THE STATISTICAL ANALYSIS OF GENETIC SEQUENCING AND RARE VARIANT ASSOCIATION STUDIES

Eugene Urrutia

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics in the Gillings School of Global Public Health.

Chapel Hill
2013

Approved by:

Michael Wu

Yun Li

Wei Sun

Donglin Zeng

William Valdar

ABSTRACT

Eugene Urrutia: The Statistical Analysis of Genetic Sequencing and Rare Variant Association Studies
(Under the direction of Michael Wu)

Understanding the role of genetic variability in complex traits is a central goal of modern human genetics research. So far, genome wide association tests have not been able to discover SNPs that explain a large proportion of the heritability of disease. It is hoped that with the advent of accessible DNA sequencing data, investigators can uncover more of the so-called missing heritability. The added information contained in sequencing data includes rare variants, that is, minor alleles whose population frequency is low.

We examine several existing region based rare variant association tests including burden based tests and similarity based tests and show that each is most powerful under a certain set of conditions which is unknown to the investigator. While some have proposed tests that combine the features of several existing tests, none as yet has provided a test to combine the features of all existing tests. Here, we propose one such test under the framework of the SKAT test, and show that it is nearly as powerful as the most appropriately chosen test under a range of scenarios.

Existing methods do not allow for missing values in the covariates. Standard use of complete case analysis may yield misleading results, including false positives and biased parameter estimates. To address this problem, we extend an existing maximum likelihood strategy for accommodating partially missing covariates to the SKAT framework for rare variant association testing. This results in a test with high power to identify genetic regions associated with quantitative traits while still providing unbiased estimation and correct control of type I error when covariates are missing at random. Since the framework is generic, we also consider the application of this approach to epigenetic data.

A wide range of variable selection approaches can be applied to isolate individual rare variants within a region, yet there has been little evaluation of these approaches. We examine key methods for prioritizing individual variants and examine how these procedures perform with respect to false positives and power via application to simulated data and real data.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction and Overview

In modern human genetics, it is desired to know whether genetics play a role in phenotype, for example the presence or absence of a disease. So far, genome wide association tests have not been able to discover SNPs that explain a large proportion of the heritability of disease. It is hoped that with the advent of accessible DNA sequencing data, investigators can uncover more of the so-called missing heritability. The added information contained in sequencing data includes rare variants, that is, minor alleles whose population frequency is low. This is in contrast to microarray technology which typically includes common single nucleotide polymorphisms whose minor allele frequency (MAF) are relatively high. Rare variants associated with disease have already been reported.

Statistical considerations need to be made to adjust to rare variant association testing. Power decreases substantially when applying common variant methodology to rare variants. The signal is lower due to fewer minor alleles present in a given study. Also, multiple comparison corrections are a concern since the number of variants is increased dramatically.

To address these concerns, investigators have adapted a region based approach to rare variant association testing. In this approach, all variants of a region, such as a single genomic exome, are tested as a group. Collapsing the data and testing only the cumulative effect, this addresses the low signal concern by amplifying over several variants and the multiple comparison correction concern by substantially decreasing the number of tests performed. In paper 1, we examine several existing methods including burden based tests and similarity based tests and show that each is most powerful under a certain set of conditions which is

unknown to the investigator. While some have proposed tests that combine the features of several existing tests, none as yet has provided a test to combine the features of all existing tests. Here, we propose one such test under the framework of the SKAT test, and show that it is nearly as powerful as the most appropriately chosen test under a range of scenarios.

It is of prime importance for investigators to consider important covariate information when performing genetic sequencing studies. If individual characteristics such as demographics, age, gender, or lifestyle, is ignored, many false positive results may be discovered which will not hold up under subsequent study. Fortunately, most of the widely used statistical procedures for rare variants are able to accommodate covariates. Methods have been developed to account for missing genotype via imputation or allele dosages. However, existing methods do not allow for missing covariates. In the case of missing covariates, misleading results may be obtained if proper adjustments are not provided. For example, if the data are missing at random, and only complete observations are used in the analysis, then there is a great danger of biased parameter estimation. In paper 2, we examine the properties of complete case, single/multiple imputation, and maximum likelihood when covariates are MCAR and MAR. We use an existing maximum likelihood strategy via iteratively reweighted least squares and apply it to the SKAT framework for rare variant association testing. This results in a test that maximizes power while still providing unbiased estimation and correct control of type I error under the condition of missing covariates under MAR.

Finally, once a region of interest has been identified, subsequent analysis is required to prioritize and select the individual variants that drive the association. By restricting analysis to a single region, the problem of finding individual associated variants becomes much less high-dimensional and much more tractable. While most rare variant tests can only identify regions associated with complex traits, variable selection procedures are well adapted to the identification of specific variants that are responsible for the regional association. We discuss several methods of variable selection, including univariable linear regression, multivariable linear regression with backward/forward selection, penalized linear regression, and stability selection. In paper 3, we examine key methods for prioritizing individual variants and examine how these procedures perform with respect to false positives and power via application to

simulated data and real data. Furthermore, we consider the direct use of forward selection in conjunction with SKAT and show that this method is highly competitive and can often select truly causal variants.

In the review of the current literature, we describe the following:

1. Statistical methods of testing whole genetic sequencing regions
2. Statistical methods for working with missing covariates
3. Statistical methods of selecting rare genetic variants within a specific genetic region

We follow with our own contributions to rare variant association testing:

1. Multiple kernel SKAT unified framework for rare variant association testing
2. Maximum likelihood based procedure for rare variant sequencing data with missing covariates
3. Evaluation of variable selection methods for selection of individual rare variants in sequencing studies

Chapter 2

Literature Review

2.1 Statistical methods of testing whole genetic sequencing regions

2.1.1 Heritability of disease

In genetic association testing, it is desired to know whether genetics play a role in phenotype, for example the presence or absence of a disease. Heritability, the inheritance of phenotypes such as disease resulting from genetic information alone, can be estimated using family based studies (McNeill et al., 2004; Dwyer et al., 1999). For example, identical twins separated at birth have identical genetic information and randomly associated environmental factors, while random pairs of persons have random genetic similarity and randomly associated environmental factors. Linear mixed modeling of an outcome can estimate the variance due to heretability versus that due to environment.

Over the past two decades, genome wide association studies (GWAS) have used DNA and RNA microarray technology to find specific genetic variants that represent the heritability of disease. Microarrays today typically measure the DNA/RNA concentration of 100,000-1,000,000 single nucleotide polymorphisms (SNPs), that is, common genetic variants with a minor allele frequency of 5% or greater. However, so far, genome wide association tests have not been able to discover SNPs explaining a large proportion of the expected heritability of disease (Eichler et al., 2010; Kaiser, 2012). Scientists have theorized that items such as interactions between genetic variants, prior biological information, gene pathway information, and better use of demographics could lead to better elucidation of heritability (Manolio et al., 2009; Zuk et al., 2012) . In this paper, we discuss another potential gain in the search for the

missing heritability: the technological advance of whole genome DNA sequencing.

The added information contained in DNA sequencing data includes rare variants, that is, minor alleles whose population frequency is low, well below the 5% threshold of common variants used in microarray studies. Already, rare variants associated with disease have been reported (Cohen et al., 2006; Walsh et al., 2008; Nejentsev et al., 2009).

We begin by briefly discussing the methods previously used in GWAS and then discuss the adaptations used in sequencing association studies.

2.1.2 Statistical methods of testing genome-wide association studies

The most popular statistical method of GWAS is regression applied to case-control or quantitative trait data (Hunter et al., 2007; Yeager et al., 2007; Thomas et al., 2008; Scott et al., 2007). Demographics such as gender, race, and age are controlled for and p-values are adjusted for multiple comparisons. Chi-squared test stratified for discrete covariates can be used but is impractical for covariates in comparison to logistic regression. For continuous phenotypes such as blood pressure, linear regression with the identity link is used similarly to logistic regression.

However, statistical considerations need to be made to adjust to rare variant association testing. Power decreases substantially when applying common variant methodology to rare variants. The signal is lower due to less minor alleles present in a given study. Also, multiple comparison corrections are a concern since the number of variants is increased dramatically, from the order of thousands to the order of billions. To address these concerns, investigators have adapted a region based approach to rare variant association testing. In this approach, all variants of a region, such as a single genomic exome, are tested as a group. Collapsing the data and testing only the cumulative effect addresses the low signal concern by amplifying over several variants, and the multiple comparison correction concern by substantially decreasing the number of tests performed.

2.1.3 Burden based sequence association tests

One class of region based methods is the burden-based class of tests. In the cohort allelic sum test and combined multivariate collapsing test (CAST/CMC) the genetic information of a region for an individual is collapsed to a single binary variable which takes the value 1 if the person has at least one rare variant present in the region and 0 otherwise (Morgenthaler and Thilly, 2007; Li and Leal, 2008). In a slight variation, the count collapsing method, the summary variable takes the value of the total number of rare variants present in the region of an individual (Morris and Zeggini, 2010). Additionally, one may wish to place a higher weight on variants which are rarer, and this is done in the weighted count collapsing method (Madsen and Browning, 2009). The burden-based rare variant association tests are similar in that they sum over the rare variant genetic information. Thus, they are most powerful when the effects of the variants are all in the same direction, that is, all are deleterious or all are protective. Power is decreased when effects are in opposite directions.

Assuming continuous outcome (y_i) , the above models are described below and solved using linear regression:

CAST/CMC:

$$y_i = \alpha X_i + \beta I(\sum_{j=1}^{p} z_{ij} > 0) + \epsilon_i$$

Count Collapsing:

$$y_i = \alpha X_i + \beta \sum_{j=1}^{p} z_{ij} + \epsilon_i$$

Weighted Count Collapsing:

$$y_i = \alpha X_i + \beta \sum_{j=1}^{p} z_{ij} w_j + \epsilon_i$$

where X_i are covariates including intercept; z_{ij} is the number of rare alleles present at locus j in individual i and takes the value 0, 1, or 2; and w_j is an assigned weight, typically higher for the rarer variables.

Additionally, several tests within the burden-based class have been proposed to address specific concerns. Liu and Leal (2010) have proposed the kernel-based adaptive cluster

(KBAC) to address statistical problems associated with misclassification of variant functionality, causality, and polymorphism status, and also with gene interactions. Using the cumulative minor-allele test (CMAT), Zawistowski et al. (2010) have broadened the scope to the application of low-coverage sequencing and imputation data, as well as population stratification. Bhatia (2009) proposed RARECOVER in order to take advantage of a subset within a region being more associated with a phenotype. Still, though, none of the methods mentioned yet address the concern of variants within a region having both deleterious and protective effects.

Investigators have developed and adapted new strategies to address this concern. Han and Pan (2010a) introduced the data-adapted sum (aSum) test which incorporates both marginal (univariable) analysis and common association strength to detect both protective and deleterious effects. Ionita-Laza et al. (2011) introduced another novel strategy, the replication-based strategy, to achieve the same. Li et al. (2010a), with their weighted haplotype and imputation-based tests (WHaIT), added imputation capabilities to the protective/deleterious model.

2.1.4 Similarity based sequence association tests

Others have proposed another set of tests called similarity-based methods. In this class the question is asked whether individuals who are genetically similar are also phenotypically similar. Neale et al. (2011) adapted the C-alpha score test to evaluate change in variance of the allele frequency rather than change in the mean of the allele frequency in cases compared to controls. Under the null hypothesis of no genetic association with outcome, distribution of counts of rare alleles should follow the binomial distribution. By testing variance rather than net effect, the test is powerful to detect genetic association when the effects of the variants are not all in the same direction.

For each of m variants, the C-alpha test statistic contrasts the observed variance with the expected variance of rare allele count in cases. The test assumes the binomial distribution

and null hypothesis of equal distribution among cases and controls:

$$T = \frac{1}{\sqrt{c}} \sum_{j=1}^{m} \left[(z_j - n_j p_0)^2 - n_j p_0 (1 - p_0) \right]$$

where z_j is the total rare allele count for variant j in cases only, n_j is the total rare allele count for variant j in both cases and controls, p_0 is the proportion of cases out of total subjects, and c is a standardization term.

2.1.5 Sequence Kernel Assocation Test

Another similarity-based method, the sequence kernel association test (SKAT), includes the flexibility to custom define what is genetic similarity through a kernel function (Wu et al., 2011). The result is an n by n matrix of pair wise genetic similarity which appears very much like a correlation matrix. SKAT is our preferred method because it offers a general framework that is adaptable to almost any scenario while retaining power when the kernel is chosen appropriately. It is also flexible in that covariates can be accomodated without the use of permutation.

SKAT is based on a semi-parametric model:

$$y_i = x_i \beta + h(z_i) + \epsilon_i$$

$h(.)$ is defined by the kernel function $K(.,.)$. In general two popular kernel functions are the dth polynomial kernel and the gaussian kernel.

Dth polynomial kernel:

$$K(z_1, z_2) = (z_1^T z_2 + \rho)^d$$

where d indexes the order of polynomial and ρ is an index parameter

When d = 1, this is equivalent to a linear function space with first-order basis functions: $\{z_1, z_2, \ldots, z_m\}$. When d = 2, this is equivalent to a quadratic function space with second-order basis functions: $\{z_j, z_j z_{j'}\}(j, j' = 1, \ldots, m)$.

Gaussian kernel:

$$K(z_1, z_2) = \exp\{-||z_1 - z_2||^2/\rho\}$$

where ρ is the scale parameter.

The Gaussian kernel is equivalent to the radial basis functions.

The kernel for the default SKAT test uses weights equal to the $\beta(1,25)$ distribution evaluated at the study-wide frequency of the particular minor allele. This produces greater power when the rarest alleles have the most effect on the outcome:

Default SKAT kernel:

$$K = ZW(ZW)^T$$

where Z is the full genetic design matrix and W is a diagonal matrix of weights

Expanded, the default SKAT kernel takes the form:

$$K_{SKAT} = \begin{bmatrix} w_1 z_{11} & w_2 z_{12} & \dots & w_m z_{1m} \\ w_1 z_{21} & w_2 z_{22} & \dots & w_m z_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ w_1 z_{n1} & w_2 z_{n2} & \dots & w_m z_{nm} \end{bmatrix} \begin{bmatrix} w_1 z_{11} & w_2 z_{12} & \dots & w_m z_{1m} \\ w_1 z_{21} & w_2 z_{22} & \dots & w_m z_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ w_1 z_{n1} & w_2 z_{n2} & \dots & w_m z_{nm} \end{bmatrix}^T$$

Thus, $K(z_1, z_2) = \sum_{j=1}^{m} w_j^2 z_{1j} z_{2j}$. It is clear that $K(z_1, z_2)$ approaches 1 when there is high genetic similarity, approaches -1 when there is great genetic dissimilarity, and is close to 0 otherwise. This similarity is weighted toward rare variants.

There are several ways to estimate the parameters β and h. LSKM estimates by minimizing a scaled penalized likelihood function.

LSKM minimizes:

$$J(h, \beta) = -\frac{1}{2}\sum_{i=1}^{n}\{y_i - x_i^T\beta - h(z_i)\}^2 - \frac{1}{2}\lambda||h||^2$$

where λ is a tuning parameter which determines the flexibility of the model. When λ=0, the model interpolates the data, while when $\lambda = \infty$, the model fits the linear model without $h(z)$. By the representer theorem, $h(z)$ can be represented by $\sum_{i=1}^{n} \alpha_i K(\cdot, z_i)$ so that the likelihood

function can be written:

$$J(\alpha, \beta) = -\frac{1}{2}\sum_{i=1}^{n}\{y_i - x_i^T\beta - \sum_{j=1}^{n}\alpha_i K(z_i, z_j)\}^2 - \frac{1}{2}\lambda\alpha^T K\alpha$$

with solutions

$$\hat{\beta} = \{X^T(I + \lambda^{-1}K)^{-1}X\}^{-1}X^T(I + \lambda^{-1}K)^{-1}y$$

$$\hat{\alpha} = \lambda^{-1}(I + \lambda^{-1}K)(y - X\hat{\beta})$$

However, Liu et al. (2007a) argue that the usefulness of the LSKM is limited due to the high computing cost of estimating λ and lack of literature on estimating ρ and σ^2. Thus, it is preferable to use the linear mixed model.

Linear mixed model:

$$y = x\beta + h + \epsilon$$

where h are random effects with distribution $N(0, \tau K)$ and ϵ are residuals with distribution $N(0, \sigma^2 I)$. It is clear that this model is equivalent to LSKM because β and h can be derived equivalent to those from LSKM using a standard linear mixed model estimating procedure:

$$\begin{bmatrix} X^T R^{-1} X & X^T R^{-1} \\ R^{-1} X & R^{-1} + (\tau K)^{-1} \end{bmatrix} \begin{bmatrix} \beta \\ h \end{bmatrix} = \begin{bmatrix} X^T R^{-1} y \\ R^{-1} y \end{bmatrix}$$

where $R = \sigma^2 I$ and $\tau = \lambda^{-1}\sigma^2$

When we apply the kernel machine to genetic sequencing data, we are primarily interested in whether or not the entire genetic region has an effect on the outcome. This test is: $H_0 : h(z) = 0$ vs. $H_a : h(z) \neq 0$. Using the linear model framework, we can equivalently test $H_0 : \tau = 0$ vs. $H_a : \tau > 0$. The test falls on the boundary. Also, because K is not block diagonal, τ is not distributed as a mixture of χ_0^2 and χ_1^2.

Liu et al. (2007a) propose testing the hypothesis by score test using REML of linear mixed model at fixed ρ:

SKAT score test statistic:

$$Q_\tau(\hat{\beta}, \sigma^2, \rho) - tr\{P_0 K(\rho)\}$$

where

$$Q_\tau(\hat{\beta}, \sigma^2, \rho) = \frac{(y - x\hat{\beta})^T K (y - x\hat{\beta})}{2\hat{\sigma}^2}$$

and

$$P_0 = I - X(X^T X)^{-1} X$$

where $\hat{\beta}$ and $\hat{\sigma}$ are estimated under the standard linear model with covariates only.

Under the null hypothesis, the quantity $(y - x\hat{\beta})$ converges to a standard normal, thus Q, quadratic in $(y - x\hat{\beta})$, is distributed $\kappa\chi^2_\nu$, a scaled mixture of χ^2, and κ and ν are calculated using one of several methods. We typically use the moment matching method described by Liu et al. (2009), although other chi-square approximation methods are available (Satterthwaite, 1946; Davies, 1980; Duchesne and Lafaye De Micheaux, 2010).

Satterthwaite approximates the null distribution with the following:

$$\kappa = \tilde{I}_{\tau\tau}/2\tilde{\epsilon}$$

$$\tilde{\nu} = 2\tilde{\epsilon}^2/\tilde{I}_{\tau\tau}$$

where

$$\tilde{I}_{\tau\tau} = I_{\tau\tau} - I_{\tau\sigma^2} I^{-1}_{\sigma^2\sigma^2} I^T_{\tau\sigma^2}$$

$$I_{\tau\tau} = tr\{P_0 K(\rho)\}^2/2$$

$$I_{\tau\sigma^2} = tr\{P_0 K(\rho) P_0\}/2$$

$$I_{\sigma^2\sigma^2} = tr\{P_0^2\}/2$$

$$\tilde{\epsilon} = tr\{P_0 K(\rho)\}/2$$

Liu uses moment matching to approximate a non-central chi square $\chi^2_l(\delta)$, while the Davies method inverts the characteristic fuction.

To generate the p-value, Q is compared the null distribution of

$$\frac{P_0^{1/2} K P_0^{1/2}}{2}$$

When the outcome is continuous:

$$P_0 = I - \tilde{X}(\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T$$

When the outcome is case/control:

$$P_0 = D_0 - D_0 \tilde{X}(\tilde{X}^T D_0 \tilde{X})^{-1} \tilde{X}^T D_0$$

Where X is a matrix of covariates including intercept; and D_0 is a diagonal matrix of $\hat{p}_j(1-\hat{p}_j)$, where $\hat{p}_j$ is the predicted proportion of rare alleles for variant j and is estimated from logistic regression of X on Y.

2.1.6 Combination based sequence association tests

We have thus far described 5 tests used for rare variant association testing, and there are numerously many more to choose from as well. The investigator must choose one from these many options before testing the data. A second choice that the investigator must make is what will be defined as a rare variant. Choices of rare variants thresholds include variants 3% MAF, 1% MAF, or 0.5% MAF. Additionally, the investigator may want to restrict to a set of only non-synonymous mutations, or those that are biologically predicted to be "harmful" by Polyphen-2 or other software. The result is that the investigator has many tests to choose from and many groupings to choose from as well, creating a very large set of combinations.

The necessary questions are: 1) Which is the most powerful test to use for a given data set, and 2) Which is the best grouping to test? The answer to those questions requires a priori knowledge of which variants are causal and what is their effect size and direction. However, knowing this information would make testing unnecessary. As a solution, one may choose to apply all tests and grouping and report the best p-value, but this clearly leads to inflated

type I error.

A final class of gene sequence association tests attemps to solve the problem by combining several tests at once in order to have power in a range of scenarios. The variable threshold test (Price et al., 2010), for example, starts with a foundation of the score test based on the likelihood function. However, instead of picking a fixed threshold of say 3% minor allele frequency, they select a range of different minor allele frequency thresholds. The score test is computed at each threshold, and a final p-value is found through permutation, so that type I error is conserved.

Optimal tests for rare variant effects in sequencing association studies (SKAT-O) (Lee et al., 2012), on the other hand, tests over a range of tests that spans from the count test to the SKAT test. That is, it tests on one hand that effect sizes and directions of the various rare alleles are perfectly corrlelated, and also the other hand that there is no correlation in effect sizes. Scenarios in between are tested as well. Thus, while K_{SKAT} may be written:

$$K_{SKAT} = Z_w Z_w^T$$

SKAT-O can be expanded as:

$$K_{SKAT-O} = Z_w R_\rho Z_w^T$$

Where Z_w is the weighted minor allele frequency design matrix, and R_ρ is the corellation matrix indexed by ρ where:

$$R_\rho = (1-\rho)I + \rho 11^T$$

Lee et. al use the minimum p-value as the test statistic and the final p-value is found by integrating the distribution function of the null mixture of χ^2 distribution.

An additional combination test was introduced by Lin and Tang (2011). The general framework allows not only test a range of MAF thresholds, but is also capable of handling covariates without the need for permutation.

Lin's score statistic is:

$$U_k = \sum_{i=1}^{n} \left(Y_i - \frac{e^{\hat{Y}_i}}{1+e^{\hat{Y}_i}} \right) \xi_k^T Z_i V_k^{-1/2}$$

Method	Comments
CAST/CMC	powerful when effects of equal size/direction
Count	powerful when effects of equal size/direction
Weighted Sum Test	powerful when effects of equal size/direction
C-Alpha	best when effects of different size/direction
SKAT	best when effects of different size/direction
VT	powerful over range of MAF thresholds
SKAT-O	powerful for both equal or different size/direction
EREC	powerful for both equal or different size/direction

Table 2.1: Summary of commonly used statistical methods of testing whole genetic sequencing regions

where $\hat{Y}_i$ is estimated under the null (covariates only), ξ_k is a kernel specific weight function, and V_k is the kernel specific variance of the score statistic.

Lin shows that the for the optimal kernel choice, $\xi_j = \beta_j$. To attempt to achieve optimality, he introduced estimated regression coefficients (EREC).

EREC:

$$\xi_j = \hat{\beta}_j \pm \delta$$

Where δ is a given constant, and $\hat{\beta}_j$ is the regression coefficient estimated from the data.

ξ_j will converge to β_j if δ decreases to 0 as the sample size n increases to N

2.2 Statistical methods for working with missing covariates

2.2.1 Mechanisms of missingness

Covariates are important to all statistical anlaysis. They can increase power when properly used, and can lead to inflated type I error when improperly used or ignored. In this section, we discuss the methods used to address partially missing covariates.

Assume we have the following data set where x_2 but not x_1 has missingness. It is

convenient to assign the value r_i to 1 if x_{2i} is observed and to 0 if x_{2i} is unobserved (N/A) .

Y	X_1	X_2	R
y_1	x_{11}	x_{21}	1
y_2	x_{12}	x_{22}	1
y_3	x_{13}	N/A	0
y_4	x_{14}	x_{24}	1
y_5	x_{15}	N/A	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_n	x_{1n}	x_{2n}	1

There are, in general, three missing data mechanisms, that is, three underlying models which predict which data points are missing. Data missing completely at random (MCAR) is randomly missing, that is, not predictable by any observed or unobserved data points. Data missing at random (MAR) may be missing in a way predictable by observed data points, but is not predictible by unobserved data points. Data not missing at random (NMAR) may be predictable by data observed or unobserved.

To summarize with our data set, again assuming x_2 but not x_1 has missingness:
Missing completely at random:

$$R \perp f(y, x_1, x_2)$$

Missing at random:

$$R \perp f(x_2|y, x_1)$$

Not missing at random:

$$R = f(x_2|y, x_1)$$

where $f(x_2|y, x_1)$ is a non-trival function

2.2.2 Complete case

Complete case is a very simple method used where all observations with at least one missing covariate are excluded from analysis. Complete case transforms the incomplete data set to a complete data set which is more convenient to work with:

Complete case:

Y	X_1	X_2	R
y_1	x_{11}	x_{21}	1
y_2	x_{12}	x_{22}	1
y_3	x_{13}	N/A	0
y_4	x_{14}	x_{24}	1
y_5	x_{15}	N/A	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_n	x_{1n}	x_{2n}	1

$\longrightarrow$

Y	X_1	X_2
y_1	x_{11}	x_{21}
y_2	x_{12}	x_{22}
y_4	x_{14}	x_{24}
$\vdots$	$\vdots$	$\vdots$
y_n	x_{1n}	x_{2n}

It is clear, however, that complete case will result in reduced power due to the decrease in sample size. Additionally, statistical inference on the full data using only the complete case is invalid under MAR and NMAR. Bias is likely to result (Little and Rubin, 1987; Knol et al., 2010).

2.2.3 Single and multiple imputation

Single imputation (SI) attempts to recover observations by filling in the missing covariate with any number of prespecified pseudo-observations. The fill-in may be the mean or it may be a random value based on the empirical distribution of the covariates. A posterior mean may also be used conditional upon observation specific data and covariates.

Single imputation using mean to fill in missing observations:

Y	X_1	X_2	R
y_1	x_{11}	x_{21}	1
y_2	x_{12}	x_{22}	1
y_3	x_{13}	N/A	0
y_4	x_{14}	x_{24}	1
y_5	x_{15}	N/A	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_n	x_{1n}	x_{2n}	1

$\longrightarrow$

Y	X_1	X_2
y_1	x_{11}	x_{21}
y_2	x_{12}	x_{22}
y_3	x_{13}	$\overline{x}_2$
y_4	x_{14}	x_{24}
y_5	x_{15}	$\overline{x}_2$
$\vdots$	$\vdots$	$\vdots$
y_n	x_{1n}	x_{2n}

Multiple imputation (MI) is a variation on single imputation. In multiple imputation, many data sets are generated from the one original set, and in each set, missing values are replaced with random values based on the emperical distribution. The statistical model is applied independently to each created set. Finally, the predicted coefficients are averaged across the imputed data sets to generate the final coefficients.

Multiple imputation using random values based on the empirical distribution:

Y	X_1	X_2	R
y_1	x_{11}	x_{21}	1
y_2	x_{12}	x_{22}	1
y_3	x_{13}	N/A	0
y_4	x_{14}	x_{24}	1
y_5	x_{15}	N/A	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_n	x_{1n}	x_{2n}	1

$\longrightarrow$

Y	X_1	X_2	R
y_1	x_{11}	x_{21}	1
y_2	x_{12}	x_{22}	1
y_3	x_{13}	$\overline{x}_2 + \epsilon$	0
y_4	x_{14}	x_{24}	1
y_5	x_{15}	$\overline{x}_2 + \epsilon$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_n	x_{1n}	x_{2n}	1

Y	X_1	X_2	R
y_1	x_{11}	x_{21}	1
y_2	x_{12}	x_{22}	1
y_3	x_{13}	$\overline{x}_2 + \epsilon$	0
y_4	x_{14}	x_{24}	1
y_5	x_{15}	$\overline{x}_2 + \epsilon$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_n	x_{1n}	x_{2n}	1

Where $\epsilon \sim N(0, \hat{\sigma}^2_{x_2})$ assuming x_2 is normally distributed.

Imputation has an advantage over complete case in that data is not thrown away. This clearly will increase power simply due to increased sample size. The two examples shown are valid under MCAR (Little and Rubin, 1987). Imputation is unbiased under MAR only when full likelihood posterior distribution used to fill-in missing data.

2.2.4 Maximum likelihood

Maximum likelihood (ML) can also be used to avoid the problem of discarding data. Here, a distribution is given to the missing data and the resulting likelihood is integrated across all

possible values of the missing covariate.

Maximum likelihood:

when x_2 is observed:

$$p(y_i, x_{2i}, r_i|x_{1i}, \beta, \alpha, \omega) = p(y_i|x_{1i}, x_{2i}, \beta)p(x_{2i}|x_{1i}, \alpha)p(r_i|y_i, x_{1i}, \omega)$$

while for missing x_2 (continuous):

$$p(y_i, r_i|x_{1i}, \beta, \alpha, \omega) = p(r_i|y_i, x_{1i}, \omega) \int_{x_{2i}} p(y_i|x_{1i}, x_{2i}, \beta)f(x_{2i}|x_{1i}, \alpha)dz_i$$

or for missing x_2 (discrete):

$$p(y_i, r_i|x_{1i}, \beta, \alpha, \omega) = p(r_i|y_i, x_{1i}, \omega) \sum_{x_{2i}} p(y_i|x_{1i}, x_{2i}, \beta)p(x_{2i}|x_{1i}, \alpha)$$

In summary, the log-likelihood is:

$$r_i \log \left[p(y_i, x_{2i}, r_i|x_{1i}, \beta, \alpha, \omega)\right] + (1 - r_i) \log \left[p(y_i, r_i|x_{1i}, \beta, \alpha, \omega)\right]$$

which can be solved throught either the Newton-Raphson method or by the Expectation-Maximization (EM) algorithm.

ML leads to unbiased results under MAR if the model is correctly specified. This is a clear advantage over complete case and over most cases of imputation. An additional advantage of ML over imputation is that ML produces the same result each time, while MI (as well as SI with errors) leads to differing results because of the variability of the imputed data.

2.2.5 Weighted maximum likelihood for data with missing covariates

Ibrahim (1990) proposed a weighted maximum likelihood to solve the above problem. This is very helpful when the above likelihood does not lend itself to a closed form. In Ibrahim's method, missing observation are expanded to multiple pseudo-observed observations and weighted according to their posterior probability of being observed. In the following

simple example, the covariate is dichotomous:

Y	X_1	X_2	R
y_1	x_{11}	x_{21}	1
y_2	x_{12}	x_{22}	1
y_3	x_{13}	N/A	0
y_4	x_{14}	x_{24}	1
y_5	x_{15}	N/A	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_n	x_{1n}	x_{2n}	1

$\longrightarrow$

Y	X_1	X_2	w
y_1	x_{11}	x_{21}	1
y_2	x_{12}	x_{22}	1
y_3	x_{13}	0	p_{i0}
y_3	x_{13}	1	p_{i1}
y_4	x_{14}	x_{24}	1
y_5	x_{15}	0	p_{i0}
y_5	x_{15}	1	p_{i1}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_n	x_{1n}	x_{2n}	1

where $p_{i0} = 1 - p_{i1} = P(x_2 = 0|y_i, x_{1i})$; and generally for missing x_{2i}:

$$w_i = \frac{p(y_i|x_{1i}, x_{2i}, \beta)p(x_{2i}|x_{1i}, \alpha)}{\sum_{x_{2i}} p(y_i|x_{1i}, x_{2i}, \beta)p(x_{2i}|x_{1i}, \alpha)}$$

while $w_i = 1$ for non-missing x_{2i}.

Following Ibrahim, the expressions take the form of a weighted complete data log-likelihood based on $N = \sum_{i=1}^{n} k_i$ observations, where k_i is the number of distinct covariate patterns for observation i. Thus, iteratively reweighted least squares (IRLS) is used in conjunction with the Newton Raphson algorithm to solve for β and α. This Newton-Raphson algorithm is considerably more convenient when maximum likelihood cannot be solved in closed form; there is no sum or integral.

When the missing covariate is continuous, Ibrahim et al. (2004) suggest approximating $f(x_{2i}|x_{1i})$ by a discrete distribution and then monte carlo is used to select L distinct points from the distribution along with the corresponding probabilities. Weights are then generated similar to the discrete method, and the weighted complete data log-likelihood is evaluated in the same way.

Method	Missingness	Advantage
Complete Case	MCAR	simple to implement
Imputation	MCAR, MAR	simple to implement, uses all data
Maximum likelihood	MCAR, MAR	consistent results with correctly specified model uses all data
Weighted ML by IRLS	MCAR, MAR	consistent results with correctly specified model convenient form uses all data

Table 2.2: Summary of methods to account for missing covariates. Imupation valid under MAR only when full likelihood posterior distribution used to fill-in missing data.

2.3 Statistical methods of selecting rare genetic variants within a genetic region

2.3.1 Variable Selection

Variable selection is the practice of selecting the subset of variables which best predicts the outcome. In most situations, this is beneficial to simplify a statistical model for a number of reasons. It is a simpler model to explain to others. The best and simplest model has the least variabilty in a subsequent data set. Also, during data collection, it is less costly to record fewer variables.

In our particular setting, we first discover a genomic region that is believed to be associated with the outcome by utilizing a region-based tests. It has now become necessary to find which of the specific variants within the region are the ones responsible for the association. It is the general inherent belief that some genetic variants are detrimental, some fewer are protective, and that many have absolutely no effect at all. Because of the prior belief that many variants have zero effect, we practice variable selection to find the variants that do have an effect or association and that are predictive of the outcome in a future data set. This is the second step in genetic sequence association testing.

Assuming exponential family:

$$g(y) = \alpha X + \beta Z$$

$$z_j \in S \Leftrightarrow \beta_j \neq 0$$

We wish to identify all $z_j : z_j \in S$ because they are the variants that are predictive

of the outcome in a future data set. It is true that some variants not belonging to S may be associated with the outcome through colinearity with predictive variants. These variants would also be helpful in pointing us toward a true biological phenomenon.

Many variable selection procedures are particularly well suited, as they assume most of the variables have no effect and a small subset of variable may have a non-zero effect on the outcome. This leads to easier model interpretation and greater power to detect effects.

Common examples of variable selection procedures include the Lasso (Tibshirani, 1996) and forward or backward stepwise subset selection. Many other simple statistical procedures could be applied toward variable selection as well. For example, one could apply a standard procedure and apply a pre-specified cutoff for effect size or p-value.

2.3.2 Univariable linear model

The simplest variable selection procedure is the univariable model. Here each genetic variant is tested for marginal association independently of the other variants. One may use generalized linear model or generalized linear mixed model to generate a p-value associated with each variant and then apply a multiple comparison correction to generate a list of statistically significant associations.

2.3.3 Multivariable linear model

Another classic way of testing is the multivariable model where all variants are tested together in a single model and then multiple comparison correction is applied to each of the p-values. One main difference between univariate and multivariate is that the multivariate may miss some variants that are masked due to high corellation with another variant in the model. The univariate does not suffer from this problem.

The advantage of the multivariate model, though, is that it is possible to use forward and backward stepwise selection. However, forward and backward stepwise selection should be used with caution as the theory is not developed. P-values and coefficients should be interpreted liberally.

2.3.4 Penalized linear regressions

A special class within the multivariable linear model are the penalized regressions. Among these, the Lasso is particularly attractive because it inherently sets the effect size of most the variables to zero, and the remaining few to non-zero. It is also computationally fast. The penalization term is customarily optimized by 10-fold cross-validation.

Lasso penalizes the sum of absolute values of regression coefficients:

$$\sum_{i=1}^{n} ||g(y_i) - X_i\beta||^2 + \lambda \sum_{j=1}^{m} ||\beta_j||^1$$

The following are other penalized regressions, which also tend to limit the number of variables in the model:

1. Akaike information criterion (AIC) (Akaike, 1974), and Bayesian information criterion (BIC) (Schwarz, 1978) penalize the number of parameters in the model, thus clearly performs variable selection:

$$\sum_{i=1}^{n} ||g(y_i) - X_i\beta||^2 + \lambda \sum_{j=1}^{m} ||\beta_j||^0$$

where λ is a constant for AIC, and λ is proportional to the sample size for BIC.

2. Ridge regression (Hoerl and Kennard, 1970) penalizes the sum of squares of regression coefficients and thus scales parameters instead of scaling to zero.

$$\sum_{i=1}^{n} ||g(y_i) - X_i\beta||^2 + \lambda \sum_{j=1}^{m} ||\beta_j||^2$$

3. Elastic Net (Zou and Hastie, 2005), a combination of ridge and lasso, penalizes both the the absolute value and square of regression coefficients, thus can perform both selection and scaling depending on weight of λ_1 and λ_2:

$$\sum_{i=1}^{n} ||g(y_i) - X_i\beta||^2 + \lambda_1 \sum_{j=1}^{m} ||\beta_j||^2 + \lambda_2 \sum_{j=1}^{m} ||\beta_j||^1$$

2.3.5 Consistent LASSO-based procedures

Fan and Li (2001) contend that a good penalty function should should produce an estimator with the following properties : Unbiasedness, sparsity, and continuity. The Lasso estimates are sparse and continuous, and Lasso is in fact the only sparse and continuous penalty within the family of $\lambda|\beta|^q$, for some q. That is, q=1 is the only q which produces sparse and continuous estimates. AIC and BIC achieve sparsity but are not continuous in β. Ridge regression is continuous but do not achieve sparsity. However, Lasso is not unbiased, as large coefficients are estimated with a biased shift toward 0 equal to a constant. Fan and Li (2001) in turn propose the smoothly clipped absolute deviation (SCAD) penalty that has all three of the desirable qualities.

SCAD:

$$p_\lambda(\beta; a) = \begin{pmatrix} \lambda|\beta|, & |\beta| <= \lambda \\ -(\beta^2 - 2a\lambda|\beta| + \lambda^2)/[2(a-1)], & \lambda < |\beta| <= a\lambda \\ (a+1)\lambda^2/2, & |\beta| > a\lambda \end{pmatrix}$$

for some $a > 2$ and $\lambda > 0$

Additionally, the adaptive Lasso by Zou (2006), through adaptive weighting, achieves consitent, unbiased estimates.

Adaptive Lasso

$$p_\lambda(\theta) = \lambda \sum_{j=1}^{m} w_j ||\beta_j||^1$$

with $w_j = 1/|\hat{\beta}_j|^\gamma$ estimated under ordinary least squares with $\gamma > 0$.

Finally, Bolasso, through bootstrap (Bach, 2008; Chatterjee and Lahiri, 2011), achieves asymptotically consistent, unbiased estimates. The algorithm begins by applying Lasso to m bootstraped samples of the data. The union of variables with non-zero coefficients in at least one bootrapped Lasso are compiled. These variables only are modeled using non-penalized regression for final parameter estimation.

2.3.6 Stability Selection

Although Lasso has the attractive property of shrinking coefficients to 0, it has a disadvantage in that there is no way to control type I error. Meinshausen and Bühlmann (2010) proposed stability selection to address this concern. In their procedure, B $\lfloor n/2 \rfloor$ subsamples out of n total observations are selected and applied to Lasso.

$$\hat{p}_{k,n/2,B} = \frac{1}{B} \sum_{b=1}^{B} I(k \in \hat{S}_{n/2,b})$$

The variable is selected as significant if $\hat{p}_{k,n/2,B} \geq \pi_{thr}$ (we set to 0.75) proportion of the B subsamples applied to Lasso.

They obtain a bound on family-wise error by making two assumptions: 1. Selection procedure no worse than guessing; and 2. All non-associated variants selected with equal likelihood.

$$FWER = \frac{1}{2\pi_{thr} - 1} \frac{q_{\Lambda}^2}{m^2}$$

where q_{Λ} is the number of variables selected by Lasso. The weak assumptions result in a bound that is quite conservative. Current research (Shah and Samworth, 2012) is directed toward tightening this bound.

Method	Comments
Univariate	well developed theory
Multivariate	well developed theory, possible masking due to correlation
Forward/Backward Selection	undeveloped theory
AIC/BIC	variable selection
Lasso	inherently shrinks many effects to zero, but no type I error
Ridge Regression	scales parameters
Ridge Regression	scales and shrinks parameters
Adaptive Lasso	asymptotically consistent
SCAD	asymptotically consistent
Bolasso	asymptotically consistent
Stability Selection	type I error with Lasso

Table 2.3: Summary of methods of variable selection

Chapter 3

Rare Variant Testing Across Methods and Thresholds Using the Multi-Kernel Sequence Kernel Association Test (MK-SKAT)

3.1 Introduction

Identification of genetic variants influencing complex phenotypes and disease is a major goal of modern human genetics research. So far, despite the success of genome wide association studies (GWAS)(Hindorff et al., 2009), newly discovered trait-associated genetic variants still fail to explain a large proportion of the heritability of complex traits (Eichler et al., 2010). It is hoped that with the advent of accessible DNA sequencing technology (Margulies et al., 2005; Mardis, 2008; Ansorge, 2009), investigators can uncover more of the so-called missing heritability. Some of the added information contained in sequencing data includes rare variants, that is variants with minor alleles whose population frequency is low. This contrasts with microarray technology which typically focuses on common variants that have relatively high minor allele frequency (MAF). Rare variants associated with disease have already been reported (Cohen et al., 2006; Walsh et al., 2008; Nejentsev et al., 2009). However, important distinctions between the analysis of common variants and rare variants must be made (Carvajal-Carmona, 2010). Most importantly, the standard analysis of common variants focuses on analysis of each individual variant, one-by-one. Yet, power decreases with lower MAF such that standard approaches for common variants are vastly underpowered for analysis of rare variants. Also, multiple comparison corrections are a concern since the number of variants is dramatically larger.

To address the limitations of using standard analytical approaches for variants, investigators

have turned to region based approaches for rare variant association testing. In this class of approaches, all variants within a region, typically a biologically meaningful unit such as a single gene or an exon, are simultaneously considered together. The cummulative effect of the entire group of variants, or more often a subgroup of the variants (e.g. those with MAF <1%), is assessed for association with the phenotype. Grouping the variants and testing only the cumulative effect addresses the low signal concern by amplifying across several variants. It also addresses the multiple comparison correction concern by substantially decreasing the number of tests performed. A wide range of methods have beeen developed with varying characteristics and underlying principles (Morgenthaler and Thilly, 2007; Li and Leal, 2008; Morris and Zeggini, 2010; Madsen and Browning, 2009; Neale et al., 2011; Wu et al., 2011).

Despite the sucess of current approaches for rare variant testing (Cohen et al., 2006; Walsh et al., 2008; Nejentsev et al., 2009), a number of practical concerns have arisen. In particular, given the wide range of testing approaches which are optimized toward different scenarios, it is unclear which method to use for any particular data set. Furthermore, it is unclear which strategy to use for grouping variants, e.g. grouping variants with MAF <3% vs <1%, within a region. Unfortunately, the answer to both questions depends on the underlying true state of nature which is unknown prior to analysis. Knowledge on this would preclude need for analysis. Selecting the "best" (often most significant) result after conducting analyses using multiple methods or multiple group strategies would lead to severely inflated type I error and increased false positives. Although some recent work has been done on omnibus testing across different grouping strategies (Price et al., 2010; Lin and Tang, 2011) or across different testing approaches (Lee et al., 2012), few methods consider both the testing approach and the grouping strategy simultaneously.

To address this problem, we propose the multi-kernel sequence kernel association test (MK-SKAT). In this article, we show that many commonly used testing approaches are equivalent to particular cases of the sequence kernel association test (SKAT). SKAT is a similiarity based analysis approach for rare variant testing wherein pair-wise similarity between individuals based on their rare variant profiles is measured via a kernel function and then compared to pair-wise similarity in phenotype. Specifically, the currently used methods

are equivalent to versions of SKAT using different kernel functions. We further show that different choices of grouping strategies are also equivalent to using the SKAT with different kernel functions. Consequently, the question of selecting a test to use as well as selecting a grouping strategy reduces to the problem of selecting an appropriate kernel function. This equivalence then leads to natural application of perturbation based procedures for omnibus testing across multiple kernels (and accordingly multiple grouping and rare variant testing approaches) (Wu et al., 2013). We conduct computer simulations and a real data applicaton to validate our approach and show that our proposed method loses a small amount of power when compared to the optimal grouping and testing approach, but offers considerably more power over poor choices.

The remainder of this paper is organized as follows. In the next section, we first review the generic SKAT method and describe how different testing approaches and different groupings all correspond to SKAT under different kernels. We then present the proposed MK-SKAT approach for testing across different tests and groupings. We show results from some representative simulation studies and from an illustration of our approach on real data. We conclude with a brief discussion.

3.2 Methods

Within this article, we describe our methodology within the context of analyzing a single gene region. However, the approach can be applied to multiple regions separately, with appropriate control for multiple comparisons. We let y_i denote the phenotype for the i^{th} individual in the study ($i = 1, \ldots, n$), and $\mathbf{X}_i$ be a vector of environmental or demographic variables for which we would like to adjust. For dichotomous phenotypes we let $y_i = 0$ or 1 for controls and cases, respectively. For each given region, we let $\mathbf{Z}_i$ be the vector of genetic variants within the region coded under the additive model. The objective is to test for an association between y and all the variants in $\mathbf{Z}$ or a subset of the variants in $\mathbf{Z}$ while adjusting for $\mathbf{X}$. We let $\mathcal{G}$ denote the indices of the variants within $\mathbf{Z}$ that we would like to test. For instance $\mathcal{G}$ may be the indices of the variants with MAF $< 1\%$ or the nonsynonymous variants. In doing so, one may select a subset of the variants in the region to test or one may test a

subset of the variants. Clearly, restricting attention to the truly causal variants would result in the higest power; however, which variants are causal is unknown. At the same time, there are a range of tests to choose from. Determining which group of variants to test and which test to use poses a grand challenge for geneticists.

In this section, we first review the SKAT method and draw connections between SKAT and several other important tests. We describe how the questions of which test to use and which variants to test can be recast as a question of kernel choice. We then develop the MK-SKAT to construct an omnibus test that simultaneously considers multiple tests and grouping strategies.

3.2.1 Connections between SKAT and other Methods

SKAT

SKAT is a similarity based test that operates by comparing pair-wise genotypic similarity between individuals to pair-wise phenotypic similarity, with correlation suggestive of association. Mathematically, SKAT uses the linear model for quantitative traits

$$y_i = \alpha_0 + \mathbf{X}_i'\boldsymbol{\alpha} + h(\mathbf{Z}_{\mathcal{G}_i}) + \varepsilon_i$$

and the logistic model for case/control studies

$$\text{logit}P(y_i = 1) = \alpha_0 + \mathbf{X}_i'\boldsymbol{\alpha} + h(\mathbf{Z}_{\mathcal{G}_i})$$

where α_0 is an intercept term, $\boldsymbol{\alpha}$ is the vector of regression coefficients for the covariates, and ε_i has mean zero and variance σ^2. The variants of interest $\mathbf{Z}_{\mathcal{G}_i}$ for the i-th individual are related to the outcome only through the function $h(\cdot)$ which is a general function lying in a functional space generated by a positive definite kernel function $K(\cdot,\cdot)$. Intuitively, $K(\mathbf{Z}_{\mathcal{G}_i}, \mathbf{Z}_{\mathcal{G}_{i'}})$ measures similarity between i-th and i'-th individuals in the study based on $\mathbf{Z}_{\mathcal{G}}$, the variants of interest. This function fully specifies the relationship between the variants and the outcome. If one sets $K(\mathbf{Z}_{\mathcal{G}_i}, \mathbf{Z}_{\mathcal{G}_{i'}}) = \mathbf{Z}_{\mathcal{G}_i}'\mathbf{Z}_{\mathcal{G}_{i'}}$, which is the linear kernel, then this implies

that the function $h(\mathbf{Z}_{\mathcal{G}_i}) = \sum_{j\in\mathcal{G}} \beta_j Z_{ij}$, i.e. $h(\cdot)$ is linear and the outcome depends on the variants in a linear manner. By specifying a different kernel, one may specify an alternative model. Under the default SKAT parameters, $K(\mathbf{Z}_{\mathcal{G}_i}, \mathbf{Z}_{\mathcal{G}_{i'}}) = \sum_{j\in\mathcal{G}} w_j^2 Z_{ij} Z_{i'j}$ where w_j is equal to a the beta probability density function with parameters 1 and 25 evaluated at the MAF for the j-th variant. Also by default, $\mathcal{G}$ is set to be the entire group of both common and rare variants within a region. This corresponds to a linear model but with additional up-weighting for the effect of rarer variants.

To test the effect of the rare variants under SKAT corresponds to testing $H_0 : h(\mathbf{Z}_{\mathcal{G}}) = 0$. Defining the kernel matrix, $\mathbf{K}$, to be the n-by-n matrix with i, i'-th term equal to $K(\mathbf{Z}_{\mathcal{G}_i}, \mathbf{Z}_{\mathcal{G}_{i'}})$, for quantitative traits, we construct the variance component score statistic

$$Q = \frac{(\mathbf{y} - \widehat{\mathbf{y}})'\mathbf{K}(\mathbf{y} - \widehat{\mathbf{y}})}{\widehat{\sigma}^2}$$

where $\widehat{\mathbf{y}} = \widehat{\alpha}_0 + \mathbf{X}\widehat{\boldsymbol{\alpha}}$ with $\widehat{\alpha}_0$, $\widehat{\boldsymbol{\alpha}}$, and $\widehat{\sigma}$ estimated under H_0. For dichotomous traits, we can construct a similar score statistic

$$Q = (\mathbf{y} - \widehat{\mathbf{y}})'\mathbf{K}(\mathbf{y} - \widehat{\mathbf{y}})$$

where $\widehat{\mathbf{y}} = \text{logit}^{-1}(\widehat{\alpha}_0 + \mathbf{X}\widehat{\boldsymbol{\alpha}})$ and $\widehat{\alpha}_0$, $\widehat{\boldsymbol{\alpha}}$ are again estimated under H_0. To obtain a p-value for significance, asymptotically, $Q \sim \sum \lambda_j \chi_1^2$ is a mixture of chi-squared distributions, with weights λ_j equal to the eigenvalues of $\mathbf{P}_0^{1/2}\mathbf{K}\mathbf{P}_0^{1/2}$ where $\mathbf{P}_0 = \mathbf{D} - \mathbf{D}\mathbf{X}(\mathbf{X}'\mathbf{D}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}$ with $\mathbf{D} = \mathbf{I}$ for quantitative traits and $\mathbf{D} = diag\{\widehat{y}_i(1 - \widehat{y}_i)\}$ for dichotomous traits. This null distribution can be approximated using moment matching approaches (Liu et al., 2009) or exact methods (Davies, 1980).

Existing Methods and Grouping Strategies as Special Cases of the SKAT

A wide range of region-based analysis approaches of rare variants have been proposed. Generally, however, they tend to fall within two classes: burden-based approaches and similarity-based approaches. Burden-based tests generally operate by collapsing the rare variants within a region into a single value using (possibly weighted) averaging and then

testing for association by regressing the phenotype on the collapsed variable or applying appropriate permutation-based approaches. Letting $\mathcal{G}$ denote the indices of the rare variants over which we would like to collapse, then the cohort allelic sum test (CAST) and combined multivariate collapsing (CMC) collapses the genetic variants within a region to a single binary variable

$$C_i = I\left(\sum_{j\in\mathcal{G}}^{p} Z_{ij} > 0\right)$$

which is an indicator for whether the i^{th} individual has any rare variants within the region. In a slight variation, the count-based collapsing method computes the collapsed variable as

$$C_i = \sum_{j\in\mathcal{G}}^{p} Z_{ij}$$

which is the total number of rare variants within the region. To place a higher weight on variants which are rarer, the weighted count collapsing method collapses the variants in $\mathcal{G}$ into

$$C_i = \sum_{j\in\mathcal{G}}^{p} w_j Z_{ij}$$

where w_j is a weight for the j^{th} variant which is inversely related to the MAF for the j^{th} variant. To test whether the rare variants are related to the phenotype, the outcome is regressed on the collapsed variable and possible covariates using the models

$$y_i = \alpha_0 + \mathbf{X}_i'\boldsymbol{\alpha} + \beta_C C_i + \varepsilon_i$$

or

$$\text{logit}P(y_i = 1) = \alpha_0 + \mathbf{X}_i'\boldsymbol{\alpha} + \beta_C C_i$$

for quantitative and dichotomous traits, respectively. Testing for the rare variant effect then corresponds to testing $H_0 : \beta_C = 0$ which can be done using a standard 1-df test. The burden-based rare variant association tests are similar in that they sum over all of the rare variant genetic information. Thus, they are most powerful when the effects of the variants are truly

associated with the outcome and with common direction of effect, that is, all variants are deleterious or all variants are protective. Power is lost when effects are opposite in directions or non-causal variants are included in $\mathcal{G}$.

Similarity-based tests were proposed to address the power loss due to variants with opposing effects. This class includes SKAT, and compares pair-wise similarity between individuals in terms of their genotype values to pair-wise similarity in phenotype, with correlation suggestive of association. Also included within this class is the C-alpha test which tests for an over-dispersion of the variance resulting from a rare variant effect rather than a change in the mean effect. By testing variance rather than net effect, the test is powerful to detect genetic association when the effects of the variants are not all in the same direction.

It has been previously noted that individual tests are equivalent to SKAT under particular kernel functions(Wu et al., 2011; Lee et al., 2012). For example, the C-alpha test is equivalent to SKAT using the kernel function $K(\mathbf{Z}_{\mathcal{G}_i}, \mathbf{Z}_{\mathcal{G}_{i'}}) = \sum_{j\in\mathcal{G}} Z_{ij}Z_{i'j}$. Further, each of the burden based methods operate by using a univariable summary of the rare variants in $\mathcal{G}$ such that the outcome is a simple linear function of the collapsed variable C_i. Therefore, each of the CAST/CMC, count-based collapsing, and weighted count-based collapsing can be viewed as SKAT with a linear kernel constructed based on the collapsed variable. Thus we have the following tests and corresponding kernels:

- (Default) SKAT: $K(\mathbf{Z}_{\mathcal{G}_i}, \mathbf{Z}_{\mathcal{G}_{i'}}) = \sum_{j\in\mathcal{G}} w_j Z_{ij}Z_{i'j}$
- C-alpha: $K(\mathbf{Z}_{\mathcal{G}_i}, \mathbf{Z}_{\mathcal{G}_{i'}}) = \sum_{j\in\mathcal{G}} Z_{ij}Z_{i'j}$
- CAST (Binary Collapsing): $K(\mathbf{Z}_{\mathcal{G}_i}, \mathbf{Z}_{\mathcal{G}_{i'}}) = I\left(\sum_{j\in\mathcal{G}}^{p} Z_{ij} > 0\right) I\left(\sum_{j\in\mathcal{G}}^{p} Z_{i'j} > 0\right)$
- Count-Based Collapsing: $K(\mathbf{Z}_{\mathcal{G}_i}, \mathbf{Z}_{\mathcal{G}_{i'}}) = \left\{\sum_{j\in\mathcal{G}}^{p} Z_{ij}\right\}\left\{\sum_{j\in\mathcal{G}}^{p} Z_{i'j}\right\}$
- Weighted Count-Based Collapsing: $K(\mathbf{Z}_{\mathcal{G}_i}, \mathbf{Z}_{\mathcal{G}_{i'}}) = \left\{\sum_{j\in\mathcal{G}}^{p} w_j Z_{ij}\right\}\left\{\sum_{j\in\mathcal{G}}^{p} w_j Z_{i'j}\right\}$

Given that many individual tests reduce to SKAT under different kernel, then the problem of choosing a particular test reduces to the problem of choosing a particular kernel.

We have, thus far, focused on testing the variants in a particular group, $\mathcal{G}$. In practice however, one must also choose, a priori, a group of variants to test. For example, one may

apply each of the tests to all of the variants in the region or one could restrict the variants of interest to just the variants with <3% MAF, < 1% MAF, or <0.5% MAF, depending on how one wishes to define "rare". Additionally the investigator may want to restrict to a set of only non-synonymous variants or those that are predicted to be "harmful" by Polyphen-2 (Adzhubei et al., 2010) or other software for predicting function. Use of different choices of variants can easily be translated into a problem of kernel choice by simply restricting $\mathcal{G}$ to be different sets of variants. For example, we can define $\mathcal{G}^{3\%}$ to be the variants with MAF $<$ 3% and $\mathcal{G}^{0.5\%}$ to be the variants with MAF $< 0.5\%$. Then if we are interested in the C-alpha test, we can apply it to the variants with MAF $< 3\%$ or $< 0.5\%$ by constructing the kernels $K(\mathbf{Z}_{\mathcal{G}_i^{3\%}}, \mathbf{Z}_{\mathcal{G}_{i'}^{3\%}}) = \sum_{j \in \mathcal{G}^{3\%}} Z_{ij} Z_{i'j}$ and $K(\mathbf{Z}_{\mathcal{G}_i^{0.5\%}}, \mathbf{Z}_{\mathcal{G}_{i'}^{0.5\%}}) = \sum_{j \in \mathcal{G}^{0.5\%}} Z_{ij} Z_{i'j}$, respectively and test using the usual SKAT procedure. Therefore, it follows that the problem of choosing which group of variants to test also reduces to the problem of choosing a particular kernel.

3.2.2 Multi-Kernel Sequence Kernel Association Test

The questions facing researchers interested in rare variant analysis are first, which is the most powerful test to use for a given data set, and second, which is the best group of variants to test within a particular region? As noted earlier, these questions can be reduced to a question of kernel choice: which kernel, from among a group of candidates, will yield highest power? Despite transforming the problem, the answer to this question requires prior knowledge of which variants are causal and what is their effect size and direction, knowledge which is rarely available (since this would preclude the need for analysis). As a solution, one may choose to test under all candidate kernels and report the best p-value, but this clearly leads to inflated type I error. However, by exploiting the connections between SKAT and other tests, we propose a solution that incorporates many tests and groupings but conserves type I error through the use of perturbation.

Our proposed unifying method, the multiple kernel SKAT (MK-SKAT), simultaneously several test and variant grouping choices at once and constructs an omnibus test. The idea behind the approach is that it constructs kernels based on each candidate test and grouping approach. For example, one may test using CAST, count-based collapsing, C-alpha, and the

defaul SKAT with 3 grouping strategies per test (MAF <3%, <1%, or <0.5%) for a total of 12 combinations corresponding then to 12 candidate kernels. MK-SKAT then conducts an omnibus test across all of the candidate kernels, by applying SKAT with each of the kernels, taking the minimum p-value, and then applying perturbation base techniques to correct for having taking the minimum p-value. A single p-value is reported. This represents a simplified version of the omnibus testing strategy of Wu et al. (2013).

The intuition behind the procedure is that asymptotically $\widehat{\sigma}^{-1}(y_i - \widehat{y}_i)$ will be approximately normal such that we can replace it with a simulated normal random variable. Using the same simulated normals for each candidate kernel allows for capture of the correlation between tests. The full MK-SKAT procedure is as follows:

1. For each combination of candidate testing procedure and each candidate grouping procedure, construct a corresponding kernel matrix, $\mathbf{K}_\ell$, to obtain a total of L candidate kernels.

2. Using each candidate kernel, $\mathbf{K}_\ell$, obtain a corresponding score statistic as Q_ℓ and p-value for significance p_ℓ.

3. Find the minimum p-value: $p_{\min} = \min_{1 \leq \ell \leq L} p_\ell$

4. For $\ell \in 1, \ldots, L$, compute $\mathbf{\Lambda}_\ell = diag(\lambda_{\ell,1}, \ldots, \lambda_{\ell,m_\ell})$, and $\mathbf{V}_\ell = [\mathbf{v}_{\ell,1}, \mathbf{v}_{\ell,2}, \ldots, \mathbf{v}_{\ell,m_\ell}]$ where $\lambda_{\ell,1} \geq \lambda_{\ell,2} \geq \ldots \geq \lambda_{\ell,m_\ell}$ are the m_ℓ positive eigenvalues of $\mathbf{P}_0^{1/2}\mathbf{K}_\ell\mathbf{P}_0^{1/2}$ with corresponding eigenvectors $\mathbf{v}_{\ell,1}, \mathbf{v}_{\ell,2}, \ldots, \mathbf{v}_{\ell,m_\ell}$

5. Generate $\mathbf{r}^* = [r_1^*, r_2^*, \ldots, r_m^*]'$ with each $r_j^* \sim N(0,1)$. Note that $m = \max_{1 \leq \ell \leq L} m_\ell$ is the maximum number of nonzero eigenvalues across the candidate kernels and may be less than n.

6. For each $\ell \in 1, \ldots, L$, rotate $\mathbf{r}^*$ using the eigenvectors to generate $\mathbf{r}_\ell^* = \mathbf{V}_\ell \mathbf{r}^*$.

7. Can then compute $Q_\ell^* = \mathbf{r}_\ell^{*\prime} \mathbf{\Lambda}_\ell \mathbf{r}_\ell^*$ for each ℓ and obtain a corresponding p-value, p_ℓ^*, by comparing Q_ℓ^* to the distribution function estimated for Q_ℓ and obtain the upper tail probability exceeding Q_ℓ^*. We set $p^* = \min_{1 \leq \ell \leq L} p_\ell^*$.

8. Repeat (5)-(7) B times to obtain $p^*_{(1)}, p^*_{(2)}, \ldots, p^*_{(B)}$ for some large number B.

9. The final p-value for significance is estimated as

$$p = B^{-1} \sum_{b=1}^{B} I(p^*_{(b)} \leq p_{mlin})$$

It is important to note that direct use of the p-value is necessary rather than using the maximum score statistic since the raw score statistics have different degrees of freedom.

Although this strategy also generates a monte carlo p-value, there are two advantages. First, covariates and variants can be correlated. In contrast, in order for permutation to be valid, the variants must be uncorrelated with the covariates. Second, the MK-SKAT procedure is more computationally efficient since the computation now relies only on generating and then rotating m normal random variables while all other parameters remain the same. In contrast, permutation requires complete re-estimation of the kernel matrices, $\mathbf{P}_0$ matrices, eigendecompositions, and distribution parameters.

This method assumes nested kernels. Although CAST is not nested, being non-linear in nature, the rarity of genetic variants being considered allows the kernel to be considered approximately linear. Additionally, MK-SKAT is conservative, so any anit-conservativeness resulting from the approximation is mitigated.

We note that this procedure is closely related to the general perturbation procedure previously used for testing across multiple kernels Wu et al. (2013). However, because each of the kernels used in this scenario for rare variant analysis is essentially a generalization of a weighted linear kernel, then they all lie within a common column space thereby simplifying the procedure.

3.2.3 Simulations

We conducted a series of computer simulations to verify that the proposed MK-SKAT procedure is valid in terms of controling type I error and has reasonable power compared to the individual tests across which the MK-SKAT is combining.

Type I Error

To demonstrate that the proposed methods are valid tests, in terms of protecting type I error, we conducted a series of simulations under null models for both continuous and dichotomous traits. We used a coalescent model to simulate a region with 100 variants in 10^4 haplotypes with LD structure representative of a European population (Schaffner et al., 2005). Eighty-five of the simulated variants had a true MAF less than 3% and 80 had a MAF less than 1%. We then paired haplotypes to simulate $n = 1000$ or 2000 diploid individuals. For type I error simulations, we simulated quantitative outcomes for each individual without regard to the genotype values under the null model:

$$y_i = 0.5X_{i1} + 0.03X_{i2} + \varepsilon_i$$

where $X_{i1} \sim ber(0.506)$, $X_{i2} \sim N(29.2, 21.1)$, and $\varepsilon_i \sim N(0, 1)$. For dichotomous outcomes, we simulated $n/2$ cases and $n/2$ controls from the null logistic model:

$$\text{logit}P(y_i = 1) = -4.2 + 0.5X_{i1} + 0.03X_{i2}$$

where $X_{i1} \sim ber(0.506)$ but $X_{i2} \sim N(0, 1)$.

In total, we simulated 10^5 data sets as described. We applied the MK-SKAT testing procedure to each data set. Specifically, we considered four different testing procedures: CAST, count-based collapsing, the C-alpha, and SKAT tests. We also considerd three different grouping strategies: we set the rare variant grouping, $\mathcal{G}$, equal to the variants with MAF $< 0.5\%$, variants with MAF $< 1\%$, and variants with MAF $< 3\%$. Under the equivalence with SKAT, this yielded a total of 12 different candidate kernels. We estimated the type I error rate at the 0.05 level of 1) SKAT with each individual kernel, 2) MK-SKAT conditional on a particular testing procedure (i.e. we assumed a fixed test while considering multiple groupings), 3) MK-SKAT conditional on a particular grouping strategy (i.e. we assumed a fixed grouping while considering multiple tests), and 4) MK-SKAT testing across all twelve candidate kernels.

Power

We also assessed the power of the MK-SKAT procedure under three different simulation settings. For each setting, we again simulated haplotypes for a region containing 100 variants as in the type I error simulations. These were then paired to generate $n = 1000$ individuals. Then we simulated outcomes under the alternative model for quantitative traits:

$$y_i = 0.5X_{i1} + 0.03X_{i2} + \boldsymbol{\beta}'\mathbf{Z}_i^c + \varepsilon_i$$

and for dichotomous traits:

$$\text{logit}P(y_i = 1) = -4.2 + 0.5X_{i1} + 0.03X_{i2} + \boldsymbol{\beta}'\mathbf{Z}_i^c$$

X_{i1}, X_{i2} and ε_i were as before, but $\mathbf{Z}_i^c$ were the genotypes of the causal variants and $\boldsymbol{\beta}$ were the corresponding regression coefficients which varied across simulation settings. For dichotomous outcomes $n/2$ subjects were sampled as cases with the remaining $n/2$ set as controls.

Under Setting 1, we considered a quantitative outcome with 50% of the variants with true population MAF $< 1\%$ randomly selected to be causal. All causal variants were given the same effect with $\beta = 0.5$. Since a large proportion of the variants were causal and they all had the same effect, this scenario favored the burden approaches and particularly count based collapsing.

Setting 2 again examined quantitative traits and was identical to Setting 1 except the effects of the causal variants were equal to -0.5 and 0.5 with equal probability. Since the causal variants had opposing effects, this scenario favored the similarity based tests.

Setting 3 differed from Settings 1 and 2 in that it examineed the case where the outcome was dichotomous. Of the variants with true MAF $< 3\%$, 20% were randomly selected to be causal. All causal variants were again given equal effect size of $\beta = 0.5$.

We emphasize that these simulations were not intended to serve as a comprehensive comparison of the methods across scenarios nor to understand when individual tests and

		C-alpha	SKAT	CAST	Count	MK-SKAT
n = 1000	0.5%	0.048	0.047	0.050	0.049	0.048
	1%	0.048	0.049	0.049	0.050	0.050
	3%	0.048	0.049	0.051	0.051	0.051
	MK-SKAT	0.050	0.051	0.051	0.051	0.051
n = 2000	0.5%	0.049	0.049	0.050	0.050	0.052
	1%	0.047	0.047	0.050	0.050	0.051
	3%	0.047	0.047	0.050	0.049	0.051
	MK-SKAT	0.052	0.051	0.052	0.051	0.050

Table 3.1: Type I error simulation results for quantitative traits. Each cell in the table corresponds to the type I error of SKAT using a kernel constructed based on the testing procedure at the top of the table and the grouping strategy at the left of the table. Rows and columns labeled MK-SKAT correspond to the omnibus tests across tests (with fixed group) and across groupings (with fixed test). The cells with both rows and columns labeled MK-SKAT correspond to the omnibus test across all test and groupings.

grouping strategies are optimal (since this depends on the true state of nature, which is unknown in any real data). Instead, these simulations serve to understand how MK-SKAT behaves relative to the best method and grouping strategy.

3.3 Results

3.3.1 Type I Error and Power

Type I error simulation results for quantitative traits and dichotomous traits are shown in Table 3.1 and Table 3.2, respectively. For quantitative traits, individual methods as well as MK-SKAT appropriately controlled the type I error at the $\alpha = 0.05$ level. However, for dichotomous traits, the C-alpha test and SKAT test tended to be conservative, reflectiing previous results (Wu et al., 2011). Thus, MK-SKAT tests were conservative as well.

Results of the power analysis for the 3 settings are shown in Tables 3.3 through 3.5. In Setting 1 (Table 3.3), the count kernel applied to the variants with MAF <1% performed the best, followed closely by the CAST kernel applied to the same grouping. This was not surprising considering they were best adapted to the true model in which all effects have the same size and direction, and only rare variants with MAF <1% are sampled to be causative. The MK-SKAT which tested over all 12 kernels had power slightly less than the most powerful single kernel. The results of the MK-SKAT testing across all 4 tests at the 1% MAF threshold

		C-alpha	SKAT	CAST	Count	MK-SKAT
n = 1000	0.5%	0.033	0.032	0.051	0.050	0.042
	1%	0.042	0.040	0.050	0.049	0.045
	3%	0.046	0.044	0.050	0.050	0.046
	MK-SKAT	0.039	0.037	0.052	0.051	0.044
n = 2000	0.5%	0.041	0.041	0.050	0.050	0.047
	1%	0.046	0.046	0.050	0.050	0.049
	3%	0.047	0.047	0.050	0.050	0.050
	MK-SKAT	0.047	0.045	0.051	0.051	0.047

Table 3.2: Type I error simulation results for dichotomous traits. Each cell in the table corresponds to the type I error of SKAT using a kernel constructed based on the testing procedure at the top of the table and the grouping strategy at the left of the table. Rows and columns labeled MK-SKAT correspond to the omnibus tests across tests (with fixed group) and across groupings (with fixed test). The cells with both rows and columns labeled MK-SKAT correspond to the omnibus test across all test and groupings.

group showed power would be nearly equivalent to the most powerful single kernel as well. Also, if one tested the count kernel over the 3 groupings, power would be conserved.

In Setting 2, power was dramatically decreased for the count and CAST kernels compared to Setting 1 (Table 3.4). This was due to the true model having bidirectional genetic effect on the outcome. Some rare variants increased the outcome, while some decreased the outcome. Compared to Setting 1, power was reduced for C-alpha and linear weighted kernels, but not to the same extent as count and CAST. C-alpha and linear weighted kernels applied to the variants with MAF <1% performed the best in Setting 2. MK-SKAT testing over all 12 kernels displayed power somewhat less than the most powerful single kernel, but much greater than any of the CAST or count kernels. If one applied MK-SKAT over the three groupings of the linear weighted kernel, power would be nearly equivalent to the most powerful single kernel. This setting clearly showed the adaptability of the MK-SKAT method under variation in the genotype/phenotype structure.

Setting 3 compared power between methods for a dichotomous outcome (Table 3.5). The linear weighted kernel applied to the variants with MAF <3% performed the best. They were best adapted to the true model where only 20% of the variants were truly causal, and rare variants with MAF <3% were sampled as causative. MK-SKAT testing over all 12 kernels had power slightly greater than the most powerful single kernel, though this is likely to be within

		C-alpha	SKAT	CAST	Count	MK-SKAT
n = 1000	0.5%	0.43	0.43	0.64	0.66	0.64
	1%	0.74	0.76	0.84	0.85	0.86
	3%	0.47	0.64	0.63	0.63	0.71
	MK-SKAT	0.69	0.72	0.81	0.85	0.84
n = 2000	0.5%	0.70	0.71	0.85	0.87	0.87
	1%	0.92	0.93	0.98	0.98	0.98
	3%	0.76	0.89	0.88	0.88	0.92
	MK-SKAT	0.92	0.93	0.97	0.98	0.97

Table 3.3: Power results for Setting 1. Each cell in the table corresponds to the power of SKAT using a kernel constructed based on the testing procedure at the top of the table and the grouping strategy at the left of the table. Rows and columns labeled MK-SKAT correspond to the omnibus tests across tests (with fixed group) and across groupings (with fixed test). The cells with both rows and columns labeled MK-SKAT correspond to the omnibus test across all test and groupings.

		C-alpha	SKAT	CAST	Count	MK-SKAT
n = 1000	0.5%	0.37	0.37	0.10	0.12	0.32
	1%	0.63	0.65	0.17	0.23	0.57
	3%	0.39	0.54	0.13	0.16	0.46
	MK-SKAT	0.60	0.63	0.16	0.23	0.55
n = 2000	0.5%	0.68	0.69	0.15	0.17	0.61
	1%	0.87	0.88	0.26	0.36	0.84
	3%	0.63	0.80	0.17	0.23	0.72
	MK-SKAT	0.87	0.89	0.27	0.36	0.83

Table 3.4: Power results for Setting 2. Each cell in the table corresponds to the power of SKAT using a kernel constructed based on the testing procedure at the top of the table and the grouping strategy at the left of the table. Rows and columns labeled MK-SKAT correspond to the omnibus tests across tests (with fixed group) and across groupings (with fixed test). The cells with both rows and columns labeled MK-SKAT correspond to the omnibus test across all test and groupings.

		C-alpha	SKAT	CAST	Count	MK-SKAT
n = 1000	0.5%	0.26	0.26	0.31	0.32	0.33
	1%	0.53	0.55	0.52	0.50	0.59
	3%	0.73	0.78	0.69	0.69	0.78
	MK-SKAT	0.77	0.79	0.72	0.73	0.80
n = 2000	0.5%	0.52	0.53	0.47	0.48	0.57
	1%	0.75	0.77	0.70	0.69	0.78
	3%	0.84	0.88	0.82	0.80	0.88
	MK-SKAT	0.90	0.91	0.85	0.86	0.91

Table 3.5: Power results for Setting 3. Each cell in the table corresponds to the power of SKAT using a kernel constructed based on the testing procedure at the top of the table and the grouping strategy at the left of the table. Rows and columns labeled MK-SKAT correspond to the omnibus tests across tests (with fixed group) and across groupings (with fixed test). The cells with both rows and columns labeled MK-SKAT correspond to the omnibus test across all test and groupings.

the range of monte carlo error. If one applied MK-SKAT to the three groupings using either the linear weighted or C-alpha kernel, power would nearly equivalent to the most powerful single kernel.

Overall, results show that while protecting type I error, the MK-SKAT can achieve power close to using the optimal test and grouping strategy. While there is generally some modest loss in power relative to the best choice, the proposed omnibus tests offer considerably better power than poor choices and represent a reasonable compromise. If one is able to restrict attention to a particular group of variants based on prior information or to a particular testing procedure based on hypotheses of the underlying model, then power can be further increased by restricting the MK-SKAT to fewer tests or fewer groupings.

3.3.2 Data Analysis

We examined the performance of our proposed method on a real data set. Briefly, we examined a single candidate gene containing 86 variants of which the majority had allele frequency less than 3%. Eight variants were non-synomymous and two were predicted to be harmful. The candidate gene was sequenced in 2000 individuals. In addition to genotype information, we had 42 separate outcomes traits and additional demographic covariates including age, gender and the top five eigenvalues of genetic variability. We performed analysis

to find whether our candidate gene had association with any of the 42 outcome traits.

To illustrate our method, we considered testing using CAST, count based collapsing, weighted count based collapsing, the C-alpha, and the default SKAT. For groupings, we considered using all of the variants in the region, the variants with MAF <3%, variants with MAF <1%, variants with MAF <0.5%, nonsynonymous variants, and variants predicted to be harmful. In total we considered 27 different kernels based on combinations of the test choice and grouping choice — the CAST, count based collapsing, and weighted count based collapsing were not applied to all of the variants. In addition to applying SKAT with each of the candidate kernels, we also applied the MK-SKAT testing across all 27 kernels.

Analysis results are presented in Figure 3.1. Overall, for many traits, using different methods and different groupings resulted in very different results in terms of significance. In general, MK-SKAT tended to yield results slightly less significant than those using the best kernel (choice of test and grouping strategy), but MK-SKAT still performed considerably better than poor choices of kernels.

3.4 Discussion

In analysis of genetic rare variants, given the difficulties associated with selecting a test and selecting a particular group of variants to test, MK-SKAT allows investigators to agnostically consider several different, popular, testing approaches as well as several different ways of thresholding the variants. Although there is some loss of power compared to the best single test and best grouping, the power is still considerably higher than when using a poor choice of test or a poor choice of grouping strategy. And type I error is conserved.

Restriction of the MK-SKAT to a smaller set of possible kernels (i.e. smaller set of tests or groupings) can yield higher power if the considered kernels are closer to the best test and grouping strategy. If such information is available, such as through previous studies of common variants within the region or through bioinformatics knowledge, we strongly encourage investigators to directly restrict interest to a smaller group of candidate kernels. On the other hand, in the absence of reliable prior knowledge, we recommend consideration of a wide range of kernels. Importantly, if kernels are very similar to one another, then the

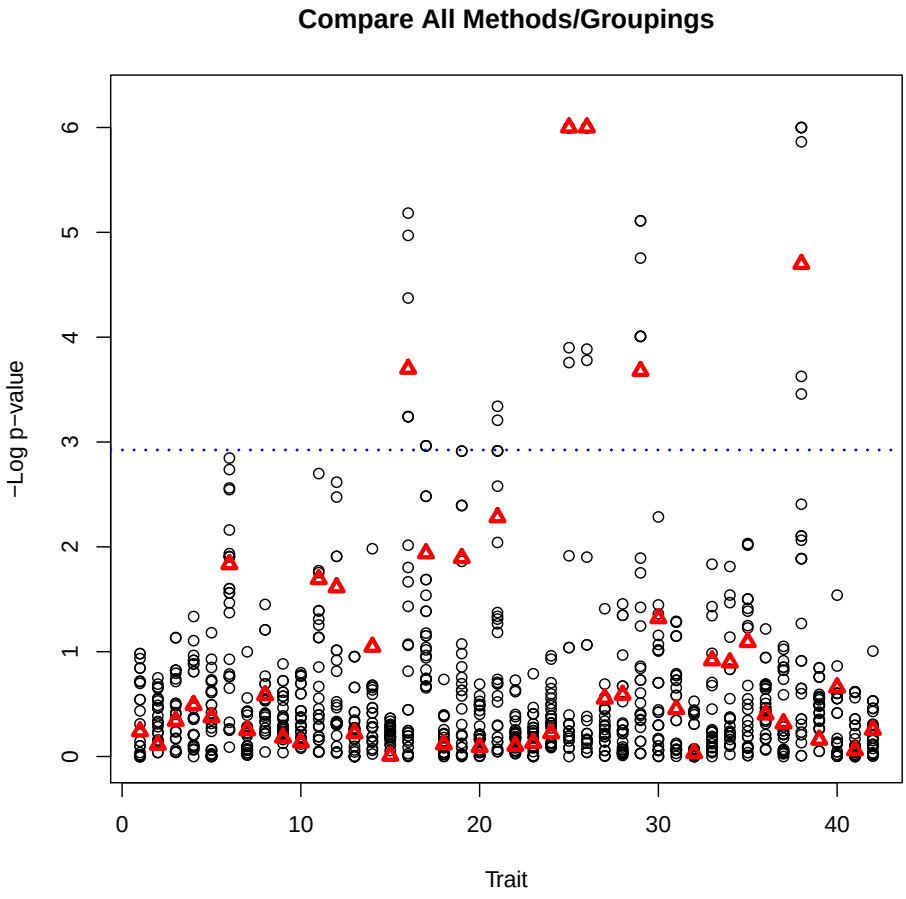


Figure 3.1: Real data analysis results. Each column of circles corresponds to the p-values from analyzing a different trait while each circle represents the p-value from a different kernel. The triangle indicates the p-value from applying MK-SKAT to all of the kernels. p-values have been truncated at 10^{-6}. The blue line indicates the bonferroni significance level.

perturbation procedure will accomodate the correlation and will not penalize the significance as much as if the considered kernels are more different.

Interestingly, while several methods are special cases of SKAT, some other methods are special cases of the MK-SKAT. The variable threshold test (Price et al., 2010) is equivalent to MK-SKAT when the kernels under consideration are based on a single testing approach with only the variable grouping being varied. However, we note that use of perturbation still offers computational advantage over the threshold test. Similarly, the SKAT-O method (Lee et al., 2012) is equivalent to MK-SKAT in which the variable grouping is fixed but one is considering a range of linear combinations of SKAT and collapsing kernels.

Further methods may also fall within the MK-SKAT framework, but although many popular tests can be considered using MK-SKAT, there are certainly many useful tests that fall outside. For example, tests that use the outcome information in order to estimate weights for variants (Ionita-Laza et al., 2011; Hoffmann et al., 2010; Han and Pan, 2010b) cannot be applied. While these tests still can be considered special cases of SKAT, the kernel is now estimated using the outcome such that standard asymptotics for SKAT and the perturbation based techniques for MK-SKAT cannot be used to obtain p-values. Further statistical work is needed in order to allow the MK-SKAT procedure to encompass these methods.

Chapter 4

Accommodating Partially Missing Covariates in the Sequence Kernel Association Test for Rare Variants

4.1 Introduction

A major focus of current human genetic research lies in the identification of genetic variants which influence disease and other complex phenotypes. Although genome wide association studies (GWAS) have found important associations between individual genetic variants and complex traits (Hindorff et al., 2009), much of the heritability is still left to be discovered (Eichler et al., 2010). DNA sequencing data promises to uncover a greater proportion of the herditability of complex traits (Margulies et al., 2005; Mardis, 2008; Ansorge, 2009), since sequencing allows for genotyping of not only the common single nucleatide polymorphisms (SNP) but also rarer genetic variants, that is genetic variants with minor alleles whose population frequency is lower. There is belief that rare variants can have larger effects on traits and a number of rare variants associated with disease have been reported (Cohen et al., 2006; Walsh et al., 2008; Nejentsev et al., 2009).

The interest in rare variants has spurred considerable research into new statistical and computational methods for testing the association between rare variants and complex traits. Since approaches for testing individual variants are often underpowered, region based testing, wherein the cumulative effect of multiple rare variants (such as within a gene) on an outcome is evaluated, has become the standard strategy. A wide range of region based tests have been developed with varying attributes. A key feature of many of these tests is the ability to accommodate covariate information. Within the context of genetic association studies, for

both common and rare variants, adjustment for covariates such as ancestry, age, gender, and environmental variables (Laird et al., 2000; Gauderman, 2003; Lunetta et al., 2000; Purcell et al., 2007) is essential in order to gaurd against confounding and prevent identification of spurious findings (Little and Rubin, 1987). Covariate adjustment can also result in improved power through reduction of the residual standard error.

While many popular rare variant association methods can control for potential confounders, difficulties arise if one or some of the covariates are partially missing on some individuals. Missing covariate information can arise through a range of processes including issues with the data collection process or due to design considerations, e.g. when a variable is measured on only a subset of individuals due to cost. Currently, a common approach for dealing with missing covariate information in rare variant studies is complete case analysis through case deletion, in which individuals with missing covariates are dropped from the analysis. Unfortunately, such approaches are problematic. In particular, using complete case observations only for partially missing covariates results in loss of power due to reduction in sample size. This is particularly troublesome for studies of rare variants if the subject with missing covariate information also is one of the few indviduals who have the particular variant. Furthermore, if the data are missing at random (MAR), such that missingness depends on the observed covariates, and only complete observations are used in the analysis, then there is a great danger of biased parameter estimation and potential difficulties in controlling type I error (Little and Rubin, 1987; Knol et al., 2010).

Recognizing the potential for misleading results or reduced power due to missing covariate information, in this paper, we consider the problem of partially missing covariates within the context of genetic sequencing studies of rare variants and develop an approach for testing the effect of rare genetic variants on a quantitative trait in the pressence of covariates that are MAR. Specifically, we focus attention on the popular sequence kernel association test (SKAT) (Wu et al., 2011), a region based test of rare variant effects in which pair-wise similarity in trait value between study subjects is compared to pair-wise similarity in rare variant profiles, with correlation suggestive of association. We extend SKAT accommodate missing data via use of a standard maximum likelihood based strategy for missing data based on the approach

of Ibrahim (1990). In particular, maximization of the likelihood can proceed via iteratively re-weighted least squares (IRLS) such that we can use the weighted linear model at convergence and apply those weights to SKAT.

Our objectives for gene sequence association study when covariates are partially missing and MAR are threefold: First, we would like to estimate the effects of covariates without bias and with high efficiency (low variance estimator). Second we would like to conserve SKAT type I error. Finally we are interested in maximizing SKAT power. Results of our simulation studies we show that complete case fails to estimate the effects of covariates unbiasedly and loses power as the proportion of missingess increases. In comparison, maximum likelihood acheives unbiased estimation of the effects of covariates, controls type I error, and retains power in comparison to oracle. Power is particularly improved in scenarios where the missingness proportion is large.

Our work restricts attention to the scenario in which only covariates may be partially missing at random. We do not consider the case in which variant information is missing — standard imputation techniques can accurately impute genotypes or at least dosages. Furthermore, we restrict attention to the MAR case and do not consider the case of not missing at random (NMAR) as general solutions for accommodating NMAR data remain elusive and require examination on a study-by-study basis. We also note that although our work focuses on SKAT, due to the close relationship between SKAT and other method such as burden tests or the C-alpha method, our approach can be seamlessly used for other testing procedures as well.

The remainder of this chapter is organized as follows. In the next section, we review the general SKAT method and then review the likelihood based approach for accommodating missing covariate information. We then describe how one can adapt SKAT to accommodate missing covariates. We examine the type I error rate and power of our approach, in comparison to complete case analysis, through a series of computer simulations. We conclude with a brief discussion.

4.2 Methods

For simplicity, we describe our work within the context of testing the association between the rare variants within a single region on a quantitative (continuous) trait, with the understanding that the approach can be applied genome-wide with appropriate adjustment for multiple comparisons. We denote the quantitative outcome for the i^{th} individual in the study as y_i $(i = 1, \ldots, n)$. The p variants within the region are in $\mathbf{Z}_i = [Z_{i1}, Z_{i2}, \ldots, Z_{ip}]'$ and the vector of covariates are denoted by $\mathbf{X}_i$. The objective is test for the effect of $\mathbf{Z}_i$ on y_i while adjusting for $\mathbf{X}_i$ with the additional complexity that variables within $\mathbf{X}_i$ may be missing for some individuals.

Here, we first describe the SKAT method for association testing in the scenario in which there is no missingness in $\mathbf{X}$ and then describe the likelihood based framework we are operating under. We then present our proposed extension of SKAT that accommodates missingness, focusing on the scenario in which only a single dichotomous covariate has missingness.

4.2.1 SKAT

SKAT is a similarity based test that operates by comparing pair-wise genotypic similarity between individuals to pair-wise phenotypic similarity, with correlation suggestive of association. To relate the variants and the covariates to the outcome, SKAT uses the semiparametric model

$$y_i = \beta_0 + \mathbf{X}_i'\boldsymbol{\beta} + h(\mathbf{Z}_i) + \varepsilon_i$$

where β_0 is an intercept, $\boldsymbol{\beta}$ is a vector of regression coefficients for the covariates, and ε_i is an error term with mean zero and variance σ^2. Within the kernel machine framework, the variants of interest $\mathbf{Z}_i$ for the i-th individual are related to the outcome only through the function $h(\cdot)$ which is a general function lying in a functional space generated by a positive definite kernel function $K(\cdot, \cdot)$. Intuitively, $K(\mathbf{Z}_i, \mathbf{Z}_{i'})$ measures similarity between i-th and i'-th individuals in the study based on $\mathbf{Z}$, the variants of interest.

The function $h(\cdot)$ fully specifies the relationship between the variants and the outcome:

if one sets $K(\mathbf{Z}_i, \mathbf{Z}_{i'}) = \mathbf{Z}_i'\mathbf{Z}_{i'}$, which is the linear kernel, then this implies that the function $h(\mathbf{Z}_i) = \sum_{j=1}^{p} \alpha_j Z_{ij}$ for some coefficients α_j, i.e. $h(\cdot)$ is linear and the outcome depends on the variants in a linear manner. By specifying a different kernel, one may specify an alternative model.

Under the default SKAT parameters, $K(\mathbf{Z}_i, \mathbf{Z}_{i'}) = \sum_{j=1}^{p} \theta_j Z_{ij} Z_{i'j}$ where θ_j is equal to a the beta probability density function with parameters 1 and 25 evaluated at the MAF for the j-th variant. This corresponds to a linear model but with additional up-weighting for the effect of rarer variants.

To test the effect of the rare variants under SKAT corresponds to testing $H_0 : h(\mathbf{Z}) = 0$. This can be done by exploiting the connection between kernel machine methods and linear mixed models. In particular, defining the kernel matrix, $\mathbf{K}$, to be the n-by-n matrix with i, i'-th term equal to $K(\mathbf{Z}_i, \mathbf{Z}_{i'})$, we can treat $h(\mathbf{Z})$ as a vector of subject specific random effects with mean 0 and variance $\tau\mathbf{K}$. Then whether $h(\mathbf{Z}) = 0$ corresponds exactly to testing whether $\tau = 0$. This is done by construction of the variance component score statistic

$$Q = \frac{(\mathbf{y} - \widehat{\mathbf{y}})'\mathbf{K}(\mathbf{y} - \widehat{\mathbf{y}})}{\widehat{\sigma}^2}$$

where $\widehat{\mathbf{y}} = \widehat{\beta}_0 + \mathbf{X}\widehat{\boldsymbol{\beta}}$ with $\widehat{\beta}_0$, $\widehat{\boldsymbol{\beta}}$, and $\widehat{\sigma}$ estimated under H_0.

To obtain a p-value for significance, asymptotically under the null, $Q \sim \sum \lambda_j \chi_1^2$ which is a mixture of chi-squares distributions, with weights λ_j equal to the eigenvalues of $\mathbf{P}_0^{1/2}\mathbf{K}\mathbf{P}_0^{1/2}$ where $\mathbf{P}_0 = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. This null distribution can be easily approximated using moment matching approaches or exact methods allowing for rapid p-value computation.

SKAT has been successfully applied in many studies, but unfortunately, it cannot accommodate missing covariates. Further developments are necessary.

4.2.2 Regression with Partially Missing Covariates

In many genetic association studies, including studies of rare variants, study subjects with missing covariate information are simply omitted from the analysis through complete case (CC) analysis which restricts analysis to a smaller subset of individuals on which complete

covariate information is observed. Although CC analysis is operationally simple, and in some studies the proportion and mechanisms of missingness do not dramatically influence the results, CC has the disadvantage of lower power due to reduced sample size (Little and Rubin, 1987). It is also biased under MAR scenarios and has decreased power compared to full-data methods (Knol et al., 2010). Thus, we consider the use of a full maximum likelihood based approach which is fitted via iteratively re-weighted least squares (IRLS). We will compare both CC analysis and our proposed strategy to an oracle procedure, that is an idealized scenario where the missing covariates are known.

As earlier, we again assume that the quantitative outcome for each individual is given as y_i and $\mathbf{X}_i$ are the covariates. However, without loss of generality and for ease of notation in our exposition, we assume that there are only two covariates X_{i1} and X_{i2} and that X_{i2} is a dichotomous variable that may be missing in some individuals but that X_{i1} is observed for all $i = 1, \ldots, n$ subjects. We further let R_i be an indicator for whether or not X_{i2} is observed for the i^{th} subjects ($R_i = 1$ if X_{i2} is observed). We will later describe extensions and accommodation of multiple missing covariates and continuous covariates. As noted previously, we assume that $\mathbf{Z}_i$ is observed without missingness.

Under the null model, in which the variants do not influence the outcome, we have

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

where ε_i is again a normal error term with mean zero and variance σ^2. Furthermore, we assume that

$$\text{logit}P(X_{i2} = 1) = \text{logit}\mu_i = \alpha_0 + X_{i2}\alpha$$

for some coefficients α_0 and α and the indicator of missingness is related to the outcome and the covariates through the logistic model

$$\text{logit}P(R_i = 1) = \text{logit}\eta_i = \omega_0 + X_{i1}\omega_1 + y_i\omega_y$$

for some coefficients ω_0 and ω_1 and ω_y. This way X_2 is MAR since $R|y, X_1$ is independent

of X_2. Under this model, we can then use maximum likelihood to accommodate the missing covariates without omission of samples.

Full-data Maximum likelihood

Maximum likelihood (ML) can be used to avoid the problem of discarding data which leads to reduced power and possible bias. Here, a distribution is given to the missing data and the resulting likelihood is integrated across all possible values of the missing covariate. When X_{i2} is observed, the full likelihood can be written as a product of conditional likelihoods.

$$p(y_i, X_{i2}, R_i|X_{i1}, \beta, \alpha, \omega) = p(y_i|X_{i1}, X_{i2}, \beta)p(X_{i2}|X_{i1}, \alpha)p(R_i|y_i, X_{i1}, \omega)$$

If X_{i2} is continuous, the probability must be integrated across all possible values of the missing X_{i2}. The probability for R can be evaluated outside the integral when the missingness occurs at random.

$$p(y_i, R_i|X_{i1}, \beta, \alpha, \omega) = p(R_i|y_i, x_{i1}, \omega) \int_{x_i} p(y_i|X_{i1}, X_{i2}, \beta) f_{X_{i2}}(x_i|X_{i1}, \alpha) dx_i$$

If X_{i2} is discrete, the probability must be summed across all possible values of the missing X_{i2}. Again, the probability for R can be evaluated outside the integral when missingness occurs at random.

$$p(y_i, R_i|X_{i1}, \beta, \alpha, \omega) = p(R_i|y_i, X_{i1}, \omega) \sum_{x_i} p(y_i|X_{i1}, X_{i2}, \beta)p(X_{i2} = x_i|X_{i1}, \alpha)$$

In summary, the full-data log-likelihood is:

$$\sum_{i=1}^{n} R_i \log\left[p(y_i, X_{i2}, R_i|X_{i1}, \beta, \alpha, \omega)\right] + (1 - R_i) \log\left[p(y_i, R_i|X_{i1}, \beta, \alpha, \omega)\right]$$

which can be solved throught either the Newton-Raphson method or by the Expectation-Maximization (EM) algorithm.

ML leads to unbiased results under MAR if the model is correctly specified. This is a clear

Y	X_1	X_2	R
y_1	X_{11}	X_{12}	1
y_2	X_{21}	X_{22}	1
y_3	X_{31}	N/A	0
y_4	X_{41}	X_{42}	1
y_5	X_{51}	N/A	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_n	X_{n1}	X_{n2}	1

$\longrightarrow$

Y	X_1	X_2	w
y_1	X_{11}	X_{12}	1
y_2	X_{21}	X_{22}	1
y_3	X_{31}	0	p_{i0}
y_3	X_{31}	1	p_{i1}
y_4	X_{41}	X_{42}	1
y_5	X_{51}	0	p_{i0}
y_5	X_{51}	1	p_{i1}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_n	X_{n1}	X_{n2}	1

Figure 4.1: Data augmentation using the approach of Ibrahim (1990) involves expanding each observation with missingness based on values that the missing variable can take. Here we assume that X_2 is dichotomous.

advantage over complete case and over most cases of imputation. An additional advantage of ML over imputation is that ML produces the same result each time, while most cases of imputation lead to differing results because of the variability of the imputed data.

Full-data maximum likelihood by Iteratively Reweighted Least Squares

Ibrahim (1990) proposed a weighted maximum likelihood to solve the above maximum likelihood. This is very helpful when the above likelihood does not lend itself to a closed form. Under Ibrahim's method, missing observation are expanded to multiple pseudo-observed observations and weighted according to their posterior probability of being observed (Fig. 4.1).

Following Ibrahim, the expressions take the form of a weighted complete data log-likelihood based on $N = \sum_{i=1}^{n} k_i$ observations, where k_i is the number of distinct covariate patterns for observation i. Thus, iteratively reweighted least squares (IRLS) is used in conjuction with the Newton Raphson algorithm to solve for covariate effects (β) and parameters of the distribution of missing covariate (α). This Newton-Raphson algorithm is considerably more convenient than the previous because it lacks the sum or integral.

The algorithm described by Ibrahim is as follows:

1. Augment missing data to weighted pseudo-observed data as described above. Each missing observation is augmented so that each possible realization of X_2 is represented

by one psuedo-observation. Those pseudo-observations are weighted (w) according the thier posterior probability of being observed.

where $p_{i0} = 1 - p_{i1} = P(X_{i2} = 0|y_i, X_{i1})$; and generally for missing X_{i2}:

$$w_i = P(X_{i2}|y_i, X_{i1}) = \frac{p(y_i|X_{i1}, X_{i2}, \boldsymbol{\beta}, \sigma^2)p(X_{i2}|X_{i1}, \boldsymbol{\alpha})}{\sum_{X_{i2}} p(y_i|X_{i1}, X_{i2}, \boldsymbol{\beta}, \sigma^2)p(X_{i2}|X_{i1}, \boldsymbol{\alpha})}$$

while $w_i = 1$ for non-missing X_{i2}

2. Use Newton Raphson on augmented psuedo-complete data, using complete case estimates as starting estimates. This form of Newton-Raphson with a pseudo full-likelihood provide tractible iterations in comparison with the difficult iterations involved with missing data. There are no sums or integrals to differentiate.

$$\begin{bmatrix} \boldsymbol{\beta}^{(t+1)} \\ \sigma^{2(t+1)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\beta}^{(t)} \\ \sigma^{2(t)} \end{bmatrix} +$$

$$\begin{bmatrix} \sum \mathbf{X}_i^T w_i \mathbf{X}_i & \frac{\sum \mathbf{X}_i^T w_i (y_i - \mathbf{X}_i\boldsymbol{\beta})}{\sigma^2} \\ \frac{\sum \mathbf{X}_i^T w_i (y_i - \mathbf{X}_i\boldsymbol{\beta})}{\sigma^2} & \frac{-\sum w_i}{2\sigma^2} + \frac{\sum w_i (y_i - \mathbf{X}_i\boldsymbol{\beta})^2}{(\sigma^2)^2} \end{bmatrix}^{-1} \begin{bmatrix} \sum \mathbf{X}_i^T w_i (y_i - \mathbf{X}_i\boldsymbol{\beta}) \\ -\frac{\sum w_i}{2} + \frac{\sum w_i (y_i - \mathbf{X}_i\boldsymbol{\beta})^2}{2(\sigma^2)} \end{bmatrix}$$

$$\boldsymbol{\alpha}^{(t+1)} = \boldsymbol{\alpha}^{(t)} + \left[\sum \mathbf{X}_i^T w_i \mathbf{X}_i \mu_i (1 - \mu_i)\right]^{-1} \sum \mathbf{X}_i^T w_i (X_{i2} - \mu_i)$$

3. Update w. With each subsequent iteration, the posterior probability of observating a psuedo-observation converges toward a more probable full-information estimate. The initial posterior probability is based on complete case only, which is biased assuming MAR.

$$w_i^{(t+1)} = \frac{p(y_i|X_{i1}, X_{i2}, \beta^{(t+1)}, \sigma^{2,(t+1)})p(X_{i2}|X_{i1}, \alpha^{(t+1)})}{\sum_{X_{i2}} p(y_i|X_{i1}, X_{i2}, \beta^{(t+1)}, \sigma^{2,(t+1)})p(X_{i2}|X_{i1}, \alpha^{(t+1)})}$$

4. Repeat 2 and 3 until convergence. Under MAR, the estimates at convergence are unbiased.

The objective then is to adapt this approach within the context of SKAT in order to

accommodate missing covariates for rare variant association testing. Since the likelihood is fit using a sequence of weighted linear regressions, the idea will be to use augment the data (including rare variants) and then using the working linear model at convergence and we apply the weights to SKAT and assess the significance of the rare variants.

4.2.3 Accommodating Missing Covariate Information in Tests of Rare Variants

The objective of our work is to allow for inclusion of subjects with partially missing covariates in studies of rare variants, in contrast to current strategies which focus on complete case analysis. To do this, we will employ the maximum likelihood approach implemented via the IRLS approach of Ibrahim (1990) within the context of rare variant analysis. We do this primarily within the context of SKAT, but we also discuss extensions to alternative rare variant testing procedures.

SKAT with Missing Covariates

Original development of the kernel machine testing framework using mixed models, the overaraching framework for SKAT, was done within the context of quantitative outcomes. Analysis of dichotomous and other types of outcome variables was done by using likelihood based models which can be fit via a sequence of weighted linear models, IRLS. Testing for non-quantitative outcomes then proceeds by utilization of the working linear model at convergence. Since we are proposing to use a full likelihood based approach for accommodation of missing covariates which can be fitted using IRLS, we also see that, at convergence, the working model is essentially just a weighted least squares regression. Thus, SKAT can also be applied using this working linear model.

Using the augmented versions of $\mathbf{y}$, $\mathbf{X}$ and $\mathbf{Z}$ containing N instead of n observations, we can again generate a kernel matrix $\mathbf{K}$ with (i, i')-th element equal to $K(\mathbf{Z}_i, \mathbf{Z}_{i'}) = \sum_{j=1}^{p} \theta_j Z_{ij} Z_{i'j}$ where θ_j is as defined earlier and $\mathbf{Z}$ now denotes the augmented matrix of genotype values. Then $\mathbf{K}$ is now $N \times N$ since it is generated from the augmented $\mathbf{Z}$. Using the augmented

data, we can construct a SKAT score statistic

$$Q_w = \frac{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T \mathbf{W}\mathbf{K}\mathbf{W}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{\hat{\sigma}_w^2}$$

where $\widehat{\boldsymbol{\beta}}$ is estimated under the null model which is fit via IRLS and $\mathbf{W} = diag(w_i)$. The estimate for σ^2 is given by

$$\hat{\sigma}_w^2 = \frac{1}{n-p} \sum_{i=1}^{n} \left(\sum_{X_{i2}=0,1} (w_i | X_{i2})(y_i - \mathbf{X}_i \beta) \right)^2 .$$

Overall, Q_w is similar in form to the original SKAT score statistic except the augmented observations are weighted by their contribution to the model based on the probability of the missing value.

To obtain a p-value for significance, asymptotically, Q_w is a mixture of chi-squares distributions, with weights λ_{jw} equal to the eigenvalues of $\mathbf{P}_{0W}^{1/2}\mathbf{K}\mathbf{P}_{0W}^{1/2}$; where $\mathbf{P}_{0W} = \mathbf{W} - \mathbf{W}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}$ for quantitative traits. Again, this can be approximated using moment matching approaches (Satterthwaite, 1946; Liu et al., 2009) or exact methods (Davies, 1980).

As noted, the idea behind our approach is that we are simply using the weighted linear model at convergence from maximization of the likelihood function for accommodating missing covariates. More intuitively, the original SKAT statistic essentially boils down to a quadratic form of the residuals estimated under the null model. Missing covariates prevent estimation of the null residuals, consequently we are simply obtaining unbiased estimates of the results using a likelihood based approach and then plugging these into the SKAT score statistic with accommodation for the correlation between residuals and for the weighted augmentation.

Othere Rare Variant Testing Approaches with Missing Covariates

Although the emphasis of our work is on using SKAT for rare variant testing, our framework can also be easily applied to other rare variant tests. In particular, we have noted in the previous Chapter that many other tests are equivalent to SKAT using particular kernels. For example, the count based collapsing method for testing rare variants corresponds to SKAT

using the kernel function

$$K(\mathbf{Z}_i, \mathbf{Z}_{i'}) = \left(\sum_{j=1}^{p} Z_{ij}\right)\left(\sum_{j=1}^{p} Z_{i'j}\right)$$

were the $\mathbf{Z}$ are again assumed to have been augmented. Then to allow for missing covariates under the count based collapsing method, we need only replace the usual SKAT kernel matrix in Q_w with a kernel matrix constructed based on the collapsing method. To use a different test which is equivalent to SKAT under a particular kernel, we need only change the kernel matrix.

4.2.4 Continuous Missing Covariates and Multiple Missing Covariates

For simplicity, we have presented our method under a simple scenario in which we have only two covariates of which one is completely observed and the other is partially missing. If we have multiple covariates which are completely observed, then the earlier results hold except we simply treat X_{i1} as a vector. However, we have further assumed that the missing covariate is dichotomous and that only one of the covariates is partially missing. In this section we discuss how we can relax these assumptions. We emphasize that this only reflects the estimation under the null model and thus the overall rare variant testing procedure remains the same except that we need to use an alternative approach to estimate the weights for the observations with missing covariates and then everything else remains as earlier.

When the missing covariate is continuous, we can still use the same approach as earlier except that $f(X_{i2}|X_{i1})$ is approximated by a discrete distribution and monte carlo is used to select L distinct points from the distribution along with the corresponding probability Ibrahim et al. (2004). Weights are then generated similar to the dichotomous case, and the weighted complete data log-likelihood is evaluated in the same way as before with

$$w_i = P(X_{i2}|y_i, X_{i1}) = \frac{p(y_i|X_{i1}, X_{i2}, \beta, \sigma_y^2) f(X_{i2}|X_{i1}, \alpha, \sigma_x^2)}{\int p(y_i|X_{i1}, X_{i2}, \beta, \sigma_y^2) f(X_{i2}|X_{i1}, \alpha, \sigma_x^2) dX_{i2}}$$

$$\begin{bmatrix} \alpha^{(t+1)} \\ \sigma_x^{2(t+1)} \end{bmatrix} = \begin{bmatrix} \alpha^{(t)} \\ \sigma_x^{2(t)} \end{bmatrix} +$$

$$\begin{bmatrix} \sum X_{i1}^T w_i X_{i1} & \frac{\sum X_{i1}^T w_i (X_{i2} - X_{i1}\alpha)}{\sigma_x^2} \\ \frac{\sum X_{i1}^T w_i (X_{i2} - X_{i1}\alpha)}{\sigma_x^2} & \frac{-\sum w_i}{2\sigma_x^2} + \frac{\sum w_i (X_{i2} - X_{i1}\alpha)^2}{(\sigma_x^2)^2} \end{bmatrix}^{-1} \begin{bmatrix} \sum X_{i1}^T w_i (X_{i2} - X_{i1}\alpha) \\ -\frac{\sum w_i}{2} + \frac{\sum w_i (X_{i2} - X_{i1}\alpha)^2}{2(\sigma_x^2)} \end{bmatrix}.$$

These values can be plugged in to obtain the SKAT score statistic again and a corresponding p-value.

The approach can also be easily extended to the scenario in which there are multiple missing covariates. For example, suppose that there are three covariates X_1, X_2 and X_3 and that X_2 and X_3 are missing for some individuals in the study. Under our assumptions, we can extend the likelihood such that the distribution of X_3 is conditional upon X_1 and X_2. In particular, we use the null model

$$y_1 = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$$

where all variables are as before except that we have an additional covariate X_{i3} which is missing in some individuals. Then we assume that

$$X_2 | X_1, \alpha \sim Ber\left(\mu = \frac{e^{X\alpha}}{1 + e^{X\alpha}}\right);\ r_2 | y, x_1, \omega \sim Ber\left(\eta = \frac{e^{X\omega}}{1 + e^{X\omega}}\right)$$

$$x_3 | x_1, x_2, \delta \sim Ber\left(\mu = \frac{e^{X\delta}}{1 + e^{X\delta}}\right);\ r_3 | y, x_1, x_2, \gamma \sim Ber\left(\eta = \frac{e^{X\gamma}}{1 + e^{X\gamma}}\right).$$

Weights are then generated in the same way as before using the likelihood with the additional covariates and correspond to the posterior probability of being observed at a particular value.

$$w_i = P(X_{i2}, X_{i3} | y_i, X_{i1}) = \frac{p(y_i | X_{i1}, X_{i2}, X_{i3}, \beta, \sigma^2) p(X_{i2} | X_{i1}, \alpha) p(X_{i3} | X_{i1}, X_{2i}, \delta)}{\sum_{X_{i2}} p(y_i | X_{i1}, X_{i2}, X_{i3}, \beta, \sigma^2) p(X_{i2} | X_{i1}, \alpha) p(X_{i3} | X_{i1}, X_{i2}, \delta)}.$$

4.3 Results

4.3.1 Type I Error Simulations

We conducted a series of simulations to assess the type I error of our method as compared to complete case analysis using scenarios similar to those of Wu et al. (2011). Specifically, we simulated a population of 10,000 haplotypes on a region of approximately 5 kb long containing 100 genetic variants using a coalescent model (Schaffner et al., 2005) with parameters set to mimic real genetic data from a population of European ancestry.

To generate a single simulated data set, we then randomly selected and paired haplotypes to generate genetic information on n diploid individuals. For each individual, we simulated two covariates with $X_1 \sim N(0,1)$ and $X_2 \sim Ber(\mu_i)$ with $\mu_i = 0.5X_{i1}$. Then the outcome for each individual was generated as

$$y_i = 1 + 0.5X_{i1} + 0.5X_{i2} + \varepsilon_i$$

where $\varepsilon_i \sim N(0,1)$. Note that the outcome does not depend on $\mathbf{Z}$. Each X_{i2} was set to be missing with probability η_i, where

$$\text{logit}\eta_i = \omega_0 + \omega_1 y_i + \omega_2 X_{i1}$$

with $\boldsymbol{\omega}$ tuned to achieve a particular degree of total missingness.

We considered scenarios in which X_2 was missing in 5%, 15%, 30% or 60% of the individuals. We also allowed n to vary as 500, 1000, and 2500. For each percentage of missingness in X_2 and sample size, we simulated 1000 data sets. We applied SKAT using complete case analysis and SKAT using IRLS to each simulated data set. We also considered applying SKAT under the oracle: that is we applied SKAT assuming that we knew the true value of X_2. While this is impossible in practice, it provides a reference to which we can compare our results. For each method, the type I error was estimated as the proportion of p-values less than 0.05.

The estimated type I error results are given in Table 4.1. Overall, each method controls

		Percent Missing			
Method	n	60%	30%	15%	5%
CC	500	0.042	0.050	0.031	0.051
	1000	0.040	0.050	0.056	0.049
	2500	0.054	0.054	0.055	0.043
Oracle	500	0.046	0.046	0.046	0.046
	1000	0.040	0.040	0.040	0.040
	2500	0.053	0.053	0.053	0.053
IRLS	500	0.047	0.045	0.045	0.048
	1000	0.039	0.040	0.042	0.037
	2500	0.052	0.055	0.050	0.057

Table 4.1: Type I error simulation results at the $\alpha = 0.05$ level comparing SKAT using complete case (CC), SKAT with IRLS to accommodating missing values (IRLS), or SKAT assuming that the missing values are known (Oracle).

the type I error rate. We also found that the coefficients for the covariates were estimated with no bias under the oracle and maximum likelihood based methods (not shown), though this was not the case for the complete case analysis. That the type I error rate is nearly controlled for the complete case approach is surprising, given the bias, but may be due to the fact that the SKAT method tends to be conservative in many cases.

4.3.2 Power Simulations

We also examined the power of the proposed approach in comparison to the oracle procedure and to complete case analysis.

Using the same strategy as before, we simulated genotype information and covariates in the same manner as in the type I error simulations. However, since we then simulated outcomes under the alternative model in which the rare variants within the region influence the outcome. Specifically, we simulated the outcome y_i as

$$y_i == 1 + 0.5X_{i1} + 0.5X_{i2} + \boldsymbol{\beta}^c \mathbf{Z}_i^c \varepsilon_i$$

where $\mathbf{Z}_i^c$ denotes the genotypes of the causal variants and $\boldsymbol{\beta}^c$ are the regression coefficients for the causal variants. Here, the causal variants were randomly selected as 5% of the variants with true MAF $< 3\%$. The effect for the j^{th} causal variant was given as $\beta_j^c 0.4|\log_1 0q_j|$, where

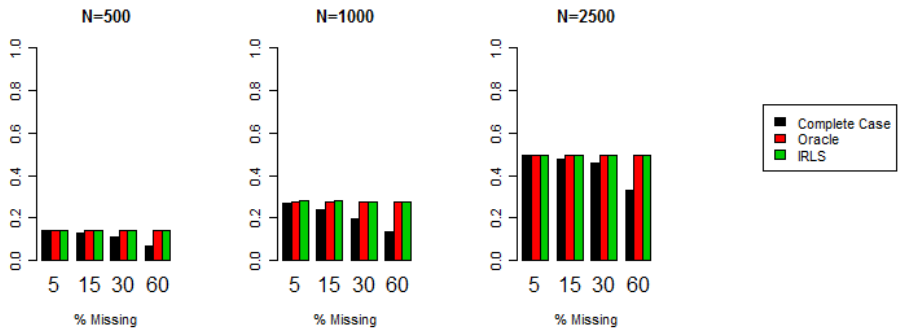


Figure 4.2: Power simulation results comparing SKAT using complete case (CC), SKAT with IRLS to accommodating missing values (IRLS), or SKAT assuming that the missing values are known (Oracle).

q_j is the true MAF of the j^{th} causal variant. This allows rarer variants to have strong effects on the outcome. Other parameters within the model are as before: we again considered $n =$ 500, 1000, and 2500 and missingness percentages for X_2 of 5%, 15%, 30% or 60%, and for each sample size and proportion of missingness, we again simulated 1000 data sets. We applied the SKAT under complete case, under the oracle, and under our proposed IRLS based approach to each data set and estimated the power in each scenario as the proportion of p-values less than the stringent $\alpha = 10^{-6}$ which reflects a level on the order of genome-wide significance.

The power results are shown in Figure 4.2. Results show that total power increases for all methods as sample size increases. However, for fixed sample size, as the proportion of missingness increases, complete case analysis loses power due to reduction in sample size (60% missingness when $n = 1000$ leads to power that is comparable to no missingness and with $n = 500$). On the otherhand, the proposed application of SKAT under the IRLS framework to accommodate missingness maintains power that is close to the oracle procedure, even when the missingness proportion is high. Interestingly, when the proportion of missingness is modest, e.g. 5%, the loss in power is not large for complete case analysis, suggesting that under some scenarios complete case analysis may not be terrible. Though, as the proportion of missingness goes up, the relative performance is much worse.

4.4 Discussion

Controlling for potential confounders and the effects of demographic and environmental covariates is important for sequencing association studies of rare variants in order to prevent spurious associations and can also improve power through reduction of the standard error. However, missing covariate data sometimes occurs and little has been done to accommodate the missing covariates. We have proposed a strategy based on full maximum likelihood using IRLS that can accommodate rare variants within the context of SKAT. We show through simulations that the approach conserves type I error while maintaining power close to the oracle. In contrast, complete case analysis, a standard approach for treatment of partially missing covariates, results in reduced power as the proportion of missingness gets large. These properties support the use of IRLS when SKAT is to be used in the presence of partially missing covarates.

Our proposed strategy shares many of the advantages of SKAT based tests. In particular, as a score test, the null model, which is where all adjustments for missingness are made, needs to be fit only once. This reduces the computational expense in genome wide experiments. Similarly, a p-value can be directly estimated without the need for monte carlo methods. However, since SKAT is closely related to several different methods, the proposed approach can also be directly applied to conduct several other tests including the CAST method, the count based collapsing method, and the C-alpha test, by simply switching the kernel function used to measure similarity between subjects based on their rare variants. Using IRLS and maximum likelihood to accommodate rare variants with still preserve the properties of each of these tests such that collapsing approaches will still be more powerful when the majority of variants function unidirectionally and SKAT and C-alpha will still be more powerful when the variants function bi-directionally. The approach can also be easily used within the context of other tests which are not exactly equivalent to SKAT under a single kernel, such as the variable threshold test (Price et al., 2010) and SKAT-O (Lee et al., 2012).

In our current work, we have focused on scenarios where there is only a single dichotomous covariate with considerable missingness. We have discussed the inclusion of continuous

covariates and multiple missing covariates, but all of this is within the context of quantitative outcomes. The approach can, in principle, be applied within the context of dichotomous (i.e. case-control) outcomes, but further development is needed. Another area requiring further research is inclusion of missing data within the variants; we find that this is, generally, less problematic since current imputation techniques are comparable to likelihood based procedures for common variants, but whether this still holds for rare variants is unclear and warrants more research.

Chapter 5

Kernel Machine Testing using Maximum Likelihood by IRLS for Gene Level Analysis of Methylation Data with Missing Covariates

5.1 Introduction

Large scale epigenome wide association studies (EWAS) Rakyan et al. (2011), in which the DNA methylation at hundreds of thousands of CpGs along the genome can be simultaneously measured across a large number of samples Bibikova et al. (2011); Sandoval et al. (2011), have resulted in the identification of differentially methylated CpGs associated with differences with a range of outcomes and contions Joubert et al. (2012); Shen et al. (2013); Heyn et al. (2013, 2012). These discoveries can provide a breadth of information from fundamental insights into the mechanisms underlying complex disease and to potential biomarkers for diagnosis or prognosis Laird et al. (2003); Attar (2012). However, despite the successes, analysis of EWAS remains challenging Bock (2012).

Standard analysis of EWAS proceeds via individual CpG analysis wherein the association between each CpG and an outcome variable (e.g. disease state, environmental exposure, etc.) is assessed one-by-one, followed by adjustment for multiple comparisons. Any CpGs surviving this correction are called differentially methylated and followed for validation and interpretation. However, this approach suffers a number of limitations (Subramanian et al., 2005; Wu and Lin, 2009). First, the need to correct for large number of multiple corrections can lead to low power such that nothing meets the criteria for signifiance. Alternatively, too many features are called significant leading to difficulties in interpretation. Individual feature analysis also fails to allow for capture of multi-feature or interactive effects. More generally,

such approaches have been found to yield poor reproducibility. An alternative to individual CpG analysis is to use multi-CpG analysis in which we group multiple CpGs together, such as those lying within a gene region, and test their cumulative effect on the outcome. Following similar principles as in gene expression and genetic association studies (Subramanian et al., 2005; Goeman et al., 2005; Wessel and Schork, 2006; Liu et al., 2007b, 2008; Tzeng and Zhang, 2007; Wang et al., 2007, 2010; Schaid et al., 2011), multi-CpG analysis can be used to overcome many of the limitations surrounding individual CpG analysis.

A particular approach that can be used for multi-CpG analysis is the kernel machine regression test which was initially proposed for gene expression data (Liu et al., 2007b, 2008) but has been also extended to analysis of SNPs (Kwee et al., 2008; Wu et al., 2010) and rare variants (Wu et al., 2011). Briefly, the approach is built upon a semi-parametric model within the kernel machine framework (Cristianini and Shawe-Taylor, 2000) in which the effects of a group of features of interest (e.g. genes in a pathway, SNPs in a region, etc.) are modeled nonparametrically and while some simple confounding covariates are adjusted for parametrically. A score test is used to test for an association between the outcome and the nonparametrically modeled group of features while linearly adjusting for the covariates. A key advantage of the kernel machine framework is the non-parametric modeling of the multi-feature effects. The approach can be directly applied to EWAS data in which the CpGs are grouped at the gene level.

An example of a study in which kernel machine regression based multi-CpG testing is useful is a recently conducted study of child birth weight in which epigenetic profiling of cord blood from approximately 1000 new-born infants was conducted within the Norwegian Mother and Child Birth Cohort (MoBa). In addition to methylation measurements at 485,000 CpG sites within 20,000 genes, for each subject in the study, a wide range of potential confounders including demographic variables and maternal behavior, diet, and environmental exposure data during pregnancy were collected. The goal was to identify associations between methylation at the gene level and birth weight while adjusting for the confounding variables. One particular confounder of interest is maternal vitamin D exposure which has been hypothesized to be linked to a range of birth outcomes and is a potential

confounder for birth weight. Unfortunately, vitamin D was measured in only a subset of the individuals such that the value is missing for a substantial number of subjects. Since each of the genes contains multiple CpG sites and the outcome is continuous, least square kernel machine regression is a natural analytic strategy, but the inability of kernel machine methods (as well as other multi-CpG tests) to accommodate partially missing covariate information poses a significant challenge.

To overcome the difficulties associated with gene level analysis of the MoBa epigenetic study of birth weight, we will consider using the method developed within the previous chapter. Although the development of the work was within the context of rare variant analysis, the overarching framework is generic and can also be applied in the present setting where we are interested in testing the effect of multiple CpGs instead of multiple rare variants. Despite this, there are a number of important differences between methylation values and rare variants. For example, methylation is typically measured as a continuous percentage (which is often logit transformed to be approximately normal), the number of CpGs within a gene is typically modest, and the correlation between adjacent CpGs is higher (whereas rare variants have low correlation due to their rarity). Therefore, this chapter involves investigation of the utility of the previous work on kernel machine testing with missing covariate information for gene level analysis of DNA methylation data.

The remainder of the chapter is organized as follows. Since the methods have been presented within the previous chapter, we do not repeat that here. Instead, in the next section, we directly proceed with simulation studies to examine the use of kernel machine test with missing covariates within the context of DNA methylation analysis. Specifically, we will compare the use of complete case analysis with the proposed method in terms of controling type I error and power. We then apply the proposed method to the motivating MoBa study of birth weight. We conclude with a very brief discussion.

5.2 Simulations

Since the structure of DNA methylation data is inherently different from rare variant data, we conducted simulations to ensure that the approach is also valid for continuous predictors

(the CpGs within a gene) and to examine the empirical power of our approach in comparison to the simpler complete case analysis strategy. We compare both complete case analysis and our proposed approach to the oracle procedure, i.e. kernel machine testing with the covariate value treated as known.

For our simulations, we let y_i be a continuous outcome for the i^{th} subject in the study $(i = 1, \ldots, n)$ and is simulated as

$$y_i = -3 + 0.5X_{i1} - 0.5X_{i2} + \mathbf{Z}_i'\boldsymbol{\xi} + \varepsilon_i$$

where $\mathbf{Z}_i$ is the vector of CpG methylation values within the gene with corresponding coefficients $\boldsymbol{\xi}$ and $\varepsilon_i \sim N(0, 1)$. Here, we let covariate X_{i1} follow a standard normal and covariate X_{i2} follow a normal distribution with mean equal to $0.5X_{i1}$ and variance σ_x^2. Since our interest is in examining missingness in the covariates, we allow X_{i2} to be missing for some individuals — for simplicity we assume that missingness is restricted to X_{i2}. We let r_i be the indicator of whether X_{i2} is observed and we assume that the probability that X_{i2} is observed (i.e. $r_i = 1$) depends on y_i and X_{i1} such that we have

$$\text{logit}P(r_i = 1) = \eta = \omega_0 + \omega X_{i1}$$

Thus, X_{i2} is considered to be missing at random (MAR) since $r_i|y_i, X_{i1}$ is independent of X_{i2}.

5.2.1 Type I Error

We first test type I error by applying SKAT to simulated data sets where methylation within the gene ($\mathbf{Z}$) has no effect on the outcome, i.e. $\xi = 0$.

For each individual, we simulated methylation data as a vector of 30, possibly correlated, normal random variables: $\mathbf{Z}_i \sim MVN(0, \boldsymbol{\Sigma})$. Although methylation is measured as a proportion between 0 and 1, it is often logit transformed to approximate normality. We considered five different correlation structures for the methylation values: independence, compound symmetry with $\rho = 0.15$, autoregressive with $\rho = 0.9$, block autoregressive with 10 blocks

with $\rho = 0.9$ within each block, and block compound symmetry with 10 blocks and $\rho = 0.15$ within each block. We tuned ω such that percent of missing covariate X_2 was 5%, 15%, 30% or 60%. We also considered sample sizes of $n = 500$ and $n = 1000$.

For each choice of correlation structure, sample size, and percent missingness, we simulated 10,000 data sets. For each data set, we conducted a complete case analysis using kernel machine testing under a linear kernel to assess the cumulative effects of the simulated CpGs on the outcome. We also applied the proposed kernel machine method with accommodating for missing covariates to each data set and for comparison, we also considered the oracle procedure in which we pretend that we knew the missing value. The type I error rate for each method was the proportion of p-values less than α level, where we considered several diffferent possible levels.

Table 5.1 shows type I error for the case of independent methylation data. Type I error was conserved for all 3 methods. However Figure 5.1 shows that type I error is not conserved for complete case analysis under certain correlation structures in methylation data. Specifically, type I error is well beyond acceptable limits when methylation is correlated in an autoregressive or compound symmetric fashion. Type I error inflation increases as missingness increases. Interestingly, under a block correlation structure improves type I error, but complete case still exceed limits in some cases. Type I error is conserved by using the oracle procedure or by using maximum likelihood by IRLS for all methylation correlation structures and sample sizes, even at more modest α levels. These results are different from what we observed with regard to rare variant analysis where type I error appeared to be conserved for complete case analysis, but this may be due to the fact that rare variants have near spherical correlation due to their low allele frequencies.

5.2.2 Power Simulations

We further assessed the power of the proposed kernel machine test using ML with IRLS to accommodate missing covariates. We also compared the usage of complete case analysis and the oracle procedure. As in the type I error simulations, we considered the same correlation structures and sample sizes. We also simulated covariates as in the type I error

	%mis	5%	5%	5%	15%	15%	15%
	α level:	0.0500	0.0050	0.0005	0.0500	0.0050	0.0005
	cc	0.0442	0.0030	0.0002	0.0434	0.0035	0.0005
	oracle	0.0447	0.0028	0.0004	0.0447	0.0028	0.0004
	irls	0.0444	0.0027	0.0004	0.0443	0.0030	0.0003
n=500	%mis	30%	30%	30%	60%	60%	60%
	α level	0.0500	0.0050	0.0005	0.0500	0.0050	0.0005
	cc	0.0426	0.0031	0.0003	0.0422	0.0029	0.0002
	oracle	0.0447	0.0028	0.0004	0.0447	0.0028	0.0004
	irls	0.0441	0.0031	0.0004	0.0412	0.0030	0.0002

	%mis	5%	5%	5%	15%	15%	15%
	α level	0.0500	0.0050	0.0005	0.0500	0.0050	0.0005
	cc	0.0464	0.0046	0.0002	0.0539	0.0044	0.0004
	oracle	0.0467	0.0045	0.0004	0.0467	0.0045	0.0004
	irls	0.0475	0.0041	0.0004	0.0478	0.0042	0.0004
n=1000	%mis	30%	30%	30%	60%	60%	60%
	alpha	0.0500	0.0050	0.0005	0.0500	0.0050	0.0005
	cc	0.0538	0.0055	0.0005	0.0559	0.0054	0.0006
	oracle	0.0467	0.0045	0.0004	0.0467	0.0045	0.0004
	irls	0.0460	0.0046	0.0003	0.0473	0.0044	0.0004

Table 5.1: Estimates of type I error in the application of kernel machine testing with complete case (cc) treatment of missing data, with oracle knowledge of the missing covariate values, and with ML by IRLS based analysis. Estimates are based on 10,000 simulated null model data sets under different sample sizes (n), signifiance levels (α), and percentage missingness (%mis). CpGs are uncorrelated here.

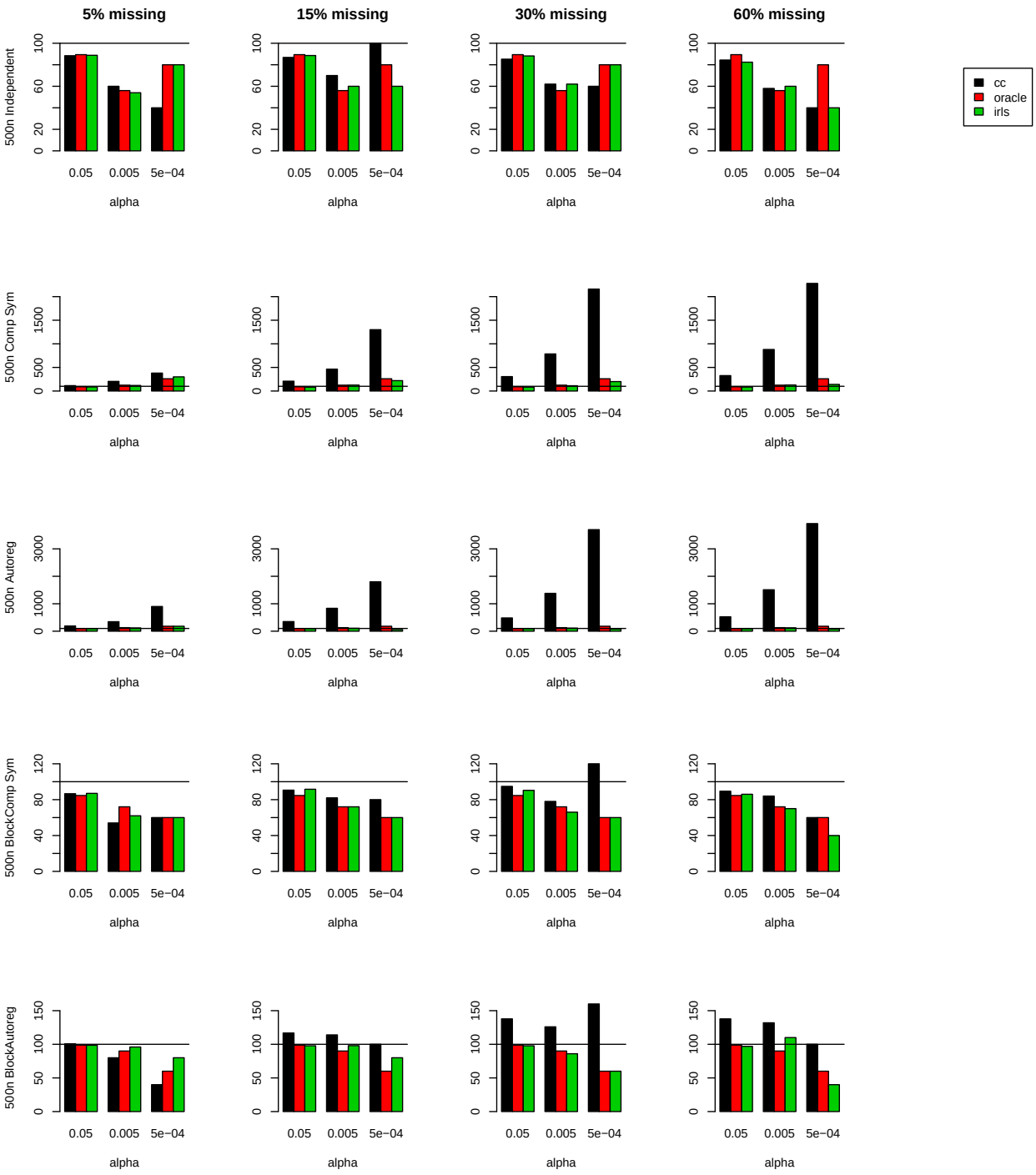


Figure 5.1: Scaled estimates of type I error in the application of kernel machine testing with complete case (cc) treatment of missing data, with oracle knowledge of the missing covariate values, and with ML by IRLS based analysis. Horizontal line indexes the ideal type I error level (alpha) and scaled to 100. Estimates are based on 10,000 simulated null model data sets under different signifiance levels, percentage missingness, and correlation structures. Sample size is fixed at $n = 500$.

simulations including missingness patterns, but now we allow the outcome y_i to depend on the methylation markers $\mathbf{Z}$. In particular, we randomly selected a single CpG to be causal with effect size ξ which differed depending the correlation structure: $\xi = 0.045$ when we used an independent or block compound symmetric correlation structure, $\xi = 0.025$ when we used a block autoregressive structure, and $\xi = 0.015$ when we used a compound symmetric or autoregressive structure. Power was estimated based on 10,000 simulations for each correlation structure and missingness pattern.

Power results are presented in Figure 5.2 and show that power of complete case decreases substantially as percent missingness increases, falling to zero as missingness approaches 60%. On the contrary, oracle and using kernel machine testing using ML with IRLS to accommodate missingness exhibit power that is very similar throughout levels of missingness. Using ML with IRLS does exhibit a modest decrease in power compared to oracle but this is considerably better than complete case analysis, despite the fact that complete case analysis does not well control type I error under some scenarios. These results are generally consistent across the different correlation structures in methylation data. Qualitatively similar results were observed for $n = 500$ and are not presented.

5.3 Application to Epigenetic Study of Birth Weight

We applied the kernel machine testing approach with ML by IRLS to accommodate missing covariate data and also kernel machine testing under complete case analysis to the motivating epigenetic study of birth weight.

Infant birth weight is an important variable related a child's subsequent development and health. Consequently, it is of great interest to understand the factors influencing a child's birth weight, including genomic factors. The MoBa epigenetic study of birth weight aimed to identify genes with methylation levels associated with differences in birth weight in infants. To this end, cord blood from over 1100 Norwegian infants in the MoBa cohort was obtained. The cohort has been described elsewhere. Following quality control, birth weight, covariates, and CpG methylation information was available on 1069 individuals. CpGs within 20,631 genes were available for analysis. The objective of the analysis is to examine each gene, one-

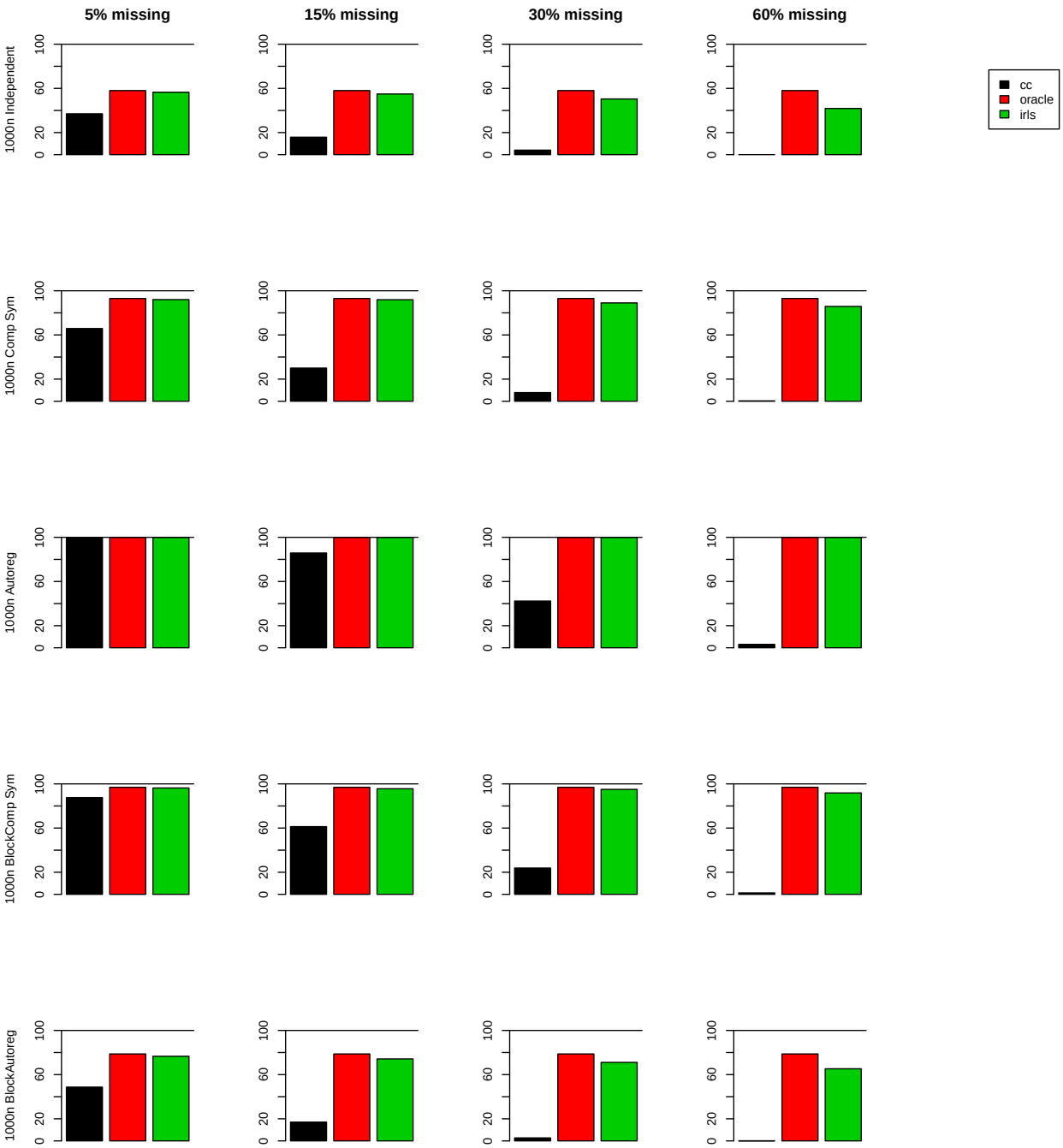


Figure 5.2: Scaled power estimates for kernel machine testing with complete case (cc) treatment of missing data, with oracle knowledge of the missing covariate values, and with ML by IRLS based analysis. Estimates are based on 10,000 simulated null model data sets under different signifiance levels, percentage missingness, and correlation structures. The effect size depended on the correlation structure to avoid saturation. Sample size is fixed at $n = 1000$.

at-a-time, and test for the cumulative effect of the CpGs within the gene on birth weight while adjusting for possible confounders.

Twelve different covariates related to birth weight were included in the analysis as possible confounders. Five of the twelve were partially missing in some individuals, but four of these five had 6 or fewer missing observations and so subjects missing in any of these four variables were removed from analysis, resulting in 9 total observations removed. The fifth covariate, vitamin D, had 123 missing values (11.6%) among the remaining 1060 observations which resulted from failure to collect this information due to expense. Our simulations show power loss when using complete case analysis coinciding with such a high percent of missingness. Thus, we apply our maximum likelihood by IRLS methodology to this partially missing covartiate. Vitamin D is a continuous covariate, thus we discretize the distribution, with 15 evenly spaced breaks. We standardized each of the covariates as well as the birth weight outcome prior to analysis.

We first estimated the covariate effects, comparing complete case to ML without consideration for the epigenetic data (i.e. under the null). Estimates of covariate effects were similar overall for the two methods (Table 5.2). Gestastional age showed the largest difference since it had the largest effect on the outcome.

We then conducted our primary analysis by applying kernel machine testing with ML by IRLS to accommodate the missing vitamin D levels to each of the 20,631 groups of CpGs, defined based on being in the same gene. Overall, following Bonferroni correction, 12 genes were associated with birth weight. In contrast, if we were to apply complete case analysis, reducing the sample size, then only three genes would have been found to be associated with birth weight. The genes are show in Table 5.3. All genes discovered by complete case analysis were also included in the list found by ML by IRLS reinforcing our simulation results indicating that complete case analysis often leads to reduced power.

5.4 Discussion

Our results have showed that using maximum likelihood by iteratively reweighted least squares is an attractive approach for multi-CpG association testing in the presence of partially

	CC	ML
Intercept	-0.01	0.00
Infant Sex	-0.14	-0.14
Cotinine	-0.05	-0.06
Gest. Age	2.19	2.46
Gest. Age^2	-1.69	-1.96
Parity = 1	0.19	0.19
Parity = 2	0.18	0.18
Parity $\geq$ 3	0.11	0.11
Maternal Age	-0.11	-0.09
$\log_1 0$ Folate	0.08	0.08
Asthma	0.03	0.03
Preeclampsia	-0.06	-0.07
Vitamin D	0.01	0.01

Table 5.2: Estimates of covariate effects on birth weight. The two procedures used are complete case and maximum likelihood by iteratively reweighted least squares

	Raw p		Corrected p	
	CC	IRLS	CC	IRLS
COBRA1	7.48E-05	2.20E-06	1.000	0.045
ENDOD1	7.37E-05	1.85E-06	1.000	0.038
FADS2	1.32E-05	4.24E-07	0.272	0.009
GRK6	1.54E-04	1.35E-06	1.000	0.028
GUCY1B2	1.13E-06	4.07E-07	0.023	0.008
KLF9	1.05E-06	2.08E-06	0.022	0.043
MBOAT4	9.19E-08	1.06E-08	0.002	<0.001
MNDA	1.43E-04	1.55E-06	1.000	0.032
SDPR	2.53E-06	2.76E-07	0.052	0.006
STAR	3.88E-06	7.50E-08	0.080	0.002
TRIM8	6.14E-05	1.37E-07	1.000	0.003
ZNF498	3.35E-05	5.96E-08	0.690	0.001

Table 5.3: Raw and Bonferroni Corrected p-values for the top results from the real data analysis. Kernel machine testing with maximum likelihood via IRLS is denoted by IRLS. Complete case analysis with kernel machine testing is denoted by CC.

observed covariates. When compared to a hypothetical oracle procedure where the covariates are not missing, power is only modestly decreased while type I error is conserved. On the contrary, our results confirm previous studies showing that complete case can lead to biased results and loss of power. Under certain correlation structures, complete case analysis can also lead to substantially inflated type I error, though this was not observed within our analysis of the MoBa epigenetic study of birth weight.

Chapter 6

Evaluation of Statistical Methods for Prioritization and Selection of Individual Rare Variants in Sequence Association Studies

6.1 Introduction

Despite the success of array based genome wide association studies (GWAS) of common variants in identifying genetic variants associated with a range of traits and diseases, such as Crohn's diseaseWTCCC (2007), type I and type II diabetesWTCCC (2007), lung cancerLandi et al. (2009); Li et al. (2010b), as well as many other traits (Hindorff et al., 2009), discovered variants explain only a modest proportion of heritability (Eichler et al., 2010). A portion of the missing heritability may be explained by rare genetic variants, that is variants with low minor allele frequencies (MAF) which were difficult to study in the past. However, recent advances in high-throughput sequencing technologySchuster (2008) have now enabled large scale studies examining uncommon gene variants through sequencing association studies which promise to identify rare genetic variants that further explain the heritability of complex traits.

Achieving the promises of sequencing studies has proven challenging. In particular, it is believed that analysis of these studies has been hindered by the low power of existing analysis methods for GWAS when applied to study rare variants. Consequently, a range of statistical methods have been developed for association testing in sequencing studiesLi and Leal (2008, 2009); Madsen and Browning (2009); Price et al. (2010); Neale et al. (2011); Yi and Zhi (2011); Wu et al. (2011). While there are important differences among the methods, they generally share the common strategy of focusing on region based testing which aims to

assess the cumulative effect of multiple rare variants in a region on the trait value. "Region" generally refers to a group of variants within a particular region of the genome (e.g. a gene), but the definition can be expanded to encompass any group of variants of interest. Aggregating information across multiple variants improves the power to identify regions that are associated with particular traits. However, because the tests focus on examining the joint effect of multiple uncommon variants, current methods cannot be used to conduct fine mapping to identify individual causal variants.

While detecting trait associated regions is important, subsequent evaluation of the contributions of individual causal variants within a gene region is crucial to achieving a comprehensive understanding of how genetic variation affects disease etiology and complex trait architecture. Pinpointing, or even prioritizing, individual variants would aid researchers interested in conducting in-depth functional analyses to interpret association results biologically. This is necessary to obtain clues as to the biological mechanisms underlying the relationship between the genetic factors and the observed trait phenotypes and better identify possible diagnostic and therapeutic options.

Although there has only been limited development of statistical methods for prioritizing individual variants, many methods commonly used for common variants can also be applied within the context of rare variants. Currently, the most common statistical tool used in the identification of individual common variants is marginal regression analysis in which the association between the trait value and each variant is examined, one-by-one. The method is ubiquitous in GWAS and other high dimensional data types with individual variants surviving some corrections for multiple comparisons considered to be of interest. An advantage of the approach is that method includes ability to control type I error when used in conjuction with multiple comparison correction. However, marginal analysis does not allow for the assessment of the individual variants while in the pressence of other variants. Furthermore, the power of standard tests for individual variants is tied to the MAF such that these methods may be underpowered for testing individual rare variants, though there has been some recent evidence that this power loss is over-stated.

As an alternative to marginal analysis, one may also consider methods that are built on

multivariable regression models such as sparse penalized regression approaches. In particular, variable selection procedures such as the Lasso (Tibshirani, 1996) operate under the multivariable regression framework but have the ability to performing simultaneous estimation and variable selection through inclusion of a penalty in the regression loss function. Sparse penalized regression methods have been widely applied within the context of analyzing common genetic variants for the purposes of fine mapping (Wu et al., 2009; Hoggart et al., 2008; He and Lin, 2010) and have also been proposed for analyzing rare genetic variants Zhou et al. (2010). Unfortunately, the main drawback of Lasso and its derivatives (Fan and Li, 2001; Zou, 2006) is that type I error control has not yet been established and some of these methods may over-select such that non-trait related loci are also included within the final models.

Due to the irregularity of the limiting distribution, standard methods for inference may not be appropriate within the context of sparse regression models. However, split sample resampling based methods such as stability selection (Meinshausen and Bühlmann, 2010) and a number of other related approaches (Valdar et al., 2012) have been proposed to enable error control. These methods have been applied within the context of analyzing common genetic variants (Alexander and Lange, 2011; Eleftherohorinou et al., 2011) and may also be useful for rare variants.

Although the properties of these procedures have been well studied, they have not been evaluated in the context of sequencing data and rare variant selection. In this paper we compare several of these methods in terms of their ability to correctly identify specific variants within a region that are associated with the disease status or other outcome while minimizing false positives. We also consider a simple approach which is based on forward selection in a multvariable regression model in which we sequentially add variants to the model and generate p-values conditional upon covariates and previously selected variants. Within these methods, we can also consider weighting by MAF, and the use of prior biological information on the ability to detect predictive genetic variants. Polyphen-2 (Adzhubei et al., 2010) and SIFT (Ng and Henikoff, 2003), for example, are becoming increasingly useful biological predictive tools. We compare the performance of each of these methods under three different criteria. First, we compare the methods with respect to selection of truly trait associated variants

(true positives) and incorrect selection of non-associated variants (false positives). Second we compare the ability to detect true postives indexed by minor allele frequency (MAF). Some methods may have better ability to identify associated variants that are relatively common or that are relatively rare. Finally, we compare ability of each method to rank order the individual variants into a relative list of importance.

6.2 Methods

In this paper, we focus on sequencing studies considering continuous traits. We assume that the study population consists of n unrelated subjects and we further assume that we are interested in identifying the causal variants within a single region.

For the ith subject ($i = 1, \cdots, n$), let y_i denote the value of the quantitative trait. $\mathbf{X}_i = (x_{i1}, x_{i2}, \cdots, x_{iq})^T$ denotes the covariates which can be either continuous or discrete and $\mathbf{Z}_i = (z_{i1}, z_{i2}, \cdots, z_{ip})^T$ denotes the genotype values of the p variants within the sequenced regions. We will assume an additive genetic model such that z_{ij} is the number of the minor alleles of the jth SNP, but we emphasize that our approach can also easily accommodate a dominant or recessive genetic model by simply changing the coding for z_{ij}.

Since we are focusing on quantitative traits, we employ a linear model defined by

$$y_i = \beta_0 + \boldsymbol{\beta}^T \mathbf{X}_i + \boldsymbol{\gamma}^T \mathbf{Z}_i + \epsilon_i \tag{6.1}$$

where β_0 is the intercept, $\boldsymbol{\beta}$ is the coefficient vector for the covariates, and $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \cdots, \gamma_p)^T$ is the coefficient vector for the p variants. The error term ϵ_i is assumed to have mean 0 and variance σ^2.

Since not all of the variants within the region are anticipated to be related to the outcome, the objective is to identify the variants Z_{ij} with corresponding $\gamma_j \neq 0$ such that the variant influences the outcome. In this section, we examine several different approaches that can be used to do this.

6.2.1 Marginal Analysis

The most commonly used method in genetic association studies is the marginal analysis such as by ordinary least squares for quantitative traits. Under this approach, each variant is tested separately for association with an outcome such as a disease while accomodating covariates such as environment or demographics. In particular, the effect of the j^{th} variant is evaluated assuming the model

$$y_i = \beta_0 + \boldsymbol{\beta}'\mathbf{X}_i + \gamma_j z_{ij} + \varepsilon_i.$$

A 1-df test can be used to test whether γ_j differs from zero. Then to achieve variable selection, the p-value for the association of the individual variant can be compared to a prespecified level. If it is below the level, then it is selected. If one wishes to control the type I error, then the threshold can be based on a pre-specified α-level to control family wise error rate (FWER) or adjusted to control the Benjamini-Hochberg false discovery rate (FDR) or some other criterion.

A common belief is that rare variants are more likely to influence trait values. Similarly, bioinformatics tools are now able to provide some prediction as to whether individual variants influence trait values. Weighting can be used to incorporate this prior knowledge and belief within this setting. One possible approach is to use the weighted FDR approach of Genovese et al. (2006). For example, if one wishes to incorporate MAF information into the selection process, using FDR we can multiple the p-value for each variant by the MAF of the corresponding variant which has been normalized by the total MAF to generate a weighted p-value

$$p_{w,k} = p_k * w_k = p_k * (\sum_{k=1}^{p} MAF_k)^{-1} MAF_k.$$

Since the arithmetic mean of the weights is equal 1, then FDR is conserved.

6.2.2 Lasso Based Methods

Another popular tool in variable selection in genetic association studies is the Lasso. Briefly, the Lasso operates under Model 6.1 and estimates the regression coefficients for the covariates $\boldsymbol{\beta}$ and the variants $\boldsymbol{\gamma}$ using the L_1 penalized loss function

$$\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}} = \underset{\boldsymbol{\beta},\boldsymbol{\gamma}}{\operatorname{argmin}} \, ||\mathbf{y} - \beta_0 - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\gamma||_2^2 + \lambda \sum_{j=1}^{p} |\gamma_j|.$$

The inclusion of the L_1 penalty allows for sparse estimation of the γs, i.e. for some of the γs to be estimated as exactly zero, when λ is large. λ is typically selected by grid search combined with cross validation or optimization of some criterion such as generalized cross validation (GCV) or BIC. For the purposes of this article, we use the AIC criterion (Akaike, 1974) for selection of λ.

Since all of the variants are considered simultaneously, it allows for some accommodating of correlation between variants and evaluation of the variants in the pressence of others. In the Lasso, there are no p-values to report, but rather simply the estimate of the variants' effect.

Within the context of the Lasso, it is also possible to use weighting to incorporate prior biological knowledge as with marginal analysis. In particular, we can use variant specific weights to adjust the penalty for each individual variant such that we estimate $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ as

$$\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}} = \underset{\boldsymbol{\beta},\boldsymbol{\gamma}}{\operatorname{argmin}} \, ||\mathbf{y} - \beta_0 - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\gamma||_2^2 + \lambda \sum_{j=1}^{p} w_j|\gamma_j|$$

where w_j is a prior weight that is related to the prior belief of the importance of the j^{th} variant. Larger values of w_j effectively increase the penalty for the j^{th} variant such that it is more likely to be shrunken to zero. Under the adaptive Lasso (Zou, 2006) sets w_j to $|\hat{\gamma}_j|^{-1}$ where $\widehat{\gamma}_j$ is some prior estimate for γ_j usually unpenalized least squares estimate, which allows for consistent variable selection under some conditions. Instead of using an initial estimate, we can also incorporate prior knowledge based on which variants are more likely to be causal. Thus, we can use weights that make it more likely for rare variants to be selected.

Specifically, for the j^{th} variant, we set the weight to be equal to the MAF for the j^{th} variant: since we believe variants with lower MAF are more likely to be important, this reduces the corresponding penalty for the variant making it more likely to be selected.

6.2.3 Stability Selection

In part because the Lasso does not control type I error and is believed to often over-select (leading to many false positives), Meinshausen and Bühlmann (2010) have proposed the stability selection procedure which is based on resampling the data and applying Lasso to allow for control of the expected number of false positives (the Per-Family Error Rate, PFER). Variants that are frequent selected across resamples are more likely to be true associated with the trait.

Operationally, stability selection proceeds by:

1. Randomly sample $n/2$ subjects from the total of n subjects in the study.

2. Apply the Lasso (or the weighted Lasso) using only the sampled $n/2$ subjects to select variants related to the outcome, but instead of optimizing a particular criterion, we select q, a prespecified number, of variants. Let $S_{n/2}$ denote the set of selected variants in the particular subsample.

3. Repeat the previous two steps B times for some large number B.

4. For each variant k, compute the selection probability as the proportion of times that the variant is selected across subsamples

$$\hat{p}_{k,n/2,B} = \frac{1}{B}\sum_{b=1}^{B} I(k \in \hat{S}_{n/2,b})$$

5. Variants whose selection probability is greater than a prespecified threshold (τ) are reported as selected.

By choosing different values of τ, stability selection can control the PFER by making two generally reasonable assumptions: 1) exchangeability, that is, all noncausal variants have

equal chance of selection; and 2) the using Lasso procedure is no worse than guessing. The PFER increases as the number of selected variables per subsample increases, and decreases with increasing total variables and increasing the threshold τ at which to accept a variable. The specific bound is given as

$$PFER \leq \frac{1}{2\tau - 1}\frac{q^2}{p}.$$

We note that this is an upper bound and that in practice, the observed PFER is usually substantially lower if the data and procedure are adequate due to the assumptions used in developing the bound.

6.2.4 Forward Selection

The final method we examine is based on simple forward selection. Although penalized regression methods have become quite popular in the statistical literature, forward selection is stll commonly used within many applied scenarios. Forward selection is applied by first testing individual variants for association with the outcome in the presence of all covariates. The single variant most highly associated with the outcome is then selected. We then include the selected variant as a covariate and then test each of the remaining variants for association with the outcome. The single variant most highly associated with the outcome, conditional on the covariates and the first selected variant, is again selected. This is repeated until all variants that are at all significantly associated with the outcome is added to the model.

Operationally, for testing the effect of additional variants while conditioning on previously selected variants and covariates, we use the score test implemented within SKAT. Since we are only testing a single variant, this score test is essentially equivalent to a standard 1-df score test within a multivariable regression model.

6.3 Simulations

We conducted a series of simulation studies to examine the relative performance of the considered methods for prioritizing and selecting individual variants that may be responsible for driving region level associations.

To simulate real sequencing data, we first generated a population of 10,000 haplotypes on a region of approximately 10 kb in length, containing 200 variants, using a coalescent model (Schaffner et al., 2005) calibrated to reflect a population of European ancestry. Using this population of haplotypes, we randomly selected $2n$ haplotypes and paired them to generate n diploid individuals. The vector of additively coded genotypes for the i^{th} simulated individual are given as $\mathbf{Z}_i$. We then simulated an outcomes for each of the $i = 1, \ldots, n$ individuals using the model

$$y_i = 0.5x_{i1} + 0.5x_{i2} + \boldsymbol{\gamma}'\mathbf{Z}_i^c + \varepsilon_i$$

where $x_{i1} \sim N(0, 1)$, $x_{i2} \sim ber(0.5)$ and ε_i. $\mathbf{Z}_i^c$ are the genotypes of the causal variants which were randomly selected as 20% of the variants with true MAF less than 1% in the simulated population. The coefficients for the causal variants are $\boldsymbol{\gamma}$ with γ_j, the coefficient for the j^{th} causal variant, set equal to $r_j 0.4|log_{10}MAF_j|$ where r_j is -1 and 1 with probabilities 0.2 and 0.8, respectively.

Since we are sometimes provided with prior biological knowledge concerning whether individual variants are actually causal, we also considerd simulation of scores reflecting prior knowledge. To mimic scenarios in which we have informative prior knowledge, we simulated scores from a Beta(2.5, 0.25) distribution for causal variants and we simulated scores from a Beta(0.25, 2.5) distribution for non-causal variants. These scores are meant to behave similarly to scores from Polyphen-2 or SIFT. We also considered some scenarios in which the prior knowledge is of poor quality and in these anti-informative settings, the distributions for causal and non-causal variants were reversed.

We considered several different scenarios based on different sample sizes and whether or not prior knowledge was useful. For each scenario, we simulated 1000 data sets. We applied each of the considered methods to each of the data sets to try to identify the individual causal variants. A number of different metrics for assessing the methods were considered.

6.3.1 Evaluative Metrics

We compare the methods in three ways. First, we examine number of true positives and false positives in relation to the total number of causal variants which which observed to have at least one minor allele in the data set. It is desirable to capture a large proportion of observed causal rare variants, but also to have a low number of false positives. Second we examine the number of true positives broken down by the minor allele frequency of the population from which the samples are drawn. This will give a picture of relative advatage by method with respect to the rarity of the variant. It may be helpful for the investigator to know the relative advantages of the methods prior to analysis. For example, the invistagor may be interested in more common variants because of higher population penetration. Or, contrarily, the investigator may be interested in rarer variants because they may be more potent.

Finally, we compare the methods in their ability to prioritize the variants in order of importance. Suppose further investigation requires a list of 5, 10 or 20 candidate variants. Analysis which automatically prioritizes will be ready to generate an imformative list. For our study, we use p-values to rank marginal anlaysis and SKAT forward ranks. We rank by magnitude of effect estimate for the Lasso. Stability selection ranks by the estimated selection probability. The formula we use produces a "rankscore" which ranges from 0 (none of the top 20 ranked variants are causal) to 1 (all top 20 ranked variants are causal). The formula gives higher priority to top ranked variants by weighting the top ranked variant 20, second 19, until the lowest 1, with the final score scaled to produce range 0 to 1. Rankscore is calculated as:

$$c\sum_{r=1}^{l}(l+1-r)I(z_r \in S)$$

where z_r represents the r ranked variant. l is the length of the list and S is the group of true associated variants. c is the scaling factor $(\sum_{r=1}^{l} r)^{-1}$ which ensures a score between 0 and 1.

6.3.2 Results

Table 6.1 displays results with under the scenario in which set let $n = 1000$ and no prior knowledge was available beyond allele frequency. From the table, we see that Lasso methods captured a greater proportion of the causal variants. Here there were an average of 15.5 observed causal rare variants across simulations, representing the maximum number of true positives that any approach can find. The Lasso methods captured on average close to 11 of these, while any other methods failed to catch more than 5. However the drawback is that Lasso methods have no way to control type I error and tend to vastly over-select. This was clearly seen in the high number of false positives selected through AIC.

Examination of true postives by MAF shows that gains in power were more prevalent in the lower range of MAF for the Lasso method, especially the very rare variants with MAF less than 0.1% where Lasso based methods capture 8 to 9 compared to 2 to 3 in the other methods. Finally, Lasso based methods, and also SKAT forward selection, showed the greatest ability to correctly arrange variants by level of importance, since the rank score was high.

Now, among the Lasso methods, adaptive and naive Lasso show better ability than weighted Lasso. This may have been partially due to the fact that the LASSO requires standardization of the design, and thus since standard deviation is clearly already related to MAF, is automatically adjusted for MAF. Further weighting is unecessary or perhaps harmful according to these results.

Among methods which had lower false positive rates, forward selection using SKAT had the most desireable outcomes. In fact SKAT forward selection dominated both marginal analysis and stability selection in that it had both more TP and fewer FP. SKAT forward had rank score on par with Lasso based methods and higher than the other error-controlled methods.

Weighting as a general strategy seemed to range from uncertain gain to unhelpful. It is not clearly shown in any method that weighting by MAF increased TP while lowering FP.

Examining the effect of sample size on abilty of the methods to detect causal variants shown in figure 6.1, we saw patterns seen for the default 1000 sample size persist in the other

	TP	FP	TP by MAF 1% to 0.5%	0.5% to 0.1%	$\leq 0.1\%$	rank.score
Marginal	2.3	0.7	0.5	1.0	0.9	0.42
Marginal-w	2.8	0.8	0.4	0.9	1.6	0.53
Lasso	10.9	15.2	0.6	2.3	8.1	0.62
Lasso-w	11.2	27.8	0.3	1.3	9.6	0.52
Adaptive Lasso	11.6	16.0	0.5	2.2	8.8	0.61
Stability Selection τ =0.6	4.9	3.5	0.5	2.0	2.4	0.49
Stability Selection τ =0.7	3.1	1.0	0.5	1.6	1.0	0.49
SKAT forward	3.2	0.3	0.5	1.2	1.5	0.63
SKAT forward-w	4.5	0.8	0.5	1.2	2.8	0.63
Mean observed	15.5	68.2	0.6	2.6	12.3	

Table 6.1: Comparison of methods in ability to correctly identify causal variants for default simulation setting. Measures of comparison include true positives, false positives, and true postitives indexed by minor allele frequency. Additionally presented is a rankscore, which measures ability to informatively order variants by level of importace, with 1 meaning all 20 top ranked variants are causal, and 0 meaning none are causal.

sample sizes as well. While all methods improve detection with increased sample size, stabilty selection showed much more variability due to sample size, with poor results in the $n = 500$ setting.

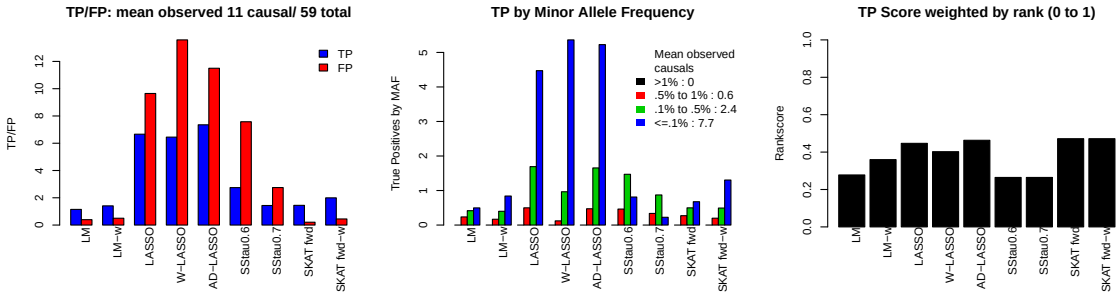
Finally, we examined the effect of prior information. All methods except stability selection gained from informative prior information. All methods clearly performed poorly under anti-informative prior information. This indicates that inclusion of prior knowledge can significantly improve analyses, but inclusion of unreliable knowledge or knowledge that goes against the truth can result in decreased ability to identify causal variants.

6.4 Data Analysis

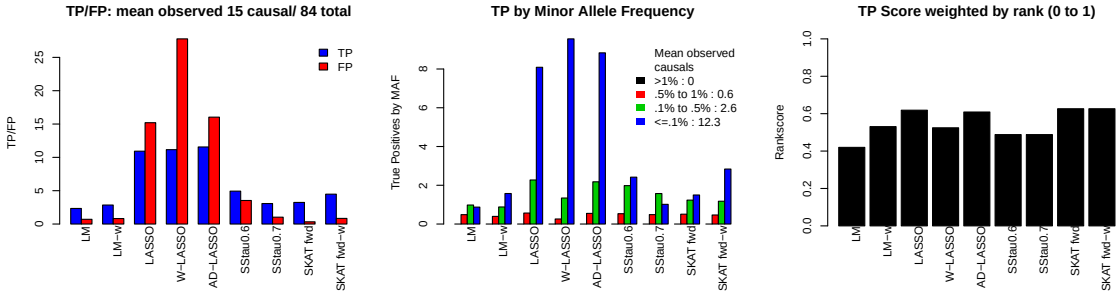
6.4.1 Overview

We applied the four methods to a real data set, in which we wished to find genetic variants associated with a quantitative trait related to lung function across 1898 individuals. The trait was continuous and the data set contained 8 additional covariates. Within the region, there were 86 genetic variants, of which 17 had MAF over 1%, 2 between 0.5% and 1%, 16 between 0.1% and 0.5%, and 51 less than 0.1%. We applied the methods as before and reported the

$n = 500$



$n = 1000$



$n = 2000$

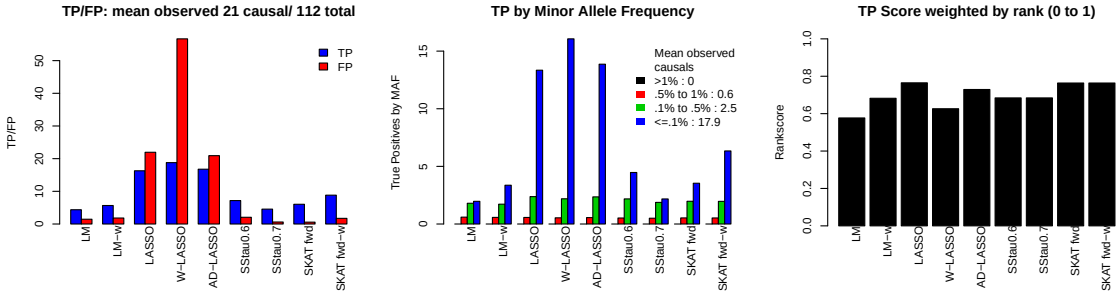
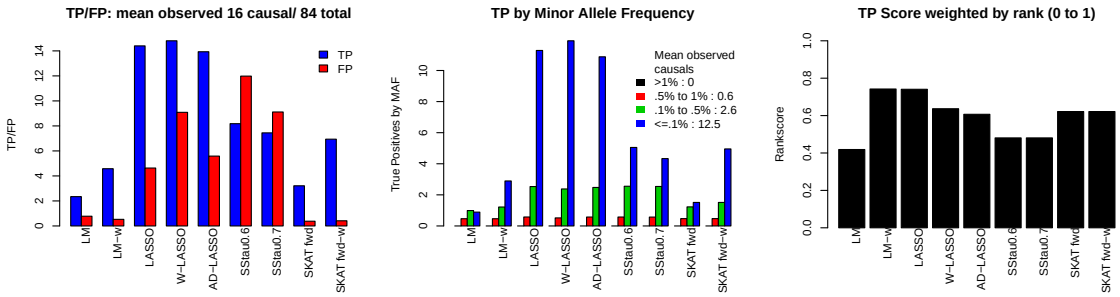
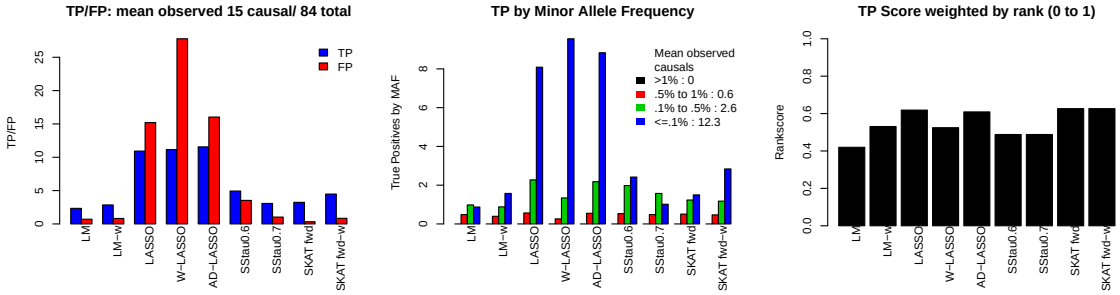


Figure 6.1: Simulation results for varied sample size. Left column compares methods by true postives and false postives, with total observed causal variants and total variants noted for comparison. Middle column compares methods by true postives with respect to minor allele frequency, with total observed variant by MAF noted for comparison. Right column compares methods by their ability to order variants by rank of importance, with 0 worst and 1 perfect.

Informative Prior



No Prior



Anti-Informative Prior

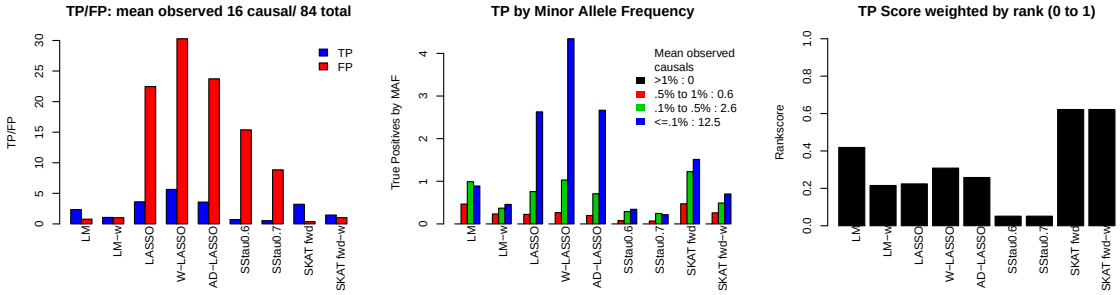


Figure 6.2: Simulation results for varied prior information. Left column compares methods by true postives and false postives, with total observed causal variants and total variants noted for comparison. Middle column compares methods by true postives with respect to minor allele frequency, with total observed variant by MAF noted for comparison. Right column compares methods by their ability to order variants by rank of importance, with 0 worst and 1 perfect.

	Total	$> 1\%$	1% to 0.5%	0.5% to 0.1%	$\leq 0.1\%$
Marginal	1	0	0	1	0
Marginal-w	2	0	0	1	1
LASSO	3	0	0	1	2
LASSO-w	16	0	0	1	15
Adaptive LASSO	5	0	0	1	4
Stability Selection τ =0.6	2	1	0	0	1
Stability Selection τ =0.7	1	0	0	0	1
SKAT forward	1	0	0	1	0
SKAT forward-w	2	0	0	1	1
Mean observed	86	17	2	16	51

Table 6.2: Real data application: Comparison of methods in number of variants identified as being associated with homeostatic model assessment levels. Measures of comparison include total selected variants, and selected variants indexed by minor allele frequency.

number of variants selected.

6.4.2 Results

The Lasso based methods selected several variants (Table 6.2). Marginal regression found 1 variant without weighting, and 2 with weighting. The naive Lasso found 3 association, weighted Lasso found 16, while adaptive Lasso found 5. Stability selection found 2 and 1 at τ equal to 0.6 and 0.7 respectively. SKAT forward selection found 1 variant without weighting, and 2 with weighting. Overall, results are similar to what we anticipated based on the simulation results.

6.5 Discussion

Lasso methods provide the power to detect the greatest number of associated variants, but lack type I error control and consequently result in large numbers of false positives. Neither adaptive Lasso nor weighting by minor allele frequency seem to increases power or reduce false positives.

Among methods which provide adequate error control SKAT forward selection provides the greater power than marginal analysis. These methods are useful when the desired result is a "clean" list of variants with prespecified number of false postivies. When relative order of importance is of greatest concern Lasso and SKAT forward selection provide the greatest

precision.

Stability selection, which applies observation resampling to Lasso, provides crude error control, but performs poorly under moderate sample size. It is dominated by SKAT forward selection in both true positives and false positives and cannot be recommended.

All methods benefit greatly from the use of informative prior information, such as that provided by Polyphen-2 or SIFT.

APPENDIX: SUPPLEMENTARY MATERIAL FOR CHAPTER 4

A1 Derivation of IRLS Newton Raphson: Partially Observed Discrete Covariate

Here we derive the Newton Raphson algorithm to solve IRLS weighted maximum likelihood for β, σ^2, α (It is not necessary to estimate ω)

$$wl(y, x_2|x_1, \beta, \sigma^2, \alpha) = wl(y|x_1, x_2, \beta, \sigma^2) + wl(x_2|x_1, \alpha)$$

$$= -\frac{\sum w_i}{2}\log\sigma^2 - \frac{\sum w_i}{2}\log 2\pi - \frac{\sum w_i(y_i - X_i\beta)^2}{2\sigma^2}$$

$$+\sum w_i x_{2,i} X_i\alpha - \sum w_i \log(1 + e^{X_i\alpha}) =$$

$$\frac{dl}{d\beta} = \frac{\sum X_i^T w_i(y_i - X_i\beta)}{\sigma^2}$$

$$\frac{dl}{d\sigma^2} = -\frac{\sum w_i}{2\sigma^2} + \frac{\sum w_i(y_i - X_i\beta)^2}{2(\sigma^2)^2}$$

$$\frac{d^2l}{d\beta^2} = \frac{-\sum X_i^T w_i X_i}{\sigma^2}$$

$$\frac{d^2l}{d\beta d\sigma^2} = \frac{-\sum X_i^T w_i(y_i - X_i\beta)}{(\sigma^2)^2}$$

$$\frac{d^2l}{d(\sigma^2)^2} = \frac{\sum w_i}{2(\sigma^2)^2} - \frac{\sum w_i(y_i - X_i\beta)^2}{(\sigma^2)^3}$$

Thus:

$$\begin{bmatrix} \beta^{(t+1)} \\ \sigma^{2(t+1)} \end{bmatrix} = \begin{bmatrix} \beta^{(t)} \\ \sigma^{2(t)} \end{bmatrix} +$$

$$\begin{bmatrix} \sum X_i^T w_i X_i & \frac{\sum X_i^T w_i(y_i - X_i\beta)}{\sigma^2} \\ \frac{\sum X_i^T w_i(y_i - X_i\beta)}{\sigma^2} & \frac{-\sum w_i}{2\sigma^2} + \frac{\sum w_i(y_i - X_i\beta)^2}{(\sigma^2)^2} \end{bmatrix}^{-1} \begin{bmatrix} \sum X_i^T w_i(y_i - X_i\beta) \\ -\frac{\sum w_i}{2} + \frac{\sum w_i(y_i - X_i\beta)^2}{2(\sigma^2)} \end{bmatrix}$$

and

$$\frac{dl}{d\alpha} = \sum w_i x_{2,i} - \sum X_i^T w_i \frac{e^{X_i\alpha}}{1 + e^{X_i\alpha}}$$

$$\frac{d^2l}{d\alpha^2} = -\sum X_i^T w_i X_i \left(\frac{e^{X_i\alpha}}{1+e^{X_i\alpha}} - \frac{(e^{X_i\alpha})^2}{(1+e^{X_i\alpha})^2} \right) = -\sum X_i^T X_i \left(\frac{e^{X_i\alpha}}{(1+e^{X_i\alpha})^2} \right)$$

Thus:

$$\alpha^{(t+1)} = \alpha^{(t)} + \left[\sum X_i^T w_i X_i \mu_i (1-\mu_i) \right]^{-1} \sum X_i^T w_i (x_{2,i} - \mu_i)$$

A2 Derivation of IRLS Newton Raphson: Partially Observed Continuous Covariate

Here we derive the Newton Raphson algorithm to solve IRLS weighted maximum likelihood for $\beta, \sigma_y^2, \alpha, \sigma_x^2$ (It is not necessary to estimate ω). We assume that $\mathbf{X}_1$ is fully observed and X_2 partially observed.

$$wl(y, x_2 | x_1, \beta, \sigma_y^2, \alpha, \sigma_x^2) = wl(y | x_1, x_2, \beta, \sigma_y^2) + wl(x_2 | x_1, \alpha, \sigma_x^2)$$

$$= -\frac{\sum w_i}{2} \log \sigma_y^2 - \frac{\sum w_i}{2} \log 2\pi - \frac{\sum w_i (y_i - X_i\beta)^2}{2\sigma_y^2}$$

$$-\frac{\sum w_i}{2} \log \sigma_x^2 - \frac{\sum w_i}{2} \log 2\pi - \frac{\sum w_i (x_{i,2} - X_{i,1}\alpha)^2}{2\sigma_x^2} =$$

$$\frac{dl}{d\beta} = \frac{\sum X_i^T w_i (y_i - X_i\beta)}{\sigma_y^2}$$

$$\frac{dl}{d\sigma_y^2} = -\frac{\sum w_i}{2\sigma_y^2} + \frac{\sum w_i (y_i - X_i\beta)^2}{2(\sigma_y^2)^2}$$

$$\frac{d^2l}{d\beta^2} = \frac{-\sum X_i^T w_i X_i}{\sigma_y^2}$$

$$\frac{d^2l}{d\beta d\sigma_y^2} = \frac{-\sum X_i^T w_i (y_i - X_i\beta)}{(\sigma_y^2)^2}$$

$$\frac{d^2l}{d(\sigma_y^2)^2} = \frac{\sum w_i}{2(\sigma_y^2)^2} - \frac{\sum w_i (y_i - X_i\beta)^2}{(\sigma_y^2)^3}$$

Thus:

$$\begin{bmatrix} \beta^{(t+1)} \\ \sigma_y^{2(t+1)} \end{bmatrix} = \begin{bmatrix} \beta^{(t)} \\ \sigma_y^{2(t)} \end{bmatrix} +$$

$$\begin{bmatrix} \sum X_i^T w_i X_i & \frac{\sum X_i^T w_i (y_i - X_i\beta)}{\sigma_y^2} \\ \frac{\sum X_i^T w_i (y_i - X_i\beta)}{\sigma_y^2} & \frac{-\sum w_i}{2\sigma_y^2} + \frac{\sum w_i (y_i - X_i\beta)^2}{(\sigma_y^2)^2} \end{bmatrix}^{-1} \begin{bmatrix} \sum X_i^T w_i (y_i - X_i\beta) \\ -\frac{\sum w_i}{2} + \frac{\sum w_i (y_i - X_i\beta)^2}{2(\sigma_y^2)} \end{bmatrix}$$

and similarly:

$$\begin{bmatrix} \alpha^{(t+1)} \\ \sigma_x^{2(t+1)} \end{bmatrix} = \begin{bmatrix} \alpha^{(t)} \\ \sigma_x^{2(t)} \end{bmatrix} +$$

$$\begin{bmatrix} \sum X_{i,1}^T w_i X_{i,1} & \frac{\sum X_{i,1}^T w_i (x_{i,2} - X_{i,1}\alpha)}{\sigma_x^2} \\ \frac{\sum X_{i,1}^T w_i (x_{i,2} - X_{i,1}\alpha)}{\sigma_x^2} & \frac{-\sum w_i}{2\sigma_x^2} + \frac{\sum w_i (x_{i,2} - X_{i,1}\alpha)^2}{(\sigma_x^2)^2} \end{bmatrix}^{-1} \begin{bmatrix} \sum X_{i,1}^T w_i (x_{i,2} - X_{i,1}\alpha) \\ -\frac{\sum w_i}{2} + \frac{\sum w_i (x_{i,2} - X_{i,1}\alpha)^2}{2(\sigma_x^2)} \end{bmatrix}$$

BIBLIOGRAPHY

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., and Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature methods*, 7(4):248–249.

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723.

Alexander, D. H. and Lange, K. (2011). Stability selection for genome-wide association. *Genetic epidemiology*, 35(7):722–728.

Ansorge, W. J. (2009). Next-generation dna sequencing techniques. *New biotechnology*, 25(4):195–203.

Attar, N. (2012). The allure of the epigenome. *Genome Biology*, 13:419.

Bach, F. R. (2008). Bolasso: model consistent lasso estimation through the bootstrap. ICML.

Bhatia, G. (2009). Rare variant analysis for common diseases.

Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J. M., Delano, D., Zhang, L., Schroth, G. P., Gunderson, K. L., et al. (2011). High density dna methylation array with single cpg site resolution. *Genomics*, 98(4):288–295.

Bock, C. (2012). Analysing and interpreting dna methylation data. *Nature Reviews Genetics*, 13(10):705–719.

Carvajal-Carmona, L. G. (2010). Challenges in the identification and use of rare disease-associated predisposition variants. *Current opinion in genetics & development*, 20(3):277–281.

Chatterjee, A. and Lahiri, S. N. (2011). Bootstrapping lasso estimators. *Journal of the American Statistical Association*, 106(494):608–625.

Cohen, J., Pertsemlidis, A., Fahmi, S., Esmail, S., Vega, G., Grundy, S., and Hobbs, H. (2006). Multiple rare variants in npc1l1 associated with reduced sterol absorption and plasma low-density lipoprotein levels. *Proceedings of the National Academy of Sciences of the United States of America*, 103(6):1810–1815.

Cristianini, N. and Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press.

Davies, R. B. (1980). Algorithm as 155: The distribution of a linear combination of χ 2 random variables. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 29(3):323–333.

Duchesne, P. and Lafaye De Micheaux, P. (2010). Computing the distribution of quadratic forms: Further comparisons between the liu–tang–zhang approximation and exact methods. *Computational Statistics & Data Analysis*, 54(4):858–862.

Dwyer, T., Blizzard, L., Morley, R., and Ponsonby, A.-L. (1999). Within pair association between birth weight and blood pressure at age 8 in twins from a cohort study. *Bmj*, 319(7221):1325–1329.

Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H., and Nadeau, J. H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics*, 11(6):446–450.

Eleftherohorinou, H., Hoggart, C. J., Wright, V. J., Levin, M., and Coin, L. J. (2011). Pathway-driven gene stability selection of two rheumatoid arthritis gwas identifies and validates new susceptibility genes in receptor mediated signalling pathways. *Human molecular genetics*, 20(17):3494–3506.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Gauderman, W. J. (2003). Candidate gene association analysis for a quantitative trait, using parent-offspring trios. *Genetic epidemiology*, 25(4):327–338.

Genovese, C. R., Roeder, K., and Wasserman, L. (2006). False discovery control with p-value weighting. *Biometrika*, 93(3):509–524.

Goeman, J., Oosting, J., Cleton-Jansen, A., Anninga, J., and Van Houwelingen, H. (2005). Testing association of a pathway with survival using gene expression data. *Bioinformatics*, 21(9):1950.

Han, F. and Pan, W. (2010a). A data-adaptive sum test for disease association with multiple common or rare variants. *Human Heredity*, 70(1):42–54.

Han, F. and Pan, W. (2010b). A data-adaptive sum test for disease association with multiple common or rare variants. *Human heredity*, 70(1):42–54.

He, Q. and Lin, D.-Y. (2010). A variable selection method for genome-wide association studies. *Bioinformatics (Oxford, England)*, pages 1–8.

Heyn, H., Carmona, F. J., Gomez, A., Ferreira, H. J., Bell, J. T., Sayols, S., Ward, K., Stefansson, O. A., Moran, S., Sandoval, J., et al. (2013). Dna methylation profiling in breast cancer discordant identical twins identifies dok7 as novel epigenetic biomarker. *Carcinogenesis*, 34(1):102–108.

Heyn, H., Li, N., Ferreira, H. J., Moran, S., Pisano, D. G., Gomez, A., Diez, J., Sanchez-Mut, J. V., Setien, F., Carmona, F. J., et al. (2012). Distinct dna methylomes of newborns and centenarians. *Proceedings of the National Academy of Sciences*, 109(26):10522–10527.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., and Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, 106(23):9362–9367.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Hoffmann, T. J., Marini, N. J., and Witte, J. S. (2010). Comprehensive approach to analyzing rare genetic variants. *PLoS One*, 5(11):e13584.

Hoggart, C. J., Whittaker, J. C., Iorio, M. D., and Balding, D. J. (2008). Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genetics*, 4(7):e1000130.

Hunter, D., Kraft, P., Jacobs, K., Cox, D., Yeager, M., Hankinson, S., Wacholder, S., Wang, Z., Welch, R., Hutchinson, A., et al. (2007). A genome-wide association study identifies alleles in fgfr2 associated with risk of sporadic postmenopausal breast cancer. *Nature genetics*, 39(7):870–874.

Ibrahim, J. G. (1990). Incomplete data in generalized linear models. *Journal of the American Statistical Association*, 85(411):765–769.

Ibrahim, J. G., Chen, M.-H., and Lipsitz, S. R. (2004). Monte carlo em for missing covariates in parametric regression models. *Biometrics*, 55(2):591–596.

Ionita-Laza, I., Buxbaum, J. D., Laird, N. M., and Lange, C. (2011). A new testing strategy to identify rare variants with either risk or protective effect on disease. *PLoS genetics*, 7(2):e1001289.

Joubert, B. R., Håberg, S. E., Nilsen, R. M., Wang, X., Vollset, S. E., Murphy, S. K., Huang, Z., Hoyo, C., Midttun, Ø., Cupul-Uicab, L. A., et al. (2012). 450k epigenome-wide scan identifies differential dna methylation in newborns related to maternal smoking during pregnancy. *Environ Health Perspect*, 120:1425–31.

Kaiser, J. (2012). Genetic influences on disease remain hidden. *Science*, 338(6110):1016–1017.

Knol, M. J., Janssen, K. J., Donders, A. R. T., Egberts, A. C., Heerdink, E. R., Grobbee, D. E., Moons, K. G., and Geerlings, M. I. (2010). Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: an empirical example. *Journal of clinical epidemiology*, 63(7):728–736.

Kwee, L., Liu, D., Lin, X., Ghosh, D., and Epstein, M. (2008). A powerful and flexible multilocus association test for quantitative traits. *The American Journal of Human Genetics*, 82(2):386–397.

Laird, N. M., Horvath, S., and Xu, X. (2000). Implementing a unified approach to family-based tests of association. *Genetic epidemiology*, 19(S1):S36–S42.

Laird, P. W. et al. (2003). The power and the promise of dna methylation markers. *Nature Reviews Cancer*, 3:253–266.

Landi, M. T., Chatterjee, N., Yu, K., Goldin, L. R., Goldstein, A. M., Rotunno, M., Mirabello, L., Jacobs, K., Wheeler, W., Yeager, M., Bergen, A. W., Li, Q., Consonni, D., Pesatori, A. C., Wacholder, S., Thun, M., Diver, R., Oken, M., Virtamo, J., Albanes, D., Wang, Z., Burdette, L., Doheny, K. F., Pugh, E. W., Laurie, C., Brennan, P., Hung, R., Gaborieau, V., McKay, J. D., Lathrop, M., McLaughlin, J., Wang, Y., Tsao, M.-S., Spitz, M. R., Wang, Y., Krokan, H., Vatten, L., Skorpen, F., Arnesen, E., Benhamou, S., Bouchard, C., Metsapalu, A., Vooder, T., Nelis, M., Välk, K., Field, J. K., Chen, C., Goodman, G.,

Sulem, P., Thorleifsson, G., Rafnar, T., Eisen, T., Sauter, W., Rosenberger, A., Bickeböller, H., Risch, A., Chang-Claude, J., Wichmann, H. E., Stefansson, K., Houlston, R., Amos, C. I., Fraumeni, J. F., Savage, S. a., Bertazzi, P. A., Tucker, M. a., Chanock, S., and Caporaso, N. E. (2009). A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. *American Journal of Human Genetics*, 85(5):679–91.

Lee, S., Wu, M. C., and Lin, X. (2012). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, 13(4):762–775.

Li, B. and Leal, S. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *The American Journal of Human Genetics*, 83(3):311–321.

Li, B. and Leal, S. (2009). Discovery of rare variants via sequencing: implications for the design of complex trait association studies. *PLoS Genetics*, 5(5):e1000481.

Li, Y., Byrnes, A. E., and Li, M. (2010a). To identify associations with rare variants, just whait:¡ i¿ w¡/i¿ eighted¡ i¿ h¡/i¿ aplotype¡ i¿ a¡/i¿ nd¡ i¿ i¡/i¿ mputation-based¡ i¿ t¡/i¿ ests. *The American Journal of Human Genetics*, 87(5):728–735.

Li, Y., Sheu, C.-C., Ye, Y., de Andrade, M., Wang, L., Chang, S.-C., Aubry, M. C., Aakre, J. a., Allen, M. S., and Chen, F. (2010b). Genetic variants and risk of lung cancer in never smokers: a genome-wide association study. *The Lancet Oncology*, 11(4):321–330.

Lin, D.-Y. and Tang, Z.-Z. (2011). A general framework for detecting disease associations with rare variants in sequencing studies. *The American Journal of Human Genetics*, 89(3):354–367.

Little, R. J. and Rubin, D. B. (1987). *Statistical analysis with missing data*, volume 4. Wiley New York.

Liu, D., Ghosh, D., and Lin, X. (2008). Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. *BMC Bioinformatics*, 9(1):292.

Liu, D., Lin, X., and Ghosh, D. (2007a). Semiparametric Regression of Multidimensional Genetic Pathway Data: Least-Squares Kernel Machines and Linear Mixed Models. *Biometrics*, 63(4):1079–1088.

Liu, D., Lin, X., and Ghosh, D. (2007b). Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. *Biometrics*, 63(4):1079–1088.

Liu, D. J. and Leal, S. M. (2010). A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. *PLoS genetics*, 6(10):e1001156.

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 & Data Analysis*, 53(4):853–856.

Lunetta, K. L., Faraone, S. V., Biederman, J., and Laird, N. M. (2000). Family-based tests of association and linkage that use unaffected sibs, covariates, and interactions. *The American Journal of Human Genetics*, 66(2):605–614.

Madsen, B. and Browning, S. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genetics*, 5(2):e1000384.

Manolio, T., Collins, F., Cox, N., Goldstein, D., Hindorff, L., , D., McCarthy, M., Ramos, E., Cardon, L., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753.

Mardis, E. (2008). Next-generation dna sequencing methods. *Annu. Rev. Genomics Hum. Genet.*, 9:387–402.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-J., Chen, Z., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380.

McNeill, G., Tuya, C., and Smith, W. (2004). The role of genetic and environmental factors in the association between birthweight and blood pressure: evidence from meta-analysis of twin studies. *International journal of epidemiology*, 33(5):995–1001.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473.

Morgenthaler, S. and Thilly, W. (2007). A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutation Research*, 615(1-2):28–56.

Morris, A. and Zeggini, E. (2010). An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genetic Epidemiology*, 34(2):188–193.

Neale, B. M., Rivas, M. A., Voight, B. F., Altshuler, D., Devlin, B., Orho-Melander, M., Kathiresan, S., Purcell, S. M., Roeder, K., and Daly, M. J. (2011). Testing for an unusual distribution of rare variants. *PLoS genetics*, 7(3):e1001322.

Nejentsev, S., Walker, N., Riches, D., Egholm, M., and Todd, J. A. (2009). Rare variants of ifih1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science*, 324(5925):387–389.

Ng, P. C. and Henikoff, S. (2003). Sift: Predicting amino acid changes that affect protein function. *Nucleic acids research*, 31(13):3812–3814.

Price, A. L., Kryukov, G. V., de Bakker, P. I., Purcell, S. M., Staples, J., Wei, L.-J., and Sunyaev, S. R. (2010). Pooled association tests for rare variants in exon-resequencing studies. *American journal of human genetics*, 86(6):832.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J., et al. (2007). Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575.

Rakyan, V. K., Down, T. A., Balding, D. J., and Beck, S. (2011). Epigenome-wide association studies for common human diseases. *Nature Reviews Genetics*, 12(8):529–541.

Sandoval, J., Heyn, H., Moran, S., Serra-Musach, J., Pujana, M. A., Bibikova, M., and Esteller, M. (2011). Validation of a dna methylation microarray for 450,000 cpg sites in the human genome. *Epigenetics*, 6(6):692–702.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics bulletin*, 2(6):110–114.

Schaffner, S. F., Foo, C., Gabriel, S., Reich, D., Daly, M. J., and Altshuler, D. (2005). Calibrating a coalescent simulation of human genome sequence variation. *Genome research*, 15(11):1576–1583.

Schaid, D., Sinnwell, J., Jenkins, G., McDonnell, S., Ingle, J., Kubo, M., Goss, P., Costantino, J., Wickerham, D., and Weinshilboum, R. (2011). Using the gene ontology to scan multilevel gene sets for associations in genome wide association studies. *Genetic Epidemiology*.

Schuster, S. (2008). Next-generation sequencing transforms today's biology. *Nature*, 200(8):16–18.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Scott, L., Mohlke, K., Bonnycastle, L., Willer, C., Li, Y., Duren, W., Erdos, M., Stringham, H., Chines, P., Jackson, A., et al. (2007). A genome-wide association study of type 2 diabetes in finns detects multiple susceptibility variants. *Science*, 316(5829):1341–1345.

Shah, R. D. and Samworth, R. J. (2012). Variable selection with error control: Another look at stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Shen, J., Wang, S., Zhang, Y.-J., Wu, H.-C., Kibriya, M. G., Jasmine, F., Ahsan, H., Wu, D. P., Siegel, A. B., Remotti, H., et al. (2013). Exploring genome-wide dna methylation profiles altered in hepatocellular carcinoma using infinium humanmethylation 450 beadchips. *Epigenetics*, 8(1):0–1.

Subramanian, A., Tamayo, P., Mootha, V., Mukherjee, S., Ebert, B., Gillette, M., Paulovich, A., Pomeroy, S., Golub, T., Lander, E., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545.

Thomas, G., Jacobs, K., Yeager, M., Kraft, P., Wacholder, S., Orr, N., Yu, K., Chatterjee, N., Welch, R., Hutchinson, A., et al. (2008). Multiple loci identified in a genome-wide association study of prostate cancer. *Nature genetics*, 40(3):310–315.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Tzeng, J. and Zhang, D. (2007). Haplotype-based association analysis via variance-components score test. *The American Journal of Human Genetics*, 81(5):927–938.

Valdar, W., Sabourin, J., Nobel, A., and Holmes, C. C. (2012). Reprioritizing genetic associations in hit regions using lasso-based resample model averaging. *Genetic epidemiology*, 36(5):451–462.

Walsh, T., McClellan, J. M., McCarthy, S. E., Addington, A. M., Pierce, S. B., Cooper, G. M., Nord, A. S., Kusenda, M., Malhotra, D., Bhandari, A., et al. (2008). Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science Signalling*, 320(5875):539.

Wang, K., Li, M., and Bucan, M. (2007). Pathway-based approaches for analysis of genomewide association studies. *The American Journal of Human Genetics*, 81(6):1278–1283.

Wang, K., Li, M., and Hakonarson, H. (2010). Analysing biological pathways in genome-wide association studies. *Nature Reviews Genetics*, 11(12):843–854.

Wessel, J. and Schork, N. (2006). Generalized genomic distance-based regression methodology for multilocus association analysis. *The American Journal of Human Genetics*, 79(5):792–806.

WTCCC (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–78.

Wu, M., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, 89(1):82–93.

Wu, M., Maity, A., Lee, S., Simmons, E., Harmon, Q., Lin, X., Engel, S., Molldrem, J., Armistead, P., et al. (2013). Kernel machine snp-set testing under multiple candidate kernels. *Genetic epidemiology*, 37(3):267–275.

Wu, M. C., Kraft, P., Epstein, M. P., Taylor, D. M., Chanock, S. J., Hunter, D. J., and Lin, X. (2010). Powerful SNP-set analysis for case-control genome-wide association studies. *The American Journal of Human Genetics*, 86(6):929–42.

Wu, M. C. and Lin, X. (2009). Prior biological knowledge-based approaches for the analysis of genome-wide expression profiles using gene sets and pathways. *Statistical methods in medical research*, 18(6):577–593.

Wu, T., Chen, Y., Hastie, T., Sobel, E., and Lange, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25(6):714.

Yeager, M., Orr, N., Hayes, R., Jacobs, K., Kraft, P., Wacholder, S., Minichiello, M., Fearnhead, P., Yu, K., Chatterjee, N., et al. (2007). Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nature genetics*, 39(5):645–649.

Yi, N. and Zhi, D. (2011). Bayesian analysis of rare variants in genetic association studies. *Genetic Epidemiology*, 35(1):57–69.

Zawistowski, M., Gopalakrishnan, S., Ding, J., Li, Y., Grimm, S., and Zöllner, S. (2010). Extending rare-variant testing strategies: analysis of noncoding sequence and imputed genotypes. *American journal of human genetics*, 87(5):604.

Zhou, H., Sehl, M. E., Sinsheimer, J. S., and Lange, K. (2010). Association Screening of Common and Rare Genetic Variants by Penalized Regression. *Bioinformatics (Oxford, England)*, 26(19):2375–2382.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.

Zuk, O., Hechter, E., Sunyaev, S., and Lander, E. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences*, 109(4):1193–1198.