Hypothesis Testing and Model Selection for Complex Data

Hypothesis Testing and Model Selection for Complex Data

A dissertation presented

by

Sixing Chen

to

The Department of Biostatistics

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Biostatistics

Harvard University

Cambridge, Massachusetts

May 2017

Hypothesis Testing and Model Selection for Complex Data

Abstract

In this dissertation, we propose methodology for hypothesis testing in statistical genetics and model selection in networks. In chapters 1 and 2, we introduce new methods to tackle difficulties in hypothesis testing for sequencing association studies brought on by advancement of sequencing technology. In chapter 3, we introduce a flexible framework for mechanistic network model selection, which is an area of the networks literature with a dearth of work.

In chapter 1, we aim to test for association in a case-control sequencing studies, where the case-control status is completely confounded by the quality of the sequencing data. Such a situation can arise when one combines next generation sequencing data from cases with publicly available sequencing data (using an older platform) from controls. We propose a regression calibration-based method and consider maximum-likelihood for conducting an association study with the aligned reads from cases and controls. The methods allow for adjusting for non-confounding covariates as well as confounders in some situations. Both methods control type I error and have comparable power to analysis conducted using the true genotype with sufficiently high but different sequencing depths. The regression calibration method allows for analysis with the naive variance estimate and standard software under certain circumstances.

In chapter 2, we present a method for sparse signal detection for association between a set of SNPs that contain rare variants and a binary phenotype. Such settings are common in the increasingly abundant whole genome sequencing analyses. Traditional single SNP tests with rare variants are subject to poor power. Thus, methods that test for association by aggregating the test statistics of multiple rare variants together in a genetic region are popu-

lar. These existing methods for rare variant analysis, such as SKAT, have good power when the signals are dense in the set of SNPs tested, but can have poor power when the signals are sparse. In contrast, thresholding methods for signal detection, such as higher criticism and Berk-Jones methods, have good power in the presence of sparse signals. However, they rely on the single SNP test statistics to behave well as normally distributed asymptotically. The normality assumption of the individual test statistics does not hold in the presence of rare variants for binary phenotypes and yields incorrect type I error rates. Our proposed rare variant higher criticism approach for sparse signal detection has higher power than the existing aggregating methods and allows weighting of the SNPs, with the correct size.

In chapter 3, we propose a procedure for mechanistic network model selection. Our proposal aims to address the dearth of work on model selection for mechanistic network models. Such models describe network growth and evolution over time starting from simple microscopic mechanisms. Along with statistical models, which are probabilistic models for the final observed network, they are two prominent paradigms for modeling network structure. In comparison to statistical models, mechanistic models are easier to incorporate domain knowledge with, to study effects of interventions and to sample from, but typically have intractable likelihoods. To handle this intractability, our procedure makes use of the flexible Super Learner framework and borrows aspects from Approximate Bayesian Computation.

Contents

List of Figures

List of Tables

Acknowledgments

I am much indebted to my dissertation advisor, Xihong Lin, for all her mentoring, support, patience, and willingness to challenge me. I am very grateful to JP Onnela, who is a member of my dissertation committee, for guiding me in exploring an entirely new area of research so late into my studies. I would like to thank the other members of my dissertation committee, Tianxi Cai and Brent Coull, for their help and fresh point of view in critiquing my work. I am also thankful for all the help from current and former members of Xihong's research group. I would also like to thank Laura Balzer for her indispensable help on the third chapter of my dissertation. Graduate school can be tough at times, and I could not have done it without my friends, my classmates, as well as our wonderful administrative staff. Last but not least, I want to express my gratitude to my family for their unending empathy, support, and love.

Analysis in Case-Control Sequencing Association Studies with Different Sequencing Depths

Sixing Chen, Xihong Lin

Department of Biostatistics

Harvard School of Public Health

1.1 Introduction

Next-generation sequencing (NGS) is becoming more and more favored as the tool to obtain genetic information for investigators (Goldstein et al., 2013). Despite the recent advances, it can still be prohibitively expensive to conduct a well-powered study entirely with NGS. To make most efficient use of available resources, one could combine data available publicly, such as data from 1000 genomes (1000 Genomes Project Consortium, 2012), or from previous studies with data sequenced for the current study. This increases available sample size and can lead to more efficient use of resources. To combine data from multiple sources for one study, however, one needs to take care in order to avoid biases from estimating allele frequencies from the varied data quality, especially with called genotypes (Kim et al., 2011).

This work is motivated by small case-only sequencing data sets, specifically from the NHLBI GO-ESP lung cohorts exome sequencing project. The exome sequencing project involves a discovery sample of whole exome sequencing of 89 acute lung injury (ALI) cases. Findings from the discovery sample were validated in a larger sample, but the discovery sample, containing information across all of the exome, is not used for association analysis and does not contain healthy controls. If combined with an external control sample (e.g. 1000 genomes), the combined sample may be well powered to detect associations from a much larger portion of the exome. However, the data quality between the phase 1 1000 genomes data, with average coverage 4, and the ALI data set, with average coverage over 70, is quite stark.

To circumvent the aforementioned difference between quality of NGS data and previously available data and confounding related to data quality and the platforms used (Nielsen et al., 2011), such as read depth, investigators often use NGS to generate data for variant discovery and genotype a larger sample size on the identified variants for association analysis (Liu and Leal, 2012; Longmate et al., 2010; Sanna et al., 2011). These study designs prevents

type-I error inflation, but can be conservative and the sample sizes obtained for the discovery phase is not used for the actual association analysis. There are also methods that tackle this issue by incorporating the quality of the genetic information via the likelihood. Skotte et al. (2012) introduced such a method, based on the score test statistic, and has some advantages over analysis using called genotype. This method only considers the prospective setting and assumes non-differential misclassification. In this setting, it refers to the sequencing quality is the same between cases and controls. Derkach et al. (2014) introduced the robust variance score method, which also incorporates the same information by using the expectation of the true genotype given the observed reads. Both methods take a score approach and do not compute regression coefficients. Derkach's method does consider the retrospective setting, but is unable to control for confounders.

The regression calibration (RC) method we introduce here is a little different from the traditional setting since we are dealing with misclassification instead of measurement error of the covariates (Carroll et al., 2006; Rosner et al., 1989). The only exception to this is Spiegelman et al. (2000), where a linear mean was assumed for binary covariates. In the traditional setting, the data consists of two parts. The first consists of subjects with only observations with error of the unknown covariates. The second, the validation set, consists of subjects with the true underlying covariates observed in additional to the observations with error. The validation set is used to estimate parameters of the joint distribution of the true covariates and the observations with error. In Carroll's method, one computes the conditional expectation of the true covariates given the observations with error, then regresses over the conditional expectation in place of the true covariates. In Rosner's method, one gets naive regression parameters by regressing over the observations with error, then adjusts the naive estimates accordingly. Our method compares more directly to that of Carroll.

Much like RC, maximum likelihood in the traditional misclassification setting has the same

data structure. The validation set's contribution to the likelihood involves extra parameters for the joint distribution of the underlying covariates and the observations with error. In the traditional setting, maximum likelihood requires the specification of the complete likelihood (Carroll et al., 1993).

In the sequencing setting, a validation set is not typically possible, but there is intrinsic information about the correlation between the sequencing reads and the underlying genotype from the coverage and error rates of the reads that enable our RC method. Due to this unique data structure and what we observe, we do not need to estimate the parameters of the underlying error distribution to compute the desired conditional expectations. Under balanced case-control sampling, even with differing data quality, the RC method allows the use of naive variance estimate to give close to correct type-I error, allowing for ease of analysis with standard software packages. This is not generally true in the traditional setting, which will usually use the sandwich estimator or bootstrap for testing (Carroll et al., 2006). The RC estimate is biased, but the bias is small when read depth is sufficiently high and error rate is sufficiently low. This is due to something similar to the small measurement error assumption in the traditional setting (Carroll et al., 2006).

The presence of the coverage and error rates of the reads also enables our maximum likelihood method (ML) without a validation set. Due to the presence of the quality information, the likelihood can now be factored into two parts, one that involves parameters of association only and one that involves the aforementioned extra parameters. The validation set typically provides the latter part of this likelihood factorization and can now be factored out and ignored. The ML approach we introduce gives unbiased estimate of the effect size under retrospective sampling without covariates. However, in the presence of non-confounding covariates, there is an induced correlation between the genotype and the non-confounding covariates under the alternative. When ignored, this leads to a bias in the estimates, though

the bias is small when the effect size is small.

In addition to comparison with regression calibration and maximum likelihood methods in the traditional setting, there is previous work on the setting where subjects have only replicates of the observations with error, but no validation set. These works are relevant since that is what the sequencing data structure entails. This setting has been previously explored with measurement error via regression calibration (Carroll et al., 2006), and with misclassification via maximum likelihood (Liu and Liang, 1991). The latter method assumes non-differential misclassification and makes distributional assumptions about the joint distribution of the true covariates and observations with error.

The methods we introduce control type-I error when using all available sample size, i.e. combining data sequenced for the current study and previously available data. This allows one to focus all resources on obtaining data used for the association analysis as well as easy reusage of data collected from previous studies. We will focus on the case when all cases are sequenced on one platform and controls on another so that matching for confounders related to data quality would defeat the purpose of combining data from multiple sources. The VCF files from each data source give information required to compute the expectation of the genotype given the observed reads. This approach accounts for the uncertainty and differing quality of genetic information in the cases and controls. Our approach does not assume non-differential misclassification, unlike the approaches in Liu and Liang (1991), and Skotte et al. (2012), gives comparable power to analysis done with the true genotype, when read depth is reasonably high, and allow for adjusting of non-confounding covariates to increase power as well as of binary confounders. In most settings, RC and ML have similar power, and in the differentiated ones, RC has higher power.

The next sections are organized as followed. In section 2, we lay out the assumptions

made about the distribution of the data as well as the reads and introduce the RC and ML methods without covariates, with non-confounding covariates, confounders, as well as population stratification. In sections 3 and 4, we specify the structure of the simulations used to evaluate performance as well as the filtering applied to the combined ALI exome sequencing and 1000 genomes data set pre-analysis. In section 5, we present the results of our simulations and data analysis, and conclude with a discussion in section 6.

1.2 Methods

We assume we have independent and identically distributed (iid) subjects over i and that we have a biallelic locus for which we know the major and minor allele. We also assume the following for the marginal distribution of the data, where p is the minor allele frequency, the outcome Y has conditional distribution $\text{logit}\left(p_i\right) = \text{logit}\left(P\left(Y_i = 1|G_i\right)\right) = \beta_0 + \beta_1 G_i$ and $G \sim \text{bin}\left(2, p\right)$ is the genotype of the locus of interest. If there are additional covariates $\mathbf{X} \perp G$, then $\text{logit}\left(p_i\right) = \text{logit}\left(P\left(Y_i = 1|G_i, \mathbf{X}_i\right)\right) = \beta_0 + \beta_1 G_i + \boldsymbol{\beta}_2^T \mathbf{X}_i$.

We assume that our source of genetic information for each subject is the observed reads $\tilde{\mathbf{G}}_i$ and the quality information for each read. Say that subject i has reads $\tilde{G}_{ij}$ for $j \in \{1, \ldots, d_i\}$, where d_i is the read depth for subject i. For read $\tilde{G}_{ij}$, there is an associated quality score, which can be mapped to misclassification rate π_{ij} for this read, where $\pi_{ij} = P\left(\tilde{G}_{ij} = G_i\right)$. For each subject, we observe $\tilde{G}_{ij}$ and π_{ij}, for $j \in \{1, \ldots, d_i\}$, from the reads directly. We can assume some distribution for d_i and π_{ij}. The setting we are interested in, where the data quality depends on the case control status, corresponds to the distributions of d_i and π_{ij} depending on the case control status.

With this information, we can compute the conditional distribution of the observed reads, $\tilde{\mathbf{G}}_i$, given the true genotype, d_i and $\boldsymbol{\pi}_i$ as $P\left(\tilde{\mathbf{G}}_i|d_i, \boldsymbol{\pi}_i, G_i = g\right) = \prod_j P\left(\tilde{G}_{ij}|d_i, \pi_{ij}, G_i = g\right)$, with the assumption that the individual reads are independent given the read depth, mis-

classification rates and the true genotype. We compute the conditional expectation of the genotype as $E\left[G_i|\tilde{\mathbf{G}}_i, d_i, \boldsymbol{\pi}_i\right] = \sum_g g \times P\left(G_i = g|\tilde{\mathbf{G}}_i, d_i, \boldsymbol{\pi}_i\right)$, where $P\left(G_i|\tilde{\mathbf{G}}_i, d_i, \boldsymbol{\pi}_i\right)$ is constructed from $P\left(\tilde{\mathbf{G}}_i|G_i, d_i, \boldsymbol{\pi}_i\right)$ and an estimate of the distribution of the genotype, $P(G_i = g)$. $P(G_i = g)$ is estimated from the full sample (cases and controls) with the EM algorithm (McKenna et al., 2010; Skotte et al., 2012), where the marginal likelihood of the reads, $\prod_i P\left(\tilde{\boldsymbol{G}}_i|d_i, \boldsymbol{\pi}_i\right) = \prod_i \left[\sum_g P\left(\tilde{\boldsymbol{G}}_i|d_i, \boldsymbol{\pi}_i, G_i = g\right) \times P(G_i = g)\right]$, is maximized over $P(G_i = g)$. Note that in the retrospective setting, we are implicitly conditioning on being sampled everywhere.

1.2.1 Regression Calibration without Non-Confounding Covariates

The difference in RC in this setting and in the traditional setting mainly differs in how the conditional expectations are computed. Typically, one has measurements with error, $\tilde{Z}_1 \ldots \tilde{Z}_m$, for unobserved quantity Z, and assumes that the conditional distribution of $Z|\tilde{Z}$ is index by some parameter $\boldsymbol{\theta}$. Then, one conducts a validation study where one obtains measurement with error and ascertains the true quantity for each subject. An estimate of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}$, is obtained from the validation study. From this estimate, one computes the conditional expectation $E\left[Z|\tilde{\boldsymbol{Z}}, \hat{\boldsymbol{\theta}}\right]$ and proceeds to regress over this quantity. In the sequencing setting, there is no way better way to ascertain the true genotype than the sequencing reads themselves, so a reliable validation study is not typically possible. Fortunately, there is already information about the conditional distribution of $\tilde{G}|G$ in the observed $\boldsymbol{\pi}_{ij}$. From this, we are able to compute the desired conditional expectation $E\left[G_i|\tilde{\mathbf{G}}_i, d_i, \boldsymbol{\pi}_i\right]$ without a validation set.

After computing $E\left[G_i|\tilde{\mathbf{G}}_i, d_i, \boldsymbol{\pi}_i\right]$, we proceed to use it in place of G in logistic regression. The regular MLE in this case would be $\hat{\beta}_0$ and $\hat{\beta}_1$ which maximizes $\prod_i \frac{\exp(\beta_0 Y_i + \beta_1 G_i Y_i)}{1+\exp(\beta_0 + \beta_1 G_i)}$ in terms of β_0 and β_1. The regression calibration estimator $\hat{\boldsymbol{\beta}}_{RC} = \begin{bmatrix} \hat{\beta}_{0E} & \hat{\beta}_{1E} \end{bmatrix}^T$ maximizes $\prod_i \frac{\exp\left(\beta_{0E} Y_i + \beta_{1E} E\left[G_i|\tilde{\mathbf{G}}_i, d_i, \boldsymbol{\pi}_i\right] Y_i\right)}{1+\exp\left(\beta_{0E} + \beta_{1E} E\left[G_i|\tilde{\mathbf{G}}_i, d_i, \boldsymbol{\pi}_i\right]\right)}$ in terms of β_{0E} and β_{1E}. The resulting estimator, $\hat{\boldsymbol{\beta}}_{RC}$, can be

framed within a set of four estimating equations, two for the regression calibration estimators and two for the genotype distribution. Under the null, the estimating equations for $\hat{\boldsymbol{\beta}}_{RC}$ is unbiased when evaluated at $\beta_{0E} = \text{logit}\left(P\left(Y_i = 1 | S_i = 1\right)\right)$ and $\beta_{1E} = 0$ (shown in appendix). We can proceed to invoke Z-estimation theory and estimate the variance of the estimators with the sandwich estimator. The test statistic is form by dividing $\hat{\boldsymbol{\beta}}_{RC}$ by the sandwich variance estimate and will have the correct size when compared to its asymptotic distribution, the standard normal.

Denote $p_g = P\left(G_i = g\right)$, then the individual contribution to the estimating equations stated above is:

$$
\begin{aligned}
\psi^{(i)}_{\beta_{0E}} &= Y_i - \text{expit}\left(\beta_{0E} + \beta_{1E} E\left[G_i | \tilde{\boldsymbol{G}}_i, d_i, \pi_i\right]\right) \\
\psi^{(i)}_{\beta_{1E}} &= E\left[G_i | \tilde{\boldsymbol{G}}_i, d_i, \pi_i\right]\left[Y_i - \text{expit}\left(\beta_{0E} + \beta_{1E} E\left[G_i | \tilde{\boldsymbol{G}}_i, d_i, \pi_i\right]\right)\right] \\
\psi^{(i)}_{p_0} &= \frac{P\left(\tilde{\boldsymbol{G}}_i | G_i = 0, d_i, \pi_i\right) - P\left(\tilde{\boldsymbol{G}}_i | G_i = 2, d_i, \pi_i\right)}{\sum_g p_g P\left(\tilde{\boldsymbol{G}}_i | G_i = g, d_i, \pi_i\right)} \\
\psi^{(i)}_{p_1} &= \frac{P\left(\tilde{\boldsymbol{G}}_i | G_i = 1, d_i, \pi_i\right) - P\left(\tilde{\boldsymbol{G}}_i | G_i = 2, d_i, \pi_i\right)}{\sum_g p_g P\left(\tilde{\boldsymbol{G}}_i | G_i = g, d_i, \pi_i\right)}
\end{aligned}
$$

Denote the bun and meat matrices from this set of estimating equations $\boldsymbol{A}$ and $\boldsymbol{B}$, with asymptotic limits $\mathbb{A}$ and $\mathbb{B}$. The naive information matrix of the logistic regression fit for RC is $\boldsymbol{A}_\beta$, a submatrix of $\boldsymbol{A}$, with asymptotic limit $\mathbb{A}_\beta$, a submatrix of $\mathbb{A}$. Under the null, when the case control sampling is balanced, the sub matrices of the meat and bun matrices corresponding to the regression calibration estimator have the same limit. Under these assumptions, it can be shown that the naive variance estimator overestimates the true variance (shown in appendix), i.e. the element corresponding to β_{1E} of $\mathbb{A}^{-1}\mathbb{B}\mathbb{A}^{-1}$ $\leq$that of $\mathbb{A}_\beta^{-1}$. So the naive variance estimator, from inverting the "observed information" like

regular logistic regression, i.e. inverting $\boldsymbol{A}_\beta$, also gives the correct level when the sampling is balanced, though it is conservative.

Simulation results suggests that if the distribution of the read depth between cases and controls are not too dissimilar or that both cases and controls have reasonable read depth, then the conservativeness is low, and the loss of power is minor under the alternative. In fact, when the minor allele frequency is low and the sample size is not sufficiently large, the convergence of the test statistic computed with the sandwich estimator is insufficient. In this case, the naive variance estimator will perform better than the sandwich estimator.

1.2.2 Regression Calibration with Non-Confounding Covariates

If we wish to adjust for non-confounding covariates $\mathbf{X} \perp G$ with RC, we first need to compute the conditional expectation $E\left(G_i|\tilde{\mathbf{G}}_i, d_i, \boldsymbol{\pi}_i\right)$ as before. Note that this conditional expectation does not take $\boldsymbol{X}$ into account. The reason is that $\boldsymbol{X}$ is non-confounding and is assumed to be independent of G. The RC estimates $\hat{\boldsymbol{\beta}}_{RC} = \begin{bmatrix} \hat{\beta}_{0E} & \hat{\beta}_{1E} & \hat{\boldsymbol{\beta}}_{2E} \end{bmatrix}^T$ maximizes the logistic regression "likelihood":

$$\prod_i \frac{\exp\left(\beta_{0E} Y_i + \beta_{1E} E\left[G_i|\tilde{\mathbf{G}}_i, d_i, \boldsymbol{\pi}_i\right] Y_i + \boldsymbol{\beta}_{2E}^T \boldsymbol{X}_i Y_i\right)}{1 + \exp\left(\beta_{0E} + \beta_{1E} E\left[G_i|\tilde{\mathbf{G}}_i, d_i, \boldsymbol{\pi}_i\right] + \boldsymbol{\beta}_{2E}^T \boldsymbol{X}_i\right)}$$

in terms of β_{0E}, β_{1E} and $\boldsymbol{\beta}_{2E}$. Just as the case without covariates, the corresponding set of estimating equations are unbiased when evaluated at $\beta_1 = 0$. Hypothesis testing can once again proceed with the corresponding sandwich estimator, or the naive variance estimator when sampling is balanced.

1.2.3 Maximum Likelihood without Non-Confounding Covariates

Just as RC, maximum likelihood in a traditional misclassification setting requires a validation study. Once again, assume we have measurements with error, $\tilde{Z}_1 \ldots \tilde{Z}_m$, for unobserved

quantity Z, and that the conditional distribution of $Z|\tilde{Z}$ is index by some parameter $\boldsymbol{\theta}$. A validation study where we obtain measurements with error and ascertain the true unobserved quantity from each subject is required to get information on $\boldsymbol{\theta}$. Instead of estimating $\boldsymbol{\theta}$, however, we form a single likelihood $L_1(\boldsymbol{\beta}, \boldsymbol{\theta}) \times L_2(\boldsymbol{\theta})$, where L_1 is from the majority of data set without validation and L_2 is from the validation study. L_1 depends on $\boldsymbol{\theta}$ since each subject's contribution is averaged over the conditional distribution of $Z|\tilde{Z}$ that depends on $\boldsymbol{\theta}$, while L_2 provides the bulk of the information on $\boldsymbol{\theta}$ since Z is observed directly in the validation set.

As above stated, a validation study is not possible with genotypes, but we have information about the correlation of G and $\tilde{G}$ from the observed $\{\pi_{ij}\}$ and d_i. Say the distribution of π_{ij} and d_i is indexed by $\boldsymbol{\gamma}$, then the full likelihood can be written as $L_1'(\boldsymbol{\beta}) \times L_2'(\boldsymbol{\gamma})$. L_1' involves the outcome Y, the reads $\tilde{\boldsymbol{G}}$, and the read depth and misclassification rates. Each individual contribution to L_1' is averaged over the conditional distribution $G_i|\tilde{\boldsymbol{G}}_i, d_i, \boldsymbol{\pi}_i$ and depends on the parameter of interest $\boldsymbol{\beta}$, but not $\boldsymbol{\gamma}$. L_2' is for the marginal distribution of d_i and $\boldsymbol{\pi}_i$ that may vary based on the case control status, but does not depend on the parameter of interest $\boldsymbol{\beta}$, and can thus be factored out. Next, we specify the likelihood.

Since we are in the retrospective setting and assuming that we do not have access to the sampling fraction, the parameters related to the marginal distribution of the data, i.e. the minor allele frequency p and β_0, are not identifiable (Prentice and Pyke, 1979). From Prentice and Pyke (1979), if the sampling status only depends on the outcome, we know that we can write $P(Y_i = 1|G_i = g, S_i = 1) = \text{expit}(\beta_{0_{cc}} + \beta_1 g)$, where $\beta_{0_{cc}} = \beta_0 + log(P(S_i = 1|Y_i = 1)/P(S_i = 1|Y_i = 0))$ and S is the indicator of being sampled. Thus, we need to form the likelihood based on parameters related to the distribution of the data given being sampled, i.e. $\beta_{0_{cc}}$, $P(G_i = 0|S_i = 1)$ and $P(G_i = 1|S_i = 1)$. We can write out the likelihood as follows:

$$
\begin{aligned}
& P\left(\tilde{\mathbf{G}}_i, d_i, \boldsymbol{\pi}_i | Y_i = 1, S_i = 1\right) \\
= & \sum_{g=0}^{2} P\left(\tilde{\mathbf{G}}_i | d_i, \boldsymbol{\pi}_i, G_i = g, Y_i = 1\right) P\left(d_i, \boldsymbol{\pi}_i | G_i = g, Y_i = 1\right) \\
& \times \frac{P\left(Y_i = 1 | G_i = g, S_i = 1\right) P\left(G_i = g | S_i = 1\right)}{P\left(Y_i = 1 | S_i = 1\right)} \\
= & \sum_{g=0}^{2} P\left(\tilde{\mathbf{G}}_i | d_i, \boldsymbol{\pi}_i, G_i = g\right) P\left(d_i, \boldsymbol{\pi}_i | Y_i = 1\right) \frac{p_i \times P\left(G_i = g | S_i = 1\right)}{P\left(Y_i = 1 | S_i = 1\right)} \\
\propto & \sum_{g=0}^{2} P\left(\tilde{\mathbf{G}}_i | d_i, \boldsymbol{\pi}_i, G_i = g\right) \frac{p_i \times P\left(G_i = g | S_i = 1\right)}{P\left(Y_i = 1 | S_i = 1\right)}
\end{aligned}
$$

$$
\begin{aligned}
& P\left(\tilde{\mathbf{G}}_i, d_i, \boldsymbol{\pi}_i | Y_i = 0, S_i = 1\right) \\
\propto & \sum_{g=0}^{2} P\left(\tilde{\mathbf{G}}_i | d_i, \boldsymbol{\pi}_i, G_i = g\right) \frac{(1 - p_i) P\left(G_i = g | S_i = 1\right)}{P\left(Y_i = 0 | S_i = 1\right)}
\end{aligned}
$$

In the second line, we dropped the conditioning of $Y_i = 1$ from $P\left(\tilde{\mathbf{G}}_i | d_i, \boldsymbol{\pi}_i, G_i = g, Y_i = 1\right)$ since the distribution of $\tilde{\mathbf{G}}_i$ only depends on Y_i through d_i and $\boldsymbol{\pi}_i$. We also dropped the conditioning of $G_i = g$ from $P\left(d_i, \boldsymbol{\pi}_i | G_i = g, Y_i = 1\right)$ since the distribution of d_i and $\boldsymbol{\pi}_i$ does not depend on G_i. Then, the term $P\left(d_i, \boldsymbol{\pi}_i | Y_i = 1\right)$ does not depend on any parameters of interest and can be factored out of the likelihood. Note that this requires 4 parameters instead of 3 of the marginal distribution. We can get rid of $\beta_{0_{cc}}$ by solving the following equation, as a function of β_1, $P\left(G_i = 0 | S_i = 1\right)$ and $P\left(G_i = 1 | S_i = 1\right)$:

$$
\begin{aligned}
P\left(Y_i = 1 | S_i = 1\right) = & \operatorname{expit}\left(\beta_{0_{cc}}\right) P\left(G_i = 0 | S_i = 1\right) + \operatorname{expit}\left(\beta_{0_{cc}} + \beta_1\right) P\left(G_i = 1 | S_i = 1\right) \\
& + \operatorname{expit}\left(\beta_{0_{cc}} + 2\beta_1\right) P\left(G_i = 2 | S_i = 1\right)
\end{aligned}
$$

Although there is no closed form solution to this equation, we can numerically solve for this in each step of the fitting process. Furthermore, to avoid the problem of having estimates that are not in the bounds of the parameter space, notably $P(G_i = 0|S_i = 1)$ and $P(G_i = 1|S_i = 1)$, we can parameterize the two as:

$$\begin{aligned} P(G_i = 0|S_i = 1) &= \text{expit}(\alpha) \\ P(G_i = 0 \cup G_i = 1|S_i = 1) &= \text{expit}(\alpha + \exp(\beta)) \\ P(G_i = 1|S_i = 1) &= \text{expit}(\alpha + \exp(\beta)) - \text{expit}(\alpha) \end{aligned}$$

Testing of the null $\beta_1 = 0$ can proceed in the normal likelihood framework by dividing the estimate, $\hat{\beta}_{1_{MLE}}$, by its corresponding standard error derived from the information matrix and comparing the corresponding quotient to the normal distribution. An issue does arise in this parameterization when allele frequency is low. In this case, $P(G_i = 0 \cup G_i = 1|S_i = 1) = \text{expit}(\alpha + exp(\beta)) \approx 1$. The column of the information matrix corresponding to β contain products of the term $\text{expit}(\alpha + exp(\beta))(1 - \text{expit}(\alpha + exp(\beta)))$, which is very close to 0. Thus, this column will be of terms very to 0. Inverting the information matrix in this case may run into numerical issues. Thus, when allele frequency is low, we recommend using RC instead.

1.2.4 Maximum Likelihood with Non-Confounding Covariates

If we want to perform analysis in with non-confounding covariate X, which is marginally independent of G, we must reconsider the likelihood:

$$
\begin{aligned}
& P\left(\tilde{\mathbf{G}}_i, d_i, \boldsymbol{\pi}_i, X_i | Y_i = 1, S_i = 1\right) \\
= & \sum_{g=0}^{2} P\left(\tilde{\mathbf{G}}_i | d_i, \boldsymbol{\pi}_i, G_i = g, X_i, Y_i = 1\right) P\left(d_i, \boldsymbol{\pi}_i | G_i = g, X_i, Y_i = 1\right) \\
& \times \frac{P\left(Y_i = 1 | G_i = g, X_i, S_i = 1\right) P\left(G_i = g | X_i, S_i = 1\right) P\left(X_i | S_i = 1\right)}{P\left(Y_i = 1 | S_i = 1\right)} \\
\propto & \sum_{g=0}^{2} P\left(\tilde{\mathbf{G}}_i | d_i, \boldsymbol{\pi}_i, G_i = g, X_i, Y_i = 1\right) \frac{p_i \times P\left(G_i = g | X_i, S_i = 1\right) P\left(X_i | S_i = 1\right)}{P\left(Y_i = 1 | S_i = 1\right)}
\end{aligned}
$$

$$
\begin{aligned}
& P\left(\tilde{\mathbf{G}}_i, d_i, \boldsymbol{\pi}_i, X_i | Y_i = 0, S_i = 1\right) \\
\propto & \sum_{g=0}^{2} P\left(\tilde{\mathbf{G}}_i | d_i, \boldsymbol{\pi}_i, G_i = g, X_i, Y_i = 0\right) \frac{(1 - p_i) P\left(G_i = g | X_i, S = 1\right) P\left(X_i | S_i = 1\right)}{P\left(Y_i = 0 | S_i = 1\right)}
\end{aligned}
$$

Similar to the situation without non-confounding covariates, we drop the conditioning of X_i and $Y_i = 1$ from $P\left(\tilde{\mathbf{G}}_i | d_i, \boldsymbol{\pi}_i, G_i = g, X_i, Y_i = 1\right)$ since $\tilde{\mathbf{G}}_i$ only depends on Y_i through d_i and $\boldsymbol{\pi}_i$ and does not depend on X_i. $P\left(d_i, \boldsymbol{\pi}_i | G_i = g, X_i, Y_i = 1\right)$ also does not depend on G_i and can be factored out. Although $P\left(X_i | S_i = 1\right)$ can be factored out and ignored in the analysis, $P\left(G_i = g | X_i, S_i = 1\right)$ cannot be ignored. Even though G and X are independent marginally, if both G and X are associated with Y, i.e. $\beta_1, \beta_2 \neq 0$, then G and X are not independent conditioning on $S = 1$. Under the null, $\beta_1 = 0$, it can be shown that $P\left(G_i = g | X_i, S_i = 1\right) = P\left(G_i = g | S_i = 1\right)$. So if we replaced $P\left(G_i = g | X_i, S_i = 1\right)$ with $P\left(G_i = g | S_i = 1\right)$ in the likelihood, we have the correct model under the null and have correct size, but the model would be incorrect under the alternative and there would be some bias. In the simulations we looked at, the bias is not too severe. This problem would be avoided under prospective sampling.

1.2.5 Analysis with Confounders

So far we have only considered incorporating non-confounding covariates $\boldsymbol{X}$. Our methods can potentially incorporate confounders, but different issues arise with both. With RC, if we are able to compute the conditional expectation of the genotype given the sequencing reads and the confounders, $E\left[G_i|\tilde{\mathbf{G}}_i, d_i, \boldsymbol{\pi}_i, \boldsymbol{X}_i\right]$, we substitute it in place of the genotype in the analysis just as before. The corresponding estimating equations will have expectation 0. However, it may not be clear in practice how to compute this conditional expectation.

With ML, the likelihood remains essentially the same as the case with non-confounding covariates. The only difference being $P\left(G_i = g|\boldsymbol{X}_i, S_i = 1\right) \neq P\left(G_i = g|S_i = 1\right)$ under the null as well as the alternative. If the marginal conditional distribution of $G|X$ is correctly specified, then we can correctly specify the null likelihood. Just as with non-confounding covariate, under the null, this marginal correlation remains the same after conditioning on being sampled ($S_i = 1$), but under the alternative, this correlation changes after averaging over Y. Once again, we can specify the likelihood correctly under the null, but it will be wrong under the alternative. So we are able to perform inference, but the estimate is biased. If the support of the confounder is infinite, then it will often be unclear how to specify this conditional distribution. Instead, if the support of the confounder was finite, e.g. binary, then we can devote extra parameters for the distribution of the genotype given each possible value of the confounder and $S_i = 1$. This is feasible when the confounder only takes on very few values. This approach does work in our simulation study.

1.2.6 Population Stratification

Although it is not clear how to control for any arbitrary confounder with our method, we have found an ad-hoc solution to control for population stratification. If we control for the first few principal components from the conditional expectation of the genotype $E\left[G_i|\tilde{\mathbf{G}}_i, d_i, \boldsymbol{\pi}_i\right]$, it adequately controls for inflation from population stratification. Let $\boldsymbol{PC}_i$ be the vector

of loadings for the PCs of the ith subject. For RC, this means adding in the PCs as covariates in the logistic regression fit, i.e. maximize $\prod_i \frac{\exp\left(\beta_{0E} Y_i + \beta_{1E} E\left[G_i | \tilde{\mathbf{G}}_i, d_i, \boldsymbol{\pi}_i\right] Y_i + \boldsymbol{\beta}_{PC}^T \boldsymbol{PC}_i Y_i\right)}{1+\exp\left(\beta_{0E} + \beta_{1E} E\left[G_i | \tilde{\mathbf{G}}_i, d_i, \boldsymbol{\pi}_i\right] + \boldsymbol{\beta}_{PC}^T \boldsymbol{PC}_i\right)}$, with inference proceeding as before. For ML, the PCs are included in the specification of $\text{logit}\left(P\left(Y_i = 1 | G_i = g, \boldsymbol{PC}_i, S_i = 1\right)\right) = \beta_0 + \beta_1 G_i + \boldsymbol{\beta}_{PC}^T \boldsymbol{PC}_i$. In addition, we need to specify the conditional distribution of the genotype given the PCs as a function of the PCs, $P\left(G_i = g | \boldsymbol{PC}_i, S = 1\right) = f\left(\boldsymbol{PC}_i\right)$. Regardless of how f is specified, it most likely will be an approximation, so ideally, it should be something flexible.

1.3 Simulation Studies

The goal of the simulation studies is to study the performance of each method under varying differences in data quality. We consider (1) a situation where the distribution of both read depth and misclassification rates are similar, (2) a situation where both are quite different, and (3) one where the read depth is similar and misclassification rate is quite different. Situation (1) corresponds to using data from platforms of similar performance; situation (2) corresponds to using data from a newer platform and an older platform; and situation (3) corresponds to using data from a newer platform and an older platform with increased number of reads.

For simulation without other covariates, given fixed number of cases and controls, we first generate the genotype of cases and controls separately, based on the distribution of G given Y. Then, for each subject i, we generate the read depth d_i and a sequence of misclassification rates π_{ij} for $j \in \{1, \ldots, d_i\}$. Now, we generate the sequence of reads for individual i based on the generated read depth and misclassification rates. The read depths are generated from various normal distributions truncated from below at 1 and rounded to the nearest integer. The misclassification rate is generated from various normal distribution truncated from below at 0 and from above at 1. The means and standard deviations of the normal distributions truncated for each situation are specified in table 1.1. From the reads, we can

Table 1.1: The normal distributions (mean, standard deviation) used to generate the read depth and misclassification rate in our simulations. The three scenarios (1)-(3) are described in the first paragraph of the simulations section

		(1)	(2)	(3)
Case	Read Depth	$N(8,1)$	$N(20,2)$	$N(20,2)$
	Misclassification	$N(0.005, 0.0001)$	$N(0.005, 0.0001)$	$N(0.005, 0.0001)$
Control	Read Depth	$N(6,1)$	$N(3,1)$	$N(16,1)$
	Misclassification	$N(0.01, 0.001)$	$N(0.05, 0.01)$	$N(0.05, 0.01)$

computed the desired $P\left(\tilde{\mathbf{G}}_i|G_i\right)$, with which we can compute $E\left(G_i|\tilde{\mathbf{G}}_i\right)$ for RC or form the likelihood for ML.

For simulation with a confounding covariate, we generate the genotype, a binary confounder X, and Y conditional on G and X prospectively, then stop when we have the requisite number of cases and controls. We then can sample the desired number of cases and controls and can proceed to generate the sequence of reads. The binary confounder X has one particular mean if $G = 0$ or 1, and a different mean if $G = 2$. For simulation with population stratification, we generate two population minor allele frequencies from an ancestral allele frequency (0.2) with the Balding-Nichols model (Balding and Nichols, 1995) with $F_{ST} = 0.01$, then generate the genotype and the phenotype based on the population the subject belongs to. Then, we sample the desired number of cases and controls independently of the population the subject belongs too and generate the sequence of reads as before. The parameters used for the simulation with population stratification are listed in table 1.2.

Table 1.2: The parameter values used for size simulation with population stratification. The columns are the same as in table A.1, except the third and fourth columns now denote the ancestral minor allele frequency and the F_{ST} used to generate the two new population minor allele frequencies

n_{case}	$n_{control}$	Anc. p	F_{ST}	rd_{case}	rd_{con}	e_{case}	e_{con}	β_0	β_1
200	300	0.2	0.01	(20,1)	(3,1)	(0.005,0.0001)	(0.05,0.001)	$logit(0.2)$	0

1.3.1 Size

For simulations under the null, we generated a case-control sample from the following prospective models:

$$logit\left(P\left(Y=1|G\right)\right) = logit\left(0.2\right)$$
$$logit\left(P\left(Y=1|G,X\right)\right) = logit\left(0.2\right) + 0.2X$$
$$logit\left(P\left(Y=1|G,Z\right)\right) = logit\left(0.2\right) + Z$$

The first model refers to the null model without any other covariates, the second refers to the null model with a confounding covariate X, while the third refers to the null model with population stratification with binary population marker Z. The true genotype G is generated with minor allele frequency 0.2 in all cases, unless stated otherwise. For simulation with population stratification, we generate two population minor allele frequencies from the specified Balding-Nichols model, then proceed to generate the true genotype for both populations. After the genotype is generated, we simulate the read depth and misclassification rates d_i and π_{ij} with the distribution as specified in the three situations stated above.

The simulations include scenarios with 1:1 and 1:2 case to control sampling ratio. In each setting, we compare the performance of analysis with the true genotype with the Wald statistic versus RC and ML. RC with the naive variance estimator will only be considered when the sampling is balanced. Comparison of the p-values under the null against the uniform distribution will be done with the QQ-plot.

1.3.2 Power

For simulations under the alternative, the data is generated with the same procedure as simulations under the null, except the prospective models are now:

$$logit\left(P\left(Y=1|G\right)\right)=logit\left(0.2\right)+0.3G$$

$$logit\left(P\left(Y=1|G,X\right)\right)=logit\left(0.2\right)+0.3G+0.2X$$

$$logit\left(P\left(Y=1|G,Z\right)\right)=logit\left(0.2\right)+0.3G+Z$$

For each scenario assessed for size, we generate data with the same parameters, but with a non-zero effect for the genotype as stated above. We will compare proportion rejected at various significance levels as well as bias in the estimators for analysis with the true genotype against RC and ML. RC with naive variance is still only considered with balanced sampling since it gives incorrect size when sampling is unbalanced.

1.4 Application on ALI Exome and 1000 Genomes Data

We performed analysis with RC and ML as well as naive analysis with the called genotype on the combined data set of the discovery sample from the NHLBI ALI cohort project and phase 1 of 1000 genomes project. The discovery sample from the Acute Lung Injury cohort project consists of exome sequencing data from 89 Caucasian subjects with varying severity of lung injury. The data from phase 1 of 1000 genomes consist of exome sequencing data from 174 CEU and GBR subjects. The original ALI data set distinguishes subjects based on ventilator-free-days (VFD). Those with VFD < 2 are considered high severity, while those with VFD < 24 are considered low severity. In our analysis, we do not make this distinction, and consider all subjects with ALI subjects as cases and 1000 genomes subjects as controls.

To compute the desired $P\left(G_i|\tilde{\mathbf{G}}_i, d_i, \boldsymbol{\pi}_i\right)$, we need $P\left(\tilde{\mathbf{G}}_i|G_i, d_i, \boldsymbol{\pi}_i\right)$ from all the subjects. This information is encoded as PL in the VCF files for the ALI subjects, and as GL for the 1000 genomes subjects. To combine the data set, we matched up SNPs that are both in the 1000 genomes phase 1 exome data and the ALI exome data that have the same name

(RS number), physical location and definition of major and minor alleles. SNPs with quality (QUAL) < 30, quality by depth (QD) < 5, allele balance (AB) > 0.75, % missing $> 10\%$, or strand bias (SB) $>$ -0.1 were removed due to low quality. We also removed SNPs that exhibit low allele frequencies by filtering out SNPs that have estimated probability for genotype 0 < 0.0025 and that of genotype 1 < 0.095. Lastly, due to the lack of genotype data and small sample sizes (89 cases and 174 controls), we were unable to filter based on the exact or asymptotic Hardy-Weinberg equilibrium tests. So as a substitute, we filtered out SNPs that had estimated probability for genotype 1 > 0.5, which is the upper bound had the SNP been under Hardy-Weinberg equilibrium. 22619 SNPs remained after this filtering.

The remaining SNPs post-filtering are analyzed with RC, ML as well as the Wald test with the called genotype. Since we have unequal number of cases and controls, we did not consider RC with naive variance estimator just as in the simulations. For all three methods of analysis, we included two PCs to control for population stratification. For RC and ML, this was done with PCs from the conditional expectation of the genotype, while the naive analysis used PCs from the called genotypes.

1.5 Results

1.5.1 Size Simulations

We first look at how well the various methods control type I error under various scenarios. When the read depth and misclassification distribution are both similar between cases and controls and sampling is balanced, RC with naive and sandwich variance both control type I error well (scenario A). On the other hand, if the distribution of read depth and misclassification both differ significantly between cases and controls with balanced sampling, RC with naive variance is conservative, while RC with sandwich variance continues to perform well (scenario B). Scenario A and B are both simulated with a high MAF. If we have scenario B with a low MAF instead, then both RC with naive and sandwich variance are conservative,

Table 1.3: Summary of the scenarios considered in the size simulations. The first column is used for identification in the article; the second refers to the distributions of read depth and misclassification rate as described above; the third refers to whether the sampling is balanced; the fourth refers to whether the simulation is done with a low minor allele frequency; the last refers to the plot set in the appendix to refer to for the QQ plots

Scenario	Dep/Mis	Balanced	Low MAF	Plot Set
A	(1)	Yes	No	1
B	(2)	Yes	No	2
C	(2)	Yes	Yes	3
D	(3)	Yes	No	4
E	(2)	No	No	5

but RC with naive variance is noticeably less conservative (scenario C). Next, we consider a similar read depth distribution but significantly different misclassification distribution between cases and controls with balanced sampling. In this case, RC with both naive and sandwich variances control the size well, so a sufficient number of reads can make up for an order of magnitude of difference in misclassification rate (scenario D). Lastly, if we have scenario B with unbalanced sampling, RC with naive variance no longer gives correct type I error, but RC with sandwich variance still performs well (scenario E). In each of these scenarios, ML controls type I error well consistently. Scenario C was omitted for ML due to numerical issues stemming from low MAF. Table 1.3 gives a summary of all the scenarios considered, while table 1.4 gives a summary of the performance of each method in each scenarios. More detail and results of the simulations can be found in the appendix.

For simulation with population stratification, we generated data in the aforementioned fash-

Table 1.4: Summary of the results of size simulations. The first column is used for identification in the article; the last three columns refer to how well the method in the column name controls size in our simulations

Scenario	RC Naive	RC Sandwich	ML
A	Correct	Correct	Correct
B	Conservative	Correct	Correct
C	Conservative	Conservative	N/A
D	Correct	Correct	Correct
E	Liberal	Correct	Correct

ion and performed naive analysis with no principal components as well as analysis that controls for the first two principal components. For the ML simulations with population stratification, we specify $P\left(G_i = g|\boldsymbol{PC}_i, S_i = 1\right) = f\left(\boldsymbol{PC}_i\right)$ as a logistic model with first and second order terms as well as an interaction term:

$$f\left(\boldsymbol{PC}_i\right) = \text{expit}\left(\alpha_0 + \alpha_1 PC_{1i} + \alpha_2 PC_{2i} + \alpha_3 PC_{1i}^2 + \alpha_4 PC_{2i}^2 + \alpha_5 PC_{1i} \times PC_{2i}\right)$$

As expected, the naive analysis shows significant inflation for RC, ML, and analysis with the true genotype. After controlling for the two principal components, we can see that the control of type I error is much improved. Due to the unbalanced sampling, however, RC with naive variance still shows some inflation as expected, but RC with sandwich variance performs well. One issue that we did encounter was numerical issues in fitting the ML. This is due to additional nuisance parameters required for the distribution of the genotype given the principal components. As a result, ML does not control type I error as well as RC with sandwich variance in this case. All corresponding QQ plots are in figure 1.1.

For simulation with a binary confounder, we generate the confounder with probabilities $P\left(X = 1|G = 0 \text{ or } 1\right) = 0.5$ and $P\left(X = 1|G = 2\right) = 0.6$, depending on the value of G. We parameterized $P\left(G_i = g|X_i, S_i = 1\right)$ as

$$P\left(G_i = 0|X_i = x, S_i = 1\right) = \text{expit}\left(\alpha_x\right)$$
$$P\left(G_i = 1|X_i = x, S_i = 1\right) = \text{expit}\left(\alpha_x + \exp\left(\beta_x\right)\right) - \text{expit}\left(\alpha_x\right)$$

The parameterization is similar to that of $P\left(G_i = g|S_i = 1\right)$ above, which is without confounders, except there is a separate set of parameters for each value of X. QQ plot for the ML analysis (see appendix A.1.6 for more detail) shows that this parameterization controls size very well. Although the performance is good, this does show the limitations of this

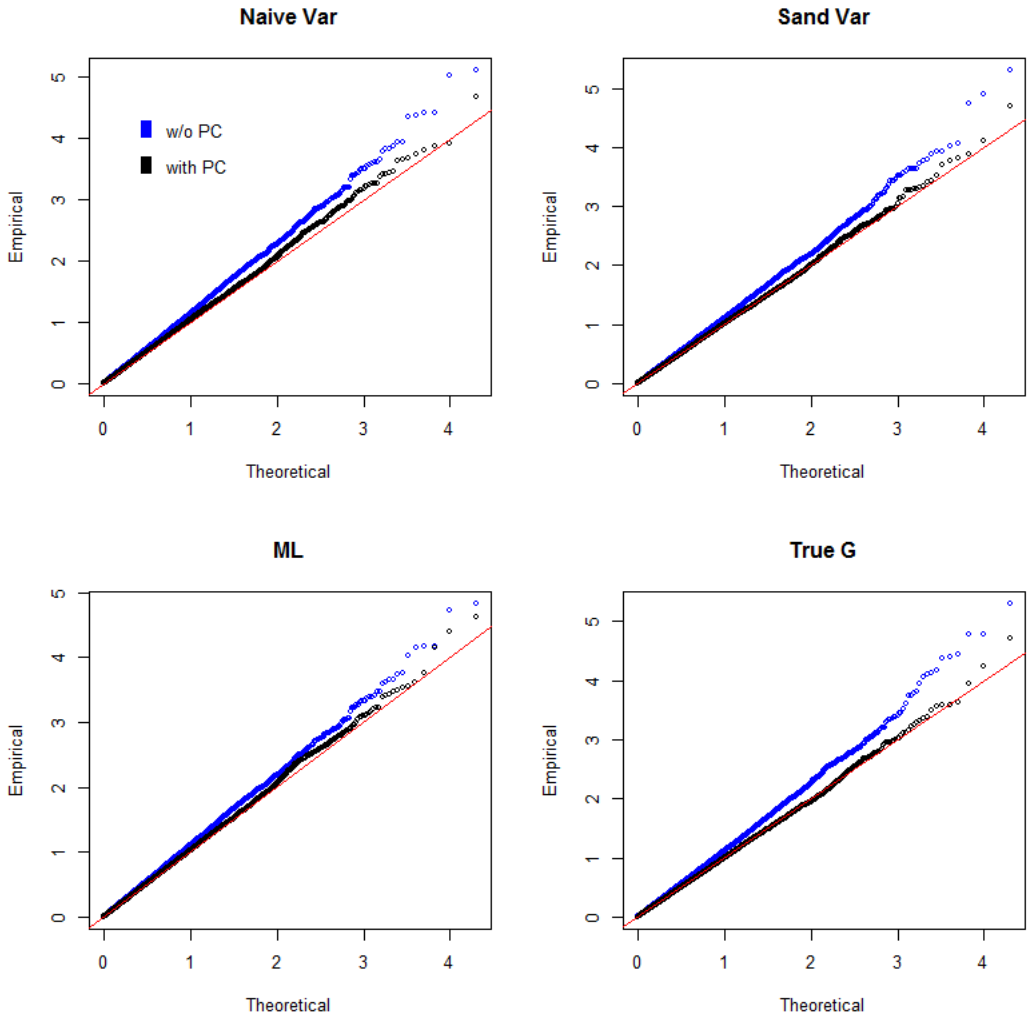

Figure 1.1: QQ plots for the size simulation with population stratification. The four plots correspond to analysis with RC and naive variance, RC and sandwich variance, ML, and with the true genotype. The blue plots correspond to the naive analysis without PCs, and the black plots correspond to analysis with PCs

approach, considering how many additional parameters were needed to fully parameterize $P(G_i = g|X_i, S_i = 1)$ for just a binary X. For more complex confounders, one would need to make additional assumptions to restrict the number of parameters.

1.5.2 Power Simulations

Next, we assess the power and bias of the methods, via simulations with the same sets of parameters used in the size assessment, but with $\beta_1 \neq 0$. Each QQ-plot in the appendix is followed by the corresponding tables of averages of the estimates and empirical power at various significance level. In each scenario, the power for the methods based on the reads is expectedly lower than that with the true genotype, but is comparable in all the cases.

In most of the scenarios, RC seems to have slightly higher power than ML. The power comparison between RC with naive and sandwich variance varies for the scenarios with balanced sampling. For scenario A, with similar distributions for read depth and misclassification distributions, power is similar for RC with naive and sandwich variance. In scenario B, in which the distribution of read depth and misclassification rate both differ significantly between cases and controls, there is small but noticeable power gain from using the sandwich variance. In scenario C, even though the distribution of read depth and misclassification rate still differ significantly, the power of RC with naive variance is slightly higher than that of RC with sandwich variance while power for both is far away from 0. This reflects the result from the null simulations where RC is less conservative with the naive variance than with the sandwich variance in scenario C. Lastly, in scenario D, read depth for both cases and controls are sufficiently high, but there is significant difference in misclassification error rates. Here, power is similar for RC with both variances.

In general, the power loss is more severe in cases where the data quality of one group is poor, such as in scenario B and E. The bias of the RC estimator follows a similar pattern,

as the bias goes from undetectable in situation (1) and (3) to noticeable but not too severe in situation (2). For ML, there is no appreciable bias even under situation (2).

1.5.3 Analysis of Combined ALI and 1000 Genomes Data

We matched the two data sets based on physical location of SNPs that had the same name and definition of major and minor alleles. SNPs with low quality, low allele frequencies and with estimated heterozygous probability greater than 0.5 were removed. We applied RC with sandwich variance, ML and naive analysis with the called genotype to the 22619 SNPs that remained after filtering. We control for population stratification with two principal components computed from the conditional expectations of the true genotype for RC and ML, or from the called genotypes for the naive analysis. For ML, we specify $P\left(G_i = g|\boldsymbol{PC}_i, S = 1\right)$ with the same model as they were in the null simulations.

The resulting QQ plots for both RC and ML look reasonably close to the 45 degree line (figure 1.2), with the tail of the QQ plot for RC showing a little more signal than ML. This reflects the results from the simulations where RC had slightly higher power than ML in most of the scenarios. On the other hand, the QQ plot for the naive analysis with called genotypes is much more liberal in comparison, with the QQ plot well above the 45 degree line. This suggests that naive analysis with such a combined data set can lead to lots of false positives.

There were no SNPs that reach genome-wide significance. There were 5 SNPs that had p-value $< 10^{-4}$ and 30 SNPs that had p-value $< 10^{-3}$ with RC, while the same numbers were 2 and 27 for ML. One of the top hits with both RC and ML is rs2943521, which belongs to the MUC5B gene. This gene is associated with mucus secretion, including lung mucus, and has been linked to other lung diseases (Seibold et al., 2011). Although the top hits are different from those in Lee et al. (2012), which analyzed the same ALI subjects, the analysis

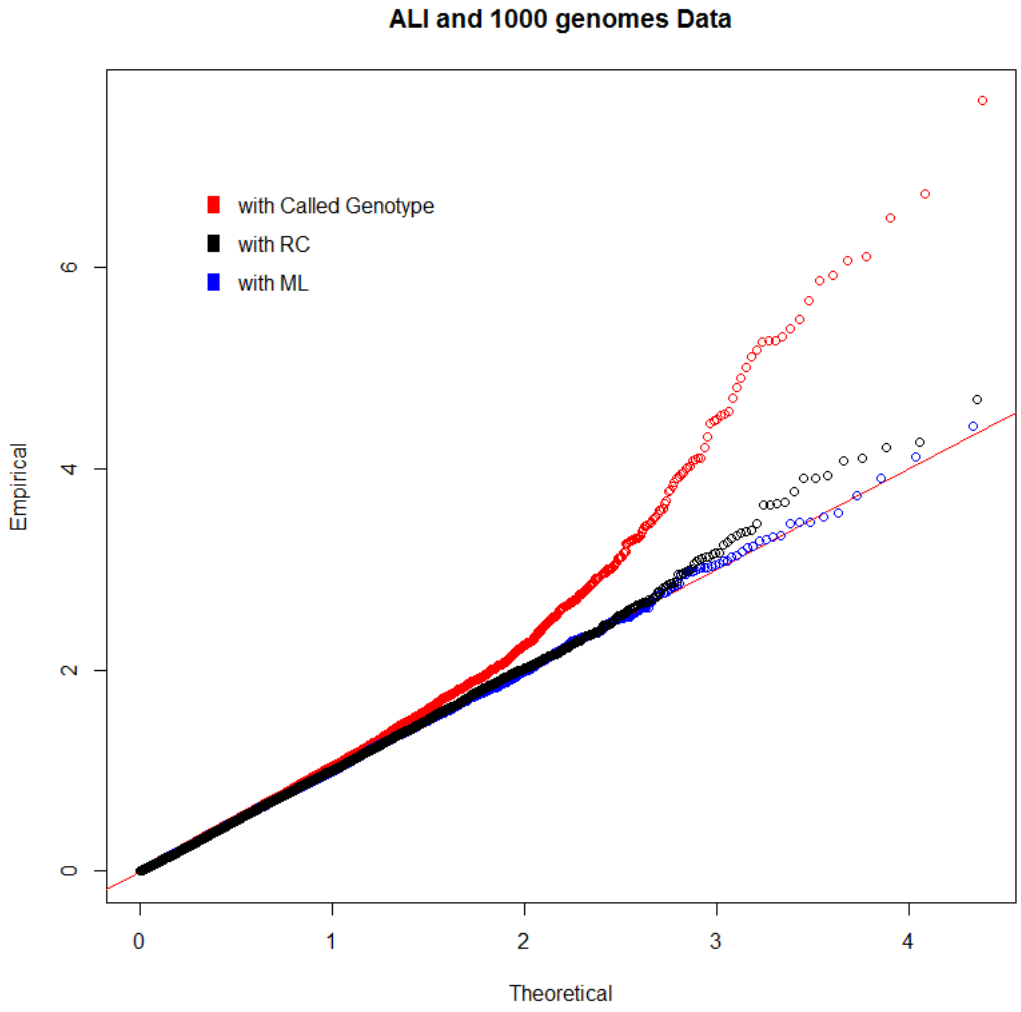

Figure 1.2: QQ plots for analysis of combined ALI and 1000 genomes data. The red plot is analysis with the called genotypes, the blue plot is analysis with ML, the black plot is analysis with RC

in that paper was from comparing subjects with severe vs mild ALI, whereas our analysis is from comparing ALI subjects vs healthy subjects. So the hypothesis being tested is different. Looking at the QQ-plots from RC and ML, we can see that the two look very similar, with RC having a few more smaller p-values in the tail.

1.6 Discussion

We propose methods for association analysis with a data set where the case-control status is completely confounded by sequencing quality while controlling for population stratification as well as some forms of confounding. Such a data set can arise where the cases and controls are sequenced on different sequencing platforms from different studies. The RC method requires the sandwich estimator for inference when case-control sampling is unbalanced, but can use the naive estimator when the sampling is balanced. The latter can be useful when speed is of importance or when convergence of the sandwich estimator is poor. However, asymptotically, the naive estimator is slightly more conservative than the sandwich estimator. When the data is of sufficient quality, the bias in the RC estimator is small. Controlling for non-confounding covariates does not affect these results. The ML estimator is unbiased under the alternative when there are no covariates, but this does not hold when there are covariates.

Although we do have a solution for controlling for population stratification, our methods do not allow for controlling for confounders in general. So extending the methods to control for general confounders is the next step. In addition, forming the sandwich estimator for RC and maximizing the likelihood for ML are a bit time consuming at the moment. Speeding up both are also of interest in the future.

The details shown in this paper are for the additive model, but extensions for the dominant and recessive model is simple. This can be done by simply substituting in the probability of

the desired homozygous genotype for the indicator of said genotype, and the above results still hold. Our methods can also be easily extended to accommodate data sets where cases and controls are both taken from multiple sources of differing quality. In our simulation studies, RC seems to perform slightly better than ML in terms of power. When sampling is balanced, the difference in control of size between the sandwich and naive estimator seems to be negligible, though the power seems to be slightly better for the sandwich estimator when convergence is sufficient. In comparison to the existing score based methods, our methods gives an estimate of the effect size and allows the use of the naive variance estimator when sampling is balanced.

With the continued advancement of sequencing technology, we have higher quality data sets available for association studies. However, rather than discarding old data sets due to confounding by data quality, our methods allow for reusing of existing data with new data for analysis.

Sparse Signal Detection in the Presence of Rare Variants and Binary Phenotype

Sixing Chen, Xihong Lin

Department of Biostatistics

Harvard School of Public Health

2.1 Introduction

With the advancement of sequencing technology, there is unprecedented access to rare variant data through large sequencing projects such as the Trans-Omics for Precision Medicine (TOPMed) of the National Heart, Lung, and Blood Institute (NHLBI) and the Genome Sequencing Program (GSP) of the National Human Genome Research Institute (NHGRI). Consequently, there is an increasing need for methods that have good power for detecting genetic association with phenotypes in the presence of rare variants. Historically, single SNP analysis through the GWAS framework has identified many SNPs that are associated with numerous phenotypes, but are typically restricted to analysis with common variants, as they are underpowered or biased with rare variants (Lee et al., 2014). To illustrate the loss of information from restricting analysis to common variants, consider the Dallas Heart Study (Romeo et al., 2009). Of the 93 SNPs in the data set, only 1 and 4 have minor allele frequency ≥ 0.05 and ≥ 0.01 respectively, while 57 are singletons. Instead, SNP-set methods that analyze multiple SNPs as a unit are popular for testing rare variants, but existing methods do not perform well with rare variants and binary phenotype when the signal is sparse (only a few SNPs in the SNP-set are associated).

As whole genome sequencing (WGS) becomes more prevalent and the number variants in a typical data set increases, the burden of Bonferroni correction, i.e. loss of power, is exacerbated and makes single SNP analysis even more impractical. With WGS data, there is access to more complete data on genetic regions, such as genes and signaling pathways. If a set of SNPs are from one genetic region, they are likely to jointly affect the phenotype, and it is more sensible to analyze them as one unit than separately. The true signals in many SNP-sets may be weak and undetectable by single SNP methods. The true signal may also be sparse, such as the FGFR2 region that shows association with breast cancer (Hunter et al., 2007). The region contains 35 SNPs, and only 4 of which show evidence of association, but no genome-wide significance. Given these shortcomings of single SNP analysis, methods

that test for association with a set of SNPs through a signal detection approach have been gaining popularity. Such methods aim to overcome issues with single SNP analysis by compiling signals across all the SNPs into a single stronger signal, though the hypothesis now being tested is a global null of no association between the phenotype and the entire SNP-set.

Several existing SNP-set analysis methods compile signals from the SNP-set via the marginal test statistic from each individual SNP. Some of these methods, especially those tailored towards sparse signals, typically rely on the marginal statistics to be well-behaving, e.g. close to normally distributed. Therefore, these methods work well for normally distributed phenotypes or for binary phenotypes with common variants, such as the aforementioned FGFR2 region. However, the marginal test statistics can behave poorly with binary phenotype and rare variants, which are more prevalent in WGS data. For example, the score statistic becomes very discrete, since the numerator and the denominator both take on only a few values. As a result, the normality assumption will no longer hold, making the approximation of its distribution difficult. In addition to weak and sparse signals as well as the presence of rare variants, one needs to deal with the linkage disequilibrium (LD) present between the SNPs being tested. LD between SNPs leads to correlation between the marginal test statistics that needs to be properly handled to avoid incorrect results. LD in common variants can lead to highly correlated test statistics, while its effect is less severe in rare variants. However, it has been shown that ignoring LD in rare variants can still lead to biased results (Neale et al., 2011). In this paper, we develop a genetic association testing method through a SNP-set approach that has good power against sparse alternatives in the presence of rare variants and binary phenotype while accounting for LD.

There are several existing methods for signal detection for SNP-set association testing that are well powered against dense alternatives, i.e. a large number of SNPs are associated. The burden tests (Li and Leal, 2008; Madsen and Browning, 2009; Morgenthaler and Thilly,

2007; Morris and Zeggini, 2010) aggregates all variants in the SNP-set into a single covariate, then tests for association between the newly formed covariate and the phenotype. Burden tests have good power when the signals are dense and in the same direction. The second condition may be especially unreasonable for a signaling pathway, which can contain both protective and risk-increasing variants in different genes. In addition to the burden tests, the sequence kernel association test (SKAT) (Wu et al., 2010, 2011) is a variance component test that assumes random effects for the SNPs in the SNP-set. The test statistic is the sum of the score statistic for testing zero variance for the random effect for each individual SNP with adjustment for LD between the SNPs, with the option to weight the SNPs individually to improve power. SKAT relies on having a dense signal or a sufficient number of SNPs to be in LD with the "true" signal SNPs for good power, but the signals need not be in the same direction. Burden tests and SKAT are both able to handle rare variants, but, when the signal in the SNP-set is weak, the aggregated quantity (of the variants or of the test statistics) mixes the small amount of signals with a large amount of noise. Thus, if the signal is sparse and LD between the signal SNPs and noise SNPs is weak, which can happen with rare variants, then SKAT and burden tests are likely to have poor power.

Methods that are designed for sparse signals include MinP (Conneely and Boehnke, 2007; Moskvina and Schmidt, 2008; Zhang and Liu, 2011) and higher criticism (Donoho and Jin, 2004). MinP computes the marginal test statistic for all SNPs in the SNP-set, then takes the most extreme one as the test statistic with adjustment for LD between the SNPs. MinP relies on "true" strong signal SNPs to have good power, which is unusual in sequencing data, and is unable to handle rare variants with binary phenotypes, due to the aforementioned issue with marginal test statistics. The original higher criticism tests the global null for a large number of marginal tests and is known to have good power against sparse alternatives. However, higher criticism relies on independence or sparse and weak correlation between the marginal test statistics as well as a large number of marginal tests (i.e. the number of SNPs

in the SNP-set) for asymptotic results. In reality, there is likely to be correlation in the marginal test statistics due to LD in the SNP-set, which will contain a fairly finite number of SNPs.

To handle the first of the two issues with higher criticism, the innovated higher criticism (iHC) (Hall and Jin, 2010) was proposed. The iHC transforms the marginal test statistics through the Cholesky decomposition of the correlation matrix and proceeds to apply higher criticism to the transformed test statistics, which are assumed to be independent. However, this approach is shown to have poor power due to the mixing of signal and noise through the transformation, since the transformed test statistics are all linear combinations of a few signal SNPs and a larger number of noise SNPs. To avoid mixing signal and noise, the generalized higher criticism (GHC) (Barnett et al., 2016) was proposed. GHC takes the marginal test statistics on the original scale and accounts for correlation directly through their correlation matrix. Thus, GHC can handle correlation between the test statistics due to LD without mixing signal and noise and retains good power against sparse alternatives, while additionally giving a means to compute analytic p-values that does not rely on a large number of SNPs in the SNP-set. Both iHC and GHC assume multivariate normality of the marginal test statistics, which is reasonable given common SNPs or normally distributed phenotypes. However, if the phenotype is binary and the SNP-set contains rare variants, then the normality assumption will be inaccurate as previously mentioned. In addition, neither methods allow for weighting of the SNPs.

To avoid this need for normality, we propose a new method for signal detection for genetic association in the presence of rare variants. The core of this method is a more meaningful way for obtaining p-values for rare variants than methods that rely on asymptotics. Typically, marginal p-values/test statistics (Fisher's exact test, score statistic) are computed for a given SNP by comparing the observed sample against more extreme possible samples for the given

SNP only. Our method computes p-values for individual SNPs that compares the observed sample against more extreme possible samples for all SNPs in the region being tested. This approach can amplify weak sparse signals against a background of numerous noise SNPs. Signals across the SNP-set are compiled to form a single statistic, while the correlation is estimated via permutation. The compilation of the signals allows for weighting of the SNPs individually to improve power. If no weighting is done, then this compilation corresponds exactly to GHC as in Barnett et al. (2016). Our approach has good power against sparse alternatives, and is able to handle rare variants and weight SNPs unlike existing methods tailored towards sparse alternatives.

The rest of the paper will be organized as follows. In section 2, we review how individual p-values/test statistics are typically computed in GWAS settings, reasons they perform poorly, and introduce our new framework for computing p-values for individual SNPs. In section 3, we discuss GHC and how it handles correlation between SNPs, and how we handle correlation within our framework. In addition, we discuss how to weight the SNPs individually. In section 4, we evaluate how well our method controls size as well as comparison of power against existing methods such as SKAT through simulation. In section 5, we analyze the aforementioned Dallas Heart Study data with our proposed method as well as existing methods, followed by concluding remarks in section 6.

2.2 Marginal P-values

Assume a sample of size n and a SNP-set with p SNPs, such as SNPs from the aforementioned FGFR2. Let $\boldsymbol{Y}$ be a $n \times 1$ vector of the phenotype and $\mathbf{G}$ a $n \times p$ matrix, where the ith row ($\boldsymbol{g}_i$) is the ith subject and the jth column ($\boldsymbol{G}_j$) is the jth SNP. Consider the model:

$$\text{logit}\left(P\left(Y_i = 1 | \boldsymbol{g}_i\right)\right) \quad = \quad \beta_0 + \boldsymbol{\beta}^T \boldsymbol{g}_i$$

The goal is to test the global null, $H_0 : \boldsymbol{\beta} = \mathbf{0}$, against the alternative, $H_A : \boldsymbol{\beta} \neq \mathbf{0}$, but tailored towards sparse alternatives when only a few elements of $\boldsymbol{\beta}$ is nonzero in the presence of rare variants. In contrast, in a traditional GWAS setting, one tests for association with each SNP separately, $H_{0j} : \beta_j = 0$, the score statistic is:

$$S_j \quad = \quad \frac{\boldsymbol{G}_j^T \left(\boldsymbol{Y} - \hat{\boldsymbol{\mu}}_0 \right)}{SE}$$

where $\hat{\boldsymbol{\mu}}_0$ is the fitted mean under the global null. Rather than testing for each H_{0j} separately with S_j, a SNP-set method may take the vector of score statistics $\boldsymbol{S}$ and compute a single test statistic for the global null $H_0 : \boldsymbol{\beta} = \mathbf{0}$. In the GWAS setting, where $\boldsymbol{G}_j$ is not sparse due to the variants being common, one can rely on classical theory and approximate the distribution of S_j with the standard normal under the null. In the sequencing setting, however, the data sets typically contain rare variants (e.g. only 3 out of 500 subjects have the minor allele), i.e. $\boldsymbol{G}_j$ is sparse, thus, the numerator of S_j becomes a sum of just a few discrete terms that are either $1 - \hat{\mu}_0$ or $-\hat{\mu}_0$. In this case, normal approximation for S_j can be inaccurate even for reasonable sample sizes due to the discreteness of S_j. Dense signal methods, such as burden and SKAT, do not rely on the assumption of multivariate normality of $\boldsymbol{S}$, unlike sparse signal methods, such as MinP and higher criticism. As a sparse signal method, GHC is no exception and assumes the vector of score statistics $\boldsymbol{S}$ behaves as a multivariate normal under the null. Thus, GHC in the current form cannot be used for data with rare variants due to violation of its assumptions.

To circumvent this issue caused by rare variants for sparse signal methods, we propose a new framework for computing p-value for individual SNPs. The core idea is to compare the measure of association of the observed sample of a given SNP against that of all possible samples for all SNPs in the region. Note the new notations $M^{(j)} = \sum_i G_{ij}$, the number of

minor alleles for the jth SNP, and $M_c^{(j)} = \sum_i G_{ij} Y_i$, the number of minor alleles in cases for the jth SNP. Consider the model with only the jth SNP for the whole sample:

$$P\left(\boldsymbol{Y}|\boldsymbol{G}_j\right) \;=\; \frac{exp\left(\beta_0 \sum_i Y_i + \beta_j \sum_i G_{ij} Y_i\right)}{\prod_i \left(1 + exp\left(\beta_0 + \beta_1 G_{ij}\right)\right)}$$

It is well known in this case that $\sum_i Y_i$ is sufficient for β_0, so conditional on $\sum_i Y_i$:

$$P\left(\boldsymbol{Y}|\sum_i Y_i = c, \boldsymbol{G}_j\right) \;=\; \frac{exp\left(\beta_j \sum_i G_{ij} Y_i\right)}{\sum_{\mathbf{Y}^*:\sum_i Y_i^* = c} exp\left(\beta_j \sum_i G_{ij} Y_i^*\right)}$$

Note that condition on $\sum_i Y_i$ and $\boldsymbol{G}_j$, $M_c^{(j)}$ is sufficient for β_j. Here, one can proceed to compute conditional p-value for hypothesis on β_j based on the null distribution of $M_c^{(j)}|\sum_i Y_i = c, \boldsymbol{G}_j$ only. However, if sparsity is high in $\boldsymbol{G}_j$, the null distribution of the p-values from $M_c^{(j)}|\sum_i Y_i = c, \boldsymbol{G}_j$ would be highly discrete also. Instead, we propose to compute the p-value based on the null distribution of $M_c^{(j)}|\sum_i Y_i = c, M^{(j)} > 0$, a weaker conditioning. $M^{(j)} > 0$ is chosen as the conditioning since it naturally corresponds to the data, which consists of SNPs which have at least one observed minor allele. Assume that the measure of association is the statistic $T = T\left(M_c^{(j)}\right)$ where smaller values are considered more extreme, and let T^o be the actual observed value of the statistic. Then, the p-value $P\left(T < T^o|\sum_i Y_i = c, M^{(j)} > 0\right)$ can be decomposed by first conditioning on $M^{(j)}$:

$$\begin{aligned}
& P\left(T < T^o|\sum_i Y_i = c, M^{(j)} > 0\right) \\
= \;& \sum_m P\left(T < T^o|\sum_i Y_i = c, M^{(j)} = m, M^{(j)} > 0\right) P\left(M^{(j)} = m|\sum_i Y_i = c, M^{(j)} > 0\right) \\
= \;& \sum_m \sum_{k:T(k)<T^o} P\left(M_c^{(j)} = k|\sum_i Y_i = c, M^{(j)} = m\right) P\left(M^{(j)} = m|\sum_i Y_i = c, M^{(j)} > 0\right)
\end{aligned}$$

This decomposition shows the resulting p-value comes from comparing the measure of association observed for the jth SNP against that of all possible samples for other individual SNPs in the region under the null. T^o, representing the observed measure of association from the jth SNP, is compared against values of T for all possible different observed number of minor alleles of each SNPs in the region (indexed by m).

The final expression requires two components, $P\left(M^{(j)} = m | \sum_i Y_i = c, M^{(j)} > 0\right)$, which can be directly estimated from the data, and $P\left(M_c^{(j)} = k | \sum_i Y_i = c, M^{(j)} = m\right)$, which is conveniently hypergeometric under the null. In comparison to the very discrete p-values directly based on the null distribution of $M_c^{(j)} | \sum_i Y_i = c, \boldsymbol{G}_j$, this formulation alleviates that issue and gives smoother p-values that are closer to $U(0,1)$ under the null. However, if there are lots of rare variants in the region, the resulting p-values are still somewhat discrete. To smooth out the remaining discreteness in the p-values computed this way, one can compute $U_j = P\left(T \leq T^o | \sum_i Y_i = c, M^{(j)} > 0\right)$ and $L_j = P\left(T < T^o | \sum_i Y_i = c, M^{(j)} > 0\right)$ for the jth SNP. Then, generate $p_j \sim U(L_j, U_j)$ as the p-value for the jth SNP.

One can interpret this framework as the construction of a null distribution of all possible 2×2 tables for each SNP in the SNP-set. The distribution goes from 2×2 tables with extremely high measures of association (more extreme) on the one end to ones with low measures (less extreme) on the other. Each 2×2 table is weighted by the product of its probability under the null conditional on the corresponding number of minor alleles and the number of SNPs in the SNP-set with the corresponding number of minor alleles. The p-value for a particular SNP is obtained by looking up where its observed 2×2 table is situated in the constructed null distribution. We noted in section 1 that this framework can amplify weak sparse signals. To illustrate this, consider a causal SNP in the SNP-set that is only moderately associated with the phenotype. If the observed 2×2 table is only compared against other possible 2×2 tables of the causal SNP, then it will likely only seem moderately associated. However,

when compared against the constructed null distribution, the observed 2×2 table of the causal SNP will appear more extreme than the bulk of the distribution, which consists of less extreme 2×2 tables that likely have bigger weights. As a result, this 2×2 table will likely be placed closer to the highly associated end of the constructed null distribution and appear more extreme than in the comparison against possible 2×2 tables of the causal SNP only. In comparison, the observed 2×2 table for a noise SNP is likely to show little to no sign of association and will likely be placed among other tables that show little signs of association in the constructed null distribution. As a result, the observed 2×2 table of a noise SNP will not appear much differently when compared against the constructed null distribution or just the possible 2×2 tables of the noise SNP alone. In the end, the signal SNPs are likely to appear more extreme in this new framework, i.e. amplified, while noise SNPs are not.

2.2.1 Partitioning the SNP-set

One unwritten premise for the sparse signal methods is that there is sufficient attenuation of the marginal test statistic/p-value of the causal SNPs such that the method is able to spot it among many other noise SNPs. This premise is typically met with common variants. However, this can be an issue for rare variants, because fewer number of observed minor alleles means a smaller range of observable imbalance in allele counts between cases and controls, leading to weaker statistical evidence for association. For example, in the extreme case when the causal SNP is a singleton, there is only two observable configurations, i.e. a single minor allele in a case or in a control. In this case, both configurations are equally likely under the null, so the marginal test statistic/p-value can give no evidence for association even if the singleton was indeed causal.

Even though the marginal p-value introduced in section 2 behaves well with rare variants, it does not circumvent this issue, as the new p-value will still show little attenuation under the alternative with extremely rare variants. The new p-value will show the greatest improvement

when the minor allele count is not extremely low so that the p-value is still able to show some attenuation under the null, but low enough so that one cannot rely on asymptotics. In light of these two considerations, we propose to use some threshold on minor allele count when testing for association on a SNP-set with rare variants. The SNP-set will be partitioned into two sets, one for SNPs with minor allele count less than the threshold and one for SNPs with minor allele count greater or equal to the threshold. A single p-value will be computed for the first set with an aggregating method such as SKAT, while marginal p-values will be computed for each SNP in the second via the framework from section 2. The resulting p-values, the single one from the first set and all the marginal ones from the second set, will be compiled into a single p-value for the whole SNP-set via a thresholding technique like GHC. If the sparse signals are in the extremely rare variants, i.e. the first set, then marginal p-values for these SNPs will have little attenuation, and one is better served by trying to pick up some power by aggregating them. If the sparse signals are in the other rare variants, i.e. the second set, then our p-values will perform well and thresholding methods should be able to pick them up.

In addition to loss of power, the extremely rare variants are also likely to produce highly correlated marginal p-values using the procedure in section 2. Our framework serves to provide a distribution for all observable 2×2 tables in the SNP-set that allows for a more precise assessment of the degree of association present in the actual observed 2×2 table for a particular SNP. If a SNP has very few minor alleles, then it can only produce a small number of observable 2×2 tables that show little evidence of association and will all be placed around the less extreme end of the distribution. As a result, the marginal p-values produced by our procedure for variants with very low minor allele counts will always be from the same part of the distribution, and thus be highly correlated. This is undesirable due to the limit in the amount of correlation GHC as well as the weighted GHC introduced below are able to handle.

2.3 Review of GHC and Handling Correlation

The original higher criticism (HC) test was developed by Donoho and Jin (2004) for testing the global null $H_0 : \boldsymbol{\beta} = \mathbf{0}$ against sparse alternatives where only a few elements of the p-dimensional $\boldsymbol{\beta}$ is nonzero. It assumes that the individual test statistics are independent and standard normally distributed under the null. For the vector of marginal test statistics $\boldsymbol{Z}$, the HC statistic is defined as:

$$HC = \sup_{t>0} \frac{S(t) - 2p\bar{\Phi}(t)}{\sqrt{2p\bar{\Phi}(t)\left(1 - 2\bar{\Phi}(t)\right)}}$$

where $S(t) = \sum_{j=1}^{p} I_{|Z_j| \geq t}$ and $\bar{\Phi}(t)$ is the survivor function of the standard normal distribution. This test rejects for large values of HC.

To accommodate correlation in $\boldsymbol{Z}$, which can arise due to LD in the context of SNP-set testing, Hall and Jin (2010) proposed the iHC which performs HC with $\boldsymbol{Z}^*$, the $\boldsymbol{Z}$ transformed with the Cholesky decomposition of the correlation matrix, in place of $\boldsymbol{Z}$. In the presence of correlation, the transformation of $\boldsymbol{Z}$ leads to mixing the signal SNPs with the noise SNPs since $\boldsymbol{Z}^*$ is a linear combination of all components of $\boldsymbol{Z}$. When the signal is sparse, these linear combinations mix just a few signal SNPs with many noise SNPs, thus masking attenuations from the few signal SNPs and leading to poor power. GHC takes a different approach to handling the correlation and keeps the same numerator as the original HC statistic. Instead of using the variance estimate for independent $\boldsymbol{Z}$, GHC directly estimate the variance of the numerator ($\hat{var}(S(t))$) given the estimate of the correlation matrix. The GHC statistic is of the form:

$$GHC = \sup_{t>0} \frac{S(t) - 2p\bar{\Phi}(t)}{\sqrt{\hat{var}(S(t))}}$$

In addition to loss of power due to mixing of signal and noise, asymptotic results for iHC requires a very large p to be accurate (Barnett and Lin, 2014). It has been shown that size of tests that use asymptotic results can be inaccurate for p as large as one million. That is an astronomical number compared to the number of SNPs in most SNP-sets of interest, e.g. a couple hundred in a signaling pathway. Asymptotic results for GHC suffers the same issue in fact. Instead of using the asymptotic approximation, Barnett et al. (2016) developed an analytic p-value computation for GHC. The formulation for the p-value computation can be written as:

$$\begin{aligned}
P\left(GHC \geq h\right) &= 1 - P\left(GHC < h\right) \\
&= 1 - P\left(\bigcap_{t>0}\left\{S\left(t\right) < h\sqrt{\hat{var}\left(S\left(t\right)\right)} + 2p\bar{\Phi}\left(t\right)\right\}\right) \\
&= 1 - P\left(\bigcap_{k=1}^{p}\left\{S\left(t_k\right) < p - k + 1\right\}\right) \\
&= 1 - \prod_{k=1}^{p} P\left(S\left(t_k\right) \leq p - k | \bigcap_{l=1}^{k-1}\left\{S\left(t_l\right) \leq p - l\right\}\right) \\
&\approx 1 - \prod_{k=1}^{p} P\left(S\left(t_k\right) \leq p - k | \left\{S\left(t_{k-1}\right) \leq p - k + 1\right\}\right)
\end{aligned}$$

The t_k are solutions to $h\sqrt{\hat{var}\left(S\left(t_k\right)\right)} + 2p\bar{\Phi}\left(t_k\right) = p - k + 1$ for $k = 1 \ldots p$. The third equality holds due to the monotone nature of $h\sqrt{\hat{var}\left(S\left(t\right)\right)} + 2p\bar{\Phi}\left(t\right)$, which allows simplifying the statement over the union over $t > 0$ to just p points. The last approximation is to say the distribution of $S\left(t_k\right)$ only depends on $S\left(t_{k-1}\right)$, rather than that of $S\left(t_l\right)$ for $l = 1 \ldots k - 1$. The conditional distribution of $S\left(t_k\right) | S\left(t_{k-1}\right) = m$ is approximated with a beta-binomial that has the first two moments matched with that of $S\left(t_k\right) | S\left(t_{k-1}\right) = m$.

In order to compile a p-value for the whole SNP-set with the new p-values, one needs to transform the p-values into statistics and estimate the correlation of said statistics. GHC assumes that the marginal test statistics are multivariate normal, so inverting via the normal

distribution function is the obvious approach. Given the vector of marginal p-values $\boldsymbol{p} = \begin{bmatrix} p_1 & \dots & p_p \end{bmatrix}^T$, define the corresponding statistic Z_j for p_j as $|Z_j| = \bar{\Phi}\left(\frac{p_j}{2}\right)$. This defines the absolute value but not the sign, which can be borrowed from conventional measures of association such as the score statistic. Z_js defined this way will reflect the observed direction of association, but it can only be done if there is imbalance in the minor alleles between cases and controls, i.e. the score statistic is non-zero. If the score statistic is 0, there is no sign to borrow, but this is evidence of no association between the SNP and the phenotype. Typically, if the test statistic is standard normal under the null, we would expect the statistic to be of each sign half of the time. So this would be accurately reflected if Z_j is assigned either sign with probability $\frac{1}{2}$ when the score statistic is 0. This procedure will be used for the p-values for every SNP in the second set of SNPs per the partitioning described in section 2.1. While the single p-value from the first set of SNPs will be inverted the same way to get the absolute value of the statistic, the sign assignment will be the same as that of the score statistic for association between $\boldsymbol{Y}$ and $\boldsymbol{C}$, where C_i is the indicator of having at least one minor allele among those SNPs in the first set. From this point on, marginal p-values/test statistics will refer to both those computed for each SNP in the second set as well as the one computed for the whole first set.

With the numerical value of the marginal test statistic $\boldsymbol{Z}$ defined, one still needs to estimate the correlation between $\boldsymbol{Z}$. In GHC, the score statistics have a closed form, and the correlation matrix of $\boldsymbol{Z}$ can be easily estimated from genotype matrix directly. The marginal test statistic/p-value introduced here has no closed form, so the correlation matrix cannot be estimated directly from the data. Instead, to estimate the correlation of $\boldsymbol{Z}$, we propose the following procedure:

1. Permute the phenotype $\boldsymbol{Y}$, which is valid under the null
2. Compute the statistics $\boldsymbol{Z}^{(b)}$ for permutations $b = \{1 \dots B\}$

3. Estimate the correlation matrix $\boldsymbol{\Sigma}$ from $\left\{\boldsymbol{Z}^{(b)}\right\}$.

This ensures that the estimate $\hat{\boldsymbol{\Sigma}}$ characterizes the correlation structure of $\boldsymbol{Z}$ under the null. If weighting of the SNPs is not desired, then one can compile a single p-value for the SNP-set with the original GHC at this point.

2.3.1 Compiling and Weighting SNPs

In order to weight the SNPs, the GHC test statistic needs to be redefined. With the original GHC statistic, the test statistic was based on $S(t) = \sum_{j=1}^{p} I_{|Z_j| \geq t}$. In order to incorporate weights, the test statistic will be based on $S_w(t) = \sum_{j=1}^{p} I_{|Z_j| \geq w_j t}$, where w_j is the weight corresponding to the jth SNP. The purpose of assigning the weights this way is that it allows one to assign lower weights to SNPs that are believed a priori to be signal SNPs, so it would be easier for the marginal test statistic of those SNPs to incur the indicator. With the new definition of $S_w(t)$, the new test statistic is also redefined:

$$GHC_w = \sup_{t>0} \frac{S_w(t) - \sum_{j=1}^{p} 2\bar{\Phi}(w_j t)}{\sqrt{\hat{var}(S_w(t))}}$$

The variance estimate $\hat{var}(S_w(t))$ for $S_w(t)$ is similar to that of $S(t)$ as from the original GHC, but with a slight modification due to the introduction of the weights (details in appendix).

With similar logic and approximations as in the original GHC, the p-value for the newly defined test statistic can be approximated as

$$P(GHC_w \geq h) \approx 1 - \prod_{k=1}^{p} P\left(S_w(t_k) \leq p-k \middle| \{S_w(t_{k-1}) \leq p-k+1\}\right)$$

The t_k here are found as solutions to $h\sqrt{\hat{var}(S_w(t_k))} + \sum_{j=1}^{p} 2\bar{\Phi}(w_j t) = p - k + 1$ for $k = 1 \ldots p$, which is quite similar to that from GHC also. The conditional distribution of

$S_w(t_k)\,|S_w(t_{k-1}) = m$ is once more approximated with a moment-matched beta-binomial distribution. The first two moments of $S_w(t_k)\,|S_w(t_{k-1}) = m$ require different approximations than those of $S(t_k)\,|S(t_{k-1}) = m$ in GHC in order to accommodate the new weights (details in appendix). Due to the approximations used, this weighted procedure has a limit to the amount of variability in the weights as well as of correlation between the marginal test statistics it can handle.

2.4 Simulation Studies

We conduct simulation studies to assess the control of type-I error as well as power of our method. The genotype data is simulated with ms, while keeping the absolute value of pairwise correlation between SNPs < 0.1. This limitation is due to the crudeness of the approximations used to approximate the moments of $S_w(t_k)\,|S_w(t_{k-1}) = m$. The phenotypes are generated in R, while the SKAT and burden p-values are computed with the R package SKAT. The GHC p-values are computed with the R package GHC, while those of the weighted GHC are computed with our own R code.

2.4.1 Null Simulations

For the size simulation, we use a minor allele count threshold of 10 and generate 50 SNPs over that threshold and 30 SNPs under that threshold for 1000 cases and 1000 controls under the null. The SNPs in the second set have minor allele counts between 10 and 40 inclusively, so that there are plenty of SNPs with minor allele counts where the framework introduced in section 2 is believed to show the greatest improvement. A marginal p-value is computed for each of the 50 SNPs over the threshold using the new framework, and a single p-value for the 30 SNPs under the threshold with SKAT. Compilation of the 51 signals is done both unweighted as well as with beta weights, $B(1, 6)$. In the latter case, the SKAT p-value for the SNPs under the threshold will use the same beta parameters for weighting. Given a distribution with density function f for the weights, the inverse of the weight for SNPs over

the threshold is f evaluated at the minor allele frequency, i.e. $\frac{1}{w_j} = f(p_j)$, while the single weight for all SNPs under the threshold is obtained the same way but with the average minor allele frequency of all such SNPs. Compilation of the p-values for the whole SNP-set is done with the original GHC if unweighted, while compilation is done with the weighted GHC introduced in section 3 if using the beta weights.

The reason the weights for the individual SNPs are assigned this way is for a more direct comparison to SKAT with the weighted linear kernel, which is the version computed in these simulations. For SKAT with this kernel, the test statistic can be written as $\boldsymbol{S}^T\mathbf{W}\boldsymbol{S}$, where $\boldsymbol{S}$ is the vector of marginal score statistics for each SNP and $\mathbf{W}$ a diagonal matrix with diagonal elements corresponding to the weights for each SNP $\left\{w_j^s\right\}$. Note the test statistic can be rewritten as $\boldsymbol{S}_w^T\boldsymbol{S}_w$, where $\boldsymbol{S}_w$ is a vector whose jth element corresponds to the product of the jth marginal test statistic and the square root of its weight, i.e. $\sqrt{w_j^s}S_j$. Thus, $\sqrt{w_j^s}$ can be seen as the weight applied on the same scale as the marginal test statistics and is assigned the value of the chosen beta density evaluated at the jth minor allele frequency in SKAT. In the weighted GHC introduced in section 3, the weights are applied to the marginal test statistics through the indicator function $I_{|Z_j|\geq w_j t}$. Rearrangement of the condition of the indicator function yields $\left|\frac{1}{w_j}Z_j\right| \geq t$, assuming $w_j > 0$, which means that $\frac{1}{w_j}$ is the weight applied on the same scale as the marginal test statistics. Thus, if the value of $\frac{1}{w_j}$ in weighted GHC is assigned the same way as $\sqrt{w_j^s}$, i.e. evaluation of the beta density at the corresponding minor allele frequency, then the weighting scheme of the two methods correspond directly.

The results over 100,000 replications are summarized in the QQ-plots in figure 2.1. Both unweighted and weighted QQ-plots adhere closely to the 45 degree line. The unweighted one shows almost no deviation until the tail where there are only a few points. The weighted one, however, shows a small amount of "wiggle" around the 45 degree line all the way through,

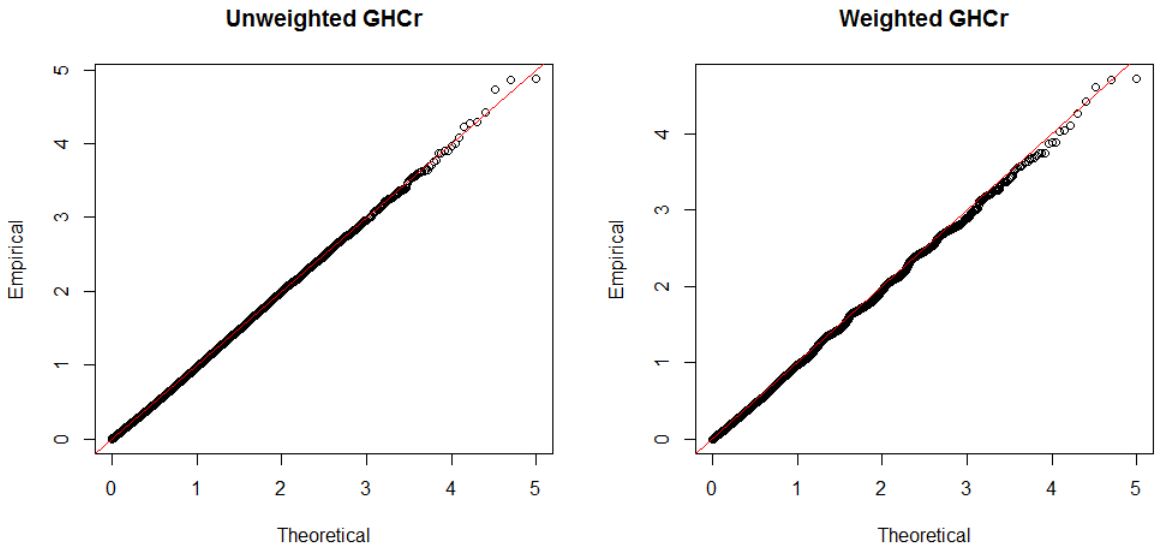

Figure 2.1: QQ-plots for unweighted GHC and weighted GHC in the null simulations

but never shows any noticeable deviation. This "wiggle" is likely caused by the crudeness of the approximations currently used to compute the conditional moments used in the p-value computation mentioned above. Both versions of GHC seem to control size well in these settings, although the performance of the weighted GHC is likely to deteriorate with more variable weights in its current form.

2.4.2 Power Simulations

Due to the aforementioned issues caused by rare variants on the marginal test statistics, no other sparse signal methods are able to handle rare variants. Instead, SKAT and burden test are used in these simulation for power comparison. We chose a variety of sparsity and effect sizes as well as ways to choose the causal SNPs to compare the power of each method. For power simulations, the genotype generation remains the same as the null simulations, but the phenotypes for the 2000 subjects are generated prospectively. Given the genotype, the phenotype is generated via a logistic model, where a given number of SNPs are selected to be the causal SNPs:

$$\text{logit}\left(P\left(Y_i = 1|\boldsymbol{g}_i\right)\right) \quad = \quad \beta_0 + \boldsymbol{\beta}^T \boldsymbol{g}_i$$

There are three regimes for selecting the causal SNPs. One, the causal SNPs are selected randomly out of the 80. Two, the causal SNPs are randomly selected from the set under the minor allele count threshold, which contains many singletons. Three, the causal SNPs are selected as the rarest out of the set over the minor allele count threshold. The purpose of the latter two is to show the utility of the procedure described in section 2.1. In the second regime, there are dense signals in the set of SNPs under the minor allele count threshold. It is unclear a priori whether aggregating such dense signals as with SKAT and burden or treating it all as a single strong signal in GHC will yield better power. In the third regime, the causal SNPs have minor allele counts where the new marginal p-values are expected to show the most improvement, so weighted GHC is expected to outperform SKAT. For the first and third causal regimes, the number of causal SNPs vary between 2, 3, and 5 out of 80, with corresponding effect size of the causal SNPs, at 1, 0.7 and 0.5. For the second causal regime, the effect sizes are determined in the same manner as the power simulations in Wu et al. (2011), where a SNP with minor allele frequency p_o is assigned effect size $c\left|log_{10}\left(p_o\right)\right|$. The number of causal SNPs vary between 15, 20, and 25 out of 80, with corresponding value of c at 0.5, 0.45, 0.4. The direction of association are always kept the same in order for burden to have reasonable power, although our method and SKAT do not benefit from this unidirectionality. The compilation of the p-values will be unweighted for all methods with the first regime for selecting causal SNPs, but will employ a $B\left(1, 6\right)$ weight for the latter two.

The power results for the various scenarios visited at $\alpha = 0.05$ are summarized in figure 2.2. For the first causal regime, the increase in power from our method is the most obvious

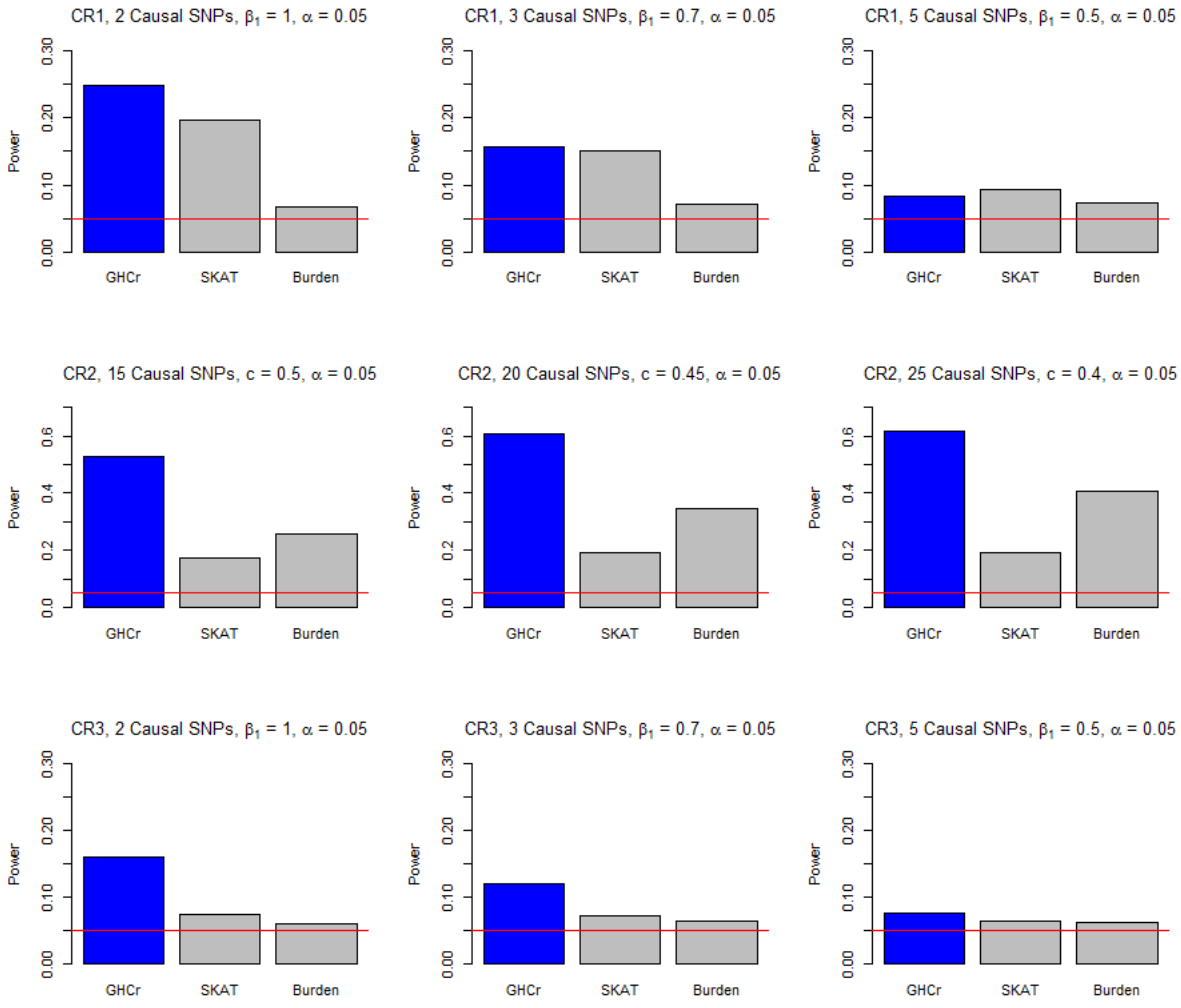

Figure 2.2: Power for all three causal regimes, at various choices of sparsity and effect size, at $\alpha = 0.05$, for our approach, SKAT, and burden, from left to right

in the sparsest case, with 2 causal SNPs and effect size $\beta_1 = 1$. With 3 causal SNPs and effect size $\beta_1 = 0.7$, the increase in performance is less pronounced, while SKAT overtakes in performance in the third setting. In the second causal regime, GHC is surprisingly outperforming both SKAT and burden by a wide margin. With the given choices of parameter values, it seems to be better to treat the signals in the set of SNPs under the threshold as a single strong signal with GHC than to aggregate all the SNPs together with SKAT and burden. Burden outperforming SKAT here is not too unexpected as all the signals are in the same direction. Though the ordering of performance may change with different choices of number of causal SNPs and/or effect sizes, these results clearly show the utility of the approach laid out in section 2.1. Lastly, in the third causal regime, GHC performs the best in all three settings. This is expected since this is where our method is expected to show the most improvement. In general, our method performs well when the sparsity is low (2 or 3 causal SNPs), and is able to retain some power when the signals are all in the extremely rare variants. However, when sparsity decreases, aggregating methods such as SKAT can perform better in some settings. Additional results at other significance levels can be found in the appendix.

2.5 Analysis of Dallas Heart Study Data

The Dallas Heart Study contains sequencing data on 93 variants in the genes ANGPTL3, ANGPTL4, and ANGPTL5 as well as log-transformed serum triglyceride levels for 3476 subjects. The subjects include both genders from three ethnicities, black, Hispanic, and white. Since our method is aimed at binary phenotypes, the serum triglyceride level was dichotomized as cases and controls. The data set is analyzed for association between the dichotomized phenotype and the SNP-set containing all variants in the data set with minor allele frequency < 0.05 (this leaves 92 of the 93 variants) with both weighted and unweighted versions of SKAT, burden, and GHC as introduced in this paper. Both versions of GHC uses the same 10 minor allele count threshold as the simulations, which puts 78 variants in the set

under the threshold and 14 variants in the set over for a total of 15 marginal statistics. The weighting scheme for each method remains the same as the simulations also. In its current form, our method is unable to control for covariates, so all three different analyses are done without covariates for comparison. Although a preliminary analysis with only gender and ethnicity as the covariates do not show any highly significant association, the analysis results should still be taken with a grain of salt and is more for the purpose of comparison between the methods.

The weighted and unweighted p-values for each method can be found in table 2.1. If the phenotype is dichotomized with the highest 10% as cases and the rest as controls, both SKAT and burden have smaller p-values than GHC. Upon further inspection of the marginal test statistics for GHC, there are 4 marginal statistics with values approximately 2 or greater with none greater than 4 for SNPs over the threshold, while the single statistic for the set of SNPs under the threshold is small and shows little sign of association. This seems to indicate several moderate signals, which is a situation where SKAT is expected to perform better. This has some loose correspondence to the third setting of the first causal regime in the power simulations, where SKAT performs better as well.

Defining the top 10% as cases is not a fixed cutoff and has no clinical interpretation. Instead, if the phenotype is dichotomized with subjects with serum triglyceride level over 150 (the clinical definition of high triglyceride) as cases and otherwise as controls, the results look opposite of the other definition, with GHC having much lower p-values than SKAT and

Table 2.1: Unweighted and weighted p-values for GHC, SKAT, and burden from analysis of the data from Dallas Heart study

Definition of cases		GHC	SKAT	Burden
Highest 10%	Unweighted	0.00197	0.00096	0.00058
	Weighted	0.00746	0.00079	0.00078
> 150	Unweighted	0.00018	0.41	0.55
	Weighted	0.00010	0.36	0.61

burden. The marginal statistics of GHC tell a difference story than before also. The largest marginal test statistic for SNPs over the threshold is only approximately 2, while the single statistic for the set of SNPs under the threshold is over 4. This seems to indicate that if the signals in the extremely rare variants combined is extremely strong, and there is little signal elsewhere, then SKAT and burden may risk this signal getting buried among the noise SNPs, but GHC may be able to pick it up as a strong single sparse signal. This definition of the phenotype only yields 11 cases, though a low number of cases does not affect the validity of our method. This situation corresponds to the second causal regime of the power simulations, where the signals are strong and all in the SNPs under the threshold. In all three settings for the second causal regime, GHC outperforms both SKAT and burden. However, the correspondence is not exact since the power of burden is boosted by the unidirectionality of the signals in the simulation, which is unlikely to be the case in the data.

The p-value of the better performing methods in either cases are similar, while GHC achieves noticeably lower p-value when it does not perform the best than SKAT and burden. Note that the results of the analysis change completely depending on the definition of the phenotype. In combination with the lack of control for covariates, these results are by no means informative in practice, but show the strength and weakness of each method in a context outside of simulation with correspondence to scenarios visited in the simulations.

2.6 Discussion

We propose a method for association testing between a SNP-set (genes, signalling pathways) and a phenotype of interest for a sequencing study setting, where rare variants are prominent. Our method allows for data sets with rare variants and binary phenotype, and is tailored towards sparse alternatives. Existing sparse signal methods, such as GHC, makes assumptions about the distribution of marginal test statistics that fail in these situations and is unable to weight the SNPs. The core of the new method is a new marginal test statistic/p-value that

requires no such assumption. This new framework compares the observed sample of a given SNP against more extreme possible samples for each SNP in the SNP-set by constructing a null distribution of all such possible samples. In comparison, typical marginal test statistics (e.g. score, Fisher's exact test) only compares the observed sample against more extreme possible samples for the SNP being tested. The constructed null distribution serves to create a finer spectrum for assessment of the observed measure of association, and to amplify weak to moderate signals. The latter is because a signal may appear weak or moderate when only compared against possible samples at the SNP being tested, but will likely appear stronger when compared against this new null distribution, which is less likely for noise SNPs. In addition, we build upon the GHC framework by incorporating weights for the SNPs to increase power, although this approach has limitation in the variability of the weights used as well as correlation among the test statistics.

In most of the scenarios visited in our simulation studies, our method has improved power over SKAT and burden in the presence of a significant portion of rare variants. Power improvement is quite noticeable when signal is not in the extremely rare variants, such as singletons. However, when the signals are in these variants, our method is still able to retain some of the power by aggregating the extremely rare variants first. In comparison with existing sparse signal methods, such as GHC and minP, our method can handle rare variants with binary phenotype and does not rely on asymptotics.

The main drawback of our method is the lack of ability to control for covariates, since it no longer allows for the hypergeometric assumption under the null. The approximations currently made to incorporate the weights are quite crude, leading to a limitation in the amount of variability in the weights as well as the correlation among the test statistics the weighted GHC is able to handle. Additionally, the use of permutation to handle LD is quite time consuming. So improved approximation for the weights, controlling for confounders,

and speeding up the software are the main focus for future work.

With the increasing abundance of whole genome sequencing data, ignoring rare variants as with GWAS becomes more and more wasteful. Existing SNP-set methods have found creative ways to deal with sparse signals, but not with rare variants when testing for association with binary phenotypes. This is a glaring hole in the landscape of SNP-set methods. Our method is a first step in attempting to address this issue.

Flexible Model Selection for Mechanistic Network Models via Super Learner

Sixing Chen[1], Antonietta Mira[2, 3], Jukka-Pekka Onnela[1]

[1]Department of Biostatistics

Harvard School of Public Health

[2]InterDisciplinary Institute of Data Science

Università della Svizzera Italiana

[3]Department of Sciences and High Technology

Università degli Studi dell'Insubria

3.1 Introduction

Many systems of scientific and societal interest can be represented as networks, and network models are used, among many other applications, to study social networks, communication patterns, scientific citations, and protein-protein interactions (Newman, 2010; Wasserman and Faust, 1994; Pastor-Satorras and Vespignani, 2007; Lusher et al., 2012; Raval and Ray, 2013).

There are (at least) two prominent paradigms to the modeling of networks, which we call the statistical approach and the mechanistic approach. In the statistical approach, one describes a model that specifies the likelihood of observing a given network, i.e., these are probabilistic models of data that take the shape of a network (Robins et al., 2007; Hoff et al., 2002; Goyal et al., 2014). In the mechanistic approach, one specifies a set of domain-specific mechanistic rules, informed by scientific understanding of the problem, that are used to grow or evolve the network over time (Barabási and Albert, 1999; Watts and Strogatz, 1998; Solé et al., 2002; Vázquez et al., 2003; Klemm and Eguiluz, 2002; Kumpula et al., 2007). For example, in the context of social networks, mechanisms of interest might include triadic closure or reciprocation of directed edges. Both modeling approaches provide distinct angles and advantages to our understanding of complex systems, and in both approaches, one is interested in learning about the connection between microscopic and macroscopic structures. Both mechanistic and statistical models are indexed by parameters and calibration/inference on those parameters sheds light on those micro/macro structures.

A particular generative mechanism in mechanistic models may seem like a strong assumption in contexts where one does not directly observe the formation of the network, and where it is difficult to study microscopic interactions in isolation of the rest of the system. For example, it might be difficult to learn about the mechanistic rules that govern the formation or dissolution of ties in in-person interactions, whereas doing so in the setting of online social networks

might be more feasible since every interaction can be recorded. In some biological networks the pairwise interactions are well understood both theoretically and experimentally, they can be studied in isolation, and these interactions are highly reproducible. For example, gene duplication is one of the main drivers of the evolution of genomes and it is well understood, and therefore perhaps not surprisingly, network models based on gene duplication were one of the first large-scale models used in systems biology (Raval and Ray, 2013).

In comparison, statistical models may be seen as a more sound approach in settings where domain specific understanding is not as readily available to guide the selection of mechanisms, or there are too many mechanisms and including all of them would not lead to insightful modeling. This is in line with the use of statistical modeling more broadly, where one of the goals might be to learn how different predictors are associated with the response, which is a problem that can be studied even if the true associations between response and predictors, let alone the underlying mechanisms, are unknown. Common statistical models have limitations in the structures they are able to accommodate (Goyal and Onnela, 2017), and fitting and sampling from some of these models can be difficult. For example, the popular class of exponential random graph models (ERGMs) appear not to be consistent under sampling (Shalizi and Rinaldo, 2013). Mechanistic models do not suffer from these limitations as much, since generation of network structures from a handful of mechanisms is usually computationally inexpensive, so it is relatively simple to sample from a particular model.

Another advantage of mechanistic models is the ease with which one can incorporate domain knowledge in the model. Since the modeler is in control of the mechanisms to include, one is able to encode relevant domain knowledge of known or hypothesized interactions between actors in the system as mechanistic rules. Duplication-divergence models in protein-protein interaction networks are good examples of this (Raval and Ray, 2013). In statistical models,

one is unable to model such interactions directly, and can only consider the types of networks structures one expects to observe due to these interactions.

While there is a very extensive literature on mechanistic models in network science, there is a dearth of work on model selection in mechanistic models (Middendorf et al., 2005). The aim of this paper is to provide a framework for model selection in mechanistic network models. For instance, given a full model which has an array of different generative potential or plausible mechanisms, we are interested in selecting between different submodels each possessing only a subset of the mechanisms of the full model. Traditional likelihood-based model selection are not applicable to most mechanistic models because in most cases their likelihood functions are not known. One of the reasons why the likelihood functions are intractable is that in mechanistic models one must consider all the possible paths to generate any one particular network realization, which leads to a combinatorial explosion save for the most trivial models. As such, one must consider likelihood-free approaches.

One recent likelihood-free approach to both inference and model selection for problems with intractable likelihoods is Approximate Bayesian Computation (ABC) (Marin et al., 2012; Sunnåker et al., 2013; Lintusaari et al., 2016). As the name suggests, this Bayesian approach aims at calibrating the model parameters by obtaining the posterior distribution for the parameters of interest. Following Bayes theorem, the posterior is obtained by combining information from the prior distribution and the observed data set. ABC inference starts by generating samples of possible parameter values from the prior distribution. For each sample from the prior, one generates a data set according to the model for the data, where the nature of the model, statistical or mechanistic, is not relevant for the purpose. Then, a set of pre-specified summary statistics is computed for each of the generated data sets as well as the observed data set. Given a distance measure for the summary statistics space, one accepts only the generated data sets whose summary statistics' distance from those of

the observed data sets are within a certain threshold. These generated data sets are deemed "close" to the observed one, and the parameters sampled from the prior corresponding to these data sets form the approximation to the posterior distribution. Model selection with ABC is similar but includes an additional layer of hierarchy for the model index.

The main difficulty in applying ABC arises when selecting the summary statistics as well as the threshold on the distance. If the selected summary statistics are sufficient for the parameters of the model and the distance threshold is zero, i.e., only data sets with exactly matching summary statistics are kept, then the resulting posterior will be exact in the limit as the number of generated samples goes to infinity (Lintusaari et al., 2016). However, should one fall short on either of these accounts, then the obtained posterior will be an approximation to the true posterior. Since likelihood-free approaches like ABC are only needed with intractable likelihoods, it will typically be difficult to find the sufficient statistics in these situations, though there is previous work on how to select good summary statistics for an ABC procedure (Prangle et al., 2014). As for the threshold on distance measure, the smaller the distance threshold, the greater the computational burden to generate a sufficient number of accepted samples. In fact, outside of using discrete summary statistics, it may be totally impractical to use a distance threshold of zero. As a result, the performance of ABC inference can suffer due to the inaccuracy of the resulting posterior. ABC model selection suffers from these same issues with reference to the model index which becomes an additional model parameter on which inference is required. Even if one were to select statistics that are marginally sufficient for the submodels, they may not be jointly sufficient for the full model and not be able to discriminate among the various models under comparison, save for some special cases. Lastly, if one does somehow manage to select all the sufficient statistics and conduct model selection with ABC, the resulting ABC Bayes factor may have no correspondence to the true Bayes factor (Robert et al., 2011).

Instead of dealing with the issues stemming from the inaccuracy of the ABC posterior, we propose a procedure for model selection that borrows the data generation from candidate models from ABC but not the Bayesian aspects, i.e., ABC without the "B". Just as in ABC, the data we generate from each candidate model will be the basis for model selection as the training data, but rather than a Bayesian approach, we propose to conduct model selection with Super Learner (Polley et al., 2011; Van der Laan et al., 2007). Originally proposed for prediction in a regression setting, and generalizable to others, Super Learner is an ensemble algorithm that makes a prediction by combining the predictions from a library of candidate algorithms. Given a particular loss function, Super Learner aims to minimize the expected loss, called the risk. The discrete Super Learner will simply pick the candidate algorithm that has the lowest cross-validated risk in the training data, whereas the full Super Learner will create the convex combination of the algorithm-specific estimates that has the lowest cross-validated risk. Given a bounded loss function, the discrete Super Learner as well as the full Super Learner have the so-called oracle property, meaning that asymptotically they perform at least as well as the optimal candidate algorithm and the optimal convex combination of the candidate algorithms, respectively (Dudoit and van der Laan, 2005; Van Der Laan and Dudoit, 2003).

Our proposed approach has similarities to that of Pudlo et al. (2015), which is a random forest-based ABC approach for model selection that is fairly robust to the choice of summary statistics. They measure performance with the prior error rate, which is the probability to select the wrong model, averaged over the prior. Our proposed approach can be seen as a generalization of that of Pudlo et al. (2015). First, the choice of performance measure is flexible and can be encoded directly into the loss function. For example, the prior error rate implicitly weighs misclassification differently for each model due to the sensitivity to the choice of prior. Should one desire a measure that does not discriminate between misclassification of different models, one can use a measure like the area under the receiver operating

characteristic curve (AUC) (Bradley, 1997; Ling et al., 2003). Second, Super Learner can make use of a host of candidate algorithms, including random forest, to perform the classification. Random forest is a very flexible algorithm, but it may not perform well in all settings. One can be more robust against this by having more candidate algorithms for Super Learner.

We can deem the random forest-based ABC approach of Pudlo et al. (2015) as a special case of our approach if we do not cross-validate and only have random forest as a candidate algorithm. In our model selection setting for mechanistic network models, we propose to use Super Learner to predict the model index for a particular network realization. Due to the intractable nature of the likelihood and the ease of generating data, model selection with mechanistic network models lends itself well to the Super Learner framework. The rest of the paper is organized as follows. In Section 2, we provide a brief overview of Super Learner as well as the procedure for model selection in the context of mechanistic network models. In addition, we introduce and motivate a simple mechanistic network model as a proof of concept, and we then use this model in our subsequent simulations. In Section 3, we lay out the details of the simulations as well as the results and evaluate the performance of Super Learner. Finally, in Section 4, we conclude with further discussions and with suggestions for future work.

3.2 Methods and Material

3.2.1 The Super Learner Framework

Given a particular loss function L, Super Learner, as introduced by Van der Laan et al. (2007), aims to minimize the risk $E[L]$ with an algorithm for prediction composed from a library or set of candidate algorithms $\{Q_l\}$. The procedure begins by partitioning the training data, with predictors $\boldsymbol{X}_t$ and outcome $\boldsymbol{Y}_t$, into V validation sets. The covariates and outcome of the vth validation set are referred to as $\boldsymbol{X}_t^v$ and $\boldsymbol{Y}_t^v$, while those of the

corresponding training set, the union of the remaining $V-1$ validation sets, are referred to as $\boldsymbol{X}_t^{-v}$ and $\boldsymbol{Y}_t^{-v}$. For the vth validation set, each candidate algorithm Q_l is trained on $\left(\boldsymbol{X}_t^{-v}, \boldsymbol{Y}_t^{-v}\right)$. The resulting trained algorithm $\hat{Q}_l^v$ is then evaluated at $\boldsymbol{X}_t^v$, giving prediction $\hat{\boldsymbol{Y}}_l^v$. After training each candidate algorithm on each training set and evaluating it on the corresponding validation set, a new data set is formed with the cross-validated predicted outcome $\left\{\hat{\boldsymbol{Y}}_l^v\right\}$ and the observed outcome $\boldsymbol{Y}_t^v$ from all validation sets. This cross-validation procedure is to prevent overfitting, and this new data set will be the basis for the final prediction algorithm.

In the case of the discrete Super Learner, the candidate algorithm with the smallest estimated cross-validate risk is chosen for final prediction. Assuming a regression setting and a squared error loss function, the cross-validated risk for the lth candidate algorithm can be estimated as $\hat{E}\left[L\left(Q_l\right)\right] = \frac{1}{V}\sum_v \frac{1}{n_v}\sum_i \left(Y_i - \hat{Y}_{l,i}^v\right)^2$, where the first summation is over the validation sets and the second is over the n_v observations in the vth validation set. In the case of the full Super Learner, the estimated risk will be minimized over all convex combinations of the candidate algorithm. In the same regression setting with squared loss and a particular convex combination $\{a_l\}$, where $\sum_l a_l = 1$ and $a_l \geq 0$, the cross-validated risk can be estimated as $\hat{E}\left[L\left(\boldsymbol{a}\right)\right] = \frac{1}{V}\sum_v \frac{1}{n_v}\sum_i \left(Y_i - \sum_l a_l \hat{Y}_{l,i}^v\right)^2$. Once the final prediction algorithm is determined, i.e. Q_{l^*} that achieves the smallest risk in the discrete Super Learner or $\{a_l^*\}$ in the full Super Learner, each candidate algorithm is refit on the entire training data in order to predict for an observation with predictors $\boldsymbol{X}_o$. Each resulting trained algorithm $\hat{Q}_l$ is then evaluated at $\boldsymbol{X}_o$, giving prediction $\hat{Y}_l^o$. The final prediction will be the $\hat{Y}_{l^*}^o$ in the discrete Super Learner, or $\sum_l a_l^* \hat{Y}_l^o$ in the full Super Learner. Figure 3.1 gives a visual representation of the Super Learner framework.

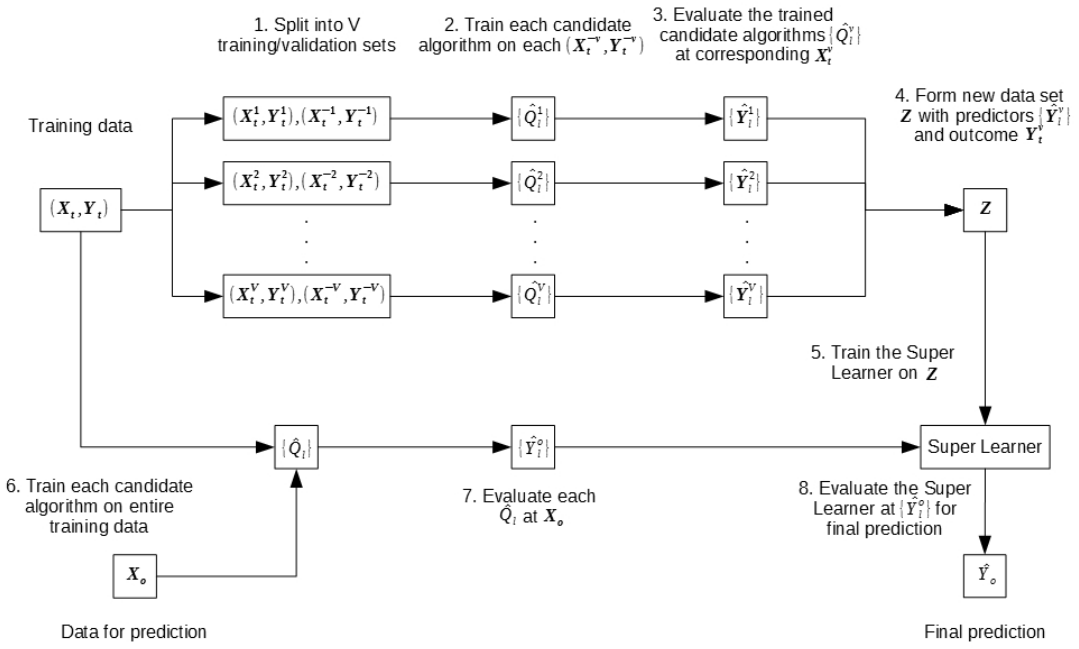

Figure 3.1: General schematic of the Super Learner framework

3.2.2 Procedure for Model Selection

Now, we introduce the procedure for mechanistic network model selection with the Super Learner framework. In this setting, the Super Learner will predict the model index based on training data generated from each candidate mechanistic network model. Before generating the training data, one needs care when choosing the parameter values for the candidate models so that it is plausible for the candidate models to generate the observed data to predict for, assuming that one of the candidate models is the true model. This can be done by choosing parameter values so that the data generated from the candidate models match the observed data based on some summary statistics. This ensures the generated data from each candidate model is at least similar to the observed data in some way. Once the parameter values are determined, one can form the training data by combining statistics computed for data generated from each candidate model as predictors with their corresponding true model index. Before going ahead with the Super Learner procedure, one needs to determine the

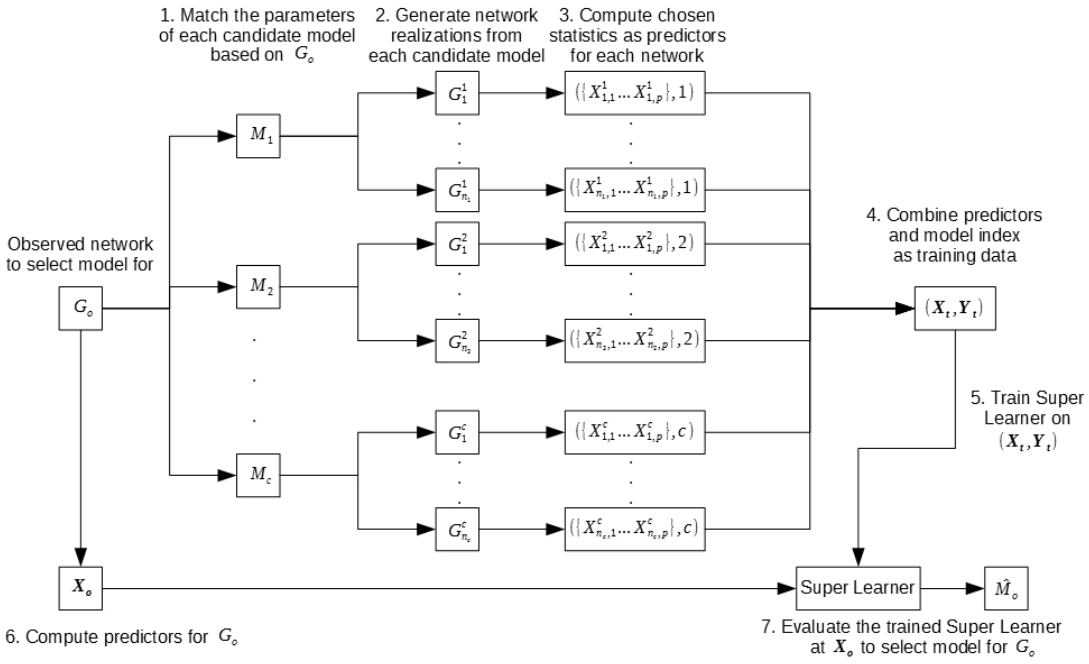

Figure 3.2: Schematic of model selection procedure with Super Learner

appropriate loss function for the setting. Since the prediction is for the model index, this is a classification problem, so a loss function like squared loss is no longer appropriate. Instead, we propose to use $1 - AUC$ as the loss function, where AUC is the area under the receiver operating characteristic curve. The AUC is an appropriate measure of the quality of the classification since it does not depend on the distribution of the model index in the data for performance evaluation. The corresponding loss function is also bounded, so the resulting Super Learner will retain the oracle property.

Algorithm 3.1 lays out the procedure we propose for model selection using Super Learner, and figure 3.2 gives a visual representation. In addition to selecting the candidate algorithms, one needs to select the all-important network summary statistics to both match on, in step 1, and to train the algorithms on, in steps 5 and 7. As previously stated in Robert et al. (2011), even if the sufficient statistics of all submodels are selected, they may not be jointly sufficient

Algorithm 3.1 Steps for the model selection with mechanistic network models via Super Learner

1. Match the parameters for all candidate models with the observed network based on relevant statistics
2. Select relevant statistics that highlight differences between models as predictors
3. ABC without the "B", generate training data for all models of interest
4. Split the training data into cross-validation sets
5. Train/evaluate all candidate algorithms on each training/validation pair based on selected predictors
6. Train Super Learner on the results from each candidate algorithm
7. Train each candidate algorithm on entire training data set
8. Classify/select model for observed network based on algorithms from steps 6 and 7

for the full model. Since it will be very difficult to obtain, sufficiency should not be the most important criterion for these statistics, but rather their ability to characterize the similarities and differences between the candidate models and thus their ability to discriminate among models. Suppose one is trying to select between a full model and one of its submodels that has one of the mechanisms of the full model turned off, then one needs to consider the characteristics of the network that the missing mechanism affects and those that it does not. The statistics chosen for matching the parameters in step 1 should reflect the characteristics that are unaffected by the missing mechanism. Conversely, those chosen as predictors in steps 5 and 7 should reflect the characteristics affected by the missing mechanism. The ability to characterize these similarities and differences will determine the performance of the algorithm and, thus, should guide the selection of the summary statistics. Though the candidate models we consider here are nested, they do not need to be in general to use this framework.

3.2.3 Model for Proof of Concept

As a proof of concept for this framework of model selection for mechanistic models, we introduce a simple mechanistic model to demonstrate its performance. The basis for the model is the classic Erdős–Rényi (ER) model (Erdős and Rényi, 1959). In the ER model, the number of nodes n is fixed, and there are two variants on how edges are placed in the graph. In one variant, sometimes called the $G(n, p)$ model, each of the $C(n, 2)$, n choose 2, possible edges are independent and included in the graph with probability p, so the number of edges in the graph has a binomial distribution. In the other variant, sometimes called the $G(n, m)$ model, the number of edges in the graph m is also fixed. In this case, the random graph has a uniform distribution over all $C(C(n, 2), m)$ possible graphs with n nodes and m edges.

Our model takes elements from both variants of the ER model. The model generates random graphs with a fixed number of nodes and edges just like the second variant of the ER model, but each edge is added one at a time with a certain probability akin to the first variant. At each step of graph generation, we select a pair of nodes that are not connected to one another uniformly from all such node pairs, and we connect the nodes with an edge with a given probability. This process is repeated until the requisite number of edges have been added. If the probability for adding each edge was always fixed, then this model would be the same as the second variant of the ER model. In our model, there is a base probability p_0 for edge placement, but two additional mechanisms are included to allow the varying of the probability. The first mechanism is triadic closure, where should the addition of the selected edge close a triangle, i.e., the selected edge with two additional existing edges form a complete subgraph between three nodes in the network, then the probability will be increased by p_1 over the base probability for adding the edge. We dub the second mechanism "triadic closure plus," where should the addition of the selected edge close more than one triangle, then the probability will be further increased by p_2 for each potentially closed triangle in

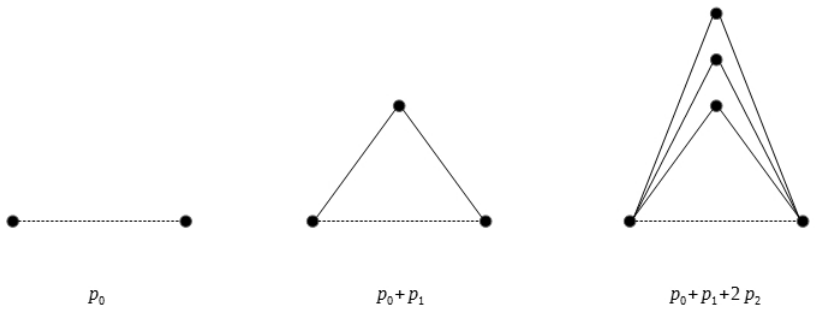

Figure 3.3: The probability to add an edge between two unconnected nodes in our model with two mechanisms when there are no closeable triangles, when there is one closeable triangle, and when there is more than one closeable triangle

excess of one. Figure 3.3 demonstrates these two mechanisms. Though this model is fairly simple, both its mechanisms can be motivated in social networks. Another example is a citation network, where a new paper that cites an existing paper A might also cite paper B if A also cites B. In a friendship network, the first mechanism corresponds to the idea that two people are more likely to become friends if they have a mutual friend, while the second mechanism further increases the likelihood for each mutual friend. As such, these mechanisms can also be related to the so-called weak ties hypothesis (Granovetter, 1973), and it has been shown, in large-scale communication networks, that a greater proportion of shared friends is associated with greater tie strength (Onnela et al., 2007).

3.3 Simulation Studies

We conduct simulation studies to assess the performance of the Super Learner framework for model selection. Super Learner will be used to select between the full model with both triadic mechanisms vs. the submodel with only the "standard" triadic closure mechanism. Networks generated from both models have 100 nodes and a base edge probability of $p_0 = 0.3$. The probability of edge placement increases by $p_1 = 0.1$ for the first triangle the edge would close and by p_2 for every subsequent triangle, and we varied the value of p_2 over the values 0.05,

0.03, 0.01, and 0.005. The number of edges is the statistic that is matched between the two models for a particular simulation and will be varied over values 500, 1000, and 2000. For a given number of edges, the variation of p_2 from 0.05 to 0.005 means that the differences caused by the additional mechanism will be more and more difficult to detect. For a given value of p_2, the variation of the number of edges from 500 to 2000 means that there will be more opportunities for the additional mechanism to manifest itself, making it easier to detect. The simulation studies will iterate through each combination of value of p_2 and edge number to see the interplay between the two.

Super Learner used for the simulations is composed of three candidate algorithms, k-nearest neighbors (KNN), support vector machine (SVM), and random forest (RF), which we discuss here briefly. These three are chosen largely for their ability to handle collinear predictors, which are often present in summary statistics for networks. Given an observed sample for classification, KNN determines which k (user-defined) samples in the training data are closest to the observed sample, based on some distance measure in the covariate space, a common choice being the Euclidean distance. The predicted class is the most frequent class among the k-nearest neighbors. Unlike KNN, which essentially formulates a new decision rule for each observation, SVM seeks to formulate a single decision rule for all classifications by separating the space of the predictors with a set of hyperplanes that segregates the space class-wise. Heuristically, a good hyperplane is one that is farthest from any sample in the training data. Once the class-wise segregation of the predictor space is complete, a new data point is classified based on the class label of the subset of the predictor space it falls in. Lastly, RF seeks to create a set of decision trees (the "forest") from the training data in order to arrive at the final prediction. To build each tree, a bootstrap sample of the training data is taken to form the root. Then, at each node, a subset of the predictors are selected, and a "best" split is determined for these predictors in order to form its daughter nodes. Typically, the quality of the split is measured by the amount of homogeneity in each daughter node.

Given an observed sample, each tree is traversed and gives a class label for the sample. An observation is then classified as the most frequent amongst all the tree-wise decisions.

Aside from the candidate algorithms, choosing appropriate predictors is an important task for the user. As discussed in the previous sections, sufficient summary statistics are difficult to obtain in all but the trivial mechanistic network models, and one should aim to use summary statistics that are likely able to characterize the differences between the candidate models. In these simulations, there are five summary statistics chosen as predictors. The first predictor is the triangle count, which is an obvious choice, since the additional mechanism in the full model will favor edges that close multiple triangles more so than those that only close one. The second is the average local clustering coefficient over all nodes. The local clustering coefficient of a node is a measure of how close its neighbors are to forming a complete graph by themselves, i.e. having every possible edge between any two neighbors. If the addition of an edge between nodes a and b will close multiple triangles, both a and b already share a set of neighbors with whom to form the potentially closed triangles. Then, without loss of generality, from the point of view of node a, the addition of the edge would mean the addition of a single neighbor, b, and the addition of multiple edges amongst its updated set of neighbors from b to those shared neighbors. In scenarios with lower total edge counts, where the degree of either a or b is likely to be smaller, this could lead to a potentially large change in the local clustering coefficient. Lastly, the additional mechanism is a rich-getting-richer scheme in terms of the degree of a node, since the more closeable triangles a pair of nodes have, the higher their existing degrees, which further leads to a higher probability of both getting an increase to their degrees with an additional edge. Thus, this mechanism is likely to affect the degree distribution in the network. As a proxy to the full degree distribution, the three quartiles (25%, 50%, 75%) of the degree distribution are included as predictors.

The mechanistic model proposed in the previous section is coded in Python and based on

the package NetworkX. The training of Super Learner, with a 5-fold cross-validation, is done with the R package SuperLearner, which contains wrappers for the chosen candidate algorithms. For a given combination of edge count and value for p_2, we generate 10,000 training samples from both the full model and the submodel as training data. Rather than using a separate sample to assess performance, the AUC of Super Learner as well as each candidate algorithm was estimated via a 10-fold cross-validation. Note that this is different from the cross-validation for Super Learner itself. For a given validation set of the 10 that the training data is first partitioned into, the 5-fold cross-validated Super Learner is trained on the remaining 9 validation sets, and then used to predict the model index for the given validation set. The AUC measure is computed for each of the 10 validation sets and averaged.

3.3.1 Simulation Results

The cross-validated estimate of the AUC for the full and discrete Super Learner, and each candidate algorithm for each scenario of the simulation studies are summarized in figure 3.4. In general, performance decreases as the value of p_2 decreases, which is no surprise as the effect of the additional mechanism diminishes as p_2 gets smaller. Performance also improves as the edge count increases, as the mechanism has more opportunities to manifest itself with more edges. In most scenarios, the discrete Super Learner has the same cross-validated AUC as that of the best performing candidate algorithm, with the full Super Learner performing a little better as expected. In these scenarios, the ordering of the performance of the candidate algorithms is likely the same across each fold in the cross-validation. Thus, the discrete Super Learner would always pick the same candidate algorithm in each fold and have the same performance as the best candidate algorithm averaged across all folds. The full Super Learner, in this case, would take a convex combination of the candidate algorithms in each fold and would perform at least as well as, and likely better than, the best performing candidate algorithm in each fold. When averaged across the cross-validation folds, the full Super Learner clearly performs better, as evidenced by the simulations. There are a few

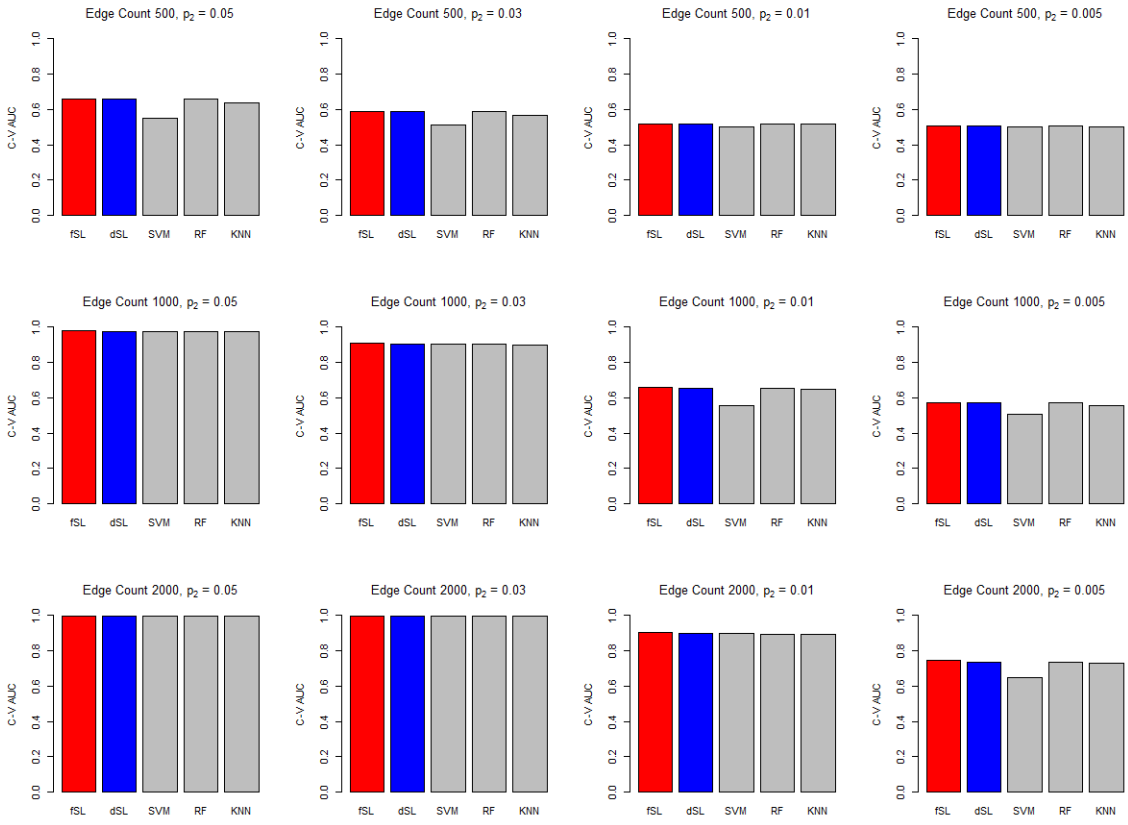

Figure 3.4: Cross-validated *AUC* for each method (full Super Learner, discrete Super Learner, support vector machine, random forest, k-nearest neighbors from left to right) in each simulation scenario

scenarios where either or both of the discrete and full Super Learner perform slightly worse than the best performing candidate algorithm. This occurs with edge count 500 and $p_2 = 0.01$ or 0.005, as well as with edge count 1000 and $p_2 = 0.01$. In each of these 3 scenarios, there are several candidate algorithms that are quite close in performance, and the ordering of their performance is likely not constant across the folds. In this case, the discrete Super Learner picks different candidate algorithms across the folds and the AUC averaged across the folds may be worse than that of the best candidate algorithm. The full Super Learner on the other hand weights different candidate algorithms most heavily across the folds, and the cross-validated AUC can also end up worse than the best candidate algorithm. This phenomenon is likely a finite sample issue, since the ordering of the performance of the candidate algorithms are likely to be constant across all folds in the limit.

The simulation results also seem to support the oracle properties of both the discrete and full Super Learner. In most scenarios, the discrete Super Learner has the same performance as the best candidate algorithm, and the full Super Learner performs slightly better. In the other scenarios, where either or both of the discrete and full Super Learner perform worse than the best candidate algorithm, they both still perform very close to the best candidate algorithm, with difference in AUC only of the order of 10^{-3}. Since the oracle properties for both Super Learners are asymptotic results, deviation from asymptotic behaviors would be expected for finite samples. The best performing candidate algorithm varies between support vector machine (SVM) and random forest (RF) across the scenarios, which is not something known a priori, but both versions of Super Learner are able to closely match or beat the best candidate algorithm. Although the differences between the best and worst performing candidate algorithms in the scenarios visited are not too big (biggest difference is about 0.1), it is still easy to see the manifestation of the oracle property, which would be even more valuable in scenarios with bigger differences. In cases when one is unlikely to fully grasp the significance of the various summary statistics as predictors or what learning

algorithm would work best due to the complex nature of the data, it would be ideal to consider multiple candidate algorithms paired with different combinations of predictors and find an optimal mix. This makes Super Learner well suited for model selection in mechanistic network models.

3.4 Discussion

We propose a procedure for model selection with mechanistic network models via the Super Learner framework. Due to the intractability of the likelihood of the typical mechanistic network models, likelihood-based model selection methods are not feasible. The Approximate Bayesian Computation (ABC) approach provides one viable means for model selection for mechanistic network models. However, an accurate ABC posterior requires one to build it from sufficient statistics, which is typically difficult to find in the case of intractable likelihoods. The lack of sufficiency and approximations required to arrive at the ABC posterior can mean the posterior distribution of the model index is inaccurate. In addition, the concatenation of the sufficient statistics of each submodel is not necessarily sufficient for the joint model. This can lead to a lack of correspondence between the ABC Bayes factor and the true Bayes factor. Rather than relying on ABC to approximate the Bayes factor for model selection, which suffers from the lack of sufficiency, we propose to use Super Learner for model selection while borrowing the generation of pseudo-data from ABC.

With training data readily generated from each candidate model, Super Learner seeks to build an optimal algorithm from a host of candidate algorithms. In this case, it seeks to build an optimal classifier from candidate algorithms to best discriminate between the candidate models with the given predictors, which are not necessarily sufficient. One is unlikely to know what classifiers paired with what predictors will perform well, but with Super Learner, one does not need to make this choice as Super Learner will try to build the optimal classifier with all that is given. However, this does not mean that the quality of the

predictors does not matter. The better the predictors are at characterizing the differences between the candidate models, the better Super Learner performs. Though the ability to characterize the differences likely correlates with sufficiency, sufficiency in and of itself should not be the criterion for choosing the predictors. Here, one can apply domain knowledge to select predictors.

The main difficulty of the proposed approach is that one needs to be sure that the candidate models can plausibly generate the observed data, assuming one of the candidate models is the true model. To do so, one needs to consider what characteristics of the data are the differences in the candidate models unlikely to affect, then set the parameters of each candidate model to match these characteristics of the observed data. This will hopefully allow the differences in the candidate models to more clearly manifest themselves and ensure that the data generated from each candidate model are similar to the observed data in some aspects. Ideally, this also means that the data generated from the true model match the observed data closely. However, there is no guarantee to this. In the current state, this part of the procedure is more an art than an algorithm.

The most immediate step for future work would be to make the matching of the parameters of the candidate models more concrete. Should this step be done incorrectly, one faces the danger of choosing the wrong model completely, even if true model is included among the candidates. Another future direction would be to assess the uncertainty in the results.

Network models see wide use in many domains, and mechanistic models allow one to easily incorporate domain knowledge. However, there is little work up to date on model selection for mechanistic network models. We propose a procedure that makes use of the Super Learner framework and leverages the ease of generating data from mechanistic models as a first step in model selection for mechanistic network models.

References

1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012.

D. J. Balding and R. A. Nichols. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. In *Human identification: The use of DNA markers*, pages 3–12. Springer, 1995.

A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 286(5439): 509–512, 1999.

I. Barnett, R. Mukherjee, and X. Lin. The generalized higher criticism for testing snp-set effects in genetic association studies. *Journal of the American Statistical Association*, (just-accepted), 2016.

I. J. Barnett and X. Lin. Analytic p-value calculation for the higher criticism test in finite d problems. *Biometrika*, 101(4):964, 2014.

A. P. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.

R. Carroll, M. Gail, and J. Lubin. Case-control studies with errors in covariates. *Journal of the American Statistical Association*, 88(421):185–199, 1993.

R. J. Carroll, D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu. *Measurement error in nonlinear models: a modern perspective*. CRC press, 2006.

K. N. Conneely and M. Boehnke. So many correlated tests, so little time! rapid adjustment of p values for multiple correlated tests. *The American Journal of Human Genetics*, 81 (6):1158–1168, 2007.

A. Derkach, T. Chiang, J. Gong, L. Addis, S. Dobbins, I. Tomlinson, R. Houlston, D. K. Pal, and L. J. Strug. Association analysis using next generation sequence data from publicly available control groups: The robust variance score statistic. *Bioinformatics*, page btu196, 2014.

D. Donoho and J. Jin. Higher criticism for detecting sparse heterogeneous mixtures. *Annals of Statistics*, pages 962–994, 2004.

S. Dudoit and M. J. van der Laan. Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Statistical Methodology*, 2(2):131–154, 2005.

P. Erdős and A. Rényi. On random graphs i. *Publ. Math. Debrecen*, 6:290–297, 1959.

D. B. Goldstein, A. Allen, J. Keebler, E. H. Margulies, S. Petrou, S. Petrovski, and S. Sunyaev. Sequencing studies in human genetics: design and interpretation. *Nature Reviews Genetics*, 14(7):460–470, 2013.

R. Goyal and J.-P. Onnela. Mechanistic and probabilistic network models. *In progress*, 2017.

R. Goyal, J. Blitzstein, and V. De Gruttola. Sampling networks from their posterior predictive distribution. *Network Science*, 2(01):107–131, 2014.

M. S. Granovetter. The strength of weak ties. *American journal of sociology*, 78(6):1360–1380, 1973.

P. Hall and J. Jin. Innovated higher criticism for detecting sparse signals in correlated noise. *The Annals of Statistics*, 38(3):1686–1732, 2010.

P. D. Hoff, A. E. Raftery, and M. S. Handcock. Latent space approaches to social network analysis. *Journal of the american Statistical association*, 97(460):1090–1098, 2002.

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

S. Y. Kim, K. E. Lohmueller, A. Albrechtsen, Y. Li, T. Korneliussen, G. Tian, N. Grarup, T. Jiang, G. Andersen, D. Witte, et al. Estimation of allele frequency and association mapping using next-generation sequencing data. *BMC bioinformatics*, 12(1):231, 2011.

K. Klemm and V. M. Eguiluz. Highly clustered scale-free networks. *Physical Review E*, 65 (3):036123, 2002.

J. M. Kumpula, J.-P. Onnela, J. Saramäki, K. Kaski, and J. Kertész. Emergence of communities in weighted networks. *Physical review letters*, 99(22):228701, 2007.

S. Lee, M. J. Emond, M. J. Bamshad, K. C. Barnes, M. J. Rieder, D. A. Nickerson, E. L. P. Team, D. C. Christiani, M. M. Wurfel, X. Lin, et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *The American Journal of Human Genetics*, 91(2):224–237, 2012.

S. Lee, G. R. Abecasis, M. Boehnke, and X. Lin. Rare-variant association analysis: study designs and statistical tests. *The American Journal of Human Genetics*, 95(1):5–23, 2014.

B. Li and S. M. Leal. 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, 2008.

C. X. Ling, J. Huang, and H. Zhang. Auc: a statistically consistent and more discriminating measure than accuracy. In *IJCAI*, volume 3, pages 519–524, 2003.

J. Lintusaari, M. U. Gutmann, R. Dutta, S. Kaski, and J. Corander. Fundamentals and recent developments in approximate bayesian computation. *Systematic Biology*, page syw077, 2016.

D. J. Liu and S. M. Leal. Seqchip: a powerful method to integrate sequence and genotype data for the detection of rare variant associations. *Bioinformatics*, 28(13):1745–1751, 2012.

X. Liu and K.-Y. Liang. Adjustment for non-differential misclassification error in the generalized linear model. *Statistics in Medicine*, 10(8):1197–1211, 1991.

J. A. Longmate, G. P. Larson, T. G. Krontiris, and S. S. Sommer. Three ways of combining genotyping and resequencing in case-control association studies. *PloS one*, 5(12):e14318, 2010.

D. Lusher, J. Koskinen, and G. Robins. *Exponential random graph models for social networks: Theory, methods, and applications*. Cambridge University Press, 2012.

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

J.-M. Marin, P. Pudlo, C. P. Robert, and R. J. Ryder. Approximate bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180, 2012. ISSN 1573-1375. doi: 10.1007/s11222-011-9288-2. URL `http://dx.doi.org/10.1007/s11222-011-9288-2`.

A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, et al. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*, 20(9): 1297–1303, 2010.

M. Middendorf, E. Ziv, and C. H. Wiggins. Inferring network mechanisms: the drosophila melanogaster protein interaction network. *Proceedings of the National Academy of Sciences of the United States of America*, 102(9):3192–3197, 2005.

S. Morgenthaler and W. G. Thilly. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (cast). *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 615(1):28–56, 2007.

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

V. Moskvina and K. M. Schmidt. On multiple-testing correction in genome-wide association studies. *Genetic epidemiology*, 32(6):567–573, 2008.

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

M. Newman. Networks: an introduction, 2010.

R. Nielsen, J. S. Paul, A. Albrechtsen, and Y. S. Song. Genotype and snp calling from next-generation sequencing data. *Nature Reviews Genetics*, 12(6):443–451, 2011.

J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.-L. Barabási. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, 104(18):7332–7336, 2007.

R. Pastor-Satorras and A. Vespignani. *Evolution and structure of the Internet: A statistical physics approach*. Cambridge University Press, 2007.

E. C. Polley, S. Rose, and M. J. Van der Laan. Super learning. In *Targeted Learning*, pages 43–66. Springer, 2011.

D. Prangle, P. Fearnhead, M. P. Cox, P. J. Biggs, and N. P. French. Semi-automatic selection of summary statistics for abc model choice. *Statistical applications in genetics and molecular biology*, 13(1):67–82, 2014.

R. L. Prentice and R. Pyke. Logistic disease incidence models and case-control studies. *Biometrika*, pages 403–411, 1979.

P. Pudlo, J.-M. Marin, A. Estoup, J.-M. Cornuet, M. Gautier, and C. P. Robert. Reliable abc model choice via random forests. *Bioinformatics*, page btv684, 2015.

A. Raval and A. Ray. *Introduction to biological networks*. CRC Press, 2013.

C. P. Robert, J.-M. Cornuet, J.-M. Marin, and N. S. Pillai. Lack of confidence in approximate bayesian computation model choice. *Proceedings of the National Academy of Sciences*, 108 (37):15112–15117, 2011.

G. Robins, P. Pattison, Y. Kalish, and D. Lusher. An introduction to exponential random graph (p*) models for social networks. *Social networks*, 29(2):173–191, 2007.

S. Romeo, W. Yin, J. Kozlitina, L. A. Pennacchio, E. Boerwinkle, H. H. Hobbs, and J. C. Cohen. Rare loss-of-function mutations in angptl family members contribute to plasma triglyceride levels in humans. *The Journal of clinical investigation*, 119(1):70–79, 2009.

B. Rosner, W. Willett, and D. Spiegelman. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Statistics in medicine*, 8(9):1051–1069, 1989.

S. Sanna, B. Li, A. Mulas, C. Sidore, H. M. Kang, A. U. Jackson, M. G. Piras, G. Usala, G. Maninchedda, A. Sassu, et al. Fine mapping of five loci associated with low-density lipoprotein cholesterol detects variants that double the explained heritability. *PLoS Genet*, 7(7):e1002198, 2011.

A. Schwartzman and X. Lin. The effect of correlation in false discovery rate estimation. *Biometrika*, 98(1):199–214, 2011.

M. A. Seibold, A. L. Wise, M. C. Speer, M. P. Steele, K. K. Brown, J. E. Loyd, T. E. Fingerlin, W. Zhang, G. Gudmundsson, S. D. Groshong, et al. A common muc5b promoter polymorphism and pulmonary fibrosis. *New England Journal of Medicine*, 364(16):1503–1512, 2011.

C. R. Shalizi and A. Rinaldo. Consistency under sampling of exponential random graph models. *Annals of statistics*, 41(2):508, 2013.

L. Skotte, T. S. Korneliussen, and A. Albrechtsen. Association testing for next-generation sequencing data using score statistics. *Genetic epidemiology*, 36(5):430–437, 2012.

R. V. Solé, R. Pastor-Satorras, E. Smith, and T. B. Kepler. A model of large-scale proteome evolution. *Advances in Complex Systems*, 5(01):43–54, 2002.

D. Spiegelman, B. Rosner, and R. Logan. Estimation and inference for logistic regression with covariate misclassification and measurement error in main study/validation study designs. *Journal of the American Statistical Association*, 95(449):51–61, 2000.

M. Sunnåker, A. G. Busetto, E. Numminen, J. Corander, M. Foll, and C. Dessimoz. Approximate bayesian computation. *PLoS Comput Biol*, 9(1):e1002803, 2013.

M. J. Van Der Laan and S. Dudoit. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples. *Technical report, Division of Biostatistics, University of California, Berkeley*, 2003.

M. J. Van der Laan, E. C. Polley, and A. E. Hubbard. Super learner. *Statistical applications in genetics and molecular biology*, 6(1), 2007.

A. Vázquez, A. Flammini, A. Maritan, and A. Vespignani. Modeling of protein interaction networks. *Complexus*, 1(1):38–44, 2003.

S. Wasserman and K. Faust. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *nature*, 393 (6684):440–442, 1998.

M. C. Wu, P. Kraft, M. P. Epstein, D. M. Taylor, S. J. Chanock, D. J. Hunter, and X. Lin. Powerful snp-set analysis for case-control genome-wide association studies. *The American Journal of Human Genetics*, 86(6):929–942, 2010.

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

Y. Zhang and J. S. Liu. Fast and accurate approximation to significance tests in genome-wide association studies. *Journal of the American Statistical Association*, 106(495):846–857, 2011.

A Supplementary Material to Chapter 1

A.1 Simulation Plots and Results

Table A.1: Parameter values used for simulations. The first column is the corresponding plot set; the second and third are the number of cases and controls; the fourth is the minor allele frequency; the fifth and sixth are the mean and variance used to generate the read depth for cases and controls; the seventh and eighth are the mean and variance used to generate the misclassification rate for cases and controls; the ninth and tenth are the intercept and effect size used for the size simulation

Plot	n_{case}	$n_{control}$	p	rd_{case}	$rd_{control}$	e_{case}	$e_{control}$	β_0	β_1
1	450	450	0.2	(8,1)	(6,1)	(0.005,0.0001)	(0.01,0.001)	$logit(0.2)$	0
2	450	450	0.2	(20,2)	(3,1)	(0.005,0.0001)	(0.05,0.001)	$logit(0.2)$	0
3	1500	1500	0.01	(20,2)	(3,1)	(0.005,0.0001)	(0.01,0.001)	$logit(0.2)$	0
4	450	450	0.2	(20,2)	(16,1)	(0.005,0.0001)	(0.05,0.001)	$logit(0.2)$	0
5	300	600	0.2	(20,2)	(3,1)	(0.005,0.0001)	(0.05,0.001)	$logit(0.2)$	0

A.1.1 Plot Set 1

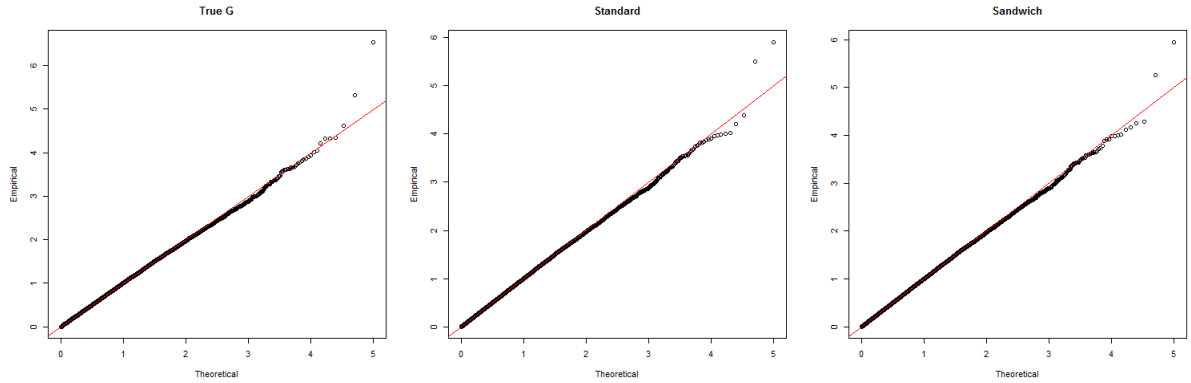

Figure A.1: QQ plots for the size simulation for analysis with the true genotype, analysis with RC and naive variance, and analysis with RC and sandwich variance

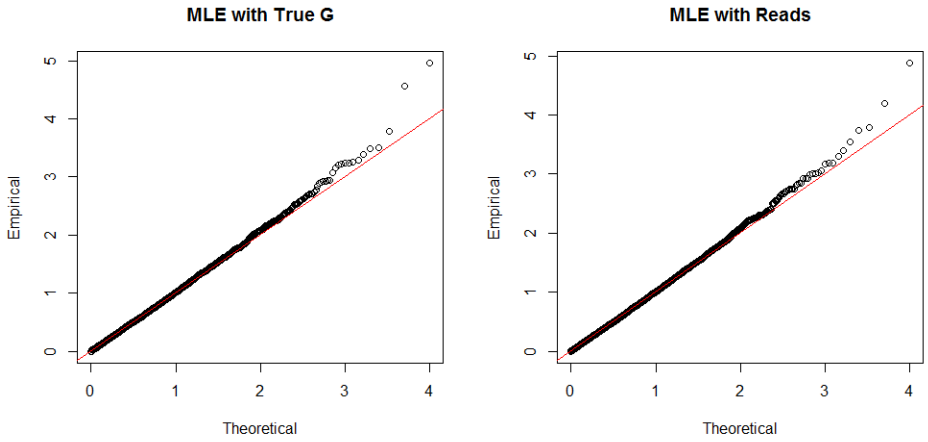

Figure A.2: QQ plots for the size simulation for analysis with the true genotype, analysis with ML

Table A.2: Power for analysis with the true genotype, analysis with RC and naive variance, analysis with RC and sandwich variance, and analysis with ML at the given effect size and various significance levels, as well as bias for RC and ML estimates for the intercept and effect size

Power $\beta_1 = 0.3$	$\alpha = 0.05$	0.01	0.001	10^{-4}
True G	0.73696	0.50322	0.23423	0.09049
RC and naive	0.71547	0.47664	0.21578	0.07913
RC and sandwich	0.71748	0.48006	0.21931	0.08253
ML	0.7082	0.469	0.2108	0.0781

Bias	$\hat{\beta}_0$	$\hat{\beta}_1$
RC	-0.1280	0.2998
ML	-0.1277	0.3000

A.1.2 Plot Set 2

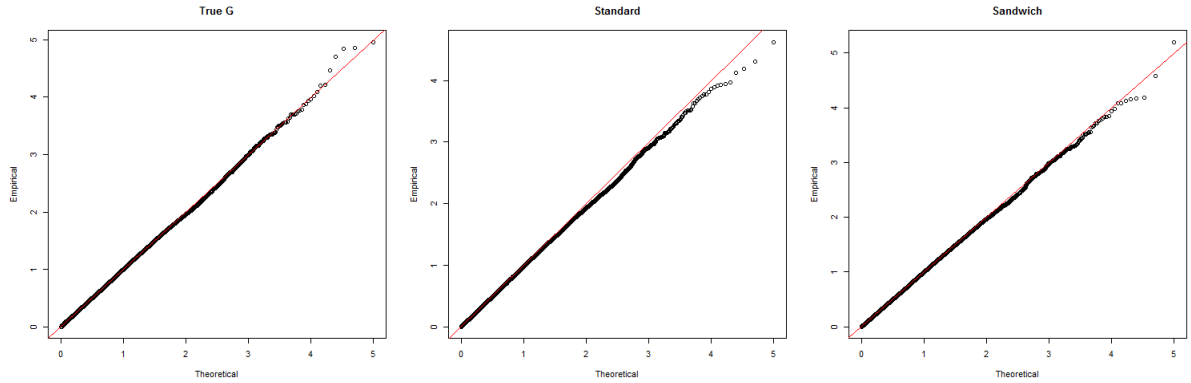

Figure A.3: QQ plots for the size simulation for analysis with the true genotype, analysis with RC and naive variance, and analysis with RC and sandwich variance

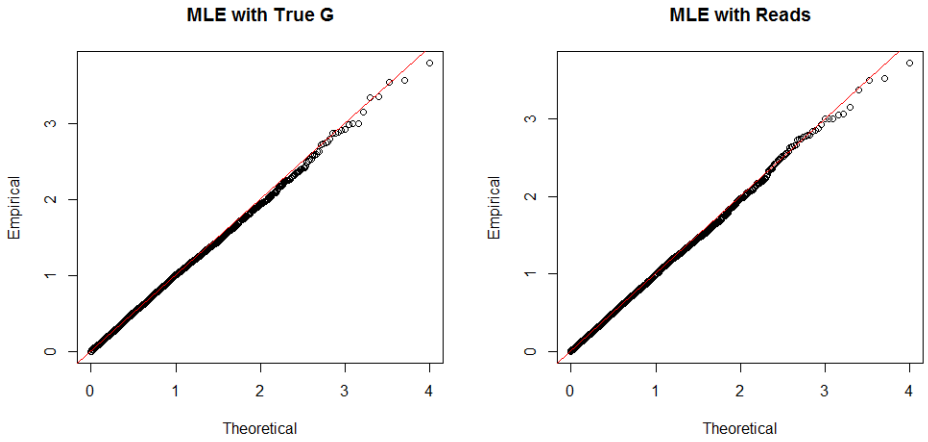

Figure A.4: QQ plots for the size simulation for analysis with the true genotype, analysis with ML

Table A.3: Power for analysis with the true genotype, analysis with RC and naive variance, analysis with RC and sandwich variance, and analysis with ML at the given effect size and various significance levels, as well as bias for RC and ML estimates for the intercept and effect size

Power $\beta_1 = 0.3$	$\alpha = 0.05$	0.01	0.001	10^{-4}
True G	0.73778	0.50354	0.2353	0.09068
RC and naive	0.63896	0.38802	0.15118	0.04852
RC and sandwich	0.65902	0.4212	0.1822	0.0651
ML	0.6457	0.3995	0.1547	0.0465

Bias	$\hat{\beta}_0$	$\hat{\beta}_1$
RC	-0.1247	0.2869
ML	-0.1281	0.3024

A.1.3 Plot Set 3

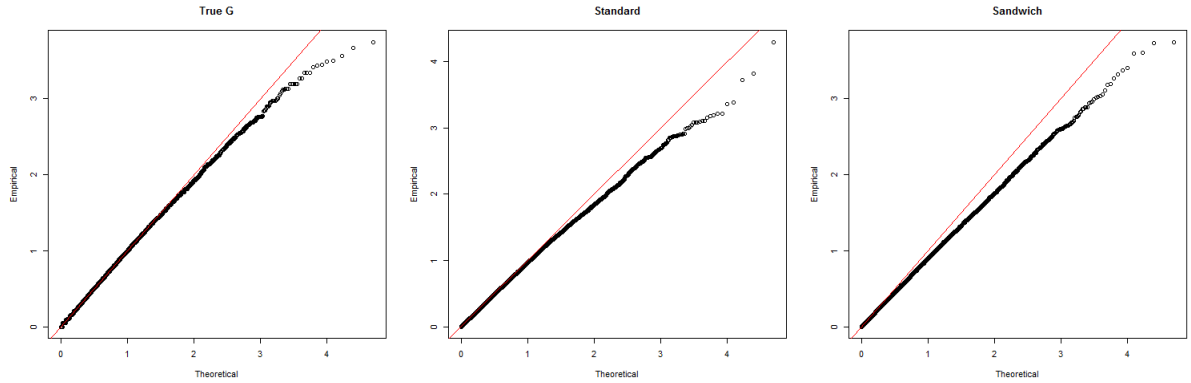

Figure A.5: QQ plots for the size simulation for analysis with the true genotype, analysis with RC and naive variance, and analysis with RC and sandwich variance

Table A.4: Power for analysis with the true genotype, analysis with RC and naive variance, and analysis with RC and sandwich variance at the given effect size and various significance levels, as well as bias for RC estimates for the intercept and effect size

Power $\beta_1 = 0.4$	$\alpha = 0.05$	0.01	0.001	10^{-4}
True G	0.36158	0.15376	0.03366	0.0051
RC and naive	0.26008	0.08596	0.01156	0.0006
RC and sandwich	0.24088	0.0851	0.01884	0.00428

Bias	$\hat{\beta}_0$	$\hat{\beta}_1$
RC	-0.0085	0.3670

A.1.4 Plot Set 4

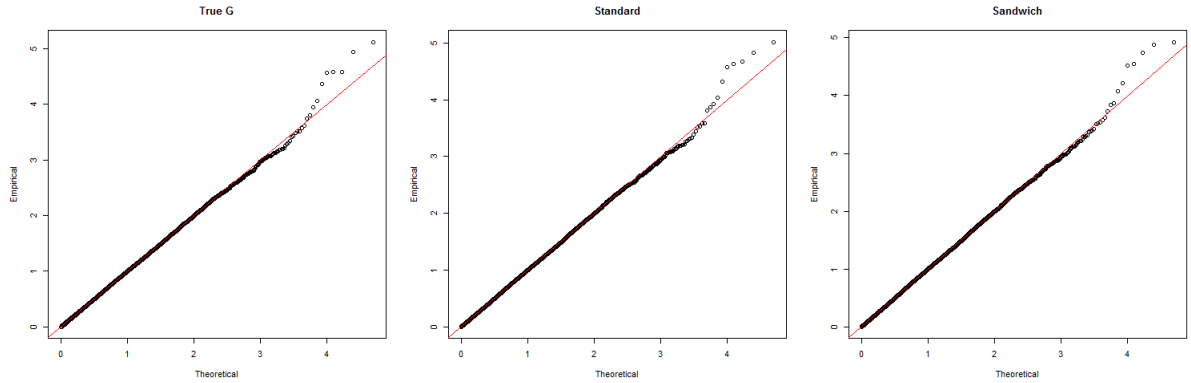

Figure A.6: QQ plots for the size simulation for analysis with the true genotype, analysis with RC and naive variance, and analysis with RC and sandwich variance

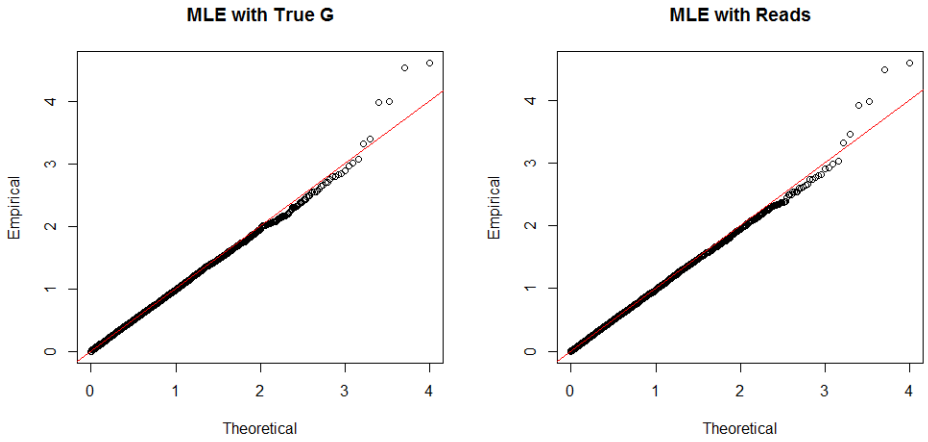

Figure A.7: QQ plots for the size simulation for analysis with the true genotype, analysis with ML

Table A.5: Power for analysis with the true genotype, analysis with RC and naive variance, analysis with RC and sandwich variance, and analysis with ML at the given effect size and various significance levels, as well as bias for RC and ML estimates for the intercept and effect size

Power $\beta_1 = 0.3$	$\alpha = 0.05$	0.01	0.001	10^{-4}
True G	0.7369	0.49962	0.23096	0.08806
RC and naive	0.73422	0.497	0.22932	0.08696
RC and sandwich	0.73442	0.49768	0.2299	0.08748
ML	0.7371	0.5022	0.2315	0.0854

Bias	$\hat{\beta}_0$	$\hat{\beta}_1$
RC	-0.1279	0.3001
ML	-0.1284	0.3015

A.1.5 Plot Set 5

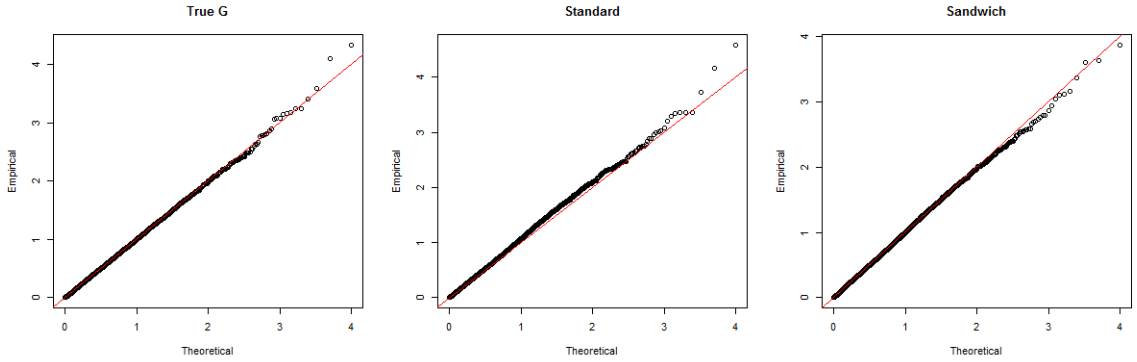

Figure A.8: QQ plots for the size simulation for analysis with the true genotype, analysis with RC and naive variance, and analysis with RC and sandwich variance

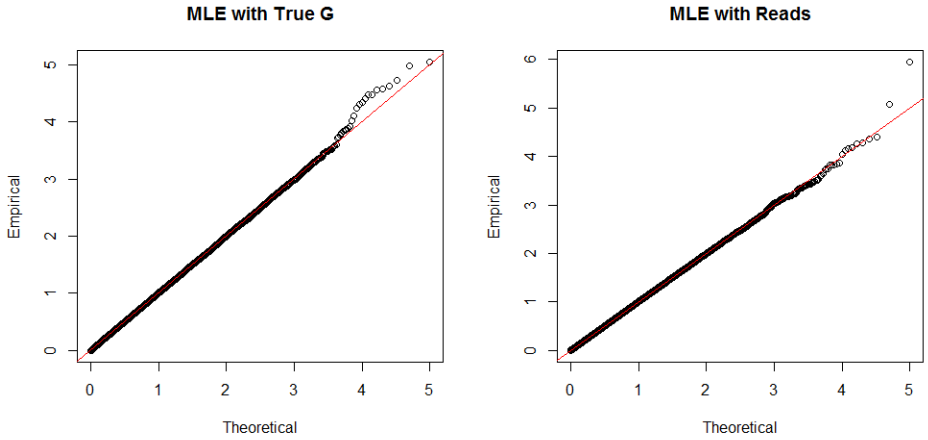

Figure A.9: QQ plots for the size simulation for analysis with the true genotype, analysis with ML

Table A.6: Power for analysis with the true genotype, analysis with RC and sandwich variance, and analysis with ML at the given effect size and various significance levels, as well as bias for RC and ML estimates for the intercept and effect size

Power $\beta_1 = 0.3$	$\alpha = 0.05$	0.01	0.001	10^{-4}
True G	0.6928	0.4503	0.2036	0.0771
RC and naive	NA	NA	NA	NA
RC and sandwich	0.6387	0.4055	0.1737	0.0625
ML	0.63042	0.38347	0.15117	0.04918

Bias	$\hat{\beta}_0$	$\hat{\beta}_1$
RC	-0.8310	0.3170
ML	-0.8218	0.3010

A.1.6 Plot Set 6

Table A.7: Parameter values used for size simulation with a binary confounder. The columns are the same as in table A.1, except the third and fourth columns now denote the mean of the binary confounder when $G = 0$ or 1 and when $G = 2$; the last column is the effect size of the confounder

n_{case}	$n_{control}$	p	p_{01}^x	p_2^x	rd_{case}	rd_{con}	e_{case}	e_{con}	β_0	β_1	β_2
300	600	0.2	0.5	0.6	(8,1)	(6,1)	(0.005,0.0001)	(0.05,0.001)	$logit(0.2)$	0	0.2

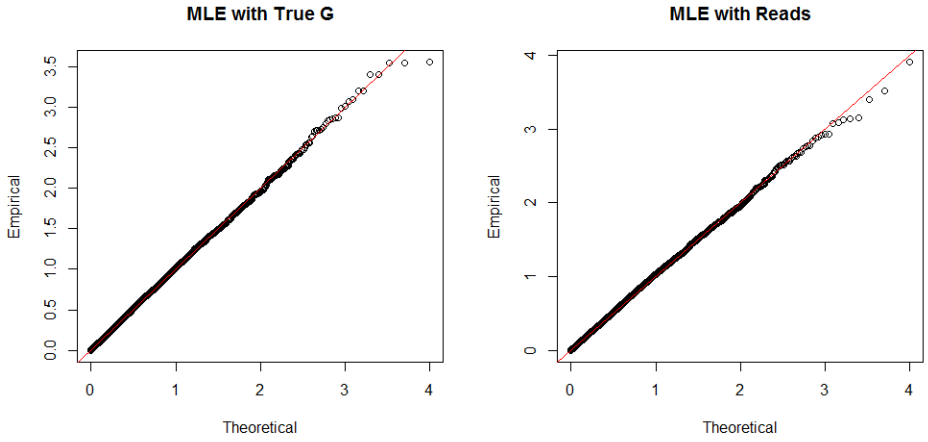

Figure A.10: QQ plots for the size simulation with a binary confounder for analysis with the true genotype, analysis with ML

Table A.8: Power for analysis with the true genotype, and analysis with ML at the given effect size and various significance levels, as well as bias for ML estimates for the intercept and effect size

Power $\beta_1 = 0.3$	$\alpha = 0.05$	0.01	0.001	10^{-4}
True G	0.6951	0.4614	0.2116	0.0802
ML	0.6625	0.4211	0.184	0.0644

Bias	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
ML	-0.9268	0.3015	0.2000

A.2 Unbiasedness of the Estimating Equation for RC

In order to show the estimating equations for $\hat{\boldsymbol{\beta}}_{RC}$ is unbiased when evaluated at $\beta_{0E} = \text{logit}\left(P\left(Y_i = 1|S_i = 1\right)\right)$ and $\beta_{1E} = 0$ under the null, we will simply show the unbiasedness of the contributions from one subject:

$$\begin{aligned}
E_{Y_i,\tilde{\boldsymbol{G}}_i,d_i,\pi_i|S_i=1}\left[\psi^{(i)}_{\beta_{0E}}\right] &= E_{Y_i,\tilde{\boldsymbol{G}}_i,d_i,\pi_i|S_i=1}\left[Y_i - \text{expit}\left(\beta_{0E} + \beta_{1E}E\left[G_i|\tilde{\boldsymbol{G}}_i,d_i,\pi_i\right]\right)\right] \\
&= E_{Y_i,\tilde{\boldsymbol{G}}_i,d_i,\pi_i|S_i=1}\left[Y_i - P\left(Y_i = 1|S_i = 1\right)\right] \\
&= E_{Y_i|S_i=1}\left[Y_i - P\left(Y_i = 1|S_i = 1\right)\right] \\
&= 0
\end{aligned}$$

$$\begin{aligned}
& E_{Y_i,\tilde{\boldsymbol{G}}_i,d_i,\pi_i|S_i=1}\left[\psi^{(i)}_{\beta_{1E}}\right] \\
=& E_{Y_i,\tilde{\boldsymbol{G}}_i,d_i,\pi_i|S_i=1}\left[E\left[G_i|\tilde{\boldsymbol{G}}_i,d_i,\pi_i\right]\left[Y_i - \text{expit}\left(\beta_{0E} + \beta_{1E}E\left[G_i|\tilde{\boldsymbol{G}}_i,d_i,\pi_i\right]\right)\right]\right] \\
=& E_{Y_i,\tilde{\boldsymbol{G}}_i,d_i,\pi_i|S_i=1}\left[E\left[G_i|\tilde{\boldsymbol{G}}_i,d_i,\pi_i\right]\left[Y_i - P\left(Y_i = 1|S_i = 1\right)\right]\right] \\
=& E_{Y_i|S_i=1}\left[E_{\tilde{\boldsymbol{G}}_i,d_i,\pi_i|Y_i,S_i=1}\left[E\left[G_i|\tilde{\boldsymbol{G}}_i,d_i,\pi_i\right]\left[Y_i - P\left(Y_i = 1|S_i = 1\right)\right]\right]\right] \\
=& E_{Y_i|S_i=1}\left[E_{\tilde{\boldsymbol{G}}_i,d_i,\pi_i|Y_i,S_i=1}\left[E\left[G_i|\tilde{\boldsymbol{G}}_i,d_i,\pi_i\right]\right] \times \left[Y_i - P\left(Y_i = 1|S_i = 1\right)\right]\right] \\
(1)\,(2) =& E_{Y_i|S_i=1}\left[E_{\tilde{\boldsymbol{G}}_i,d_i,\pi_i|Y_i}\left[E\left[G_i|\tilde{\boldsymbol{G}}_i,d_i,\pi_i,Y_i\right]\right] \times \left[Y_i - P\left(Y_i = 1|S_i = 1\right)\right]\right] \\
=& E_{Y_i|S_i=1}\left[E\left[G_i|Y_i\right] \times \left[Y_i - P\left(Y_i = 1|S_i = 1\right)\right]\right] \\
(1) =& E\left[G_i\right] \times E_{Y_i|S_i=1}\left[Y_i - P\left(Y_i = 1|S_i = 1\right)\right] \\
=& 0
\end{aligned}$$

(1) since $G \perp Y$ under the null

(2) we can drop $S_i = 1$ from the conditioning since S_i only depends Y_i, so once we condition on Y_i, S_i is independent of everything else

A.3 Show Naive Variance Overestimates the True Variance for RC with Balanced Sampling

Define matrices:

$$\boldsymbol{D} = \begin{bmatrix} X & 0 \\ 0 & X \end{bmatrix} \qquad \text{(some block diagonal matrix)}$$

$$\boldsymbol{V} = \begin{bmatrix} 0 & X \\ 0 & 0 \end{bmatrix} \qquad \text{(only top right block is nonzero)}$$

$$\boldsymbol{B} = \boldsymbol{D} + \boldsymbol{V} \qquad \text{(bun matrix)}$$

$$= \boldsymbol{D} + \boldsymbol{I}\boldsymbol{I}\boldsymbol{V}$$

$$\boldsymbol{B}^{-1} = \boldsymbol{D}^{-1} - \boldsymbol{D}^{-1}\left(\boldsymbol{I} + \boldsymbol{V}\boldsymbol{D}^{-1}\right)^{-1}\boldsymbol{V}\boldsymbol{D}^{-1}$$

$$\boldsymbol{M} = \boldsymbol{D} + \boldsymbol{V} + \boldsymbol{V}^T \qquad \text{(meat matrix)}$$

Then:

$$
\begin{aligned}
&\boldsymbol{B}^{-1}\boldsymbol{M}\left(\boldsymbol{B}^{-1}\right)^{T} \\
=&\left[\underbrace{\boldsymbol{D}^{-1}}_{A_1}-\underbrace{\boldsymbol{D}^{-1}\left(\boldsymbol{I}+\boldsymbol{V}\boldsymbol{D}^{-1}\right)^{-1}\boldsymbol{V}\boldsymbol{D}^{-1}}_{A_2}\right]\times\left[\underbrace{\boldsymbol{D}}_{B_1}+\underbrace{\boldsymbol{V}+\boldsymbol{V}^{T}}_{B_2}\right] \\
&\times\left[\underbrace{\boldsymbol{D}^{-1}}_{C_1}-\underbrace{\boldsymbol{D}^{-1}\boldsymbol{V}^{T}\left(\boldsymbol{I}+\boldsymbol{D}^{-1}\boldsymbol{V}^{T}\right)^{-1}\boldsymbol{D}^{-1}}_{C_2}\right]
\end{aligned}
$$

$$
\begin{aligned}
&\left(\boldsymbol{I}+\boldsymbol{D}^{-1}\boldsymbol{V}^{T}\right)^{-1} \\
=&\boldsymbol{I}-\left(\boldsymbol{D}+\boldsymbol{V}^{T}\right)^{-1}\boldsymbol{V}^{T}
\end{aligned}
$$

Note, can show $\left(\boldsymbol{D}+\boldsymbol{V}^{T}\right)^{-1}$ is of form $\begin{bmatrix} X & 0 \\ X & X \end{bmatrix}$ and $(\boldsymbol{D}+\boldsymbol{V})^{-1}$ is of form $\begin{bmatrix} X & X \\ 0 & X \end{bmatrix}$ with blockwise inversion.

$$
\begin{aligned}
A_1B_1C_1 &= \boldsymbol{D}^{-1}\boldsymbol{D}\boldsymbol{D}^{-1} \\
&= \boldsymbol{D}^{-1}
\end{aligned}
$$

$$
\begin{aligned}
A_1B_1C_2 &= \boldsymbol{D}^{-1}\boldsymbol{D}\boldsymbol{D}^{-1}\boldsymbol{V}^T\left(\boldsymbol{I}+\boldsymbol{D}^{-1}\boldsymbol{V}^T\right)^{-1}\boldsymbol{D}^{-1} \\
&= \boldsymbol{D}^{-1}\boldsymbol{V}^T\left(\boldsymbol{I}+\boldsymbol{D}^{-1}\boldsymbol{V}^T\right)^{-1}\boldsymbol{D}^{-1} \\
&= \boldsymbol{D}^{-1}\boldsymbol{V}^T\left(\boldsymbol{I}-\left(\boldsymbol{D}+\boldsymbol{V}^T\right)^{-1}\boldsymbol{V}^T\right)\boldsymbol{D}^{-1} \\
&= \boldsymbol{D}^{-1}\boldsymbol{V}^T\boldsymbol{D}^{-1}-\underbrace{\boldsymbol{D}^{-1}\boldsymbol{V}^T\left(\boldsymbol{D}+\boldsymbol{V}^T\right)^{-1}\boldsymbol{V}^T\boldsymbol{D}^{-1}}_{=\mathbf{0}} \\
&= \boldsymbol{D}^{-1}\boldsymbol{V}^T\boldsymbol{D}^{-1}
\end{aligned}
$$

$$
\begin{aligned}
A_1B_2C_1 &= \boldsymbol{D}^{-1}\left(\boldsymbol{V}+\boldsymbol{V}^T\right)\boldsymbol{D}^{-1} \\
&= \boldsymbol{D}^{-1}\boldsymbol{V}\boldsymbol{D}^{-1}+\boldsymbol{D}^{-1}\boldsymbol{V}^T\boldsymbol{D}^{-1}
\end{aligned}
$$

$$
\begin{aligned}
A_1B_2C_2 &= \boldsymbol{D}^{-1}\left(\boldsymbol{V}+\boldsymbol{V}^T\right)\boldsymbol{D}^{-1}\boldsymbol{V}^T\left(\boldsymbol{I}+\boldsymbol{D}^{-1}\boldsymbol{V}^T\right)^{-1}\boldsymbol{D}^{-1} \\
&= \underbrace{\boldsymbol{D}^{-1}\boldsymbol{V}^T\boldsymbol{D}^{-1}\boldsymbol{V}^T\left(\boldsymbol{I}+\boldsymbol{D}^{-1}\boldsymbol{V}^T\right)^{-1}\boldsymbol{D}^{-1}}_{=\mathbf{0}}+\boldsymbol{D}^{-1}\boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{V}^T\left(\boldsymbol{I}+\boldsymbol{D}^{-1}\boldsymbol{V}^T\right)^{-1}\boldsymbol{D}^{-1} \\
&= \boldsymbol{D}^{-1}\boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{V}^T\left(\boldsymbol{I}+\boldsymbol{D}^{-1}\boldsymbol{V}^T\right)^{-1}\boldsymbol{D}^{-1} \\
&= \boldsymbol{D}^{-1}\boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{V}^T\left(\boldsymbol{I}-\left(\boldsymbol{D}+\boldsymbol{V}^T\right)^{-1}\boldsymbol{V}^T\right)\boldsymbol{D}^{-1} \\
&= \boldsymbol{D}^{-1}\boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{V}^T\boldsymbol{D}^{-1}-\underbrace{\boldsymbol{D}^{-1}\boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{V}^T\left(\boldsymbol{D}+\boldsymbol{V}^T\right)^{-1}\boldsymbol{V}^T\boldsymbol{D}^{-1}}_{=\mathbf{0}} \\
&= \boldsymbol{D}^{-1}\boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{V}^T\boldsymbol{D}^{-1}
\end{aligned}
$$

$$
\begin{aligned}
A_2B_1C_1 &= \boldsymbol{D}^{-1}\left(\boldsymbol{I}+\boldsymbol{V}\boldsymbol{D}^{-1}\right)^{-1}\boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{D}\boldsymbol{D}^{-1} \\
&= \boldsymbol{D}^{-1}\left(\boldsymbol{I}+\boldsymbol{V}\boldsymbol{D}^{-1}\right)^{-1}\boldsymbol{V}\boldsymbol{D}^{-1} \\
&= \boldsymbol{D}^{-1}\left(\boldsymbol{I}-\boldsymbol{V}\left(\boldsymbol{D}+\boldsymbol{V}\right)^{-1}\right)\boldsymbol{V}\boldsymbol{D}^{-1} \\
&= \boldsymbol{D}^{-1}\boldsymbol{V}\boldsymbol{D}^{-1} - \underbrace{\boldsymbol{D}^{-1}\boldsymbol{V}\left(\boldsymbol{D}+\boldsymbol{V}\right)^{-1}\boldsymbol{V}\boldsymbol{D}^{-1}}_{=\mathbf{0}} \\
&= \boldsymbol{D}^{-1}\boldsymbol{V}\boldsymbol{D}^{-1}
\end{aligned}
$$

$$
\begin{aligned}
A_2B_1C_2 =& \boldsymbol{D}^{-1}\left(\boldsymbol{I}+\boldsymbol{V}\boldsymbol{D}^{-1}\right)^{-1}\boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{D}\boldsymbol{D}^{-1}\boldsymbol{V}^T\left(\boldsymbol{I}+\boldsymbol{D}^{-1}\boldsymbol{V}^T\right)^{-1}\boldsymbol{D}^{-1} \\
=& \boldsymbol{D}^{-1}\left(\boldsymbol{I}+\boldsymbol{V}\boldsymbol{D}^{-1}\right)^{-1}\boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{V}^T\left(\boldsymbol{I}+\boldsymbol{D}^{-1}\boldsymbol{V}^T\right)^{-1}\boldsymbol{D}^{-1} \\
=& \boldsymbol{D}^{-1}\left(\boldsymbol{I}-\boldsymbol{V}\left(\boldsymbol{D}+\boldsymbol{V}\right)^{-1}\right)\boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{V}^T\left(\boldsymbol{I}-\left(\boldsymbol{D}+\boldsymbol{V}^T\right)^{-1}\boldsymbol{V}^T\right)\boldsymbol{D}^{-1} \\
=& \boldsymbol{D}^{-1}\boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{V}^T\left(\boldsymbol{I}-\left(\boldsymbol{D}+\boldsymbol{V}^T\right)^{-1}\boldsymbol{V}^T\right)\boldsymbol{D}^{-1} \\
& \underbrace{-\boldsymbol{D}^{-1}\boldsymbol{V}\left(\boldsymbol{D}+\boldsymbol{V}\right)^{-1}\boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{V}^T\left(\boldsymbol{I}-\left(\boldsymbol{D}+\boldsymbol{V}^T\right)^{-1}\boldsymbol{V}^T\right)\boldsymbol{D}^{-1}}_{=\mathbf{0}} \\
=& \boldsymbol{D}^{-1}\boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{V}^T\boldsymbol{D}^{-1} - \underbrace{\boldsymbol{D}^{-1}\boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{V}^T\left(\boldsymbol{D}+\boldsymbol{V}^T\right)^{-1}\boldsymbol{V}^T\boldsymbol{D}^{-1}}_{=\mathbf{0}} \\
=& \boldsymbol{D}^{-1}\boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{V}^T\boldsymbol{D}^{-1}
\end{aligned}
$$

$$
\begin{aligned}
A_2B_2C_1 &= \boldsymbol{D}^{-1}\left(\boldsymbol{I}+\boldsymbol{V}\boldsymbol{D}^{-1}\right)^{-1}\boldsymbol{V}\boldsymbol{D}^{-1}\left(\boldsymbol{V}+\boldsymbol{V}^T\right)\boldsymbol{D}^{-1} \\
&= \underbrace{\boldsymbol{D}^{-1}\left(\boldsymbol{I}+\boldsymbol{V}\boldsymbol{D}^{-1}\right)^{-1}\boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{V}\boldsymbol{D}^{-1}}_{=\mathbf{0}} + \boldsymbol{D}^{-1}\left(\boldsymbol{I}+\boldsymbol{V}\boldsymbol{D}^{-1}\right)^{-1}\boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{V}^T\boldsymbol{D}^{-1} \\
&= \boldsymbol{D}^{-1}\left(\boldsymbol{I}+\boldsymbol{V}\boldsymbol{D}^{-1}\right)^{-1}\boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{V}^T\boldsymbol{D}^{-1} \\
&= \boldsymbol{D}^{-1}\left(\boldsymbol{I}-\boldsymbol{V}\left(\boldsymbol{D}+\boldsymbol{V}\right)^{-1}\right)\boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{V}^T\boldsymbol{D}^{-1} \\
&= \boldsymbol{D}^{-1}\boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{V}^T\boldsymbol{D}^{-1} - \underbrace{\boldsymbol{D}^{-1}\boldsymbol{V}\left(\boldsymbol{D}+\boldsymbol{V}\right)^{-1}\boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{V}^T\boldsymbol{D}^{-1}}_{=\mathbf{0}} \\
&= \boldsymbol{D}^{-1}\boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{V}^T\boldsymbol{D}^{-1}
\end{aligned}
$$

$$
\begin{aligned}
A_2B_2C_2 &= \boldsymbol{D}^{-1}\left(\boldsymbol{I}+\boldsymbol{V}\boldsymbol{D}^{-1}\right)^{-1}\boldsymbol{V}\boldsymbol{D}^{-1}\left(\boldsymbol{V}+\boldsymbol{V}^T\right)\boldsymbol{D}^{-1}\boldsymbol{V}^T\left(\boldsymbol{I}+\boldsymbol{D}^{-1}\boldsymbol{V}^T\right)^{-1}\boldsymbol{D}^{-1} \\
&= \boldsymbol{D}^{-1}\left(\boldsymbol{I}+\boldsymbol{V}\boldsymbol{D}^{-1}\right)^{-1}\boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{V}^T\left(\boldsymbol{I}+\boldsymbol{D}^{-1}\boldsymbol{V}^T\right)^{-1}\boldsymbol{D}^{-1} \\
&+ \boldsymbol{D}^{-1}\left(\boldsymbol{I}+\boldsymbol{V}\boldsymbol{D}^{-1}\right)^{-1}\boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{V}^T\boldsymbol{D}^{-1}\boldsymbol{V}^T\left(\boldsymbol{I}+\boldsymbol{D}^{-1}\boldsymbol{V}^T\right)^{-1}\boldsymbol{D}^{-1} \\
&= \mathbf{0}
\end{aligned}
$$

Finally,

$$
\begin{aligned}
&\boldsymbol{B}^{-1}\boldsymbol{M}\left(\boldsymbol{B}^{-1}\right)^{T} \\
=&\left[\underbrace{\boldsymbol{D}^{-1}}_{A_1}-\underbrace{\boldsymbol{D}^{-1}\left(\boldsymbol{I}+\boldsymbol{V}\boldsymbol{D}^{-1}\right)^{-1}\boldsymbol{V}\boldsymbol{D}^{-1}}_{A_2}\right]\times\left[\underbrace{\boldsymbol{D}}_{B_1}+\underbrace{\boldsymbol{V}+\boldsymbol{V}^{T}}_{B_2}\right] \\
&\times\left[\underbrace{\boldsymbol{D}^{-1}}_{C_1}-\underbrace{\boldsymbol{D}^{-1}\boldsymbol{V}^{T}\left(\boldsymbol{I}+\boldsymbol{D}^{-1}\boldsymbol{V}^{T}\right)^{-1}\boldsymbol{D}^{-1}}_{C_2}\right] \\
=&A_1B_1C_1-A_1B_1C_2+A_1B_2C_1-A_1B_2C_2-A_2B_1C_1+A_2B_1C_2-A_2B_2C_1+A_2B_2C_2 \\
=&\boldsymbol{D}^{-1}-\boldsymbol{D}^{-1}\boldsymbol{V}^{T}\boldsymbol{D}^{-1}+\boldsymbol{D}^{-1}\boldsymbol{V}\boldsymbol{D}^{-1}+\boldsymbol{D}^{-1}\boldsymbol{V}^{T}\boldsymbol{D}^{-1}-\boldsymbol{D}^{-1}\boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{V}^{T}\boldsymbol{D}^{-1}-\boldsymbol{D}^{-1}\boldsymbol{V}\boldsymbol{D}^{-1} \\
&+\boldsymbol{D}^{-1}\boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{V}^{T}\boldsymbol{D}^{-1}-\boldsymbol{D}^{-1}\boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{V}^{T}\boldsymbol{D}^{-1}+\boldsymbol{0} \\
=&\boldsymbol{D}^{-1}-\boldsymbol{D}^{-1}\boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{V}^{T}\boldsymbol{D}^{-1}
\end{aligned}
$$

So $\boldsymbol{B}^{-1}\boldsymbol{M}\left(\boldsymbol{B}^{-1}\right)^{T}$ is still block diagonal, since $\boldsymbol{D}^{-1}\boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{V}^{T}\boldsymbol{D}^{-1}$ is of form $\begin{bmatrix} X & 0 \\ 0 & 0 \end{bmatrix}$. In fact, if $\boldsymbol{D}$ is symmetric and full rank, then we can write $\boldsymbol{D}=\boldsymbol{D}_{*}\boldsymbol{D}_{*}^{T}$, thus $\boldsymbol{D}^{-1}\boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{V}^{T}\boldsymbol{D}^{-1}$ is at least positive semi-definite. So the diagonal of the top left block of $\boldsymbol{B}^{-1}\boldsymbol{M}\left(\boldsymbol{B}^{-1}\right)^{T}$ is less of equal to that of $\boldsymbol{D}^{-1}$.

Now we need to verify that the bun and meat matrix of the balanced regression calibration meets the above descriptions of $\boldsymbol{B}$ and $\boldsymbol{M}$. Here, $\boldsymbol{B}^{-1}\boldsymbol{M}\left(\boldsymbol{B}^{-1}\right)^{T}$ is the sandwich covariance of the regression parameters and the estimate genotype probabilities from the set of estimating equations, with the top left block corresponding to the regression parameters. The top left block of $\boldsymbol{D}^{-1}$ corresponds to the naive covariance matrix for the regression parameters.

The individual contributions to the estimating equations are as follows:

$$\begin{aligned}
\psi^{(i)}_{\beta_{0E}} &= Y_i - \text{expit}\left(\beta_{0E} + \beta_{1E} E\left[G_i|\tilde{\boldsymbol{G}}_i, d_i, \pi_i\right]\right) \\
\psi^{(i)}_{\beta_{1E}} &= E\left[G_i|\tilde{\boldsymbol{G}}_i, d_i, \pi_i\right]\left[Y_i - \text{expit}\left(\beta_{0E} + \beta_{1E} E\left[G_i|\tilde{\boldsymbol{G}}_i, d_i, \pi_i\right]\right)\right] \\
\psi^{(i)}_{p_0} &= \frac{P\left(\tilde{\boldsymbol{G}}_i|G_i = 0, d_i, \pi_i\right) - P\left(\tilde{\boldsymbol{G}}_i|G_i = 2, d_i, \pi_i\right)}{\sum_g p_g P\left(\tilde{\boldsymbol{G}}_i|G_i = g, d_i, \pi_i\right)} \\
\psi^{(i)}_{p_1} &= \frac{P\left(\tilde{\boldsymbol{G}}_i|G_i = 1, d_i, \pi_i\right) - P\left(\tilde{\boldsymbol{G}}_i|G_i = 2, d_i, \pi_i\right)}{\sum_g p_g P\left(\tilde{\boldsymbol{G}}_i|G_i = g, d_i, \pi_i\right)}
\end{aligned}$$

Furthermore:

$$\begin{aligned}
E\left[G_i|\tilde{\boldsymbol{G}}_i, d_i, \pi_i\right] &= \sum_g g \times P\left(G_i = g|\tilde{\boldsymbol{G}}_i, d_i, \pi_i\right) \\
&= \frac{\sum_g g \times p_g P\left(\tilde{\boldsymbol{G}}_i|G_i = g, d_i, \pi_i\right)}{\sum_{g'} p_{g'} P\left(\tilde{\boldsymbol{G}}_i|G_i = g', d_i, \pi_i\right)}
\end{aligned}$$

$$
\begin{aligned}
&\frac{\partial E\left[G_i|\tilde{\boldsymbol{G}}_i, d_i, \pi_i\right]}{\partial p_0} \\
=&\frac{-p_1 P\left(\tilde{\boldsymbol{G}}_i|G_i=1, d_i, \pi_i\right)\left[P\left(\tilde{\boldsymbol{G}}_i|G_i=0, d_i, \pi_i\right) - P\left(\tilde{\boldsymbol{G}}_i|G_i=2, d_i, \pi_i\right)\right]}{\left(\sum_{g'} p_{g'} P\left(\tilde{\boldsymbol{G}}_i|G_i=g', d_i, \pi_i\right)\right)^2} \\
&+\frac{\begin{array}{c}2\left[\left(\sum_g p_g P\left(\tilde{\boldsymbol{G}}_i|G_i=g, d_i, \pi_i\right)\right)\left(-P\left(\tilde{\boldsymbol{G}}_i|G_i=2, d_i, \pi_i\right)\right)\right. \\ \left.-\left(1-p_0-p_1\right) P\left(\tilde{\boldsymbol{G}}_i|G_i=2, d_i, \pi_i\right)\left(P\left(\tilde{\boldsymbol{G}}_i|G_i=0, d_i, \pi_i\right) - P\left(\tilde{\boldsymbol{G}}_i|G_i=2, d_i, \pi_i\right)\right)\right]\end{array}}{\left(\sum_{g'} p_{g'} P\left(\tilde{\boldsymbol{G}}_i|G_i=g', d_i, \pi_i\right)\right)^2}
\end{aligned}
$$

$$
\begin{aligned}
&\frac{\partial E\left[G_i|\tilde{\boldsymbol{G}}_i, d_i, \pi_i\right]}{\partial p_1} \\
=&\frac{\begin{array}{c}\left[\sum_g p_g P\left(\tilde{\boldsymbol{G}}_i|G_i=g, d_i, \pi_i\right)\right] P\left(\tilde{\boldsymbol{G}}_i|G_i=1, d_i, \pi_i\right) \\ -p_1 P\left(\tilde{\boldsymbol{G}}_i|G_i=1, d_i, \pi_i\right)\left[P\left(\tilde{\boldsymbol{G}}_i|G_i=1, d_i, \pi_i\right) - P\left(\tilde{\boldsymbol{G}}_i|G_i=2, d_i, \pi_i\right)\right]\end{array}}{\left(\sum_{g'} p_{g'} P\left(\tilde{\boldsymbol{G}}_i|G_i=g', d_i, \pi_i\right)\right)^2} \\
&+\frac{\begin{array}{c}2\left[\left(\sum_g p_g P\left(\tilde{\boldsymbol{G}}_i|G_i=g, d_i, \pi_i\right)\right)\left(-P\left(\tilde{\boldsymbol{G}}_i|G_i=2, d_i, \pi_i\right)\right)\right. \\ \left.-\left(1-p_0-p_1\right) P\left(\tilde{\boldsymbol{G}}_i|G_i=2, d_i, \pi_i\right)\left(P\left(\tilde{\boldsymbol{G}}_i|G_i=1, d_i, \pi_i\right) - P\left(\tilde{\boldsymbol{G}}_i|G_i=2, d_i, \pi_i\right)\right)\right]\end{array}}{\left(\sum_{g'} p_{g'} P\left(\tilde{\boldsymbol{G}}_i|G_i=g', d_i, \pi_i\right)\right)^2}
\end{aligned}
$$

Everything below will be evaluated at the true parameters and under the null, i.e. $\beta_1 = 0$, which induces $\beta_{1E} = 0$. In this case, $P\left(Y_i = 1|S_i = 1\right) = \text{expit}\left(\beta_{0E}\right)$.

The bottom right block of both the bun and meat matrices are the same since it is a likelihood and is trivial to show. The meat matrix is defined as:

$$\begin{bmatrix} \text{Regression parameter block} & \begin{matrix} \psi^{(i)}_{\beta_{0E}}\psi^{(i)}_{p0} & \psi^{(i)}_{\beta_{0E}}\psi^{(i)}_{p1} \\ \psi^{(i)}_{\beta_{1E}}\psi^{(i)}_{p0} & \psi^{(i)}_{\beta_{1E}}\psi^{(i)}_{p1} \end{matrix} \\ \text{Top right transpose} & \text{Genotype prob block} \end{bmatrix}$$

$$\begin{aligned}
\psi^{(i)}_{\beta_{0E}}\psi^{(i)}_{p0} =& \left[Y_i - \text{expit}\left(\beta_{0E} + \beta_{1E}E\left[G_i|\tilde{\boldsymbol{G}}_i, d_i, \pi_i\right]\right)\right] \\
&\times \frac{P\left(\tilde{\boldsymbol{G}}_i|G_i = 0, d_i, \pi_i\right) - P\left(\tilde{\boldsymbol{G}}_i|G_i = 2, d_i, \pi_i\right)}{\sum_g p_g P\left(\tilde{\boldsymbol{G}}_i|G_i = g, d_i, \pi_i\right)} \\
\psi^{(i)}_{\beta_{1E}}\psi^{(i)}_{p0} =& E\left[G_i|\tilde{\boldsymbol{G}}_i, d_i, \pi_i\right]\left[Y_i - \text{expit}\left(\beta_{0E} + \beta_{1E}E\left[G_i|\tilde{\boldsymbol{G}}_i, d_i, \pi_i\right]\right)\right] \\
&\times \frac{P\left(\tilde{\boldsymbol{G}}_i|G_i = 0, d_i, \pi_i\right) - P\left(\tilde{\boldsymbol{G}}_i|G_i = 2, d_i, \pi_i\right)}{\sum_g p_g P\left(\tilde{\boldsymbol{G}}_i|G_i = g, d_i, \pi_i\right)} \\
\psi^{(i)}_{\beta_{0E}}\psi^{(i)}_{p1} =& \left[Y_i - \text{expit}\left(\beta_{0E} + \beta_{1E}E\left[G_i|\tilde{\boldsymbol{G}}_i, d_i, \pi_i\right]\right)\right] \\
&\times \frac{P\left(\tilde{\boldsymbol{G}}_i|G_i = 1, d_i, \pi_i\right) - P\left(\tilde{\boldsymbol{G}}_i|G_i = 2, d_i, \pi_i\right)}{\sum_g p_g P\left(\tilde{\boldsymbol{G}}_i|G_i = g, d_i, \pi_i\right)} \\
\psi^{(i)}_{\beta_{1E}}\psi^{(i)}_{p1} =& E\left[G_i|\tilde{\boldsymbol{G}}_i, d_i, \pi_i\right]\left[Y_i - \text{expit}\left(\beta_{0E} + \beta_{1E}E\left[G_i|\tilde{\boldsymbol{G}}_i, d_i, \pi_i\right]\right)\right] \\
&\times \frac{P\left(\tilde{\boldsymbol{G}}_i|G_i = 1, d_i, \pi_i\right) - P\left(\tilde{\boldsymbol{G}}_i|G_i = 2, d_i, \pi_i\right)}{\sum_g p_g P\left(\tilde{\boldsymbol{G}}_i|G_i = g, d_i, \pi_i\right)}
\end{aligned}$$

The bun matrix is defined as:

$$\begin{bmatrix} \text{Regression parameter block} & \begin{matrix} -\frac{\partial \psi^{(i)}_{\beta_{0E}}}{\partial p_0} & -\frac{\partial \psi^{(i)}_{\beta_{0E}}}{\partial p_1} \\ -\frac{\partial \psi^{(i)}_{\beta_{1E}}}{\partial p_0} & -\frac{\partial \psi^{(i)}_{\beta_{1E}}}{\partial p_1} \end{matrix} \\ \mathbf{0} & \text{Genotype prob block} \end{bmatrix}$$

$$\begin{aligned}
\left.\frac{\partial \psi^{(i)}_{\beta_{0E}}}{\partial p_0}\right|_{\beta_{1E}=0} &= 0 \\
\left.\frac{\partial \psi^{(i)}_{\beta_{1E}}}{\partial p_0}\right|_{\beta_{1E}=0} &= \frac{\partial E\left[G_i|\tilde{\boldsymbol{G}}_i, d_i, \pi_i\right]}{\partial p_0}\left[Y_i - \text{expit}\left(\beta_{0E} + \beta_{1E} E\left[G_i|\tilde{\boldsymbol{G}}_i, d_i, \pi_i\right]\right)\right] \\
\left.\frac{\partial \psi^{(i)}_{\beta_{0E}}}{\partial p_1}\right|_{\beta_{1E}=0} &= 0 \\
\left.\frac{\partial \psi^{(i)}_{\beta_{1E}}}{\partial p_1}\right|_{\beta_{1E}=0} &= \frac{\partial E\left[G_i|\tilde{\boldsymbol{G}}_i, d_i, \pi_i\right]}{\partial p_1}\left[Y_i - \text{expit}\left(\beta_{0E} + \beta_{1E} E\left[G_i|\tilde{\boldsymbol{G}}_i, d_i, \pi_i\right]\right)\right]
\end{aligned}$$

Expectations for terms of the top right block of the meat matrix:

$$
\begin{aligned}
& E\left[\psi_{\beta_{0E}}^{(i)}\psi_{p_0}^{(i)}\middle|\,\beta_{1E}=0,Y_i,d_i,\pi_i\right] \\
=&\left[Y_i-P\left(Y_i=1|S_i=1\right)\right]E\left[\frac{P\left(\tilde{\boldsymbol{G}}_i|G_i=0,d_i,\pi_i\right)-P\left(\tilde{\boldsymbol{G}}_i|G_i=2,d_i,\pi_i\right)}{\sum_g p_g P\left(\tilde{\boldsymbol{G}}_i|G_i=g,d_i,\pi_i\right)}\middle|\,Y_i,d_i,\pi_i\right] \\
(1)=&\left[Y_i-P\left(Y_i=1|S_i=1\right)\right]E\left[\frac{\frac{1}{p_0}P\left(\tilde{\boldsymbol{G}}_i,G_i=0|d_i,\pi_i\right)-\frac{1}{p_2}P\left(\tilde{\boldsymbol{G}}_i,G_i=2|d_i,\pi_i\right)}{P\left(\tilde{\boldsymbol{G}}_i|d_i,\pi_i\right)}\middle|\,Y_i,d_i,\pi_i\right] \\
=&\left[Y_i-P\left(Y_i=1|S_i=1\right)\right]E\left[\frac{1}{p_0}P\left(G_i=0|\tilde{\boldsymbol{G}}_i,d_i,\pi_i\right)-\frac{1}{p_2}P\left(G_i=2|\tilde{\boldsymbol{G}}_i,d_i,\pi_i\right)\middle|\,Y_i,d_i,\pi_i\right] \\
(2)=&\left[Y_i-P\left(Y_i=1|S_i=1\right)\right] \\
&\times E\left[\frac{1}{p_0}P\left(G_i=0|\tilde{\boldsymbol{G}}_i,d_i,\pi_i,Y_i\right)-\frac{1}{p_2}P\left(G_i=2|\tilde{\boldsymbol{G}}_i,d_i,\pi_i,Y_i\right)\middle|\,Y_i,d_i,\pi_i\right] \\
=&\left[Y_i-P\left(Y_i=1|S_i=1\right)\right]\left[\frac{1}{p_0}P\left(G_i=0|d_i,\pi_i,Y_i\right)-\frac{1}{p_2}P\left(G_i=2|d_i,\pi_i,Y_i\right)\right] \\
(2)=&\left[Y_i-P\left(Y_i=1|S_i=1\right)\right]\left[\frac{1}{p_0}P\left(G_i=0|d_i,\pi_i\right)-\frac{1}{p_2}P\left(G_i=2|d_i,\pi_i\right)\right] \\
(1)=&\left[Y_i-P\left(Y_i=1|S_i=1\right)\right]\left[\frac{p_0}{p_0}-\frac{p_2}{p_2}\right] \\
=&0
\end{aligned}
$$

Note that the expectation is 0 given (Y_i, d_i, π_i), and will be 0 marginally. $E\left[\psi^{(i)}_{\beta_{0E}}\psi^{(i)}_{p_1}\right]$ can be shown to be 0 in a similar manner.

$$
\begin{aligned}
&E\left[\psi^{(i)}_{\beta_{1E}}\psi^{(i)}_{p_0}\middle|\,\beta_{1E}=0, Y_i, d_i, \pi_i\right] \\
=&\left[Y_i - P\left(Y_i=1|S_i=1\right)\right] \\
&\times E\left[E\left[G_i|\tilde{\boldsymbol{G}}_i, d_i, \pi_i\right]\frac{P\left(\tilde{\boldsymbol{G}}_i|G_i=0, d_i, \pi_i\right) - P\left(\tilde{\boldsymbol{G}}_i|G_i=2, d_i, \pi_i\right)}{\sum_g p_g P\left(\tilde{\boldsymbol{G}}_i|G_i=g, d_i, \pi_i\right)}\middle|\, Y_i, d_i, \pi_i\right] \\
(1) =&\left[Y_i - P\left(Y_i=1|S_i=1\right)\right] \\
&\times E\left[E\left[G_i|\tilde{\boldsymbol{G}}_i, d_i, \pi_i\right]\left(\frac{1}{p_0}P\left(G_i=0|\tilde{\boldsymbol{G}}_i, d_i, \pi_i\right) - \frac{1}{p_2}P\left(G_i=2|\tilde{\boldsymbol{G}}_i, d_i, \pi_i\right)\right)\middle|\, Y_i, d_i, \pi_i\right]
\end{aligned}
$$

We can similarly show that:

$$
\begin{aligned}
&E\left[\psi^{(i)}_{\beta_{1E}}\psi^{(i)}_{p_1}\middle|\,\beta_{1E}=0, Y_i, d_i, \pi_i\right] \\
=&\left[Y_i - P\left(Y_i=1|S_i=1\right)\right] \\
&\times E\left[E\left[G_i|\tilde{\boldsymbol{G}}_i, d_i, \pi_i\right]\left(\frac{1}{p_1}P\left(G_i=1|\tilde{\boldsymbol{G}}_i, d_i, \pi_i\right) - \frac{1}{p_2}P\left(G_i=2|\tilde{\boldsymbol{G}}_i, d_i, \pi_i\right)\right)\middle|\, Y_i, d_i, \pi_i\right]
\end{aligned}
$$

Now we look at the expectation of the top right block of the bun matrix. Note that $\left.\frac{\partial \psi^{(i)}_{\beta_{0E}}}{\partial p_0}\right|_{\beta_{1E}=0}$ and $\left.\frac{\partial \psi^{(i)}_{\beta_{0E}}}{\partial p_1}\right|_{\beta_{1E}=0}$ equal 0. For the other two terms:

$$
\begin{aligned}
&E\left[\left.\frac{\partial \psi^{(i)}_{\beta_{1E}}}{\partial p_0}\right| \beta_{1E}=0, Y_i, d_i, \pi_i\right] \\
&= \left[Y_i - P\left(Y_i=1|S_i=1\right)\right] E\left[\left.\frac{\partial E\left[G_i|\tilde{\boldsymbol{G}}_i, d_i, \pi_i\right]}{\partial p_0}\right| Y_i, d_i, \pi_i\right] \\
&= \left[Y_i - P\left(Y_i=1|S_i=1\right)\right] E\left[\left.\begin{array}{c}\frac{-p_1 P\left(\tilde{\boldsymbol{G}}_i|G_i=1,d_i,\pi_i\right)\left[P\left(\tilde{\boldsymbol{G}}_i|G_i=0,d_i,\pi_i\right)-P\left(\tilde{\boldsymbol{G}}_i|G_i=2,d_i,\pi_i\right)\right]}{\left(\sum_{g'} p_{g'} P\left(\tilde{\boldsymbol{G}}_i|G_i=g',d_i,\pi_i\right)\right)^2} \\ + \frac{\begin{array}{c}2\left[\left(\sum_g p_g P\left(\tilde{\boldsymbol{G}}_i|G_i=g,d_i,\pi_i\right)\right)\left(-P\left(\tilde{\boldsymbol{G}}_i|G_i=2,d_i,\pi_i\right)\right)\right. \\ \left.-\left(1-p_0-p_1\right) P\left(\tilde{\boldsymbol{G}}_i|G_i=2,d_i,\pi_i\right)\left(P\left(\tilde{\boldsymbol{G}}_i|G_i=0,d_i,\pi_i\right)-P\left(\tilde{\boldsymbol{G}}_i|G_i=2,d_i,\pi_i\right)\right)\right]\end{array}}{\left(\sum_{g'} p_{g'} P\left(\tilde{\boldsymbol{G}}_i|G_i=g',d_i,\pi_i\right)\right)^2}\end{array}\right| Y_i, d_i, \pi_i\right] \\
(1) &= \left[Y_i - P\left(Y_i=1|S_i=1\right)\right] E\left[\left.\begin{array}{c}\frac{-P\left(\tilde{\boldsymbol{G}}_i, G_i=1|d_i,\pi_i\right)\left[\frac{1}{p_0}P\left(\tilde{\boldsymbol{G}}_i, G_i=0|d_i,\pi_i\right)-\frac{1}{p_2}P\left(\tilde{\boldsymbol{G}}_i, G_i=2|d_i,\pi_i\right)\right]}{P\left(\tilde{\boldsymbol{G}}_i|d_i,\pi_i\right)^2} \\ + \frac{\begin{array}{c}2\left[P\left(\tilde{\boldsymbol{G}}_i|d_i,\pi_i\right)\left(-P\left(\tilde{\boldsymbol{G}}_i|G_i=2,d_i,\pi_i\right)\right)\right. \\ \left.-P\left(\tilde{\boldsymbol{G}}_i, G_i=2|d_i,\pi_i\right)\left(\frac{1}{p_0}P\left(\tilde{\boldsymbol{G}}_i, G_i=0|d_i,\pi_i\right)-\frac{1}{p_2}P\left(\tilde{\boldsymbol{G}}_i, G_i=2|d_i,\pi_i\right)\right)\right]\end{array}}{P\left(\tilde{\boldsymbol{G}}_i|d_i,\pi_i\right)^2}\end{array}\right| Y_i, d_i, \pi_i\right] \\
&= \left[Y_i - P\left(Y_i=1|S_i=1\right)\right] E\left[-P\left(G_i=1|\tilde{\boldsymbol{G}}_i,d_i,\pi_i\right)\left[\frac{1}{p_0}P\left(G_i=0|\tilde{\boldsymbol{G}}_i,d_i,\pi_i\right)-\frac{1}{p_2}P\left(G_i=2|\tilde{\boldsymbol{G}}_i,d_i,\pi_i\right)\right]\right. \\
&\quad + \frac{2\left(-P\left(\tilde{\boldsymbol{G}}_i|G_i=2,d_i,\pi_i\right)\right)}{P\left(\tilde{\boldsymbol{G}}_i|d_i,\pi_i\right)} \\
&\quad \left.-2P\left(G_i=2|\tilde{\boldsymbol{G}}_i,d_i,\pi_i\right)\left(\frac{1}{p_0}P\left(G_i=0|\tilde{\boldsymbol{G}}_i,d_i,\pi_i\right)-\frac{1}{p_2}P\left(G_i=2|\tilde{\boldsymbol{G}}_i,d_i,\pi_i\right)\right)\right| Y_i, d_i, \pi_i\Big] \\
&= \left[Y_i - P\left(Y_i=1|S_i=1\right)\right] E\left[\left.\begin{array}{c}\frac{2\left(-P\left(\tilde{\boldsymbol{G}}_i|G_i=2,d_i,\pi_i\right)\right)}{P\left(\tilde{\boldsymbol{G}}_i|d_i,\pi_i\right)} \\ -\left(P\left(G_i=1|\tilde{\boldsymbol{G}}_i,d_i,\pi_i\right)+2P\left(G_i=2|\tilde{\boldsymbol{G}}_i,d_i,\pi_i\right)\right)\left(\frac{1}{p_0}P\left(G_i=0|\tilde{\boldsymbol{G}}_i,d_i,\pi_i\right)-\frac{1}{p_2}P\left(G_i=2|\tilde{\boldsymbol{G}}_i,d_i,\pi_i\right)\right)\end{array}\right| Y_i, d_i, \pi_i\right] \\
&= \left[Y_i - P\left(Y_i=1|S_i=1\right)\right] E\left[\left.\frac{2\left(-P\left(\tilde{\boldsymbol{G}}_i|G_i=2,d_i,\pi_i\right)\right)}{P\left(\tilde{\boldsymbol{G}}_i|d_i,\pi_i\right)}\right| Y_i, d_i, \pi_i\right] \\
&\quad - \left[Y_i - P\left(Y_i=1|S_i=1\right)\right] E\left[\left.E\left[G_i|\tilde{\boldsymbol{G}}_i,d_i,\pi_i\right]\left(\frac{1}{p_0}P\left(G_i=0|\tilde{\boldsymbol{G}}_i,d_i,\pi_i\right)-\frac{1}{p_2}P\left(G_i=2|\tilde{\boldsymbol{G}}_i,d_i,\pi_i\right)\right)\right| Y_i, d_i, \pi_i\right] \\
&= \left[Y_i - P\left(Y_i=1|S_i=1\right)\right] E\left[\left.\frac{2\left(-P\left(\tilde{\boldsymbol{G}}_i|G_i=2,d_i,\pi_i\right)\right)}{P\left(\tilde{\boldsymbol{G}}_i|d_i,\pi_i\right)}\right| Y_i, d_i, \pi_i\right] - E\left[\left.\psi^{(i)}_{\beta_{1E}}\psi^{(i)}_{p0}\right| \beta_{1E}=0, Y_i, d_i, \pi_i\right]
\end{aligned}
$$

So, to show that $E\left[\psi^{(i)}_{\beta_{1E}}\psi^{(i)}_{p_0}\Big|_{\beta_{1E}=0}\right] = E\left[-\frac{\partial\psi^{(i)}_{\beta_{1E}}}{\partial p_0}\Big|_{\beta_{1E}=0}\right]$, it suffices to show that $E\left[\left[Y_i - P\left(Y_i=1|S_i=1\right)\right]E\left[\frac{2\left(-P\left(\tilde{\boldsymbol{G}}_i|G_i=2,d_i,\pi_i\right)\right)}{P\left(\tilde{\boldsymbol{G}}_i|d_i,\pi_i\right)}\middle| Y_i,d_i,\pi_i\right]\right]=0$:

$$
\begin{aligned}
&\left[Y_i - P\left(Y_i=1|S_i=1\right)\right]E\left[\frac{2\left(-P\left(\tilde{\boldsymbol{G}}_i|G_i=2,d_i,\pi_i\right)\right)}{P\left(\tilde{\boldsymbol{G}}_i|d_i,\pi_i\right)}\middle| Y_i,d_i,\pi_i\right]\\
(1)=&\left[Y_i - P\left(Y_i=1|S_i=1\right)\right]E\left[\frac{-\frac{2}{p_2}P\left(\tilde{\boldsymbol{G}}_i,G_i=2|d_i,\pi_i\right)}{P\left(\tilde{\boldsymbol{G}}_i|d_i,\pi_i\right)}\middle| Y_i,d_i,\pi_i\right]\\
=&-\left[Y_i - P\left(Y_i=1|S_i=1\right)\right]E\left[\frac{2}{p_2}P\left(G_i=2|\tilde{\boldsymbol{G}}_i,d_i,\pi_i\right)\middle| Y_i,d_i,\pi_i\right]\\
(2)=&-\left[Y_i - P\left(Y_i=1|S_i=1\right)\right]E\left[\frac{2}{p_2}P\left(G_i=2|\tilde{\boldsymbol{G}}_i,d_i,\pi_i,Y_i\right)\middle| Y_i,d_i,\pi_i\right]\\
=&-\left[Y_i - P\left(Y_i=1|S_i=1\right)\right]\frac{2}{p_2}P\left(G_i=2|d_i,\pi_i,Y_i\right)\\
(2)=&-\left[Y_i - P\left(Y_i=1|S_i=1\right)\right]\frac{2}{p_2}P\left(G_i=2|d_i,\pi_i\right)\\
(1)=&-\left[Y_i - P\left(Y_i=1|S_i=1\right)\right]\frac{2}{p_2}p_2\\
=&-2\left[Y_i - P\left(Y_i=1|S_i=1\right)\right]
\end{aligned}
$$

$$
\begin{aligned}
\Rightarrow&E\left[\left[Y_i - P\left(Y_i=1|S_i=1\right)\right]E\left[\frac{2\left(-P\left(\tilde{\boldsymbol{G}}_i|G_i=2,d_i,\pi_i\right)\right)}{P\left(\tilde{\boldsymbol{G}}_i|d_i,\pi_i\right)}\middle| Y_i,d_i,\pi_i\right]\right]\\
=&E\left[-2\left[Y_i - P\left(Y_i=1|S_i=1\right)\right]\right]\\
=&0
\end{aligned}
$$

This shows that $E\left[\psi^{(i)}_{\beta_{1E}}\psi^{(i)}_{p_0}\Big|_{\beta_{1E}=0}\right] = E\left[-\frac{\partial\psi^{(i)}_{\beta_{1E}}}{\partial p_0}\Big|_{\beta_{1E}=0}\right]$, similarly:

$$
\begin{aligned}
&E\left[\frac{\partial\psi^{(i)}_{\beta_{1E}}}{\partial p_1}\middle|\beta_{1E}=0,Y_i,d_i,\pi_i\right]\\
&=[Y_i-P(Y_i=1|S_i=1)]\,E\left[\frac{\partial E\left[G_i|\tilde{\boldsymbol{G}}_i,d_i,\pi_i\right]}{\partial p_1}\middle|Y_i,d_i,\pi_i\right]\\
&=[Y_i-P(Y_i=1|S_i=1)]\,E\left[\begin{array}{c}\left[\sum_g p_g P\left(\tilde{\boldsymbol{G}}_i|G_i=g,d_i,\pi_i\right)\right]P\left(\tilde{\boldsymbol{G}}_i|G_i=1,d_i,\pi_i\right)\\ -p_1P\left(\tilde{\boldsymbol{G}}_i|G_i=1,d_i,\pi_i\right)\left[P\left(\tilde{\boldsymbol{G}}_i|G_i=1,d_i,\pi_i\right)-P\left(\tilde{\boldsymbol{G}}_i|G_i=2,d_i,\pi_i\right)\right]\\ \hline \left(\sum_{g'}p_{g'}P\left(\tilde{\boldsymbol{G}}_i|G_i=g',d_i,\pi_i\right)\right)^2\end{array}\right.\\
&\quad+\left.\frac{\begin{array}{c}2\left[\left(\sum_g p_g P\left(\tilde{\boldsymbol{G}}_i|G_i=g,d_i,\pi_i\right)\right)\left(-P\left(\tilde{\boldsymbol{G}}_i|G_i=2,d_i,\pi_i\right)\right)\right.\\ \left.-(1-p_0-p_1)P\left(\tilde{\boldsymbol{G}}_i|G_i=2,d_i,\pi_i\right)\left(P\left(\tilde{\boldsymbol{G}}_i|G_i=1,d_i,\pi_i\right)-P\left(\tilde{\boldsymbol{G}}_i|G_i=2,d_i,\pi_i\right)\right)\right]\end{array}}{\left(\sum_{g'}p_{g'}P\left(\tilde{\boldsymbol{G}}_i|G_i=g',d_i,\pi_i\right)\right)^2}\right|Y_i,d_i,\pi_i\right]\\
(1)&=[Y_i-P(Y_i=1|S_i=1)]\,E\left[\begin{array}{c}P\left(\tilde{\boldsymbol{G}}_i|d_i,\pi_i\right)\frac{1}{p_1}P\left(\tilde{\boldsymbol{G}}_i,G_i=1|d_i,\pi_i\right)\\ -P\left(\tilde{\boldsymbol{G}}_i,G_i=1|d_i,\pi_i\right)\left[\frac{1}{p_1}P\left(\tilde{\boldsymbol{G}}_i,G_i=1|d_i,\pi_i\right)-\frac{1}{p_2}P\left(\tilde{\boldsymbol{G}}_i,G_i=2|d_i,\pi_i\right)\right]\\ \hline P\left(\tilde{\boldsymbol{G}}_i|d_i,\pi_i\right)^2\end{array}\right.\\
&\quad+\left.\frac{\begin{array}{c}2\left[P\left(\tilde{\boldsymbol{G}}_i|d_i,\pi_i\right)\left(-\frac{1}{p_2}P\left(\tilde{\boldsymbol{G}}_i,G_i=2|d_i,\pi_i\right)\right)\right.\\ \left.-P\left(\tilde{\boldsymbol{G}}_i,G_i=2|d_i,\pi_i\right)\left(\frac{1}{p_1}P\left(\tilde{\boldsymbol{G}}_i,G_i=1|d_i,\pi_i\right)-\frac{1}{p_2}P\left(\tilde{\boldsymbol{G}}_i,G_i=2|d_i,\pi_i\right)\right)\right]\end{array}}{P\left(\tilde{\boldsymbol{G}}_i|d_i,\pi_i\right)^2}\right|Y_i,d_i,\pi_i\right]\\
&=[Y_i-P(Y_i=1|S_i=1)]\,E\left[\frac{\frac{1}{p_1}P\left(\tilde{\boldsymbol{G}}_i,G_i=1|d_i,\pi_i\right)-\frac{2}{p_2}P\left(\tilde{\boldsymbol{G}}_i,G_i=2|d_i,\pi_i\right)}{P\left(\tilde{\boldsymbol{G}}_i|d_i,\pi_i\right)}\right.\\
&\quad-P\left(G_i=1|\tilde{\boldsymbol{G}}_i,d_i,\pi_i\right)\left[\frac{1}{p_1}P\left(G_i=1|\tilde{\boldsymbol{G}}_i,d_i,\pi_i\right)-\frac{1}{p_2}P\left(G_i=2|\tilde{\boldsymbol{G}}_i,d_i,\pi_i\right)\right]\\
&\quad\left.-2P\left(G_i=2|\tilde{\boldsymbol{G}}_i,d_i,\pi_i\right)\left(\frac{1}{p_1}P\left(G_i=1|\tilde{\boldsymbol{G}}_i,d_i,\pi_i\right)-\frac{1}{p_2}P\left(G_i=2|\tilde{\boldsymbol{G}}_i,d_i,\pi_i\right)\right)\middle|Y_i,d_i,\pi_i\right]\\
(1)&=[Y_i-P(Y_i=1|S_i=1)]\,E\left[\frac{1}{p_1}P\left(G_i=1|\tilde{\boldsymbol{G}}_i,d_i,\pi_i\right)-\frac{2}{p_2}P\left(G_i=2|\tilde{\boldsymbol{G}}_i,d_i,\pi_i\right)\middle|Y_i,d_i,\pi_i\right]\\
&\quad-[Y_i-P(Y_i=1|S_i=1)]\,E\left[E\left[G_i|\tilde{\boldsymbol{G}}_i,d_i,\pi_i\right]\left[\frac{1}{p_1}P\left(G_i=1|\tilde{\boldsymbol{G}}_i,d_i,\pi_i\right)-\frac{1}{p_2}P\left(G_i=2|\tilde{\boldsymbol{G}}_i,d_i,\pi_i\right)\right]\middle|Y_i,d_i,\pi_i\right]
\end{aligned}
$$

Again, to show that $E\left[\psi^{(i)}_{\beta_{1E}}\psi^{(i)}_{p_1}\Big|_{\beta_{1E}=0}\right] = E\left[-\frac{\partial\psi^{(i)}_{\beta_{1E}}}{\partial p_1}\Big|_{\beta_{1E}=0}\right]$, it suffices to show that $E\left[[Y_i - P(Y_i=1|S_i=1)]\, E\left[\frac{1}{p_1}P\left(G_i=1|\tilde{\boldsymbol{G}}_i,d_i,\pi_i\right) - \frac{2}{p_2}P\left(G_i=2|\tilde{\boldsymbol{G}}_i,d_i,\pi_i\right)\Big| Y_i,d_i,\pi_i\right]\right] = 0$:

$$
\begin{aligned}
&[Y_i - P(Y_i=1|S_i=1)]\, E\left[\frac{1}{p_1}P\left(G_i=1|\tilde{\boldsymbol{G}}_i,d_i,\pi_i\right) - \frac{2}{p_2}P\left(G_i=2|\tilde{\boldsymbol{G}}_i,d_i,\pi_i\right)\Big| Y_i,d_i,\pi_i\right] \\
(2) =& [Y_i - P(Y_i=1|S_i=1)] \\
&\times E\left[\frac{1}{p_1}P\left(G_i=1|\tilde{\boldsymbol{G}}_i,d_i,\pi_i,Y_i\right) - \frac{2}{p_2}P\left(G_i=2|\tilde{\boldsymbol{G}}_i,d_i,\pi_i,Y_i\right)\Big| Y_i,d_i,\pi_i\right] \\
=& [Y_i - P(Y_i=1|S_i=1)]\left[\frac{1}{p_1}P(G_i=1|d_i,\pi_i,Y_i) - \frac{2}{p_2}P(G_i=2|d_i,\pi_i,Y_i)\right] \\
(2) =& [Y_i - P(Y_i=1|S_i=1)]\left[\frac{1}{p_1}P(G_i=1|d_i,\pi_i) - \frac{2}{p_2}P(G_i=2|d_i,\pi_i)\right] \\
(1) =& [Y_i - P(Y_i=1|S_i=1)]\left[\frac{1}{p_1}p_1 - \frac{2}{p_2}p_2\right] \\
=& -[Y_i - P(Y_i=1|S_i=1)]
\end{aligned}
$$

$$
\begin{aligned}
\Rightarrow& E\left[[Y_i - P(Y_i=1|S_i=1)]\, E\left[\frac{1}{p_1}P\left(G_i=1|\tilde{\boldsymbol{G}}_i,d_i,\pi_i\right) - \frac{2}{p_2}P\left(G_i=2|\tilde{\boldsymbol{G}}_i,d_i,\pi_i\right)\Big| Y_i,d_i,\pi_i\right]\right] \\
=& E\left[-[Y_i - P(Y_i=1|S_i=1)]\right] \\
=& 0
\end{aligned}
$$

This shows that $E\left[\psi^{(i)}_{\beta_{1E}}\psi^{(i)}_{p_1}\Big|_{\beta_{1E}=0}\right] = E\left[-\frac{\partial\psi^{(i)}_{\beta_{1E}}}{\partial p_1}\Big|_{\beta_{1E}=0}\right]$.

Thus $E\left[\psi^{(i)}_{\boldsymbol{\beta}}\left(\psi^{(i)}_{\boldsymbol{p}}\right)^T\right] = E\left[-\frac{\partial\psi^{(i)}_{\boldsymbol{\beta}}}{\partial \boldsymbol{p}}\right]$.

Now we look at the block of each matrix that corresponds to the regression parameters. Under H_0, for the meat matrix looks like:

$$\left[Y_i - P\left(Y_i = 1|S_i = 1\right)\right]^2 \begin{bmatrix} 1 & E\left[G_i|\tilde{\boldsymbol{G}}_i, d_i, \pi_i\right] \\ E\left[G_i|\tilde{\boldsymbol{G}}_i, d_i, \pi_i\right] & E\left[G_i|\tilde{\boldsymbol{G}}_i, d_i, \pi_i\right]^2 \end{bmatrix}$$

The negative of the bun matrix looks like:

$$P\left(Y_i = 1|S_i = 1\right)\left[1 - P\left(Y_i = 1|S_i = 1\right)\right] \begin{bmatrix} 1 & E\left[G_i|\tilde{\boldsymbol{G}}_i, d_i, \pi_i\right] \\ E\left[G_i|\tilde{\boldsymbol{G}}_i, d_i, \pi_i\right] & E\left[G_i|\tilde{\boldsymbol{G}}_i, d_i, \pi_i\right]^2 \end{bmatrix}$$

The top left term has the same expectation is obvious to see. For the off diagonal term, we have:

$$\begin{aligned}
& E\left[\left[Y_i - P\left(Y_i = 1|S_i = 1\right)\right]^2 E\left[G_i|\tilde{\boldsymbol{G}}_i, d_i, \pi_i\right]\right] \\
=& E\left[\left[Y_i - P\left(Y_i = 1|S_i = 1\right)\right]^2 E\left[E\left[G_i|\tilde{\boldsymbol{G}}_i, d_i, \pi_i\right]|Y_i\right]\right] \\
=& P\left(Y_i = 1|S_i = 1\right)\left[1 - P\left(Y_i = 1|S_i = 1\right)\right]^2 E\left[E\left[G_i|\tilde{\boldsymbol{G}}_i, d_i, \pi_i\right]|Y_i = 1\right] \\
& + P\left(Y_i = 0|S_i = 1\right)\left[0 - P\left(Y_i = 1|S_i = 1\right)\right]^2 E\left[E\left[G_i|\tilde{\boldsymbol{G}}_i, d_i, \pi_i\right]|Y_i = 0\right] \\
(2) =& P\left(Y_i = 1|S_i = 1\right)\left[1 - P\left(Y_i = 1|S_i = 1\right)\right]^2 E\left[E\left[G_i|\tilde{\boldsymbol{G}}_i, d_i, \pi_i, Y_i = 1\right]|Y_i = 1\right] \\
& + P\left(Y_i = 0|S_i = 1\right)\left[0 - P\left(Y_i = 1|S_i = 1\right)\right]^2 E\left[E\left[G_i|\tilde{\boldsymbol{G}}_i, d_i, \pi_i, Y_i = 0\right]|Y_i = 0\right] \\
=& P\left(Y_i = 1|S_i = 1\right)\left[1 - P\left(Y_i = 1|S_i = 1\right)\right]^2 E\left[G_i|d_i, \pi_i, Y_i = 1\right] \\
& + P\left(Y_i = 0|S_i = 1\right)\left[0 - P\left(Y_i = 1|S_i = 1\right)\right]^2 E\left[G_i|d_i, \pi_i, Y_i = 0\right] \\
(2) =& P\left(Y_i = 1|S_i = 1\right)\left[1 - P\left(Y_i = 1|S_i = 1\right)\right]^2 E\left[G_i|d_i, \pi_i\right] \\
& + P\left(Y_i = 0|S_i = 1\right)\left[0 - P\left(Y_i = 1|S_i = 1\right)\right]^2 E\left[G_i|d_i, \pi_i\right] \\
(1) =& E\left[G_i\right]\left[P\left(Y_i = 1|S_i = 1\right)\left[1 - P\left(Y_i = 1|S_i = 1\right)\right]^2 + \left[1 - P\left(Y_i = 1|S_i = 1\right)\right]P\left(Y_i = 1|S_i = 1\right)^2\right] \\
=& E\left[G_i\right]P\left(Y_i = 1|S_i = 1\right)\left[1 - P\left(Y_i = 1|S_i = 1\right)\right]\left[\left[1 - P\left(Y_i = 1|S_i = 1\right)\right] + P\left(Y_i = 1|S_i = 1\right)\right] \\
=& E\left[G_i\right]P\left(Y_i = 1|S_i = 1\right)\left[1 - P\left(Y_i = 1|S_i = 1\right)\right]
\end{aligned}$$

$$
\begin{aligned}
&P\left(Y_i=1|S_i=1\right)\left[1-P\left(Y_i=1|S_i=1\right)\right]E\left[E\left[G_i|\tilde{\boldsymbol{G}}_i,d_i,\pi_i\right]\right]\\
=&E\left[G_i\right]P\left(Y_i=1|S_i=1\right)\left[1-P\left(Y_i=1|S_i=1\right)\right]
\end{aligned}
$$

So the off diagonal terms are equal, now for the bottom right term:

$$
\begin{aligned}
&E\left[\left[Y_i-P\left(Y_i=1|S_i=1\right)\right]^2E\left[G_i|\tilde{\boldsymbol{G}}_i,d_i,\pi_i\right]^2\right]\\
=&E\left[E\left[\left[Y_i-P\left(Y_i=1|S_i=1\right)\right]^2E\left[G_i|\tilde{\boldsymbol{G}}_i,d_i,\pi_i\right]^2\middle|Y_i\right]\right]\\
=&P\left(Y_i=1|S_i=1\right)E\left[\left[Y_i-P\left(Y_i=1|S_i=1\right)\right]^2E\left[G_i|\tilde{\boldsymbol{G}}_i,d_i,\pi_i\right]^2\middle|Y_i=1\right]\\
&+P\left(Y_i=0|S_i=1\right)E\left[\left[Y_i-P\left(Y_i=1|S_i=1\right)\right]^2E\left[G_i|\tilde{\boldsymbol{G}}_i,d_i,\pi_i\right]^2\middle|Y_i=0\right]\\
=&P\left(Y_i=1|S_i=1\right)\left(1-P\left(Y_i=1|S_i=1\right)\right)^2E\left[E\left[G_i|\tilde{\boldsymbol{G}}_i,d_i,\pi_i\right]^2\middle|Y_i=1\right]\\
&+\left(1-P\left(Y_i=1|S_i=1\right)\right)P\left(Y_i=1|S_i=1\right)^2E\left[E\left[G_i|\tilde{\boldsymbol{G}}_i,d_i,\pi_i\right]^2\middle|Y_i=0\right]\\
=&P\left(Y_i=1|S_i=1\right)\left(1-P\left(Y_i=1|S_i=1\right)\right)\\
&\times\left[\left(1-P\left(Y_i=1|S_i=1\right)\right)E\left[E\left[G_i|\tilde{\boldsymbol{G}}_i,d_i,\pi_i\right]^2\middle|Y_i=1\right]\right.\\
&\left.+P\left(Y_i=1|S_i=1\right)E\left[E\left[G_i|\tilde{\boldsymbol{G}}_i,d_i,\pi_i\right]^2\middle|Y_i=0\right]\right]
\end{aligned}
$$

$$
\begin{aligned}
&E\left[P\left(Y_i=1|S_i=1\right)\left[1-P\left(Y_i=1|S_i=1\right)\right]E\left[G_i|\tilde{\boldsymbol{G}}_i,d_i,\pi_i\right]^2\right]\\
=&P\left(Y_i=1|S_i=1\right)\left(1-P\left(Y_i=1|S_i=1\right)\right)\\
&\times\left[P\left(Y_i=1|S_i=1\right)E\left[E\left[G_i|\tilde{\boldsymbol{G}}_i,d_i,\pi_i\right]^2\middle|Y_i=1\right]\right.\\
&\left.+\left(1-P\left(Y_i=1|S_i=1\right)\right)E\left[E\left[G_i|\tilde{\boldsymbol{G}}_i,d_i,\pi_i\right]^2\middle|Y_i=0\right]\right]
\end{aligned}
$$

When the distribution of read depth and error rates differ between cases and controls, the value $E\left[E\left[G_i|\tilde{\boldsymbol{G}}_i, d_i, \pi_i\right]^2 \middle| Y_i\right]$ varies depending on Y_i since the variance of $E\left[G_i|\tilde{\boldsymbol{G}}_i, d_i, \pi_i\right]$ will defer between cases and controls even though the expectation of $E\left[G_i|\tilde{\boldsymbol{G}}_i, d_i, \pi_i\right]$, which is $E[G_i]$, is the same. In this case, if sampling is balanced, i.e. $P(Y_i = 1|S_i = 1) = \frac{1}{2}$, the expectation of the bottom right terms of the meat and bun matrices are still the same, otherwise, they are different. So if the distribution of read depth and error rates differ between cases and controls, the expectation of the bun and meat matrices under the null meet the description of the matrices $\boldsymbol{B}$ and $\boldsymbol{M}$ as stated above when sampling is balanced, otherwise, they do not.

(1) G_i is marginally independent of (d_i, π_i), and under H_0, G_i is independent of Y_i and S_i, so $P(G_i = g|d_i, \pi_i) = P(G_i = g) = P(G_i = g|S_i = 1) = p_g$

(2) under H_0, G_i is independent of Y_i given (d_i, π_i)

Note, DAG under H_0:

$$\begin{array}{ccccc} S & \leftarrow & Y & & G_i \\ & & \downarrow & & \downarrow \\ & & d_i, \pi_i & \rightarrow & \tilde{\boldsymbol{G}}_i \end{array}$$

B Supplementary Material to Chapter 2

B.1 Variance of $S_w(t_k)$

From Schwartzman and Lin (2011):

$$P(Z_j > t_j, Z_k > t_k) = \bar{\Phi}(t_j)\bar{\Phi}(t_k) + \phi(t_j)\phi(t_k)\sum_{n=1}^{\infty}\frac{r_{jk}^n}{n!}\mathcal{H}_{n-1}(t_j)\mathcal{H}_{n-1}(t_k)$$

where $\mathcal{H}_i(t)$ is the ith Hermite polynomial and r_{jk} is the element in the jth row and kth column of correlation matrix $\mathbf{\Sigma}$. Note that

$$P(|Z_j| > t_j, |Z_k| > t_k) = 2\left(\bar{\Phi}(t_j) - P(Z_j > t_j, Z_k > -t_k) + P(Z_j > t_j, Z_k > t_k)\right)$$

Then, the variance of S_w can be written as:

$$\begin{aligned}
var(S_w(t)) &= var\left(\sum_{i=1}^{p} I_{|Z_i|\geq w_i t}\right) \\
&= E\left[\left(\sum_{i=1}^{p} I_{|Z_i|\geq w_i t}\right)^2\right] - E\left[\sum_{i=1}^{p} I_{|Z_i|\geq w_i t}\right]^2 \\
&= E\left[\sum_{i=1}^{p} I_{|Z_i|\geq w_i t}^2 + 2\sum_{j<k=1}^{p} I_{|Z_j|\geq w_j t} I_{|Z_k|\geq w_k t}\right] - E\left[\sum_{i=1}^{p} I_{|Z_i|\geq w_i t}\right]^2 \\
&= 2\sum_{j<k=1}^{p} P(|Z_j| \geq w_j t, |Z_k| \geq w_k t) + \sum_{i=1}^{p}\bar{\Phi}(w_i t) - \left(\sum_{i=1}^{p}\bar{\Phi}(w_i t)\right)^2
\end{aligned}$$

Now the above expression can be substutited for each $P(|Z_j| \geq w_j t, |Z_k| \geq w_k t)$ term. The infinite sum can be approximated accurately with just the first few terms.

B.2 Approximation of the Moments of $S_w(t_k)\,|S_w(t_{k-1}) = m$

Let $\{\{j'\}\}$ and $\{\{j*\}\}$ both denote the collection of all size m subsets of $\{1 \dots p\}$ for notational purposes, $\overline{\{j'\}}$ and $\overline{\{j*\}}$ denote the event that $\{j'\}$ and $\{j*\}$ are the only indices that incur the indicator at t_{k-1}. We approximate the first moment of $S_w(t_k)\,|S_w(t_{k-1}) = m$ as:

$$\begin{aligned}
E\left[S_w(t_k)\,|S_w(t_{k-1}) = m\right] &= E\left[\sum_j I_{|Z_j| \geq w_j t_k} | S_w(t_{k-1}) = m\right] \\
&= \sum_{\{j'\}} E\left[\sum_j I_{|Z_j| \geq w_j t_k} | \overline{\{j'\}}\right] \times P\left(\overline{\{j'\}} | S_w(t_{k-1}) = m\right) \\
&\approx \sum_{\{j'\}} \sum_{j \in \{j'\}} \frac{\bar{\Phi}(w_j t_k)}{\bar{\Phi}(w_j t_{k-1})} \times \frac{P\left(\overline{\{j'\}}\right)}{\sum_{\{j*\}} P\left(\overline{\{j*\}}\right)} \\
&= \sum_j \frac{\bar{\Phi}(w_j t_k)}{\bar{\Phi}(w_j t_{k-1})} \times \frac{\sum_{\{j'\}: j \in \{j'\}} P\left(\overline{\{j'\}}\right)}{\sum_{\{j*\}} P\left(\overline{\{j*\}}\right)}
\end{aligned}$$

The first approximation states that the distribution of Z_j given that $j \in \{j'\}$ and $\overline{\{j'\}}$ does not depend on Z_k for $j \neq k \in \{j'\}$. The terms $\frac{\sum_{\{j'\}: j \in \{j'\}} P\left(\overline{\{j'\}}\right)}{\sum_{\{j*\}} P\left(\overline{\{j*\}}\right)}$ represent the probability that j is one of the indicators incurred at t_{k-1} given that only m indicators are incurred at t_{k-1}. One can approximate these values by first approximating $P\left(\overline{\{j'\}}\right)$ by assuming independence between $\{Z_j\}$ for each $\{j'\}$. This approach is quite accurate, but it essentially entails iterating through the powerset of $\{1 \dots p\}$, which is completely inpractical. Note that the terms $\frac{\sum_{\{j'\}: j \in \{j'\}} P\left(\overline{\{j'\}}\right)}{\sum_{\{j*\}} P\left(\overline{\{j*\}}\right)}$ sum up to m. Currently, we approximate $\frac{\sum_{\{j'\}: j \in \{j'\}} P\left(\overline{\{j'\}}\right)}{\sum_{\{j*\}} P\left(\overline{\{j*\}}\right)}$ by scaling the terms $\bar{\Phi}(w_j t_{k-1})$ up so they sum to m. This simplified the above expression to:

$$\begin{aligned}
E\left[S_w(t_k)\,|S_w(t_{k-1}) = m\right] &\approx \sum_j \frac{\bar{\Phi}(w_j t_k)}{\bar{\Phi}(w_j t_{k-1})} \times \frac{\bar{\Phi}(w_j t_{k-1})}{\sum_{j*} \bar{\Phi}(w_{j*} t_{k-1})} \times m \\
&= m \frac{\sum_j \bar{\Phi}(w_j t_k)}{\sum_{j*} \bar{\Phi}(w_{j*} t_{k-1})}
\end{aligned}$$

Note that is approximation is quite crude, and leads to the limitations we currently face with the weighted GHC. Likewise, for the second moment:

$$
\begin{aligned}
& E\left[S_w\left(t_k\right)^2 | S_w\left(t_{k-1}\right)=m\right] \\
=& E\left[\left(\sum_j I_{|Z_j| \geq w_j t_k}\right)^2 | S_w\left(t_{k-1}\right)=m\right] \\
=& E\left[\sum_j I_{|Z_j| \geq w_j t_k}^2 | S_w\left(t_{k-1}\right)=m\right]+2 \times E\left[\sum_{j<l} I_{|Z_j| \geq w_j t_k} I_{|Z_l| \geq w_l t_k} | S_w\left(t_{k-1}\right)=m\right] \\
\approx & E\left[\sum_j I_{|Z_j| \geq w_j t_k} | S_w\left(t_{k-1}\right)=m\right] \\
& +2 \times \sum_{j<l} \frac{P\left(|Z_j| \geq w_j t_k, |Z_l| \geq w_l t_k\right)}{P\left(|Z_j| \geq w_j t_{k-1}, |Z_l| \geq w_l t_{k-1}\right)} \times \frac{\sum_{\{j'\}: j, l \in\{j'\}} P\left(\overline{\{j'\}}\right)}{\sum_{\{j*\}} P\left(\overline{\{j*\}}\right)}
\end{aligned}
$$

The $\frac{\sum_{\{j'\}: j, l \in\{j'\}} P\left(\overline{\{j'\}}\right)}{\sum_{\{j*\}} P\left(\overline{\{j*\}}\right)}$ terms are approximated by scaling the $P\left(|Z_j| \geq w_j t_{k-1}, |Z_l| \geq w_l t_{k-1}\right)$ terms up so they sum to $\begin{pmatrix} m \\ 2 \end{pmatrix}$. The expression then simplifies to:

$$
\begin{aligned}
& E\left[S_w\left(t_k\right)^2 | S_w\left(t_{k-1}\right)=m\right] \\
\approx & m \frac{\sum_j \bar{\Phi}\left(w_j t_k\right)}{\sum_{j*} \bar{\Phi}\left(w_{j*} t_{k-1}\right)}+2 \times \sum_{j<l} \frac{P\left(|Z_j| \geq w_j t_k, |Z_l| \geq w_l t_k\right)}{P\left(|Z_j| \geq w_j t_{k-1}, |Z_l| \geq w_l t_{k-1}\right)} \\
& \times \frac{P\left(|Z_j| \geq w_j t_{k-1}, |Z_l| \geq w_l t_{k-1}\right)}{\sum_{j*<l*} P\left(|Z_{j*}| \geq w_{j*} t_{k-1}, |Z_{l*}| \geq w_{l*} t_{k-1}\right)} \times \begin{pmatrix} m \\ 2 \end{pmatrix} \\
= & m \frac{\sum_j \bar{\Phi}\left(w_j t_k\right)}{\sum_{j*} \bar{\Phi}\left(w_{j*} t_{k-1}\right)}+2 \begin{pmatrix} m \\ 2 \end{pmatrix} \frac{\sum_{j<l} P\left(|Z_j| \geq w_j t_k, |Z_l| \geq w_l t_k\right)}{\sum_{j*<l*} P\left(|Z_{j*}| \geq w_{j*} t_{k-1}, |Z_{l*}| \geq w_{l*} t_{k-1}\right)}
\end{aligned}
$$

B.3 Power Simulation Results

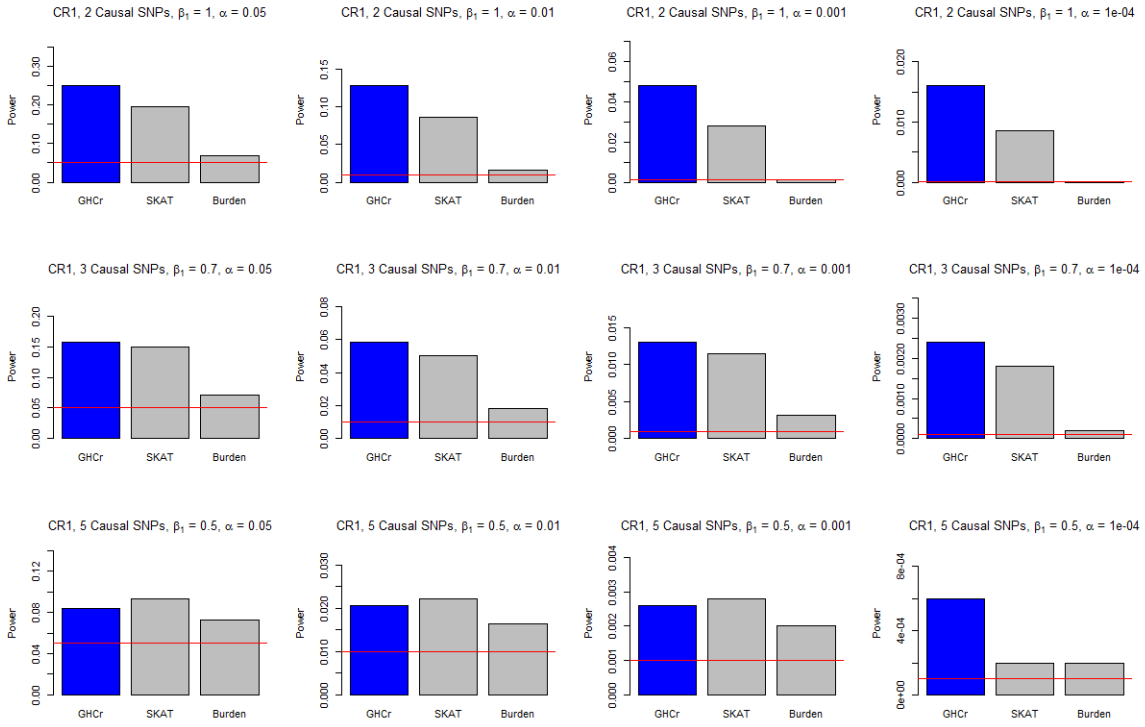

Figure B.1: Power simulation results for the first causal regime

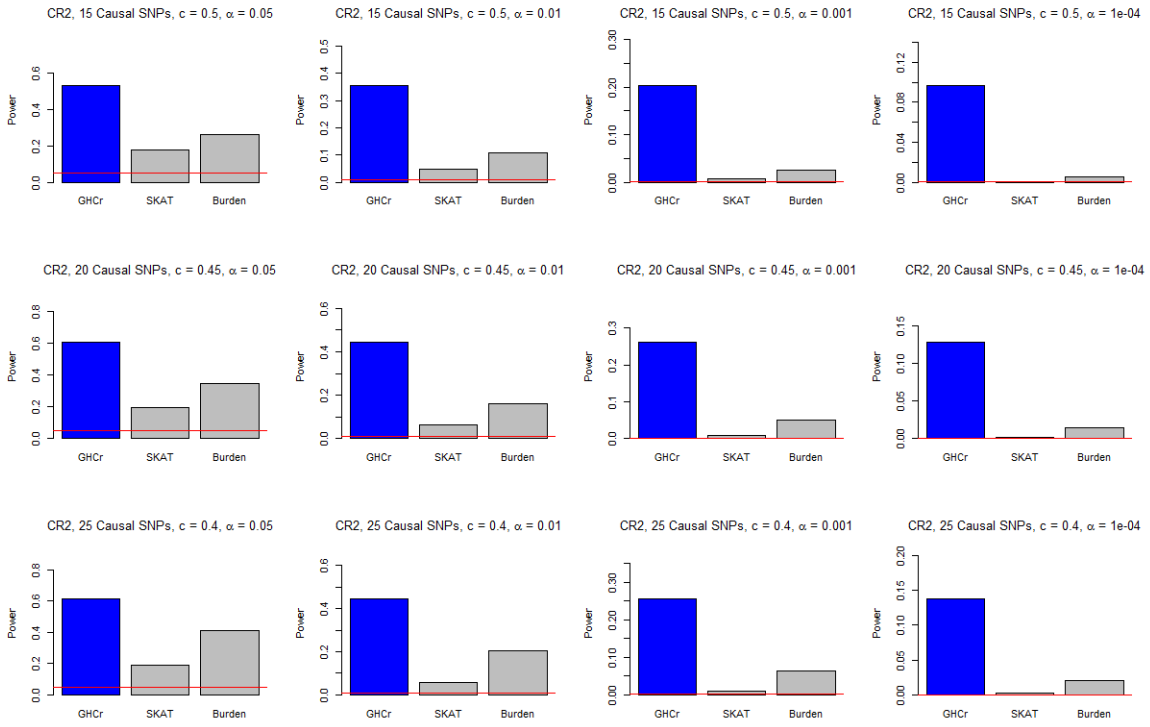

Figure B.2: Power simulation results for the second causal regime

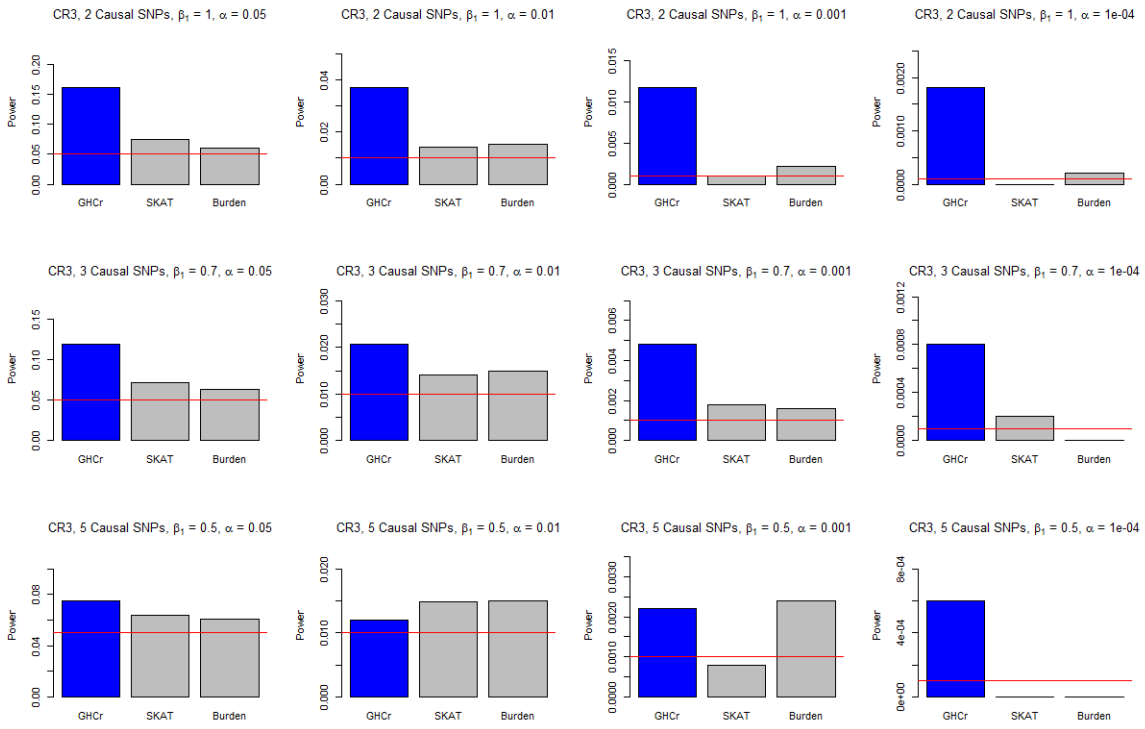

Figure B.3: Power simulation results for the third causal regime

C Supplementary Material to Chapter 3

C.1 Simulation Numerical AUC Results

Table C.1: Cross-validated AUC results from the simulations for the full Super Leaner, discrete Super Learner, support vector machine, random forest, k-nearest neighbors at various edge counts and values of p_2

		$p_2 = 0.05$	$p_2 = 0.03$	$p_2 = 0.01$	$p_2 = 0.005$
$EC = 500$	fSL	0.65898	0.58875	0.51746	0.50457
	dSL	0.65707	0.58846	0.51606	0.50493
	SVM	0.55004	0.50956	0.49892	0.50094
	RF	0.65707	0.58846	0.51814	0.50507
	KNN	0.63730	0.56502	0.51603	0.50102
$EC = 1000$	fSL	0.97887	0.90921	0.66112	0.57223
	dSL	0.97767	0.90452	0.65349	0.57040
	SVM	0.97583	0.90452	0.55492	0.50450
	RF	0.97767	0.90231	0.65542	0.57040
	KNN	0.97378	0.89997	0.64798	0.55610
$EC = 2000$	fSL	0.99951	0.99849	0.90484	0.74456
	dSL	0.99904	0.99775	0.90082	0.73649
	SVM	0.99878	0.99707	0.90082	0.64878
	RF	0.99904	0.99775	0.89611	0.73649
	KNN	0.99768	0.99544	0.89320	0.72942