W}_N^*$, $\mathbf{K}_{\rho N}$, $\mathbf{S}_N$. Then

$$\begin{aligned} Q_\rho^* &= (\mathbf{Y} - \boldsymbol{\mu})'\mathbf{W}^*\mathbf{K}_\rho\mathbf{W}^*(\mathbf{Y} - \boldsymbol{\mu})/\phi^2 \\ &= (\mathbf{Y}_N - \boldsymbol{\mu}_N)'\mathbf{W}_N^*\mathbf{S}_N\mathbf{K}_{\rho N}\mathbf{S}_N\mathbf{W}_N^*(\mathbf{Y}_N - \boldsymbol{\mu}_N)/\phi^2. \end{aligned}$$

Since $E[s_i w_i^*(y_i - \mu_i)/\phi] = E[E\{s_i w_i^*(y_i - \mu_i)/\phi|D_i\}] = E[(y_i - \mu_i)/\phi] = 0$, it follows that asymptotically

$$\mathbf{S}_N\mathbf{W}_N^*(\mathbf{Y}_N - \boldsymbol{\mu}_N)/\phi \sim MVN(\mathbf{0}, \mathbf{V}_N^{-1}),$$

where $\mathbf{V}_N^{-1} = \text{diag}\{E[s_i w_i^{*2}(y_i - \mu_i)^2/\phi^2]\}$, and

$$Q_\rho^* \sim \sum \lambda_j \chi_{1,j}^2,$$

where $(\lambda_1, ..., \lambda_m)$ are the eigenvalues of $\mathbf{V}_N^{-1/2}\mathbf{K}_{\rho N}\mathbf{V}_N^{-1/2}$. In practice, the eigenvalues can be estimated with the eigenvalues of $\widehat{\mathbf{V}}^{-1/2}\mathbf{K}_\rho\widehat{\mathbf{V}}^{-1/2}$ where $\widehat{\mathbf{V}} = \text{diag}[\widehat{\phi}^2/w_1^{*2}/(y_1 - \widehat{\mu}_1)^2, ..., \widehat{\phi}^2/w_n^{*2}/(y_n - \widehat{\mu}_n)^2]$, which only uses data from the case-control samples. To calculate the mean of Q_ρ^*, let $(\mathbf{u}_1, ..., \mathbf{u}_m)$ be the eigenvectors of $\mathbf{V}_N^{-1/2}\mathbf{K}_{\rho N}\mathbf{V}_N^{-1/2}$. Then

$$Q_\rho^* = \sum \lambda_j(\mathbf{Y}_N - \boldsymbol{\mu}_N)'\mathbf{W}_N^*\mathbf{S}_N\mathbf{u}_j\mathbf{u}_j'\mathbf{S}_N\mathbf{W}_N^*(\mathbf{Y}_N - \boldsymbol{\mu}_N)/\phi^2$$

and $E(Q_\rho^*) = \sum \lambda_j$.

A.2.3 Null distribution of small-sample IPW SKAT-O

Define $\mathbf{Z} = \widehat{\mathbf{V}}^{-1/2}\mathbf{GW}$ and $\bar{\mathbf{z}} = (\bar{z}_1, ..., \bar{z}_n)'$, where $\bar{z}_i = \sum_{j=1}^{p} z_{ij}/p$. Additionally, let $\mathbf{M} = \bar{\mathbf{z}}(\bar{\mathbf{z}}'\bar{\mathbf{z}})^{-1}\bar{\mathbf{z}}'$ and

$$\tau(\rho) = p^2\rho\bar{\mathbf{z}}' + \frac{1-\rho}{\bar{\mathbf{z}}'\bar{\mathbf{z}}}\sum_{j=1}^{p}(\bar{\mathbf{z}}'\mathbf{z}_{\cdot j})^2,$$

where $\mathbf{z}_{\cdot j}$ is the jth column of $\mathbf{Z}$. Following the same argument in Lee and others (2012), it can be shown that Q_ρ^* is equivalent to

$$(1-\rho)\kappa_1 + \tau(\rho)\kappa_2$$

where

$$\kappa_1 = (1-\rho)\bar{\mathbf{y}}'(\mathbf{I}-\mathbf{M})\mathbf{ZZ}'(\mathbf{I}-\mathbf{M})\bar{\mathbf{y}} + 2(1-\rho)\bar{\mathbf{y}}'(\mathbf{I}-\mathbf{M})\mathbf{ZZ}'\mathbf{M}\bar{\mathbf{y}}$$

and

$$\kappa_2 = \frac{\bar{\mathbf{y}}'\bar{\mathbf{z}}\bar{\mathbf{z}}'\bar{\mathbf{y}}}{\bar{\mathbf{z}}'\bar{\mathbf{z}}}.$$

It can be shown that κ_2 asymptotically follows the χ_1^2 distribution, and κ_1 is asymptotically the same as $\sum_{k=1}^{m}\lambda_k\eta_k + \zeta$ where $\{\lambda_1, ..., \lambda_m\}$ are non-zero eigenvalues of $\mathbf{Z}'(\mathbf{I}-\mathbf{M})\mathbf{Z}$, $\eta_k(k = 0, ..., m)$ are independent and identically distributed χ_1^2 random variables, and ζ satisfies the following conditions: $E(\zeta) = 0$, $\mathrm{Var}(\zeta) = 4\mathrm{trace}(\mathbf{Z}'\mathbf{MZZ}'(\mathbf{I}-\mathbf{M})\mathbf{Z})$, $\mathrm{Corr}(\sum_{k=1}^{m}\lambda_k\eta_k, \zeta) = 0$, and $\mathrm{Corr}(\eta_0, \zeta) = 0$.

From there, asymptotic p-values can be obtained through one-dimensional integration. When the sample size is small, however, the asymptotic moments of κ_1 and κ_2 can be larger than their small-sample moments. Thus, we apply a small-sample adjustment procedure to the null distributions of κ_1 and κ_2 that is similar to the adjustment procedure we applied to the null distribution of Q_ρ^* in Section 2.2.4. Specifically, we compute the small-sample variance and kurtosis of κ_1 and κ_2 and apply the moment-matching approximation to obtain their adjusted asymptotic distribution. To obtain a p-value, we apply the algorithm in Lee and others (2012) with the null distributions of κ_1 and κ_2.

A.2.4 Additional simulation results

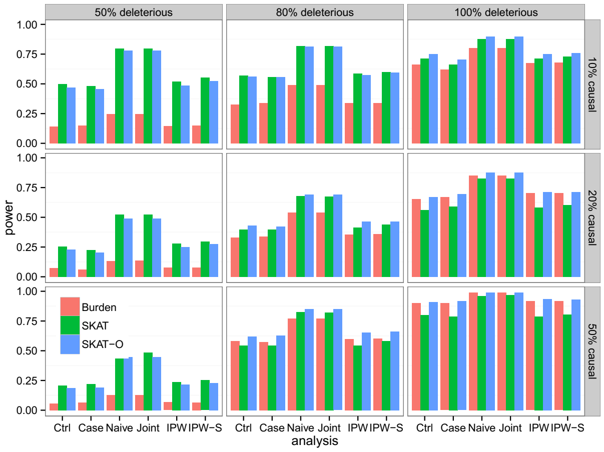

Figure A.1: Empirical power at $\alpha = 0.001$ of methods for testing an association between randomly selected 3 kb regions with a continuous secondary trait. From top to bottom, the plots consider settings in which 10%, 20% and 50% of rare variants were causally associated with the secondary trait. From left to right, the plots consider settings in which 50%/50%, 80%/20%, and 100%/0% of the causal variants were deleterious/protective. The secondary trait and variants are assumed to be not associated with the disease, i.e. $\beta_Y = 0$ and $\beta_{Gj} = 0$ for all j. Sample size is fixed at 2000 cases and 2000 controls.

A.3 Chapter 3 Appendix

A.3.1 Additional figures and tables

Table A.1: Empirical unconditional vs. estimated conditional covariance matrix of random effects for model A.

	Empirical				Estimated			
	GERP_RS	PhyloP	PhyloM	PhyloV	GERP_RS	PhyloP	PhyloM	PhyloV
GERP_RS	7.896	1.016	2.752	3.787	6.4325	0.9046	2.1486	1.9214
PhyloP	1.016	0.668	0.611	0.668	0.9046	0.6643	0.5534	0.4682
PhyloM	2.752	0.611	1.319	1.675	2.1486	0.5534	1.0784	0.8748
PhyloV	3.787	0.668	1.675	4.122	1.9214	0.4682	0.8748	1.5773

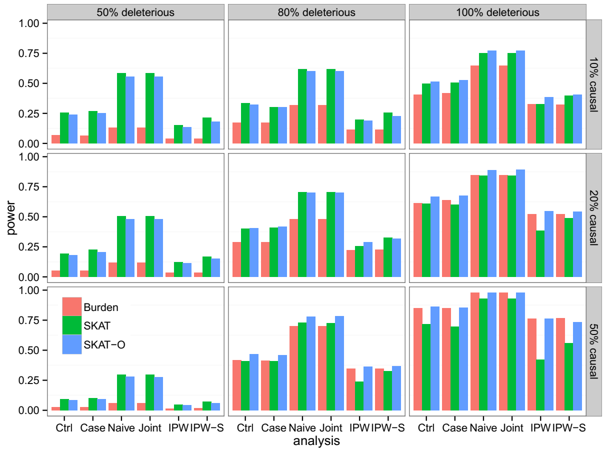

Figure A.2: Empirical power at $\alpha = 2.5 \times 10^{-6}$ of methods for testing an association between randomly selected 3 kb regions with a continuous secondary trait. From top to bottom, the plots consider settings in which 10%, 20% and 50% of rare variants were causally associated with the secondary trait. From left to right, the plots consider settings in which 50%/50%, 80%/20%, and 100%/0% of the causal variants were deleterious/protective. The secondary trait and variants are assumed to be not associated with the disease, i.e. $\beta_Y = 0$ and $\beta_{Gj} = 0$ for all j. Sample size is fixed at 1000 cases and 1000 controls.

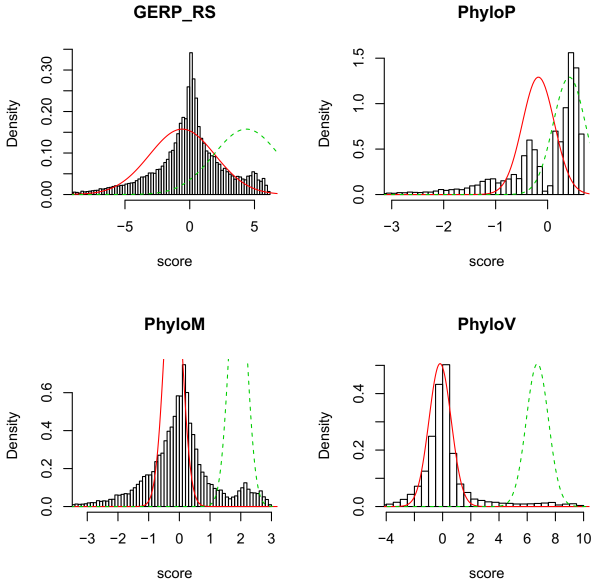

Figure A.3: Empirical distribution of functional scores for annotation set (A) and estimated conditional distributions using CAMM. Red and solid curves correspond to the estimated distributions for non-functional variants. Green and dotted curves correspond to the estimated distributions for functional variants.

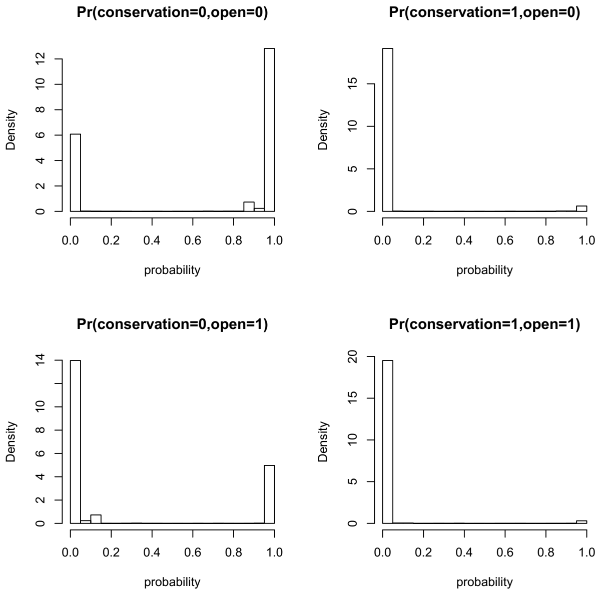

Figure A.4: Distribution of posterior probabilities for model F.

References

Abramowitz, M. and Stegun, I. A. (1987). *Handbook of Mathematical Functions.* Dover Publications, New York.

Adzhubei, I. A., Schmidt, S., Peshkin, L. et al. (2010). A method and server for predicting damaging missense mutations. *Nature Methods* **7** 248–249.

Amemiya, T. (1981). Qualitative Response Models: a survey. *Journal of Economic Literature* **29** 1483–1536.

Bansal, V., Libiger, O., Torkamani, A. et al. (2010). Statistical analysis strategies for association studies involving rare variants. *Nature Reviews Genetics* **11** 773–785.

Cirulli, E. T. and Goldstein, D. B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews Genetics* **11** 415–425.

Davies, R. B. (1980). Algorithm AS 155: the distribution of a linear combination of χ^2 random variables. *Journal of the Royal Statistical Society: Applied Statistics* **29** 323–333.

Davison, A. C. and Hinkley, D. V. (1999). *Bootstrap methods and their application.* Cambridge University Press, Cambridge.

Davydov, E. V., Goode, D. L., Sirota, M. et al. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Computational Biology* **6**.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **39** 1–38.

Fernandez-Navarro, P., Pita, G., Santamarina, C. et al. (2013). Association analysis between breast cancer genetic variants and mammographic density in a large population-based study (Determinants of Density in Mammographies in Spain) identifies susceptibility loci in TOX3 gene. *European Journal of Cancer* **49** 474–481.

He, J., Li, H., Edmondson, A. C. et al. (2012). A gaussian copula approach for the analysis of secondary phenotypes in case-control genetic association studies. *Biostatistics* **13** 497–508.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47** 663–685.

Ionita-Laza, I., McCallum, K., Xu, B. et al. (2016). A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nature Genetics* **48** 214–220.

Jiang, Y., Scott, A. J. and Wild, C. J. (2006). Secondary analysis of case-control data. *Statistics in Medicine* **25** 1323–1339.

Kelemen, L. E., Sellers, T. a. and Vachon, C. M. (2008). Can genes for mammographic density inform cancer etiology? *Nature Reviews Cancer* **8** 812–23.

Kellis, M., Wold, B., Snyder, M. P. et al. (2014). Defining functional DNA elements in the human genome. *PNAS* **111** 6131–8.

Kenny, E. E., Pe'er, I., Karban, A. et al. (2012). A genome-wide scan of Ashkenazi Jewish Crohn's disease suggests novel susceptibility loci. *PLoS genetics* **8** e1002559.

Kircher, M., Witten, D. M., Jain, P. et al. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics* **46** 310–315.

Lawley, D. N. and Maxell, A. E. (1962). Factor Analysis as a Statistical Method Author. *Journal of the Royal Statistical Society. Series D* **12** 209–229.

Lee, S., Emond, M. J., Bamshad, M. J. et al. (2012a). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *American Journal of Human Genetics* **91** 224–237.

Lee, S., Wu, M. C. and Lin, X. (2012b). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* **13** 762–775.

Lee, S. H., Wray, N. R., Goddard, M. E. et al. (2011). Estimating missing heritability for disease from genome-wide association studies. *American Journal of Human Genetics* **88** 294–305.

Li, B. and Leal, S. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *American Journal of Human Genetics* **83** 311–321.

Li, H., Gail, M. H., Berndt, S. et al. (2010). Using cases to strengthen inference on the association between single nucleotide polymorphisms and a secondary phenotype in genome-wide association studies. *Genetic Epidemiology* **34** 427–433.

Lin, D. Y. and Zeng, D. (2009). Proper analysis of secondary phenotype data in case-control association studies. *Genetic Epidemiology* **33** 256–265.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. Wiley, New York.

Liu, H., Tang, Y. and Zhang, H. H. (2009). A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. *Computational Statistics and Data Analysis* **53** 853–856.

Lu, Q., Hu, Y., Sun, J. et al. (2015). A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. *Nature Scientific Reports* **5**.

Madsen, B. E. and Browning, S. R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genetics* **5**.

Manolio, T. A., Collins, F. S., Cox, N. J. et al. (2009). Finding the missing heritability of complex diseases. *Nature* **461** 747–53.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models.* 2nd ed. Chapman and Hall, Florida.

Mensah-Ablorh, A., Lindstrom, S., Haiman, C. A. et al. (2016). Meta-analysis of rare variant association tests in multiethnic populations. *Genetic Epidemiology* **40** 57–65.

Monda, K. L., Chen, G. K., Taylor, K. C. et al. (2013). A meta-analysis identifies new loci associated with body mass index in individuals of African ancestry. *Nature Genetics* **45** 690–6.

Monsees, G. M., Tamimi, R. M. and Kraft, P. (2009). Genome-wide association scans for secondary traits using case-control samples. *Genetic Epidemiology* **33** 717–728.

Morris, A. P. and Zeggini, E. (2010). An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genetic Epidemiology* **34** 188–193.

Nagelkerke, N. J. D., Moses, S., Plummer, F. A. et al. (1995). Logistic-Regression in Case-Control Studies - the Effect of Using Independent as Dependent-Variables. *Statistics in Medicine* **14** 769–775.

Nolan, D. J., Han, D. Y., Lam, W. J. et al. (2010). Genetic adult lactase persistence is associated with risk of Crohn's Disease in a New Zealand population. *BMC Research Notes* **3** 339.

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

Price, A. L., Kryukov, G. V., de Bakker, P. I. W. et al. (2010). Pooled association tests for rare variants in exon-resequencing studies. *American Journal of Human Genetics* **86** 832–838.

Price, A. L., Patterson, N. J., Plenge, R. M. et al. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38** 904–909.

Richardson, D. B., Rzehak, P., Klenk, J. et al. (2007). Analyses of case-control data for additional outcomes. *Epidemiology* **18** 441–5.

Rose, J., Behm, F., Drgon, T. et al. (2010). Personalized smoking cessation: interactions between nicotine dose, dependence and quit-success genotype. *Molecular Medicine* **16** 1.

Rosenberg, N. A., Pritchard, J. K., Weber, J. L. et al. (2002). Genetic Structure of Human Populations. *Science* **298** 2381–2385.

Sammel, M. D., Ryan, L. M. and Legler, J. M. (1997). Latent variable models for mixed discrete and continuous outcomes. *Journal of the Royal Statistical Society, Series B* **59** 667–678.

Schaffner, S. F., Foo, C., Gabriel, S. et al. (2005). Calibrating a coalescent simulation of human genome sequence variation. *Genome Research* **15** 1576–1583.

Schifano, E. D., Li, L., Christiani, D. C. et al. (2013). Genome-wide association analysis for multiple continuous secondary phenotypes. *American Journal of Human Genetics* **92** 744–759.

Siepel, A., Bejerano, G., Pedersen, J. S. et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research* **15** 1034–1050.

Siepel, A., Pollard, K. S. and Haussler, D. (2006). New methods for detecting lineage-specific selection. In *10th International Conference on Research in Computational Molecular Biology*.

So, H. C. and Sham, P. C. (2010). A unifying framework for evaluating the predictive power of genetic variants based on the level of heritability explained. *PLoS Genetics* **6** 1–13.

Speliotes, E. K., Willer, C. J., Berndt, S. I. et al. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genetics* **42** 937–948.

Tang, H., Wei, P., Duell, E. J. et al. (2014). Axonal guidance signaling pathway interacting with smoking in modifying the risk of pancreatic cancer: A gene- and pathway-based interaction analysis of GWAS data. *Carcinogenesis* **35** 1039–1045.

Tchetgen Tchetgen, E. J. (2014). A general regression framework for a secondary outcome in case-control studies. *Biostatistics* **15** 117–128.

The ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* **489** 57–74.

Uhl, G. R., Drgon, T., Johnson, C. et al. (2010). Genome-wide association for smoking cessation success in a trial of precessation nicotine replacement. *Molecular Medicine* **16** 1.

Vachon, C. M., Van Gils, C. H., Sellers, T. A. et al. (2007). Mammographic density, breast cancer risk and risk prediction. *Breast Cancer Research* **9** 217–225.

Wen, W., Cho, Y.-S., Zheng, W. et al. (2012). Meta-analysis identifies common variants associated with body mass index in east Asians. *Nature genetics* **44** 307–11.

Wray, N. R., Yang, J., Goddard, M. E. et al. (2010). The genetic interpretation of area under the ROC curve in genomic profiling. *PLoS Genetics* **6**.

Wu, M. C., Lee, S., Cai, T. et al. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *American Journal of Human Genetics* **89** 82–93.

Zaitlen, N., Lindström, S., Pasaniuc, B. et al. (2012). Informed Conditioning on Clinical Covariates Increases Power in Case-Control Association Studies. *PLoS Genetics* **8**.

Zhang, D. and Lin, X. (2003). Hypothesis testing in semiparametric additive mixed models. *Biostatistics* **4** 57–74.