Leveraging Summary Statistics and Integrative Analysis for Prediction and Inference in Genome-Wide Association Studies

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Jack William Pattee

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor of Philosophy

Dr. Wei Pan, Adviser

July, 2020

Acknowledgements

I'd like to thank everyone who has supported me throughout my time in graduate school. I would like to particularly thank my adviser Dr. Wei Pan for his support, guidance, insight, and advice. I credit much of my growth as a scholar and as a person to his mentorship.

I would also like to thank my collaborator and committee member Dr. Weihong Tang for her advice and mentorship, and for arranging extensive collaboration opportunities for me. I deeply grateful for these opportunities. I'd also like to thank my committee members Dr. Weihua Guan and Dr. Eric Lock for their valuable time and advice. I also thank Chong Wu and Zhiyuan Xu for their advice and support, especially in the early stage of my graduate career when I was inexperienced and in need of help. Finally, I'd like to thank Dr. Beth Virnig for her mentorship and advice.

Dedication

To my parents Randall Pattee and Valerie Ruttenberg for their endless love and support, and the rich and varied opportunities that they have selflessly provided me with. To my brothers Emmett Pattee and Cyrus Pattee for teaching me the value of perspective, humility, and living in the moment.

Abstract

Genome-wide association studies (GWASs) have attained substantial success in parsing the genetic etiology of complex traits. GWAS analyses have identified many genetic variants associated with various traits, and polygenic risk scores estimated from GWASs have been used to effectively predict certain clinical phenotypes. Despite these accomplishments, GWASs suffer from some pervasive issues with power and interpretability. To address these issues, we develop powerful and novel approaches for prediction and inference on genetic and genomic data. Our approaches focus on two key elements. First is the incorporation of additional sources of genetic and genomic data. A typical GWAS characterizes the genetic basis of a trait in terms of associations between the trait and a set of single nucleotide polymorphisms (SNPs). This approach can often be underpowered and difficult to understand biologically. We can often increase power and interpretability by effectively incorporating other sources of genetic and genomic data into the single SNP analysis structure. Second is the development of methods that are widely applicable in the context of summary statistics. Many published GWAS analyses do not provide so-called individual level genetic and genomic data, and instead provide only summary statistic information. Given this, we want our methods to be able to be flexible in the context of summary statistics without the need for individual level information.

We first develop a novel approach to integrating somatic and germline information from tumors to identify genes associated with lung cancer risk. We leverage this approach to discover potentially novel genes associated with lung cancer. We then investigate the problem of estimating powerful and parsimonious models for polygenic risk scores in the context of summary statistics. We develop a set of novel methods for model estimation, model selection, and the assessment of model performance, and demonstrate their beneficial properties in extensive simulation and in application to GWASs of lung cancer, blood lipid levels, and height. Lastly, we integrate our methods for polygenic risk score estimation into a two sample two-stage least squares analysis framework to identify potentially novel endophenotypes associated with increased risk

of Alzheimer's disease. We demonstrate via simulation and real data application that our approach is powerful and effective.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Genome wide association studies (GWAS) have attained substantial success in elucidating the genetic basis of complex traits [1]. Despite these successes, GWAS suffer from issues with power and interpretability. Principal among these is the so-called missing heritability problem [2], which states that a substantial proportion of the expected heritability of complex traits cannot be characterized by current GWAS analyses. The typical GWAS approach characterizes the genetic basis of some trait in terms of associations between single nucleotide polymorphisms (SNPs) and the trait. There are several potential issues with this approach that may mitigate power. Due to the high number of SNPs in the human genome and the small effect size of an individual SNP, this SNP-by-SNP analysis approach may miss substantial signal. We also note the interpretability issue, namely that many SNPs that are identified to be statistically associated with a complex trait are located in a region of unknown functionality [3, 4]. Given the goal of parsing the biological etiology of a complex trait, the causal mechanism by which a SNP modulates complex trait risk is also of interest. This mechanism can be difficult to recover with a standard GWAS approach.

GWAS analyses have facilitated inference that has identified many genetic variants associated with complex traits [5]. GWAS have likewise facilitated the development of polygenic risk scores, which have been leveraged to effectively predict certain clinical phenotypes [6, 7]. This thesis seeks to address some of the power and interpretability issues characteristic to GWAS analyses. We develop approaches for improving risk prediction and inference in application to genetic data, and demonstrate that parsimonious

methods for risk prediction and novel approaches for association testing can be leveraged in tandem to gain greater understanding of the genetic basis of complex diseases.

Given the widespread availability of genetic and genomic data from published analyses, we consider two particular elements throughout this work. Firstly, we seek to develop methods that are applicable in the summary statistic framework, explained as follows. Given privacy concerns related to genetic data, it is often the case that results from published analyses contain only univariate association statistics between a trait (or traits) and a set of SNPs, and do not contain individual level genetic data. We seek to develop methodology that can be flexibly implemented in the context of summary statistics that requires limited or no individual level information. Secondly, we seek to incorporate multiple sources of genetic and genomic information in our approaches. Recent approaches to gene-based inference that leverage transcriptomic data have been very successful in improving the power and interpretability of GWAS analyses [8, 9]. Motivated by this, we seek to leverage multiple sources of genetic and genomic data. In particular, we often consider the integration of multiple data sources within a two sample two-stage least squares framework. In this framework, SNPs are instrumental variables, some endophenotype (such as gene expression) is the stage 1 phenotype, and some phenotype of interest (i.e. cancer status) is the stage 2 phenotype. Many recently popular methods, namely TWAS and PrediXcan, can be conceptualized as applications of the two-stage least squares framework.

In Chapter 2, we develop an extension of the transcriptome-wide association framework [8, 9] that facilitates the analysis of cancer phenotypes. In particular, we consider application to a lung cancer GWAS. Motivated by literature demonstrating that it may be most powerful to use transcriptome data from a highly relevant tissue type [10], we use gene expression measured from lung adenocarcinomas to construct functional weights. Given the complex interplay of germline and somatic features in the control of gene expression in tumors [11, 12], we develop a novel approach to the estimation of functional weights that leverages both somatic and germline data. This approach differs from a typical TWAS analysis, which generally uses gene expression measured in normal tissue to estimate funcitonal weights. We show that this approach has several beneifical properties as opposed to the naive approach, which does not consider somatic determinants of gene expression. We apply a powerful approach to association testing

that leverages functional weights in conjunction with GWAS summary statistics to a lung cancer GWAS. We find that integrating somatic and germline information from tumors facilitates the discovery of novel genes associated with lung cancer.

In Chapter 3, we develop novel methodology that facilitates model estimation and model selection for polygenic scores in the summary statistic framework. Simple methods for polygenic risk score estimation, such as thresholding [13] and pruning and thresholding [14], are generally tractable in the summary statistic framework. However, these methods do not model SNP effects under linkage disequilibrium (LD) and thus may be suboptimal approximations of the true genetic architecture. Summary based methods LDPred [15] and LassoSum [16] leverage publicly available reference data to estimate SNP effects under LD. These approaches consistently generate better prediction than methods that to not account for LD. We develop two novel penalized methods for the estimation of polygenic risk scores in the summary statistic framework. These approaches can be conceptualized as implementations of the truncated LASSO penalty [17] and the elastic net [18] in the summary statistic framework. A difficult problem is the selection of tuning parameters in the absence of individual level genetic data. To this end, we develop methods for estimating the AIC and BIC in the summary statistic framework. Lastly, we propose a measure of assessing model performance when only summary statistic information is available from the out-of-sample data. We demonstrate the beneficial properties of these methods in extensive simulation studies, and in application to GWAS analyses of blood lipid levels, lung cancer, and height. In total, these methods facilitate the estimation and selection of powerful and sparse models, and broaden the scope of summary statistic analyses.

In chapter 4, we leverage the polygenic risk score methodology described in chapter 3 in conjunction with a two-stage least squares framework to conduct powerful inference in an application to Alzheimer's disease. We consider a two sample two-stage least squares framework where both stage 1 and stage 2 data are comprised of summary statistics. In our application, the stage 1 data is a set of 1,738 heritable endophenotypes observed in the UK Biobank cohort [19], and the stage 2 data is Alzheimer's status as measured in the International Genetics of Alzheimer's Project [20]. The goal is to find potentially modifiable endophenotypes associated with increased Alzheimer's risk. We leverage our methodology from chapter 3 to estimate and select parsimonious and

powerful stage 1 models, thus better facilitating inference in stage 2. There are multiple existing approaches to this problem; in particular, we contrast our penalized regression approach to Mendelian randomization approaches [21] and an approach that leverages simple polygenic risk scores that do not account for LD [22]. Through simulation study and application to real data, we demonstrate that penalized regression methodologies have better power than other methods. We also examine the interplay between modeling choices and violations of the assumptions of two-stage least squares, in particular the difficult problem of identifying violations related to pleiotropy. We present simulation results indicating that type I error due to pleiotropy is a persistent problem in methods that leverage polygenic risk scores, and some compelling evidence that application of our pseudo AIC and pseudo BIC models may reduce some issues associated with pleiotropy.

Chapter 2

Integrating germline and somatic genetics to identify genes associated with lung cancer

2.1 Introduction

This chapter is reproduced in slightly altered form from published work [23].

Genetic variation accounts for a large proportion of the variability in many complex human traits. Despite many successes, genome wide association studies (GWAS) have failed to characterize the sources of genetic variability for many traits [5, 2]. The typical GWAS uses individual genetic variants as the unit of analysis, normally single nucleotide polymorphisms (SNPs). GWAS often suffer from low power, thus failing to identify variants associated with complex traits. This is due to the polygenic nature of complex traits, and the small effect size of single SNPs. Additionally, GWAS suffer from a large multiple testing burden. To mitigate some of the issues with SNP-based GWAS, we consider the use of gene-based association tests [24, 25, 26]. These tests aggregate SNPs across a unit of analysis, typically a gene. These tests aggregate power from multiple SNPs to help mitigate the impact of small effect size on study power. Additionally, gene-based tests improve power by reducing the multiple testing burden. There are many existing gene-based association tests, which have varying relative power

depending on the underlying genetic architecture.

The mechanism by which genetic variants modulate complex traits is not well understood. If a given SNP is statistically associated with a trait in a GWAS, but it is not located in a region with established functionality, it is not clear how the SNP affects the trait biologically [3, 27, 4]. It has been demonstrated that many SNPs affect complex traits by modulating gene expression [28, 29]. In this way, we can think of gene expression levels as an intermediate phenotype that mediates the effect of SNPs on complex traits.

Existing gene-based association test PrediXcan [9] integrates eQTL data with GWAS data to augment the power of GWASs to detect associated genes. The TWAS methodology [8] extends the PrediXcan methodology to allow for the use of GWAS summary data, obviating the need for individual-level GWAS data. These methods first model the genetic component of gene expression using an eQTL dataset for which we have gene expression data and germline SNP data, and then leverage this gene expression model to construct weighted gene-based association tests. Throughout this chapter, we will refer to weighted gene-based association tests that leverage eQTL weights as TWAS-type tests, while acknowledging here that this methodology was developed by both TWAS and PrediXcan. The TWAS and PrediXcan test statistics can be thought of as weighted Sum tests. This framework can be extended to other sum of powered score tests [30, 31], and to other mediating phenotypes in addition to gene expression [32]. It has been convincingly shown that PrediXcan and TWAS increase the power of GWAS analyses for certain complex traits, and help inform the biological mechanism by which SNPs influence the expression of complex traits.

In practice, many applications of the TWAS methodology use eQTL data from normal tissues (i.e. non-cancer) to estimate the genetic component of gene expression. In this chapter, we integrate somatic and germline information to model the germline component of gene expression in tumors, and then leverage that model to construct weighted gene-based association for cancer phenotypes. By accounting for both the somatic and germline factors of gene expression in tumors, we integrate two sources of genetic information that have typically been analyzed in parallel. This can be thought of as a test for germline genetic association that leverages multiomics data, which is an area of recent interest [33]. This approach is motivated by evidence of germline influence on

the somatic environment of cancer cells [11], and evidence of genetic interaction between germline and somatic variants in cancer cells [12]. Our method is novel in the sense that we explicitly leverage the somatic and germline characteristics of tumors to conduct gene-based association testing, facilitating the discovery of novel genes associated with cancer phenotypes.

In particular, we estimate the genetic component of gene expression for adenocarcinoma tumors using data from The Cancer Genome Atlas [34]. Tumors undergo somatic alterations during tumorigenesis, which in turn modulate the gene expression levels in the cell. Although transcript regulation in tumors is complex, it is established that copy-number variation and DNA methylation status influence transcript abundance in tumors [35, 36]. These somatic alterations may obscure the effect of germline SNPs on gene expression level in tumors. Given this, we want to isolate the effect of germline SNP variation on gene expression level while controlling for the effect of somatic alterations. We additionally control for the effect of tumor purity, which has been shown to confound genomic analyses [37]. To do this, we first regress out the effect of copy-number, DNA methylation, and tumor purity on tumor gene expression, and then treat the residual as our mediating phenotype and proceed as in TWAS. We follow a similar methodology for isolating the effect of germline mutations on tumor expression levels as outlined in existing literature [38]. We demonstrate that germline genetic variants are significantly associated with tumor expression levels after controlling for somatic alterations. We extend upon their work by leveraging these associations to construct gene-based association tests for cancer phenotypes via a novel methodology.

2.2 Methods

2.2.1 Integrating eQTL data and GWAS data

We build upon the work of Xu et al [30] in incorporating the genetic component of gene expression into gene-based association testing, which in turn builds upon the work of TWAS [8] and PrediXcan [9]. We briefly review the methodology here.

First, we review the TWAS methodology. Consider that we have an eQTL dataset comprised of, for a set of genes, gene expression level Y^*, and k SNP genotype scores with additive coding in some neighborhood around the gene $X^* = (X_1^*, \ldots, X_k^*)'$. To ease

notation, we consider the case for a single gene. Using this data, we construct a model for the genetically regulated expression (GReX). Note that we are only considering SNPs within a neighborhood around the gene boundary, and thus we only model cis-regulating SNPs. We then estimate the following linear model:

$$Y^* = \sum_{j=1}^{k} w_j X_j^* + \boldsymbol{\epsilon} \tag{2.1}$$

In TWAS, weights $(\hat{w}_1, \ldots, \hat{w}_k)$ are estimated using a prediction method, such as a Bayesian linear mixed model [39] or an elastic net [18]. In our analysis, we use the elastic net. $(\hat{w}_1, \ldots, \hat{w}_k)$ are estimated weights corresponding to the GReX for the gene. Let us also define $\hat{W} = \text{diag}(\hat{w}_1, \ldots, \hat{w}_j)$.

Now, consider that we have a GWAS dataset. We have a binary phenotype Y measured on a set of n subjects. Each subject has been genotyped on a set of p SNPs, denoted $(X_1, \ldots, X_p)$. We estimate a generalized linear model to obtain a vector of score statistics. For a binary trait such as cancer status, we estimate the following joint model:

$$\text{logit}\{P(Y = 1)\} = \beta_0 + \sum_{j=1}^{p} X_j \beta_j \tag{2.2}$$

We can use this model to estimate score statistics for each of the p SNPs. Now, consider that we want to use these score statistics to construct a gene-based test of association. We consider an arbitrary gene and a set of k SNPs in some region around the gene: $X = (X_1, \ldots, X_k)'$. We have the corresponding score vector:

$$U = \big(U_1, \ldots, U_k\big)'$$

with $U_j = X_j'(Y - \hat{\mu}^0)$, where $\hat{\mu}^0$ is the estimated mean of Y under the null hypothesis:

$$H_0 : \boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)' = \mathbf{0} \tag{2.3}$$

Given the weights $\hat{W}$, we denote the weighted score statistics as:

$$U^* = \hat{W}U = (\hat{w}_1 U_1, \ldots, \hat{w}_k U_k)'$$

Thus, we can construct the following TWAS test statistic, where we aggregate weighted score statistics across a gene:

$$T_{TWAS} = \sum_{j=1}^{k} U_j^*$$

We now consider the extension of TWAS to other sum of powered score (SPU) tests [30]. Xu et al. make the observation that TWAS is equivalent to a weighted Sum test of multiple SNPs in a generalized linear model (2.2). If we consider that $(X_1, \ldots, X_k)'$ are SNPs within a gene region for some gene, TWAS can be thought of as a weighted Sum test with weights $(\hat{w}_1, \ldots, \hat{w}_k)$ and alternative hypothesis $\beta_1 = \cdots = \beta_k = \beta_c \neq 0$. We can use the estimated GReX weights $(\hat{w}_1, \ldots, \hat{w}_k)$ for the gene to a construct a variety of weighted gene-based tests. Xu et al (2017) propose a class of weighted sum of powered score (SPU) tests, defined as follows:

$$T_{SPU(\gamma)} = \sum_{j=1}^{k} (U_j^*)^\gamma$$

The parameter γ can be chosen adaptively via the so-called aSPU test [40]. Because there is no uniformly most powerful gene-based association test, it may be beneficial to use an adaptive test. The aSPU test has been shown to maintain power given different underlying genetic architectures.

An advantage of these weighted SPU tests is that they can be applied to summary statistic data. This is desirable because individual level genetic data is often not published due to privacy concerns. We note that the asymptotic distribution of a SPU test statistic depends on the covariance matrix of the score vector. In the case where we only have summary statistic information available, we approximate this covariance matrix by using available reference data [8]. Additionally, we replace score statistics U with the Z-statistics from the summary statistic data. We define $Z_j = \hat{\beta}_j / SE_j$, where $\hat{\beta}_j$ is the estimated marginal effect size and SE_j is the standard error. We make the substitution that $U = Z$, where $Z = (Z_1, \ldots, Z_p)'$, and proceed as before. Given these adjustments, we are able to construct the weighted SPU statistic given only summary statistic data from a GWAS.

2.2.2 Integrating somatic information

In this chapter we extend the TWAS framework to include somatic information and tumor purity, thus facilitating the analysis of tumor expression data. There is evidence that the use of different tissue expression weights, and the combination of tissue expression weights, significantly affects eQTL results and TWAS association results [41, 42].

Additionally, there is evidence that using tissue expression weights from tissue relevant to the trait in question may be most informative [10]. There is recent interest in differentiating eQTLs that modulate gene expression in normal cells versus eQTLs in that modulate expression cancer cells, and evidence that some eQTLs only modulate expression in cancer cells [43]. Given this, we want to use gene expression data from adenocarcinoma tumors, because we believe that gene expression levels in tumors may be most relevant to the lung cancer phenotype. Cancer cells undergo somatic mutations and epigenetic modulation during the process of tumorigenesis. It is known that germline genetic variants and somatic alterations in regulatory regions influence transcriptional regulation [44], and there is evidence that germline and somatic components relate to drug sensitivity [45, 46]. There is evidence that lung adenocarcinoma tumors have variable levels of purity, and that tumor purity is a determining factor of gene expression [37]. Given this, we believe integrating somatic and germline information alongside tumor purity data may lead to increased power to detect cancer-associated variants.

We want to isolate the effect of germline variants on gene expression, controlling for the effects of somatic mutation and tumor purity, and perhaps increasing the power of a TWAS-type test. Modeling somatic features alongside germline variants may increase the precision of our imputation model, leading to more reliable inference. The two somatic alterations that we consider are copy number variation and methylation status, mirroring the methodology of Li et al [38]. We also consider tumor purity, as measured by the consensus of the ESTIMATE, LUMP, IHC, and ABSOLUTE methods [37]. Consider, for an arbitrary gene, we have data on expression level T, copy number variation Sc, and methylation status M. Note that M could be a vector; that is, there could be multiple methylation probes corresponding to a single gene. For each tumor sample, we have a measure of tumor purity P. We then fit the following model:

$$T = \beta_0 + \beta_1 Sc + \beta_2 M + \beta_3 P + \boldsymbol{\epsilon} \tag{2.4}$$

We use the Anscombe transform of transcript level T for variance stabilization, which mirrors the methodology of the TWAS paper [8]. We include twenty PEER factors to account for latent factors driving gene expression [47], using the published protocol outlined by Stegle et al. [48]. This provides us with a vector of estimated residuals $\hat{\boldsymbol{\epsilon}}$, the

so-called residual expression. We use the residual expression as the eQTL phenotype, and proceed as in **(2.1)** to construct weighted gene-based association tests. That is to say, for a set of SNPs X^* within a $\pm$500 kb region of some gene with expression level T and estimated residual expression $\hat{\boldsymbol{\epsilon}}$, we construct the following model:

$$\hat{\boldsymbol{\epsilon}} = \sum_{j=1}^{k} w_j X_j^* + \boldsymbol{\omega} \tag{2.5}$$

Using the weights $(\hat{w}_1, \ldots, \hat{w}_k)$, we proceed with the weighted gene-based association tests.

We note that associations discovered via the use of functional weights derived from tumor expression data should be considered an extension on existing TWAS methodology, and not a substitute. Functional weights derived using germline information from normal tissue provide a general description of the gene expression environment. We consider functional weights derived from tumors for a particular cancer phenotype to be a finely tuned mechanism for obtaining additional information about the somatic environment of tumor cells of a particular type.

2.2.3 Constructing functional weights

To construct functional weights from the tumor expression data, we used the elastic net model [18], with the estimated residual expression level $\hat{\boldsymbol{\epsilon}}$ as the phenotype, and the predictors being all SNPs within a 500kb neighborhood of the gene, denoted here as X^*. We estimate the weights as follows:

$$(\hat{w}_1, \ldots, \hat{w}_k) = argmin_{(w_1,\ldots,w_k)} \sum_{i=1}^{n} (\hat{\epsilon}_i - \sum_{j=1}^{k} x_{ij}^* w_j)^2 + \alpha\lambda \sum_{j=1}^{k} |w_j| + (1-\alpha)\lambda \sum_{j=1}^{k} w_j^2 \tag{2.6}$$

We fixed $\alpha = .5$, and determined the value of λ using 5-fold cross-validation. If there was exactly one candidate SNP in the neighborhood X^*, we estimated a univariate linear regression. We considered a list of genes with coordinates provided by UCSC, which was downloaded off of the UCSC genome browser. In our analysis, we only considered those genes for which we could construct a non-null penalized regression model: that is, at least one effect size estimate was nonzero.

2.2.4 Lung adenocarcinoma data

We used data downloaded from The Cancer Genome Atlas (TCGA) to estimate the genetic component of residual gene expression level [34]. We downloaded data from the TCGA Cancer Browser on lung adenocarcinomas. The data contains 519 tumor samples for which we have RNAseq gene expression data (Illumina HiSeqV2, version 2015-02-24). There is gene expression data for 20,530 genes. We include only observations from tumors, and furthermore only observations that are 'primary tumors', excluding samples labelled 'recurrent tumor'. Primary tumors are those grown at the anatomical site where the tumor progression began, in this case the lung. In the case where multiple samples correspond to different aliquots from the same solid tissue, we only retain one sample. We additionally limit the data to only those samples for which we have copy number and methylation data. We then exclude all individuals for which we do not have matched germline SNP data. These quality control steps bring our sample size to 445.

The mapping from CNV and methylation probe to gene was provided in the TCGA data. The CNV data is generated via the gistic2 method [49], and is thresholded such that each CNV observation takes one of the following values: $\{-2, -1, 0, 1, 2\}$. There is one CNV value corresponding to each gene. The methylation data is taken from the HumanMethylation450k chip. We only used the subset of the methylation probes from the 450k chip that are present in the HumanMethylation27K chip. The 27K chip was used to adhere more closely to published methodology. The 27k chip contains information on 27,578 CpG methylation sites mapped to 14,495 genes. There can be more than one methylation probe mapped to a given gene. For each of the 20,530 genes, we regressed out the effect of copy number variation, methylation status, and tumor purity using the model described in equation (2.4). We use the resulting vector of estimated residual expression levels $\hat{\boldsymbol{\epsilon}}$ as our intermediate phenotype.

In order to compute the GReX of the estimated residual expression levels, we need germline genetic data from matched normal tissue samples. We used germline SNP data from normal cells matched to the set of 445 adenocarcinoma primary tumor samples. The germline SNP data was genotyped via the Affymetrix SNP 6.0 platform. This data was preprocessed such that all SNPs with minor allele frequency $< .05$ were removed.

After this step, there were $\sim 625,000$ such SNPs that were within a 500 Kb region of at least one gene. We used this data to demonstrate associations between germline genetic variants and residual expression as a proof of concept in section (**2.3.2**). Note that this is not the same data that was used to estimate the GReX; that was done with imputed data, as described below.

The expression residuals derived from the TCGA data were used in conjunction with the matched germline genetic data to construct a model for the genetic component of residual gene expression. Before fitting an elastic net model (2.6), we imputed the TCGA germline SNP data to the 1000G Phase 3 V5 reference panel [50], including only individuals of European ancestry in the reference panel. We performed the imputation with the Michigan imputation server [51]. After imputation, we removed all SNPs with imputation quality score $R^2 < .8$. To ensure that the same set of SNPs is used for weight building and test statistic construction, we pruned the TCGA data down to only those SNPs contained in both the 1000G LD reference data and the lung cancer GWAS described in section **(2.2.5)**. The rationale for this is further described in section **(2.2.6)**. We then excluded all SNPs with MAF$< .01$ and HWE p-value $< 1 \times 10^{-9}$, and pruned pairwise such that no two SNPs were in linkage disequilibrium with $r^2 > .9$. This left us with a set of 696,487 SNPs. We used this set of SNPs to model the genetic component of residual gene expression.

2.2.5 GWAS data

We conducted a meta-analysis of two lung cancer GWASs to investigate marginally associated SNPs, and to obtain a set of summary statistics for our weighted gene-based analysis. The EAGLE and OncoArray GWASs were downloaded from dbGap [52]. The EAGLE study is a case control study that was conducted in northern Italy. We analyzed four subsets of the OncoArray study on lung cancer: GRU, DS-CA-MDS, HMB, and CADM. The univariate association analyses were performed in PLINK using the –logistic command to perform a logistic regression with no covariates [53, 54]. This analysis assumes additive effects for minor alleles. We did not adjust for population stratification, because data hosted on the dbGap website demonstrate that the studies are well clustered by ancestry group, with the OncoArray and EAGLE studies containing only European ancestry individuals. All together, the OncoArray studies contain 8126

cases and 6491 controls. The EAGLE study contains 1945 cases and 1991 controls. In total, there are 10,071 cases and 8,482 controls.

The EAGLE study genotype data consists of a set of 561,466 SNPs genotyped on the Illumina HapMap550v3-B array. The OncoArray study was genotyped on a set of 492,435 SNPs from the custom OncoArray genotyping chip. When we perform the weighted gene-based association tests, we need GWAS summary statistics for the set of SNPs used to build the weights; i.e. those SNPs present in the TCGA adenocarcinoma data. Since these GWASs were genotyped on different sets of SNPs than one another, and in turn were genotyped on a different set of SNPs than the TCGA data, we need to perform imputation to match the sets of SNPs.

We imputed the EAGLE data and the OncoArray data independently before conducting the meta-analysis. The EAGLE study was imputed to the 1000G Phase 3 V5 reference panel using the Michigan Imputation Server [51]. After imputation, we removed all SNPs with imputation quality score $R^2 < .8$, Hardy-Weinberg p-value $< 10^{-9}$, call rate $< 90\%$, and minor allele frequency $< .01$. We were left with around 7 million SNPs.

For the OncoArray study, we performed summary statistic based imputation using the ImpG method [55]. Before imputation, we excluded SNPs with Hardy-Weinberg p-value $< 10^{-9}$, call rate $< 90\%$, and minor allele frequency $< .01$. As our reference panel, we used the European ancestry subset of the 1000G phase 3 data, which consists of 503 individuals of European descent genotyped on 8.4 million variants. In the 1000G reference data, we excluded all SNPs with call rate $< 95\%$, Hardy-Weinberg p-value $< 10^{-9}$, and minor allele frequency $< .01$. Likewise, we excluded all individuals with missing rate greater than 10%. We then imputed the OncoArray data using the set of 503 European ancestry individuals as a reference panel. Using the imputation quality score provided by the ImpG software, we then excluded all SNPs with imputation quality score $R^2 < .8$. After this analysis was complete, we were left with a set of around 4.4 million SNPs for the OncoArray study. We note that the EAGLE study was imputed with individual level data, while the OncoArray study was imputed via the ImpG method, which uses summary statistic data. We used ImpG for the OncoArray data due to the larger size of the OncoArray study.

2.2.6 Notes on the application of gene-based tests

When constructing weighted gene-based tests, we use weights derived from the TCGA eQTL data, score statistics taken from the GWAS meta-analysis, and the 1000G reference panel to estimate the SNP covariance matrix. We limit the number of individuals in the reference panel data to those 489 individuals of European descent included in the reference panel from the TCGA package. We remove SNPs from the 1000G reference data with HWE p-value $< 10^{-9}$, call rate $< 90\%$, and minor allele frequency $< .01$. We pruned the TCGA germline SNP data as described in section **(2.2.4)**. Note that we cannot conduct weighted gene-based tests accurately unless every SNP in the TCGA germline SNP data is also contained in the reference panel and in the meta-analysis. Excluding missing SNPs from the weights, which were estimated with penalized regression, causes the other weights to be miscalibrated given that individual effect size estimates from a multivariable model are interpretable only in the context of the full set of effect size estimates. With this in mind, we limit the TCGA data to only those SNPs present in both the 1000G reference panel data and the GWAS meta-analysis. In the case where a SNP is present in the weights and the reference panel but was not present in the meta-analysis, the TWAS FUSION package imputes the score statistic of the SNP using the ImpG method with no quality control. We note that this may introduce false positives if we impute a highly significant score statistic with a low confidence level. The FUSION package has controls on the proportion of imputed SNPs in a gene and the imputation quality of SNPs in a gene, but it's not clear that they rigorously control false positives introduced by this imputation process. By tailoring weight construction to the SNPs available in the GWAS analysis, we are able to strictly control the imputation quality of SNPs, and perhaps avoid the introduction of errors to this process.

We consider weighted gene-based association tests with three different sets of weights: unweighted tests (i.e. all weights are equal to 1), weights derived from tumor expression data, and weights derived from normal tissue eQTL data as precomputed and downloaded in the TWAS FUSION package [8]. For all of the results, note that we use a Bonferroni correction at the .05 level: that is, for $\alpha = .05$, we use significance threshold α/G, where G is the number of genes in the analysis.

For different sets of weights, we apply some combination of the SPU(1), SPU(2), and aSPU tests. In the results section, we refer to the SPU(1) and SPU(2) tests as the Sum and SSU tests, respectively. For the Sum and SSU tests, we are able to obtain an asymptotic p-value. For the aSPU test, we considered $\gamma \in \{1, 2, \ldots, 6, \infty\}$. We obtained p-values via simulation, using a step-up procedure.

2.3 Results

2.3.1 Somatic features are associated with tumor expression

As a proof of concept, we want to show that copy number variation, methylation status, and tumor purity have an effect on gene expression level. If they do not, then there is no need to control for their effect. Likewise, we want to be sure that there is a germline genetic effect on the residual expression level. If there is not, we will not be able to construct useful weights for a weighted gene-based analysis. In an analysis similar to the one performed by Li et al (2013), we show that gene expression level is modulated by CNVs, methylation status, and tumor purity. Then, we conduct a univariate SNP analysis to show that residual expression levels are modulated by germline SNPs. We used data on lung adenocarcinoma tumors downloaded from TCGA and preprocessed as described in section **(2.2.4)**.

For each of the 20,169 genes with at least one nonzero gene expression value, we estimate a linear model with copy number variation, methylation status, and tumor purity as predictors, as in equation (2.4). We obtain p-values for the effect of CNV, methylation status, and tumor purity using an F-test. The joint F-test was applied when multiple probes mapped to the same gene, allowing us to aggregate power over multiple probes. The coefficient of partial determination was used to determine the amount of variation explained by CNV, methylation status, and tumor purity, respectively. Twenty PEER factors were also included as covariates, thus the effect of CNV, methylation, and tumor purity was estimated after controlling for latent confounding. There were 17,776 with CNV data. At a significance level of $\alpha = .05$, we found that 71.7% of genes have expression level significantly associated with CNV. On average, CNVs account for 11.5% of the variation in gene expression after adjusting for other covaraites. Out of the 14,158 genes with methylation data, 42.9% have expression level signifcantly associated with

methylation status. On average, methylation status accounts for 1.7% of the variation in gene expression after adjusting for other covariates. Because all tumor samples have a measure of tumor purity, we are able to determine the influence of tumor purity on gene expression for all 20,169 genes. We find that 13.5% of genes have expression level significantly associated with tumor purity. On average, tumor purity accounts for 0.4% of the variation in gene expression.

These results demonstrate that copy number variation, methylation status, and tumor purity influence gene expression levels in tumor cells. Copy number variation in particular controls a large proportion of the variation of gene expression. PEER factors model latent factors driving gene expression, which explains why tumor purity explains a small additional proportion of variation. Nevertheless, we believe that it is best practice to model tumor purity explicitly, given that the data is available and it is a known determinant of gene expression levels. Because copy number variation and methylation status are gene specific predictors, they are less likely to explain the same variance as PEER factors. This explains the larger proportion of genes that have expression level significantly associated with copy number variation and methylation status, as opposed to tumor purity.

2.3.2 Germline variants are associated with residual expression

To show that residual expression level $\boldsymbol{\epsilon}$ is modulated by germline SNPs, we conduct a single-SNP analysis. This consists of regressing the estimated residual expression level $\hat{\boldsymbol{\epsilon}}$ of each gene against each SNP within 500Kb of the coding region of the gene. As described in section **(2.2.4)**, there were roughly $625,000$ such SNPs. Note that this is not the same set of SNPs that was used to estimate functional weights. Considering that one SNP could be within 500Kb of more than one gene, there were ~ 4.1 million pairwise SNP-gene tests. We tested association using a linear model, with a single SNP in additive coding as the predictor. Using a strict Bonferroni correction where $\alpha = .05/T$ and T is the number of pairwise SNP-gene tests, we found that 5850 SNP-gene pairs showed a significant SNP effect on residual gene expression level. 634 genes had at least 1 SNP significant at the Bonferroni cutoff. Considering that the same SNP may be tested pairwise with several different genes, these tests have a nuanced correlation structure,

and thus the Bonferroni correction is likely too conservative. Using the Benjamini-Hochberg FDR correction with a significance level of .1, we found that 36,015 SNP-gene pairs showed a significant SNP effect on residual expression level, and 5034 genes that had at least 1 significant SNP. Given this, it seems well established that there is a genetic component of residual expression level. We see that there is significant overlap between genes modulated by copy number variation, methylation, tumor purity, and cis-SNPs (figure 2.1).

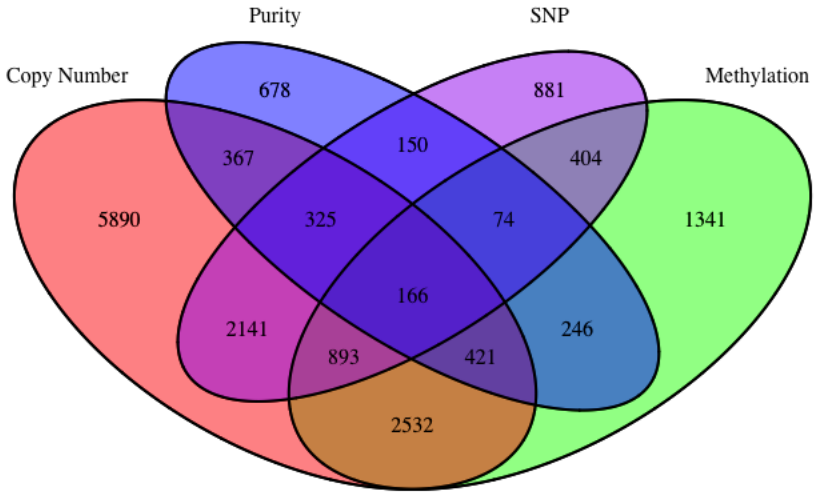

Figure 2.1: Counts of genes regulated by one of the four determinants of gene expression: copy number variation, methylation status, tumor purity, and cis-acting SNPs. Significance for cis-acting SNPs was determined using the Benjamini-Hochberg FDR correction with a significance level of .1. Significance for other determinants of gene expression was determined by an $\alpha = .05$ cutoff.

2.3.3 GWAS analyses and meta-analysis identify marginally associated variants

We conducted a meta-analysis of the two lung cancer GWASs described in section **(2.2.5)**; that is, the EAGLE study, and the subset of the OncoArray study. Only SNPs present in both studies with satisfactory imputation quality were included in

the meta-analysis. There were $\sim$ 4.2 million such SNPs. We conducted a fixed-effect meta-analysis using METAL [56]. The meta-analysis was sample size weighted, with the weight for each study being the number of cases in that study. One note on this analysis is that the p-values of the OncoArray and meta-analyzed GWAS are somewhat deflated. This is because the ImpG method penalizes off-diagonal elements of the LD matrix estimated from the reference panel as a way to compensate for overfitting [55], which has the effect of decreasing the value of imputed test statistics and thus inflating the corresponding p-values. As a consequence, we found the genomic inflation factors of the studies to be as follows: the OncoArray study has a genomic inflation factor of .95 and the meta-analysis has a genomic inflation factor of .97. The EAGLE study was imputed using individual level data, and thus does not suffer from p-value inflation. The genomic inflation factor of the EAGLE study is 1.03.

The OncoArray study shows significantly associated SNPs on chromosomes 5 and 15, while the EAGLE study shows significantly associated SNPs on chromosomes 12 and 15. The meta-analysis shows significantly associated SNPs on chromosomes 5, 6, and 15 (figure 2.2).

These results are consistent with previously identified risk loci for lung cancer [57]. We hope that, by using the information contained in the functional weights, we are able to discover significantly associated genes in regions where we do not see marginally significant SNPs.

2.3.4 Unweighted gene-based tests replicate GWAS results

As a baseline, we first present the results of the unweighted Sum and SSU tests. These tests were conducted on a set of 20,530 candidate genes, of which we could successfully estimate test statistics for 17,186. All SNPs within a $\pm$500 kb base pair region around the coding region of the gene were included, which is the case for all of the gene-based tests. This mirrors the methodology of the TWAS paper [8]. We find nineteen associated genes with the SSU and aSPU tests, and no associated genes with the Sum test (table 2.1). Given the fairly large number of SNPs considered and the assumption that causal genetic mutations are sparse, it is expected that the unweighted Sum test is underpowered in this situation. That is affirmed by the fact that no genes are identified with the Sum test. We see that the unweighted SSU test is powered to discover genes

given sparse causal effects, but is does not provide substantively new information from the meta-analyzed GWAS.

Gene	Chr	Gene Start	Gene End	Best Cis-SNP	SNP P-Val	Sum Test	SSU Test	aSPU Test
TBC1D2B	15	78287326	78369994	rs17486278	2.38E-33	2.46E-01	4.70E-13	1.00E-07
SNORA63	15	78383513	78383639	rs951266	9.94E-34	2.92E-01	1.05E-15	1.00E-07
SH2D7	15	78384926	78396393	rs951266	9.94E-34	2.91E-01	1.04E-15	1.00E-07
CIB2	15	78396947	78423877	rs951266	9.94E-34	4.82E-01	1.56E-23	1.00E-07
IDH3A	15	78441718	78462884	rs951266	9.94E-34	4.37E-01	9.75E-25	1.00E-07
ACSBG1	15	78463186	78527049	rs951266	9.94E-34	6.22E-01	5.65E-28	1.00E-07
DNAJA4	15	78556486	78574538	rs951266	9.94E-34	2.66E-01	4.30E-31	1.00E-07
WDR61	15	78575577	78591940	rs951266	9.94E-34	2.73E-01	1.19E-31	1.00E-07
CRABP1	15	78632665	78640572	rs951266	9.94E-34	1.91E-01	4.25E-34	1.00E-07
IREB2	15	78730517	78793798	rs951266	9.94E-34	2.61E-01	6.66E-36	1.00E-07
AGPHD1	15	78799905	78826012	rs951266	9.94E-34	9.52E-01	7.37E-41	1.00E-07
PSMA4	15	78832746	78841563	rs951266	9.94E-34	9.99E-01	5.98E-41	1.00E-07
CHRNA5	15	78857861	78887611	rs951266	9.94E-34	9.30E-01	2.63E-41	1.00E-07
CHRNA3	15	78887646	78913637	rs951266	9.94E-34	8.16E-01	1.88E-41	1.00E-07
CHRNB4	15	78916635	78933587	rs951266	9.94E-34	5.71E-01	1.48E-41	1.00E-07
ADAMTS7	15	79051544	79103773	rs951266	9.94E-34	1.98E-01	1.24E-43	1.00E-07
MORF4L1	15	79165122	79190081	rs951266	9.94E-34	9.97E-02	7.98E-45	1.00E-07
CTSH	15	79214091	79237420	rs951266	9.94E-34	8.57E-02	1.54E-44	1.00E-07
RASGRF1	15	79252288	79383215	rs951266	9.94E-34	2.93E-02	2.78E-43	1.00E-07

Table 2.1: Table displaying the p-values for the unweighted Sum and SSU tests, alongside the most significant SNP in the region and it's p-value. The significance cutoff is $.05/17186 \approx 2.91 \times 10^{-6}$. All genes in this table were significant under the SSU test, while none were significant under the Sum test.

We find that the only genes achieving significance after the Bonferroni correction are located on chromosome 15, near SNPs that obtained genome-wide significance in the meta-analyzed GWAS. Additionally, unweighted gene-based tests do not suggest dysregulation of the relevant gene in the same way that TWAS-type tests do. For this reason, we can say that there is a limited amount of new information to be gained from the unweighted tests.

2.3.5 New integrative method yields a novel association

Now, we use the weights derived from the residual gene expression of the TCGA lung adenocarcinoma tumors to construct weighted gene-based tests. There were 8558 genes for which we could estimate a non-null model for the GReX, thus our multiple testing correction is .05/8558. We present the results of the weighted Sum, SSU, and aSPU tests.

We find two significant genes from the weighted Sum test, four significant genes from the weighted SSU test, and four significant genes from the weighted aSPU test (Table 2.2). Three of these genes are located on chromosome 15, and were also identified as significant by the unweighted SSU test. One gene, TRIM31, is not nearby a SNP that reaches genome-wide significance, and is not identified by any other gene-based tests. In addition to aggregating information across a gene to increase power to detect associations, TWAS significance suggests mediation of lung cancer risk by the expression of these genes, indicating a mechanism that is not necessarily suggested by unweighted tests.

Gene	Chr	Gene Start	Gene End	Best Cis-SNP	SNP P-Val	Sum Test	SSU Test	aSPU Test
TRIM31	6	30070673	30080867	rs3129817	8.83E-07	1.13E-03	***3.22E-07***	***1.60E-06***
CRABP1	15	78632665	78640572	rs951266	9.94E-34	***4.61E-07***	***8.40E-24***	***1.00E-07***
AGPHD1	15	78799905	78826012	rs951266	9.94E-34	***8.08E-28***	***4.01E-31***	***1.00E-07***
CHRNA3	15	78887646	78913637	rs951266	9.94E-34	4.07E-03	***3.30E-16***	***1.00E-07***

Table 2.2: Table displaying the results for the weighted Sum, SSU, and aSPU tests using functional weights derived from the TCGA lung adenocarcinoma tumors, corrected for somatic alterations and tumor purity. The significance cutoff is $.05/8558 \approx 5.84 \times 10^{-6}$. Significant p-values are bolded and italicized.

The weighted Sum test identifies genes CRABP1 and AGPHD1 as significant. The weighted SSU and aSPU tests identify additional genes TRIM31 and CHRNA3. Of particular interest is gene TRIM31, which is not nearby a SNP of genome-wide significance, and is not identified by any other tests, weighted or unweighted. TRIM31 has been identified as an oncogene for pancreatic cancer [58], and has been identified in one study as a potential tumor suppressor for small cell lung cancer [59]. This illustrates a potential benefit of leveraging tumor-specific weights in the TWAS analysis of cancer phenotypes, namely leveraging information from genes that are differentially expressed in tumor tissue. Genes CRABP1, AGPHD1 (also called HYKK), and CHRNA3 are nearby the highly significant locus on chromosome 15, and have been identified in previous literature as potentially associated with lung cancer. AGPHD1 and CHRNA3 have been associated with lung cancer via GWAS-type analyses [60], while CRABP1 has been associated with lung cancer via an analysis of mRNA [61], providing further evidence that the dysregulation of CRABP1 affects lung cancer risk.

2.3.6 Functional weights from normal tissues yield further associations

As a point of comparison, we present the results from the analysis using weights derived from regular tissue expression data as precomputed by the authors of the FUSION package [8]. We consider analysis using the GTEx v6 whole blood weights ($N = 338$), the GTEx v6 lung weights ($N = 278$), and the YFS blood weights ($N = 1,264$). There are 2057 genes for which we have GTEx blood weights, 2937 genes for which we have GTEx lung weights, and 4700 genes for which we have YFS blood weights. In our analysis, we were able to estimate non-null test statistics for 1815 genes using the GTEx blood weights, 2605 genes using the GTEx lung weights, and 4421 genes using the YFS blood weights. Given this, each set of weights has a different multiple testing correction. We consider two competing factors that may drive the results from these methods. Firstly, we consider that weights derived from the lung tissue may be more relevant to the lung cancer phenotype than those weights derived from the whole blood cells. Secondly, we suspect that sample size plays a role in the usefulness of weights, given that we can more accurately model gene expression given a larger sample size. Thus, we may suspect that the YFS blood weights may be the most useful in finding significant genes.

The GTEx blood weights find four significantly associated genes, the GTEx lung weights find one significantly associated gene, and the YFS blood weights find six significantly associated genes (table 2.3). The GTEx Lung weights identify only gene CTSH on chromosome 15 as significant, which was also identified by the unweighted SSU test and is nearby SNPs that are highly marginally significant. This provides some evidence that the expression of CSTH in lung tissue modulates the effect of germline variation on lung cancer risk. The GTEx Blood and YFS Blood weights both identify gene PNPO as significant. Likewise, gene paralogs HLA-DQB1 and HLA-DQA1 are identified as significant by the GTEx blood weights and the YFS blood weights, respectively. This illustrates that the two blood-based weights identify similar genes. This analysis demonstrates the usefulness of the aSPU test, which uniquely identifies gene CASP8 as significant. This shows the aSPU test can provide substantively new information as opposed to the Sum and SSU tests.

All significant genes identified on chromosome 15 by the FUSION weights are nearby

Weights	Gene	Chr	Gene Start	Gene End	Best Cis-SNP	SNP P-Val	Sum Test	SSU Test	aSPU Test
GTEx Blood	CASP8	2	202098166	202152434	rs3769823	5.87E-06	4.15E-04	4.88E-05	***2.60E-05***
GTEx Blood	C4A	6	31949801	31970458	rs1150752	1.45E-04	***1.17E-05***	8.04E-05	***2.30E-05***
GTEx Blood	HLA-DQB1	6	32627244	32636160	rs9272426	2.33E-08	***1.13E-05***	4.75E-04	3.20E-05
GTEx Blood	PNPO	17	46018872	46025654	rs11079804	2.29E-06	***1.92E-05***	***1.07E-05***	3.90E-05
GTEx Lung	CTSH	15	79213400	79241916	rs2036527	4.12E-34	***3.49E-06***	***1.79E-05***	***8.40E-06***
YFS Blood	SLC12A7	5	1050499	1112150	rs2735948	3.77E-11	5.41E-05	***4.15E-08***	***1.00E-07***
YFS Blood	HLA-DQA1	6	32595956	32614839	rs9272426	2.62E-08	8.78E-02	***1.01E-05***	2.39E-05
YFS Blood	CATSPER2	15	43920701	43960316	rs524908	6.39E-07	1.11E-01	***9.08E-06***	4.80E-05
YFS Blood	IREB2	15	78729773	78793798	rs2036527	4.12E-34	***5.55E-16***	***4.72E-19***	***1.00E-07***
YFS Blood	PSMA4	15	78832747	78841604	rs2036527	4.12E-34	***1.67E-36***	***1.21E-36***	***1.00E-07***
YFS Blood	PNPO	17	46018872	46025654	rs11079804	2.29E-06	***4.33E-04***	1.04E-05	4.00E-05

Table 2.3: Table displaying the results of the weighted Sum, SSU, and aSPU tests using functional weights from the TWAS FUSION package. For the GTEx blood weights, the multiple testing correction was $.05/1815 \approx 2.75 \times 10^{-5}$. For the GTEx lung weights, the multiple testing correction was $.05/2605 \approx 1.92 \times 10^{-5}$. For the YFS blood weights, the multiple testing correction was $.05/4421 \approx 1.13 \times 10^{-5}$. Significant p-values are bolded and italicized.

a peak of highly significant SNPs, and were also identified by the unweighted SSU test. We see several novel associations that were not identified by other sets of weights. The gene PNPO has been associated with the development of other cancer types [62], but its association with lung cancer here is novel, as far as the authors are aware. We consider it's association particularly robust and interesting, given that it was identified by two different sets of weights, and is not located near a SNP of genome-wide significance.

We make a note of the issue brought up in **(2.2.6)**, namely that low quality variants must be imputed during the process of estimating weighted test statistics if the weights are not constructed in conjunction with the GWAS. Because we have already performed imputation on the lung cancer GWAS and pruned away SNPs of low imputation quality, we know that any SNPs imputed during the process of estimating weighted test statistics are of low quality. For four of the genes identified as significant by the TWAS weights, the most significant SNP in the gene locus is not present in the meta-analysis, and is imputed during the estimation of test statistics. These genes are C4A, HLA-DQA1, HLA-DQB1, and SLC12A7. This does not necessarily invalidate the significance of these genes, but may give some pause as to the robustness of the association. Quantile-quantile plots and genomic inflation factors (Appendix **A.1**) show that type I error is generally well controlled.

2.3.7 Functional weights without control for somatic features identify fewer genes

To investigate the usefulness of controlling for somatic alterations in the estimation of functional weights, we conducted an analysis using the non-residualized gene expression from the TCGA adenocarcinoma data as our eQTL phenotype. That is, for some gene with expression level T and a set of SNPs X^* within a $\pm$500 kb region of the gene, we estimate the functional weights as follows:

$$T = \sum_{j=1}^{k} w_j X_j^* + \boldsymbol{\epsilon} \tag{2.7}$$

Note that the gene expression level T has been Anscombe-transformed and adjusted for twenty PEER factors, as in the estimation of the somatic-adjusted weights described in **(2.2.2)**. Using the estimated functional weights $(\hat{w}_1, \ldots, \hat{w}_k)$, we construct weighted gene-based test statistics. For the full expression we were able to estimate 7776 non-null models for gene expression, as opposed to 8558 non-null models for the residual expression. Thus, using the residualized expression accounted for an increase of 10.1% non-null expression models. This is evidence that the residual expression level is easier to model, and thus more heritable, than the full expression. We present the significant results for the weighted Sum, SSU, and aSPU tests using the full expression weights in table 2.4. We identify one significant gene with the SSU and aSPU tests, and no significant genes with the Sum test. Given that we are able to estimate fewer non-null functional weights when we do not control for somatic features, and we are able to identify fewer associated genes with these functional weights, there is evidence that controlling for somatic features improves the usefulness of functional weights estimated from tumor data.

Gene	Chr	Gene Start	Gene End	Best Cis-SNP	SNP P-Val	Sum Test	SSU Test	aSPU Test
CHRNA5	15	78857861	78887611	rs951266	9.94E-34	3.29E-02	***1.01E-06***	***3.10E-06***

Table 2.4: Table displaying the results of the weighted sum test using functional weights derived from the full TCGA gene expression data. The significance cutoff is $.05/7776 \approx 6.43 \times 10^{-6}$. Significant p-values are bolded and italicized.

2.3.8 Comparison of different functional weights

While we expect the performance of weighted gene-based tests to be highly dependent on the functional weights used, we may expect some similarity in performance based on the tissue type used to estimate the functional weights. To compare the similarity of the different sets of weights, we compare the Sum test association statistics of the different weighted gene-based tests. It may not be informative to compare the weights directly, because the candidate SNP sets used in weight estimation are different for different sets of weights. Additionally, one set of weights may select some SNP A from a group of correlated SNPs, while a second set of weights may select a different SNP B from the same group of correlated SNPs, mostly due to random variation. Given the underlying relationship between the Sum test association statistics and genetic correlation, we believe that using the association statistics to assess the similarity of the different weights is informative. For each pairwise comparison, we used only the genes in common among the two sets of weights; that is, those genes for which we could estimate a non-null model for gene expression for both sets of weights. We display the correlation across Sum test association statistics for the different sets of functional weights in table 2.5. Additional details on the similarity of the performance of weighted Sum, SSU, and aSPU tests is located in Appendix **A.2**.

	GTEx Lung	GTEx Blood	TCGA Adjusted	TCGA Unadjusted	Unweighted
YFS Blood	.533 (994)	.705 (821)	.232 (2372)	.222 (2123)	-.008 (4346)
GTEx Lung		.782 (949)	.590 (1278)	.500 (1160)	-.020 (1801)
GTEx Blood			.421 (869)	.388 (760)	-.047 (1257)
TCGA Adjusted				.792 (5640)	-.008 (8558)
TCGA Unadjusted					-.004 (7776)

Table 2.5: Table displaying the correlation across Sum test association statistics for different sets of functional weights. The first number is the Pearson correlation coefficient r. The number in parentheses is the number of genes in common among the two sets of weights.

We find that the Sum test association statistics from the GTEx Lung and GTEx Blood weights are best correlated; this is reasonable, given that both sets of weights were generated from the same study. Compared to weights derived from normal tissue, the Sum test association statistics from the adjusted TCGA weights are best correlated with the Sum test association statistics from the GTEx Lung weights; this is reasonable, given

that the TCGA weights and GTEx Lung weights were generated from the same tissue type. The Sum test association statistics corresponding to the adjusted TCGA weights are better correlated with the Sum test association statistics corresponding to the three sets of weights derived from normal tissue than are the Sum test association statistics corresponding to the unadjusted TCGA weights. This indicates that adjusting for somatic features and tumor purity with a residualized approach better approximates the gene expression environment of normal tissue by removing variation in gene expression due to somatic features. Further evidence of this, demonstrated by imputing GTEx gene expression into TCGA, is described in Appendix **A.3**. We see that Sum test association statistics corresponding to the adjusted TCGA weights are well correlated with the Sum test association statistics from the unadjusted TCGA weights, although there is substantial difference between the two. This further demonstrates the importance of adjusting for somatic features and tumor purity. Sum test association statistics from all sets of functional weights are essentially uncorrelated with Sum test association statistics corresponding to the unweighted test, indicating that functional weights significantly drive the performance of the weighted Sum tests, as opposed to the underlying GWAS.

2.4 Discussion

In this work, we present a new methodology for integrating germline and somatic variation to model transcriptomic variation in tumors for the identification of genes associated with lung cancer. Functional weights derived from tumors may be useful for detecting novel associations, given the nuanced nature of gene expression in tumor cells. In particular, we consider data on lung adenocarcinomas downloaded from The Cancer Genome Atlas for which we have gene expression data, copy number variaton data, and methylation data from tumors, and matched germline SNP data. We also have tumor purity data, as estimated by Aran et al. [37]. We mirror existing method [38] to isolating the genetically regulated component of gene expression in tumor cells by controlling for the effect of other somatic alterations. We demonstrate that somatic alterations in copy number variation, CpG island methylation, and tumor purity affect gene expression in adenocarcinoma tumor cells. We also show that the residual expression level after controlling for somatic alterations and tumor purity is modulated by germline genetic

variants by analyzing pairwise SNP-gene associations. With this established, we model the genetic component of residual expression level that is modulated by cis-acting SNPs. Using the model for the genetic component of residual expression as weights, we construct weighted gene-based association tests, considering the Sum, SSU, and aSPU tests. We apply the weighted gene-based association tests to a meta-analysis of lung cancer genome-wide association studies, and compare our results to those obtained by applying unweighted gene-based tests, weights unadjusted for somatic features, and multiple sets of weights derived by the authors of the FUSION package.

We find that the use of functional weights derived from tumor expression data and adjusted for somatic features identify four genes associated with lung cancer. Three of these are also identified by unweighted gene-based tests, while one is uniquely identified by the residualized tumor weights. This gene, TRIM31, is not identified by other sets of functional weights, and is not nearby a SNP that achieves genome-wide significance. We note that different approaches to estimating functional weights find several different associated genes on chromosome 15; this is likely due to the effect of LD on SNPs in the region, leading to correlation among the imputed expression levels.

There are some complications in the modeling of tumor expression eQTL data that may have affected our analysis. Firstly, we note that the gene expression data from TCGA on the adenocarcinomas is extracted from tumor biopsies, which are not homogeneous samples of cancer cells. Rather, tumor biopsies are a mixture of different cell types, some of which are cancer cells. There is evidence that eQTLs function differently in different cell types, and that many eQTLs identified from the analysis of so-called "bulk tumor" data may in fact be attributable to eQTLs in non-cancer cells [43]. We present our analysis as a method for bulk tumor data that identifies novel associations, but stop short of directly attributing these associations to the effect of eQTLs in cancer cells. Additionally, there is evidence that methylation may be modulated by germline genetic variation via so-called mQTLs [63]. Our method first regresses out the effect of methylation status on gene expression and estimates functional weight using the residual expression level. This approach may introduce collider bias if the methylation status of a certain gene is under genetic control and is affected by tumorigenesis in lung cancer.

This method of constructing functional weights could be applied to other cancers.

This methodology requires the existence of tumor expression data, data on the somatic features of tumors, and matched data on germline genetic variation. The Cancer Genome Atlas hosts this data for several cancer types, thus facilitating the application of this methodology for different types of cancer.

During the revision of this paper, we became aware of another published paper that uses similar methodology to control for the confounding effect of somatic variation and tumor purity when estimating functional weights from tumors [41]. Our method was developed without knowledge of this paper, and our use of the SSU and aSPU tests (in addition to the Sum test) and application to lung cancer differ from theirs.

2.5 Data Availability

The GWAS datasets in this project were downloaded from dbGaP, with accession numbers phs000093.v2.p2 and phs001273.v2.p2. The gene expression and somatic alteration data for the TCGA lung adenocarcinomas was downloaded from the TCGA data portal, and the TCGA germline SNP data was downloaded from dbGaP. The FUSION software, corresponding weights, and reference panel were downloaded from the TWAS FUSION website: `http://gusevlab.org/projects/fusion/`. The 1000 Genomes reference data was downloaded from the 1000 Genomes project.

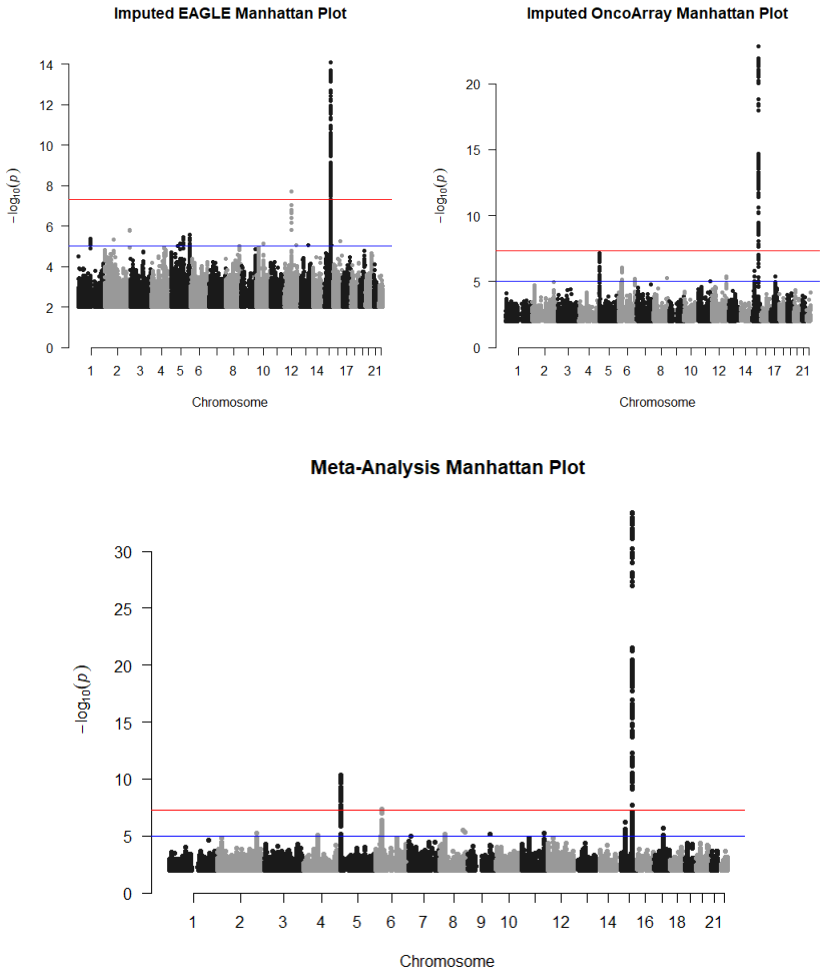

Figure 2.2: The top two panels display the marginal p-values obtained from GWAS analyses of the EAGLE and OncoArray studies, respectively. The red line denotes the threshold for genome-wide significance at 5×10^{-8}. The blue line denotes the threshold for suggestive significance at 1×10^{-5}. We see that the EAGLE studies has peaks on chromosomes 12 and 15, and the OncoArray study has peaks on chromosomes 5 and 15. The bottom panel displays the marginal p-values obtained from a meta-analysis of the OncoArray and EAGLE GWAS. We see peaks on chromosomes 5, 6, and 15. The significance peak on chromosome 12 disappears after meta-analysis.

Chapter 3

Penalized regression and model selection methods for polygenic scores on summary statistics

3.1 Introduction

The polygenic model of inheritance predicts that the genetic basis of complex phenotypes consists of small effects from thousands of genetic variants. Genome-wide association studies (GWAS) have affirmed this model, identifying many genetic variants that are associated with complex traits [1]. However, marginally associated markers explain only a limited proportion of the heritability of many traits [2]. Polygenic risk scores, defined as a linear combination of individual SNP effects, have been used to quantify the genetic component of some complex phenotypes. Polygenic risk scores estimated from GWAS have been useful for predicting some clinical phenotypes [64, 7, 6]. Polygenic risk scores have also been used to infer the genetic architecture of complex traits [65, 66]. The simple polygenic risk score is obtained by summing marginal genetic effects across all SNPs. Extensions on this method include thresholding [13], in which SNPs with marginal p-values below a certain cutoff point are excluded, and pruning and thresholding, which combines thresholding with the exclusion of highly correlated SNPs via pruning [14]. These methods use only marginal effect size estimates, and do not attempt to construct

a joint model that estimates effect sizes under linkage disequilibrium. Thus, it can be said that they do not attempt to model the true structure of the genetic effects. We propose a method for constructing polygenic risk scores that integrates marginal effect size estimates with publicly available reference panel data, which is used to estimate linkage disequilibrium. By estimating effect sizes under linkage disequilibrium, we more closely model the true structure of the genetic effects. This allows us to capture more of the genetic heritability, as shown via simulation and application to real data.

Popular methods LDPred [15], LassoSum [16], and JAMPred [67] estimate joint models that account for linkage disequilibrium. Recently published methods in this area include PRS-CS [68] and SBayesR [69]. Other methods, such as EBPRS [70], leverage the available GWAS data to estimate a distribution of SNP effect sizes that is leveraged to adjust the marginal SNP effects. These methods do not necessitate genome-level data. They use publicly available reference data and published summary statistics from GWAS. This is important because often the published results from a GWAS do not include genome-level information. Our software implements new penalized regression methods for estimating polygenic risk scores that model linkage disequilibrium. Given a reference panel and marginal SNP effects, the software constructs a joint penalized regression model. We extend upon the work of Shin et al [16], who propose using the LASSO penalty, to other penalties: namely the truncated LASSO penalty [17] and the elastic net penalty [18]. These penalties have some theoretical benefits as compared to the LASSO penalty; the TLP may induce more sparsity when the truth is sparse and produce less biased estimates, while the elastic net may handle correlated covariates more stably. The TLP also has application for valid inference that may be useful [71]. We call these methods TlpSum and ElastSum, respectively.

Additionally, we describe some criteria that can be used for model selection in the case where we do not have access to validation data. In an application where we have access to validation data, we may select the model that maximizes the correlation between the estimated polygenic risk score and the validation phenotype. In the case where we don't have access to validation data, we may want still want to perform model selection on a set of candidate polygenic risk scores. Our methodology approximates the model fitting criteria AIC and BIC in the situation where we do not have individual level data. These methods, so called 'pseudo AIC' and 'pseudo BIC', approximate the AIC and

BIC criteria for an estimated polygenic risk score given GWAS summary statistics and a reference panel. These methods extend upon the existing model selection criterion pseudovalidation [16]. Pseudovalidation controls model degrees of freedom by weighting SNPs by their local false discovery rate. This method is somewhat ad hoc, and local FDR estimation may not perform reliably when we don't have a dense set of summary statistics available. This is often the case with published summary statistics, which may only include SNPs above some marginal significance threshold. Pseudo AIC and pseudo BIC leverage the well established theory of AIC and BIC to impose a penalty on degrees of freedom. This leads the pseudo AIC and BIC to select sparser models that more accurately represent the truth, as demonstrated via simulation study in **(3.3.2)** and **(3.3.2.1)**. We also show that pseudo BIC and pseudo AIC select models with better predictive performance on out-of-sample data in certain simulation settings and in application to a large GWAS of blood lipid levels.

Lastly, we propose a metric for assessing the predictive accuracy of a polygenic risk score in the case where we have only summary statistic information on our out-of-sample data. We call this metric 'quasi-correlation'. Given an estimated polygenic risk score and a reference panel, this method allows us to estimate the predictive r^2 of the polygenic risk score as applied to an out-of-sample dataset comprised of summary statistics. Thus, we can determine which model fits best on out-of-sample data given a candidate set of polygenic risk scores. This enables us to select a validated polygenic risk score ready for use on other data. These methods allow us to use published summary statistic data of large sample size to assess the predictive accuracy of polygenic risk scores, broadening the scope of application. We demonstrate the utility of these methods by applying them to large GWAS summary statistic data on lipids in **(3.3.3)**. Applications to lung cancer and height are located in section **(3.3.4)** and Appendix **B.2**, respectively.

The central aim of this chapter is to assess the predictive performance of the methods for polygenic risk score estimation and corresponding model selection. We demonstrate that TlpSum and ElastSum often perform similarly to LassoSum as measured by predictive accuracy, but outperform LassoSum when applied to data with substantial allelic heterogeneity. We show that our proposed model selection methods, pseudo AIC and pseudo BIC, select models with better predictive performance on out-of-sample data than pseudovalidation in certain applications. A secondary but relevant concern is

the characterization of the fitted models in terms of sparsity. Given a set of models with similar predictive performance on out-of-sample data, it is often desirable to select the most parsimonious model, which is the so-called principle of parsimony. A more parsimonious model that maintains good predictive performance better facilitates interpretation and certain applications of polygenic risk scoring. One useful example is the application of polygenic risk scoring to two-stage least squares regression for causal inference [22]. In this type of applications, overparameterized PRS models may contain substantial pleiotropic effects, making causal inference difficult with violated modeling assumptions. In this chapter, we demonstrate via simulation that TlpSum, pseudo AIC, and pseudo BIC impose additional sparsity on their selected models. We discuss the results with respect to the primary aim of prediction, and characterize the selected models to possibly explain the performance difference among the various methods.

An implementation of the methods described in this chapter is provided in our R package 'penRegSum', located at `https://github.com/jpattee/penRegSum`. This package directly interfaces with PLINK for ease of computation [53, 54]. We note that the authors of the LassoSum package [16] provide extensive functionality for the estimation and selection of polygenic scores on summary data. Our package can be considered an extension on their work that is best used in conjunction with their package.

3.2 Methods

3.2.1 Penalized regression with summary statistics

Consider that we have a linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{3.1}$$

with $\mathbf{X}$ denoting an $n \times p$ design matrix, $\mathbf{y}$ denoting a vector of observed outcomes, and $\boldsymbol{\epsilon} \sim MVN(\mathbf{0}, \sigma^2\mathbf{I})$ for some σ^2. Ordinary least squares estimates are obtained by minimizing the sum of squared errors

$$f(\boldsymbol{\beta}) = SSE = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{3.2}$$

In the case where p is large and $\boldsymbol{\beta}$ may be sparse, penalized regression models can be useful. Penalized regression models introduce a penalty term to the objective function.

This penalty term is typically a function of $\boldsymbol{\beta}$, and is denoted $J(\boldsymbol{\beta})$. Additionally, consider now that $\mathbf{y}$ is a standardized response vector, and $\mathbf{X}$ is a standardized design matrix. This yields the following objective function:

$$f(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + J(\boldsymbol{\beta}) = \mathbf{y}^T\mathbf{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} + J(\boldsymbol{\beta}).$$

Shin et al [16] note that, given some approximations, penalized regression can be used to estimate polygenic risk scores in the case where only summary statistics are available. Consider that we have two separate datasets: one of summary statistics, and one of reference data. We use the summary statistic data to estimate univariate SNP effects, and the reference data to estimate the correlation matrix of the SNPs. Let us denote the standardized phenotype vector from the summary statistic data divided by $\sqrt{N}$ (the sample size of the summary statistic data) as $\mathbf{y}_s$. Let us denote the standardized summary statistic design matrix as $\mathbf{X}_s$. Let us denote the standardized SNP reference data as $\mathbf{X}_r$. We can now define the quantity

$$\mathbf{r} = \mathbf{X}_s^T\mathbf{y}_s,$$

where $\mathbf{r}$ represents the SNP-wise correlation between the SNPs and the phenotype in the summary statistic data. We also define $\mathbf{R}$, which is the correlation matrix as estimated from the reference data as

$$\mathbf{R} = \mathbf{X}_r^T\mathbf{X}_r.$$

Given these approximations, we can now define an objective function for the estimation of polygenic risk scores using summary statistic data. That objective function is as follows:

$$f(\boldsymbol{\beta}) = \mathbf{y}^T\mathbf{y} + \boldsymbol{\beta}^T\mathbf{R}\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\mathbf{r} + J(\beta).$$

Shin et al. note that this is no longer strictly a penalized regression problem, due to the use of two different design matrices $\mathbf{X}_r$ and $\mathbf{X}_s$. This may lead to unstable and non-unique solutions. They propose regularizing $\mathbf{R}$ as

$$\mathbf{R_s} = (1 - s)\mathbf{X}_r^T\mathbf{X}_r + s\mathbf{I} \tag{3.3}$$

for some $0 < s < 1$. This regularization ensures that we have an objective function in the form of a LASSO problem, as proven in previous literature [16]. Substituting $\mathbf{R}_s$

for $\mathbf{R}$ yields the following tractable objective function:

$$f(\boldsymbol{\beta}) = \mathbf{y}^T\mathbf{y} + \boldsymbol{\beta}^T\mathbf{R}_s\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\mathbf{r} + J(\boldsymbol{\beta}). \tag{3.4}$$

We now turn our attention to the penalty term $J(\boldsymbol{\beta})$. Shin et al propose using the LASSO penalty, which is a popular penalized regression method for high dimensional problems. The LASSO induces sparsity in $\boldsymbol{\beta}$, performing parameter selection and estimation simultaneously. In LASSO, the penalty term is the L1 penalty, ie $J(\boldsymbol{\beta}) = \lambda||\boldsymbol{\beta}||_1 = \lambda\sum_i|\beta_i|$, where λ is a tuning parameter selected via a model selection method. The LASSO tends to bias parameter estimates towards zero in a uniform manner. This causes biased effect size estimates. To mitigate bias issues, we propose the use of the Truncated Lasso Penalty, or the TLP [17]. The TLP can be expressed as follows: $J(\boldsymbol{\beta}, \tau) = \lambda\sum_i min(|\beta_i|, \tau)$, where λ and τ are tuning parameters determined via model selection. The TLP does not penalize effect size estimates above some threshold τ, which may decrease bias. Additionally, we propose the use of the elastic net penalty [18], namely $J(\beta) = \alpha\lambda||\boldsymbol{\beta}||_1 + (1-\alpha)\lambda||\boldsymbol{\beta}||^2$. This method has some advantages of the LASSO while retaining some advantages of ridge regression, such as stable estimation of highly correlated covariates. The application of the elastic net penalty to summary statistics is called ElastSum. We note that the use of the regularized covariance matrix approximates a sort of elastic net already. To see that, consider the expanded expression of equation (3.4):

$$f(\boldsymbol{\beta}) = \mathbf{y}^T\mathbf{y} + (1-s)\boldsymbol{\beta}^T\mathbf{R}\boldsymbol{\beta} + s\boldsymbol{\beta}^T\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\mathbf{r} + J(\beta). \tag{3.5}$$

We note that the term $s\boldsymbol{\beta}^T\boldsymbol{\beta}$ approximates the L2 penalty. Given this, we are unsure of the utility of the elastic net penalty in many cases. We also note that this may affect the TLP estimates. If $s > 0$, the objective function will function somewhat like an elastic net, meaning the TLP may not induce its characteristic sparsity.

3.2.2 Notes on the application of penalized regression

If SNPs are in high linkage disequilibrium, then it may be difficult or impossible to arrive at stable estimates for $\boldsymbol{\beta}$ given the objective function (3.4). In this case, we advise performing LD clumping on the data prior to estimating a penalized regression model. Even after clumping, it is often the case that convergence is impossible (or very

slow) unless a sufficiently large value of s is chosen. The value of s chosen depends on the sparseness of the genetic signal. For phenotypes with a sparse genetic signal, choosing an s as small as 0 may work, and choosing $s \sim .1$ should ensure good convergence and fairly sparse effect size estimates. For phenotypes with more dense signal, we recommend experimenting with larger values of s.

As Shin et al note [16], penalized regression generates effect size estimates that are not appropriately scaled. Considering that penalized regression is conducted on normalized data, we can say these estimates are scaled as correlations. If we want to use a polygenic risk score generated via penalized regression to estimate genetic risk, we need to appropriately scale our estimates. We have the following expression for the effect size estimates: $\hat{\beta}_i^{unstandardized} = \hat{\beta}_i \frac{sd(\mathbf{y})}{sd(\mathbf{X}_i)}$, where X_i is column i in the reference panel $\mathbf{X}_r$, $\mathbf{y}$ is the phenotype vector, and $\hat{\beta}_i$ are the effect size estimates produced by the penalized regression; that is, those estimates minimizing objective function (4).

Existing literature demonstrates that estimation by LD blocks improves the predictive performance of penalized regression methods applied to summary statistics [16]. We also recommend performing estimation by LD blocks, and do so in this chapter unless otherwise specified. We used LD blocks as defined by the LDetect method [72].

If some SNPs in the summary statistic data are not included in the reference data, we are not able to incorporate those SNPs into our objective function (5) as currently formulated. Simply excluding these SNPs from our analysis may result in information loss. A straightforward approach, as described by the authors of the LassoSum paper, is to treat these SNPs as though they are mutually independent. We take an identical approach here. We define $\boldsymbol{\beta}_0$ as a subvector of $\boldsymbol{\beta}$ corresponding to those SNPs missing from the reference panel, and the corresponding submatrix $\mathbf{R}_0$ of $\mathbf{R}$ as containing all zero entries. Thus, we can reformulate the objective function as follows:

$$f(\boldsymbol{\beta}) = \mathbf{y}^T\mathbf{y} + (1-s)\boldsymbol{\beta}^T\mathbf{R}\boldsymbol{\beta} + s\boldsymbol{\beta}^T\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\mathbf{r} + (1-s)\boldsymbol{\beta}_0^T\boldsymbol{\beta}_0 + J(\beta)$$

This approach to handling SNPs missing from the reference panel has been implemented in our R package.

We estimate penalized regression models on summary statistics via coordinate descent [73]. The details of this algorithm are located in Appendix **B.1**.

3.2.3 Pseudo AIC / BIC

It may be the case that we do not have access to validation data for use in model selection. In this case, it is desirable to have a model selection technique to select tuning parameters that can be applied to summary statistic data. Shin et al propose the pseudovalidation method for this purpose [16]. Pseudovalidation approximates the correlation between the predicted phenotypes and the phenotypes from the summary statistic data. One drawback of their method is that it may tend to overfit the model, as the pseudovalidation criteria will tend to increase as parameters are added to the model. This is somewhat controlled for by their weighting of marginal p-values by local FDR, but this isn't necessarily a rigorous approcah. We propose a method of estimating model fitting metrics AIC and BIC using only summary statistics and a reference panel. We believe these methods may select less overfit and therefore sparser models.

Suppose that we have trait $\mathbf{Y}$ measured on N subjects. We have data on p SNPs for each subject, giving us $N \times p$ design matrix $\mathbf{X}$. Say $\mathbf{X}, \mathbf{Y}$ are centered at zero. We assume model (1). Given this, we have the following likelihood function:

$$L = \prod_{i=1}^{N} p(y_i|x_i; \boldsymbol{\beta}, \sigma^2) \propto \sigma^{-N} \exp[(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})/(2\sigma^2)]$$

and the following log-likelihood function:

$$l = C - N * ln\sigma - \frac{1}{2\sigma^2}(\mathbf{Y}'\mathbf{Y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{Y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}).$$

C does not depend on the parameters, and so can be ignored.

Placing this problem in our summary statistic framework, we want to estimate $\mathbf{Y}'\mathbf{Y}$, $\mathbf{X}'\mathbf{Y}$ and $\mathbf{X}'\mathbf{X}$ from reference data and summary statistics. Suppose we have univariate summary statistics $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)'$ and corresponding variances $\widehat{var}(\hat{\beta}_j)$, which quantify the marginal associations between phenotype $\mathbf{y}$ and each of the p SNPs in design matrix $\mathbf{X}$. We also have reference panel $\mathbf{X}_r$. Denote the variance of a given SNP j, as estimated from the reference panel, as $\hat{s}_j^2$. We define that there are N individuals in $\mathbf{X}$ and n individuals in $\mathbf{X}_r$. Because of the differing sample sizes, we want to compare quantities that have been normalized by sample size when estimating the log-likelihood. With this in mind, we define the following approximations:

$$\frac{1}{N}\widehat{\mathbf{X}^T\mathbf{X}} = \boldsymbol{\Sigma} = \frac{1}{n}\mathbf{X}_r^T\mathbf{X}_r, \tag{3.6}$$

$$\frac{1}{N}\widehat{\mathbf{Y}'\mathbf{Y}} = (N * \hat{s}_j^2 * \widehat{var}(\hat{\beta}_j) + \hat{s}_j^2 * \hat{\beta}_j^2), \tag{3.7}$$

$$\frac{1}{N}\widehat{\mathbf{X}'\mathbf{Y}} = diag(\mathbf{\Sigma})(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)' = (\hat{s}_1^2\hat{\beta}_1, \hat{s}_2^2\hat{\beta}_2, \ldots, \hat{s}_p^2\hat{\beta}_p)'. \tag{3.8}$$

Note that $diag(\mathbf{\Sigma}) = (\hat{s}_1^2, \ldots, \hat{s}_p^2)$. In practice, we advise taking some central tendency the expression for $\frac{1}{N}\widehat{\mathbf{Y}'\mathbf{Y}}$ across the p SNPs to obtain a more accurate approximation. We have found the median to work well.

We briefly justify approximations (3.6), (3.7), and (3.8) here. Expression (3.6) simply describes the approximation of the covariance matrix by a reference panel. Expression (3.7) can be derived as follows. Consider that, for single linear regression, $\hat{\beta}_i = \frac{\sum_{j=1}^N x_{ji}y_j}{N*s_i^2}$ and $\widehat{var}(\hat{\beta}_i) = \frac{\sum_{j=1}^N (y_j - x_{ji}\hat{\beta}_i)^2}{N^2*s_i^2}$. Thus we have: $\frac{1}{N}\widehat{\mathbf{Y}'\mathbf{Y}} = N * s_i^2 * \frac{\sum_{j=1}^N (y_j - x_{ji}\hat{\beta}_i)^2}{N^2*s_i^2} + s_i^2 * \hat{\beta}_i^2$. Expanding the squared term and using the fact that $\hat{\beta}_i = \frac{\sum_{j=1}^N x_{ji}y_j}{N*s_i^2}$, we have: $N * s_i^2 * \frac{\sum_{j=1}^N (y_j - x_{ji}\hat{\beta}_i)^2}{N^2*s_i^2} = \frac{\sum_{j=1}^N (y_j - x_{ji}\hat{\beta}_i)^2}{N} = \frac{\sum_{j=1}^N y_j^2}{N} - 2\hat{\beta}_i^2 * s_i^2 + \hat{\beta}_i^2 * s_i^2$. Thus, we conclude that $\frac{1}{N}\widehat{\mathbf{Y}'\mathbf{Y}} = \frac{\sum_{j=1}^N y_i^2}{N}$. Now, we examine expression (3.8). Given that $\hat{\beta}_i = \frac{\sum_{j=1}^N x_{ji}y_j}{N*s_i^2}$, it is straightforward that $\frac{\sum_{j=1}^N x_{ji}y_j}{N} = s_i^2\hat{\beta}_i$. Expression (3.8) follows from this. Note that expressions (3.6), (3.7), and (3.8) have been derived assuming single linear regression. Given some mild assumptions and changes in interpretation, these expressions are still valid when summary statistics are estimated using multiple regression, i.e. in a GWAS that includes non-SNP covariates. Details are in Appendix **B.4**.

To estimate the log likelihood of a linear regression model, we must estimate the sum of squared errors (3.2). Additionally, we must estimate the residual variance $\tilde{\sigma}^2$. We estimate the SSE with the penalized regression estimates, which we denote $\hat{\boldsymbol{\beta}}_P$. Note that these differ from the marginal effect size estimates $\hat{\boldsymbol{\beta}}$. We estimate the residual variance with the ordinary least square estimates, denoted $\hat{\boldsymbol{\beta}}_{OLSE}$.

To estimate $\tilde{\sigma}^2$, we use the ordinary least squares estimates

$$\hat{\boldsymbol{\beta}}_{OLSE} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

For a linear regression, we have residual variance estimated as follows:

$$\tilde{\sigma}^2 = MSE = \frac{(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})}{N - q}$$

where q is the degrees of freedom. In the case of linear regression, this is equivalent to the number of parameters. We substitute $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_{OLSE}$ to yield

$$\tilde{\sigma}^2 = \frac{\mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}}{N - q}.$$

The above expression is tractable given our substitutions. Because our approximations (3.6), (3.7), and (3.8) have been normalized by sample size before comparison, the value we get from direct comparison is equivalent to a so-called "average SSE", and must be multiplied by the sample size N. This yields the following expression:

$$\hat{\tilde{\sigma}}^2 = \frac{1}{N-q}[\frac{1}{N}\widehat{\mathbf{Y}'\mathbf{Y}} - (\frac{1}{N}\widehat{\mathbf{X}'\mathbf{Y}})^T(\frac{1}{N}\widehat{\mathbf{X}'\mathbf{X}})^{-1}\frac{1}{N}\widehat{\mathbf{X}'\mathbf{Y}}] \times N. \tag{3.9}$$

A demonstration the effectiveness of estimator $\hat{\tilde{\sigma}}^2$ via simulation in Appendix **B.4**, and show that it behaves well as compared to some other plausible estimators. To calculate the SSE based on some set of estimates $\hat{\boldsymbol{\beta}}_P$, we substitute our approximations (3.6), (3.7), (3.8) and the penalized regression estimates into the following expanded expression for SSE:

$$SSE = \mathbf{Y}'\mathbf{Y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{Y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}.$$

This yields the following expression:

$$\widehat{SSE} = (\frac{1}{N}\widehat{\mathbf{Y}'\mathbf{Y}} - 2\hat{\boldsymbol{\beta}}_P'\frac{1}{N}\widehat{\mathbf{X}'\mathbf{Y}} + \hat{\boldsymbol{\beta}}_P'\frac{1}{N}\widehat{\mathbf{X}'\mathbf{X}}\hat{\boldsymbol{\beta}}_P) \times N. \tag{3.10}$$

Note that, as in the estimation of $\tilde{\sigma}^2$, we multiply by the expression for SSE by N because all of the terms in the expression are normalized.

Given this, we can express our log-likelihood, as estimated from the reference panel and summary statistics, as follows:

$$l = -\frac{1}{2\hat{\tilde{\sigma}}^2}\widehat{SSE}.$$

Consider that we have omitted constants from the above expression that do not affect the relative values of the pseudo AIC / BIC. Given the log-likelihood, we can construct the pseudo AIC and BIC as follows, which mirrors existing literature on AIC and BIC for penalized regression [74] :

$$AIC = 2k - 2l,$$

$$BIC = ln(N) * k - 2l.$$

Where k is the degrees of freedom of the model, and l is the log-likelihood. Since our penalized regression models can be thought of as a form of elastic net, we use the degrees of freedom of the ridge regression model, calculated as $df(\lambda, s) = tr[(\mathbf{X}_r'\mathbf{X}_r + \lambda s\mathbf{I})^{-1}(\mathbf{X}_r'\mathbf{X}_r)]$. If this is too intensive to calculate for data with a large number of parameters p, we can also use the degrees of freedom for the LASSO model, which is simply the number of nonzero parameter estimates.

3.2.4 Notes on the application of pseudo AIC / BIC

In the case where we have a binary phenotype, the univariate effect size estimates are typically obtained via logistic regression. We note that the theory on pseudo AIC / BIC applies only to linear regression, and is intractable for logistic regression. Thus, we want to convert the univariate logistic regression estimates to univariate linear regression estimates $\hat{\boldsymbol{\beta}}$.

Consider that we have some binary vector of phenotypes $\mathbf{y}$ and a centered design matrix $\mathbf{X}$ where each column is a SNP. We start with the condition that linear and logistic regression should give similar results; that is, for any entry x_{ij}, $\pi = p(y_i = 1|x_{ij})$ is approximately equal under linear and logistic regression. Let us denote the logistic regression estimates as $(\hat{b}_0, \hat{b}_1)'$, and the linear regression estimates as $(\hat{\beta}_0, \hat{\beta}_1)'$. Thus, we have:

$$\hat{\beta}_0 + \hat{\beta}_1 x_{ij} = \frac{1}{1 + e^{-(\hat{b}_0 + \hat{b}_1 x_{ij})}}.$$

Given that these two expressions are equivalent, we know that the terms comprising their respective Taylor expansions are also equivalent. Taking the first term of the Taylor expansion of each expression and equating them, we get the get the following equivalence:

$$\hat{\beta}_0 = \frac{1}{1 + e^{-\hat{b}_0}}.$$

Taking the second term of the Taylor expansion for each expression and equating them, we get the equivalence

$$\hat{\beta}_1 x_{ij} = \frac{e^{-\hat{b}_0}}{(1 + e^{-\hat{b}_0})^2}\hat{b}_1 x_{ij},$$

from which we have, straightforwardly,

$$\hat{\beta}_1 = \frac{e^{-\hat{b}_0}}{(1 + e^{-\hat{b}_0})^2}\hat{b}_1.$$

Note that $e^{-b_0} = \frac{p(Y=0)}{p(Y=1)}$, which is easily obtainable. It is simply the ratio of controls to cases in the phenotype vector $\mathbf{y}$. This method will work best when b_1 is small, and thus the slope of the estimated logistic function is shallow. This is almost always the case in GWAS applications given the small effect sizes of individual SNPs, so this approximation should hold.

We also need to approximate the standard error of the linear regression estimates. A derivation with a general formula and some discussion is contained Appendix **B.3**. Following this derivation, we get the expression

$$var(\hat{\beta}_1) = \Big(\frac{e^{-\hat{b}_0}}{(1+e^{-\hat{b}_0})^2}\Big)^2 var(\hat{b}_1),$$

and thus

$$SE(\hat{\beta}_1) = \Big(\frac{e^{-\hat{b}_0}}{(1+e^{-\hat{b}_0})^2}\Big) se(\hat{b}_1).$$

We assume the standard error of the logistic regression estimate $SE(\hat{b}_1)$ is contained in the summary statistic information, making this calculation straightforward. An application of our pseudo AIC/BIC methodology to binary lung cancer data is in **(3.3.4)**.

Another issue is the selection of the degrees of freedom q in the calculation of the OLSE-based $\hat{\tilde{\sigma}}^2$. In the case where $p > N$, we cannot include all univariate summary statistics in our estimation of $\hat{\tilde{\sigma}}^2$. We propose using pruning and thresholding to determine a set of independent and moderately associated SNPs, and using this set for the calculation of $\hat{\tilde{\sigma}}^2$. Some experimentation has shown the estimation of $\hat{\tilde{\sigma}}^2$ to be relatively invariant across reasonable choices of a SNP set.

When calculating $\widehat{SSE}$ and $\hat{\tilde{\sigma}}^2$ in practice, we found it useful to regularize the estimated covariance matrix. When estimating $\widehat{SSE}$ as described in (3.10) and $\hat{\tilde{\sigma}}^2$ as described in (3.9), we make the substitution described below. Note that this bears some similarity to the regularization we describe in equation (3) with $s = .2$, although it is not identical:

$$\frac{1}{n}\widehat{\mathbf{X}^T\mathbf{X}} = \frac{1}{n}\mathbf{X}_r^T\mathbf{X}_r + .2\mathbf{I}.$$

In our experience, when estimating penalized regression models via summary statistics and pseudo AIC / BIC via summary statistics on the same data, it is crucial not to reuse the same reference panel for the calculation of the polygenic risk scores and the

calculation of the pseudo AIC / BIC. Doing so leads to a sort of overfitting issue, and will badly degrade the performance of the pseudo AIC / BIC. In practice, this can be avoided by splitting the reference panel in half, and using one half for the estimation of polygenic risk scores and the other half for the estimation of model fitting metrics.

As in the estimation of polygenic risk scores via penalized regression, we recommend estimating the pseudo AIC / BIC by independent LD blocks [72]. This is relevant to the covariance matrix $\frac{1}{n}\widehat{\mathbf{X}^T\mathbf{X}}$ as estimated from the reference panel. All estimation of pseudo AIC / BIC in this chapter was done by LD blocks unless otherwise noted.

3.2.5 Quasi-Correlation

The so-called quasi-correlation is a model-fitting metric that can be used to evaluate the performance of a polygenic risk score on out-of-sample data for which we have only summary statistics. It similar to existing method SummaryAUC [75], except that quasi-correlation is relevant to continuous phenotype data. The quasi-correlation estimates the true correlation. Because the correlation between a polygenic risk score and a validation phenotype is frequently used for model selection of polygenic risk scores, we can apply the quasi-correlation for model selection purposes as well.

Now, we describe the scenario when application of quasi-correlation is useful. In this scenario, we have three datasets. Firstly, we have the 'training' dataset, with centered design matrix denoted $\mathbf{X}$ and centered phenotype denoted $\mathbf{Y}$. Using some method, such as penalized regression, we estimate a polygenic risk score. We denote this $\hat{\boldsymbol{\beta}}^P = (\hat{\beta}_1^P, \ldots, \hat{\beta}_p^P)'$. Secondly, we have the reference panel, denoted $\mathbf{X}_r$. That is, $\mathbf{X}_r$ is some centered matrix with columns corresponding to the same SNPs as those in $\mathbf{X}$. We also use $\mathbf{X}_r$ to estimate the variances of the SNPs. Let's denote this vector of estimated variances as $(\hat{s}_1^2, \ldots, \hat{s}_p^2)$. Lastly, we have the 'testing' dataset, where we want to test the accuracy of our polygenic risk score $\hat{\boldsymbol{\beta}}^P$. For this data, we have centered design matrix $\mathbf{X}_*$, centered phenotype $\mathbf{Y}_*$, and sample size n_t. Using univariate linear regression, we estimate marginal effect sizes $\hat{\boldsymbol{\beta}}_*$ for the testing data. We assume that we do not have access to either $\mathbf{X}_*$ or $\mathbf{Y}_*$, and only have $\hat{\boldsymbol{\beta}}_*$.

We want to use our polygenic risk score $\hat{\boldsymbol{\beta}}^P$ to predict phenotypes for the testing data. Then, we want to calculate the correlation between our estimated phenotypes on the testing data $\hat{\mathbf{Y}}_* = \mathbf{X}_*\hat{\boldsymbol{\beta}}^P$ and the true phenotypes $\mathbf{Y}_*$. That is, we want to estimate

the following quantity:

$$cor(\mathbf{Y}_*, \hat{\mathbf{Y}}_*) = \frac{\frac{1}{n_t}\sum_i Y_i^* \hat{Y}_i^* - (\frac{1}{n_t}\sum_i Y_i^*)(\frac{1}{n_t}\sum_i \hat{Y}_i^*)}{\sqrt{var(\mathbf{Y}_*)var(\hat{\mathbf{Y}}_*)}}.$$

We note that $(\frac{1}{n_t}\sum_i Y_i^*) = 0$, because we assume a centered $\mathbf{Y}_*$. Thus, we have the following expression:

$$cor(\mathbf{Y}_*, \hat{\mathbf{Y}}_*) = \frac{\frac{1}{n_t}\sum_i Y_i^* \hat{Y}_i^*}{\sqrt{var(\mathbf{Y}_*)var(\hat{\mathbf{Y}}_*)}}.$$

This is still not tractable, given that we cannot calculate $\hat{\mathbf{Y}}_*$ directly because we don't have access to $\mathbf{X}_*$. To obviate this, we make the following observation:

$$\frac{1}{n_t}\sum_i Y_i^* \hat{Y}_i^* = \frac{1}{n_t}\mathbf{Y}_*^T \hat{\mathbf{Y}}_* = \frac{1}{n_t}\mathbf{Y}_*^T \mathbf{X}_* \hat{\boldsymbol{\beta}}^P = \frac{1}{n_t}(\mathbf{X}_*^T \mathbf{Y}_*)^T \hat{\boldsymbol{\beta}}^P = \sum_{j=1}^{p} \hat{s}_j^2 \hat{\beta}_j^* \hat{\beta}_j^P.$$

Now, we must find a way to estimate $var(\hat{\mathbf{Y}}_*)$. To show the derivation, we introduce the notation that X_i^{*T} is a transposed column vector corresponding to a row of $\mathbf{X}_*$. We then have

$$var(\hat{Y}_*) = \frac{1}{n_t}\sum_i \big[(X_i^{*T}\hat{\boldsymbol{\beta}}^P)^2\big] - \big[\frac{1}{n_t}\sum_i (X_i^{*T}\hat{\boldsymbol{\beta}}^P)\big]^2 = A - B.$$

Now, we investigate terms A and B:

$$A = \frac{1}{n_t}\sum_i (\hat{\boldsymbol{\beta}}^P)^T X_i^* X_i^{*T} \hat{\boldsymbol{\beta}}^P = (\hat{\boldsymbol{\beta}}^P)^T (\frac{1}{n_t}\sum_i X_i^* X_i^{*T})\hat{\boldsymbol{\beta}}^P = (\hat{\boldsymbol{\beta}}^P)^T \mathbf{X}_r^T \mathbf{X}_r \hat{\boldsymbol{\beta}}^P,$$

$$B = \big[\frac{1}{n_t}\sum_i (X_i^{*T}\hat{\boldsymbol{\beta}}^P)\big]^2 = (\bar{X}_*\hat{\boldsymbol{\beta}}^P)^2 = 0.$$

Finally, we must estimate $var(\mathbf{Y}_*)$. We note that we can approximate the variance of a centered phenotype using summary statistics via equation (3.7). As in section **(3.2.3)**, we suggest taking some measure of central tendency of the p estimates of $\widehat{var}(\mathbf{Y}_*)$, such as the median. Given these approximations, we can now define the quasi-correlation in a usable form:

$$quasiCor(\mathbf{Y}_*, \hat{\mathbf{Y}}_*) = \frac{\sum_i \hat{s}_i^2 \hat{\beta}_i^* \hat{\beta}_i^P}{\sqrt{\widehat{var}(\mathbf{Y}_*)(\hat{\boldsymbol{\beta}}^P)^T \mathbf{X}_r^T \mathbf{X}_r \hat{\boldsymbol{\beta}}^P}}.$$

3.3 Results

3.3.1 Simulation study for penalized regression

Here we show the effectiveness of penalized regression in predicting quantitative phenotypes in a simulation scenario. We show the accuracy of penalized regression compared to results from similar methods LDPred [15], LDPred-Inf, pruned and thresholded (P+T) polygenic risk scoring, and simple polygenic risk scoring.

We simulated quantitative phenotypes using data from the Wellcome Trust Case Control Consortium, or WTCCC [76]. We conducted three simulations. One simulation used genotype data from chromosomes 1 and 4 ($\sim 30,000$ SNPs), which we will call simulation 1. The other simulation used used genotype data from chromosomes 1, 2, 3 and 4 ($\sim 61,000$ SNPs), which we will call simulation 2. The third simulations used SNPs from all chromosomes ($\sim 230,000$ SNPs), which we call simulation 3. The ratio of sample size of the training data to the number of SNPs has been shown to affect the predictive performance of polygenic risk scoring in previous literature [15]; this ratio differentiates simulation 1, 2, and 3. The data was comprised of 12,479 individuals. The data were split into three sets: training, which consisted of 6240 individuals, tuning, which consisted of 3119 individuals, and testing, which consisted of 3120 individuals. We pruned SNPs such that no two included SNPs were in linkage disequilibrium higher than 0.9 in order to ensure convergence. In practice, one can perform LD clumping to ensure that no two SNPs are in LD higher than 0.9. We additionally removed all ambiguous SNPs (A/T, C/G), and all SNPs with $MAF < .01$.

We simulated SNP effect sizes from the point normal model:

$$\beta_j \sim_{iid} \begin{cases} N(0, \frac{h^2}{Mp}), \text{ with probability p} \\ 0, \text{ with probability 1-p} \end{cases}$$

Where h^2 is the SNP-based heritability of the disease (0.5 in our simulation), M is the number of SNPs, and p is the fraction of causal SNPs. We used the following values of p in our simulation: $p = 0.1$, $p = 0.01$, $p = 0.001$, $p = .0005$. Simulation 1 excluded the case where $p = .0005$, and simulation 3 excluded the case where $p = .1$. We used the SNP effects to generate quantitative phenotypes under the additive model of genetic effects. Using the simulated phenotypes and the training data, we

calculated summary statistics. We used the summary statistics from the training data and LD information from the tuning data to estimate penalized regression models, using LassoSum, TlpSum, and ElastSum. With these penalized regression estimates, we generated predicted phenotypes for the tuning data set, and selected tuning parameter values that optimized the prediction r^2. We then calculated predicted phenotypes for the test data, using the optimized tuning parameter values. We then report the predictive r^2 of the testing data. We performed 20 replications for each method at each value of p.

When applying LDPred, we tuned parameter p on the tuning data, and obtained prediction r^2 from the testing data. The true value of p was contained in the set of tuning values for p, as were four other values; two larger than the true p, and two smaller. As per the recommendation of the original paper [15], we used $M/3000$ as the LD parameter, which controls the size of a sliding window of how many SNPs to consider when estimating joint effect sizes. When applying the polygenic risk score (denoted PRS), we used all marginal SNP effect size estimates from the training data. When applying the pruned and thresholded polygenic risk score (denoted PRS P+T), we first performed LD clumping in PLINK to ensure that no two SNPs were in LD $R^2 > .2$. We then implemented a p-value cutoff, where only SNPs with marginal p-value below some cutoff were included in the risk score. The p-value cutoff was treated as a tuning parameter, and determined by maximizing accuracy on the tuning data. Both the LD R^2 cutoff and the method of determining the p-value cutoff were done as in the LDPred paper [15]. The results are displayed in figures 3.1, 3.2, and 3.3.

These results suggest that the penalized regression methods have an advantage over LDPred when the ratio of SNPs to sample size N is smaller, as demonstrated by the simulation 1 results, while LDPred outperforms the penalized methods slightly as N grows, as demonstrated by the simulation 2 results. In simulation 3, the penalized regression methods have roughly equivalent predictive accuracy to LDPred. Additionally, it appears that the penalized regression methods perform comparatively better when p is smaller; that is, the signal is sparser. LDPred and the penalized regression methods outperform clumped polygenic risk scoring in all cases. The simple PRS performs poorly in most cases, except for when the fraction of causal SNPs p is large. These simulation results demonstrate that the penalized regression methods are competitive with

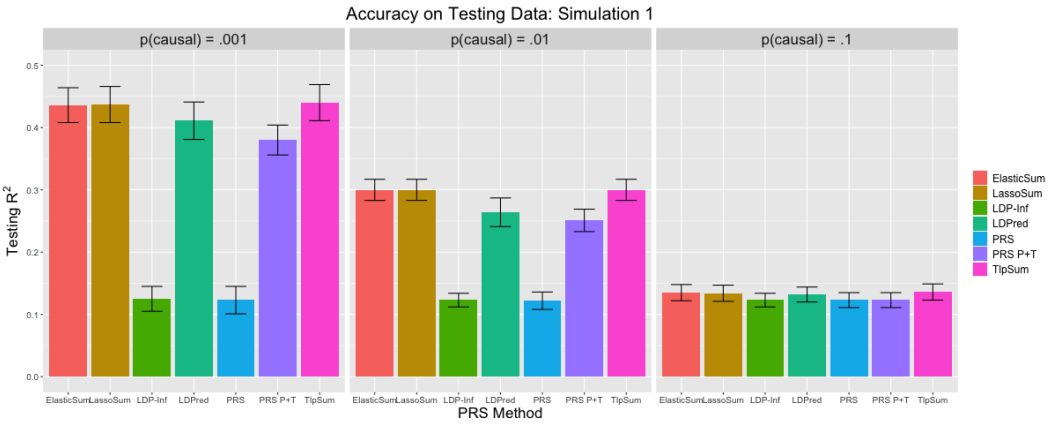

Figure 3.1: Prediction r^2 values for simulation 1. Error bars represent standard deviation for the r^2 value across 20 replications.

LDPred in all simulation settings, and outperform PRS methods that do not account for linkage disequilibrium. In this simulation structure, we do not see much difference in performance between the three penalized regression methods.

3.3.1.1 Simulating allelic heterogeneity

The penalized regression methods LassoSum, TlpSum, and ElastSum demonstrate similar predictive performance in **(3.3.1)**. Motivated by the concept of so-called 'widespread allelic heterogeneity' [77], we conduct a simulation where causal SNPs are clustered together in regions of high linkage disequilibrium. This simulates allelic heterogeneity, which is characterized by multiple SNPs within a single region (often a gene) that are causal for a trait. Under this simulation structure, we investigate the performance of the penalized regression methods, and demonstrate that TlpSum incurs modest but persistent gains in predictive accuracy as compared to LassoSum and ElastSum.

We set up the simulation as follows. We use the 'simulation 1' structure from **(3.3.1)**, with the following adjustments. Instead of simulating effect sizes from the point normal model with the probability of nonzero effect drawn independently for each SNP, we now simulate causal SNPs (i.e. SNPs with nonzero effect size) in groups of size 2 to 8. The process for simulating SNP effect sizes is described in Appendix **B.5**. We also adjusted the fraction of causal SNPs p and the SNP-based heritability

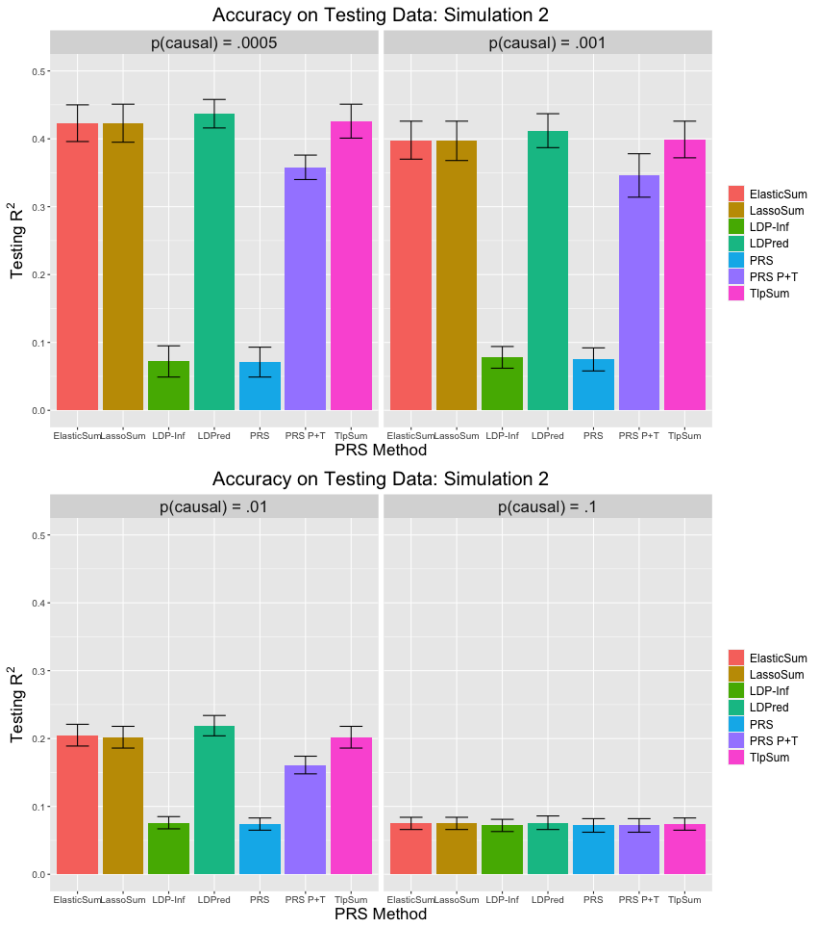

Figure 3.2: Prediction r^2 values for simulation 2. Error bars represent standard deviation for the r^2 value across 20 replications.

h^2 from **(3.3.1)**. We considered values for p of .002 and .005, and values for h^2 of .2, .5, .6. We conducted 100 replications at each simulation setting. In all four of the simulation settings considered, TlpSum had better predictive accuracy on out-of-sample data as compared to ElastSum and LassoSum. This improvement was measured to be statistically significant at $p < .05$ with a paired t-test. Figures 3.4 and 3.5 describe the performance of the TlpSum as compared to ElastSum and LassoSum across the four simulation settings. Additional results describing the relative performance of the LassoSum and the ElastSum, the results from some significance tests, and some results on predictive accuracy are located in Appendix **B.7**, in particular figure B.8 and tables

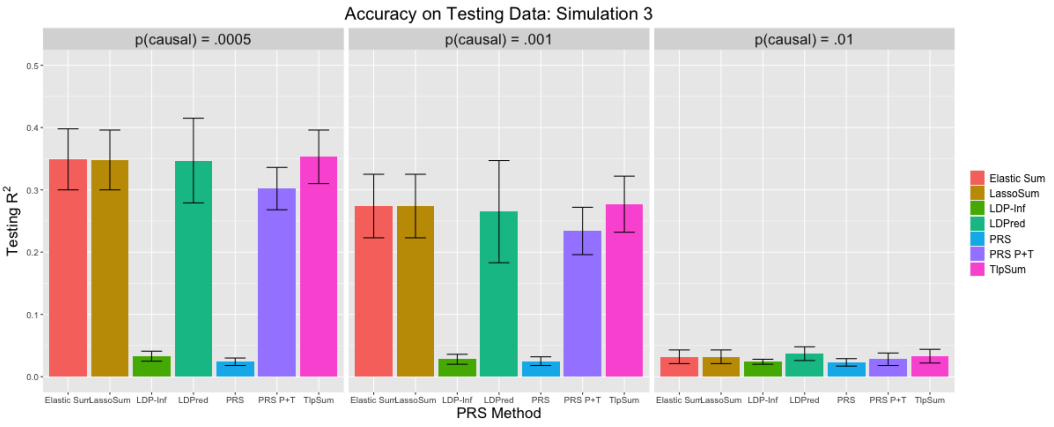

Figure 3.3: Prediction r^2 values for simulation 3. Error bars represent standard deviation for the r^2 value across 20 replications.

B.1 - B.4.

Figures 3.4 and 3.5 demonstrate the persistent advantage of the TlpSum as compared to the LassoSum and ElastSum when effect sizes are simulated under widespread allelic heterogeneity. This substantive improvement in predictive accuracy is evidence for the superior performance of the TlpSum as compared to other penalized regression methods for summary statistics in the context of widespread allelic heterogeneity.

3.3.1.2 Investigation of models fit by penalized regression

Many penalized regression methods impose a degree of sparsity on the estimated effect sizes. In particular, when the fraction of causal SNPs p is small, the proportion of nonzero estimated effects is generally also small. It is of interest to characterize this sparsity and examine how it might influence the predictive performance. In this section, we characterize the sparsity of the fitted penalized regression models. This issue bears some similarity to fine mapping, which includes methods such as CaviarBF [78] and FINEMAP [79]. We do not formulate formal hypothesis tests for variable selection in penalized regression in this chapter, and we do not seek to compare our method to the fine mapping literature.

Given that TLP does not penalize effect size estimates above a certain threshold, it

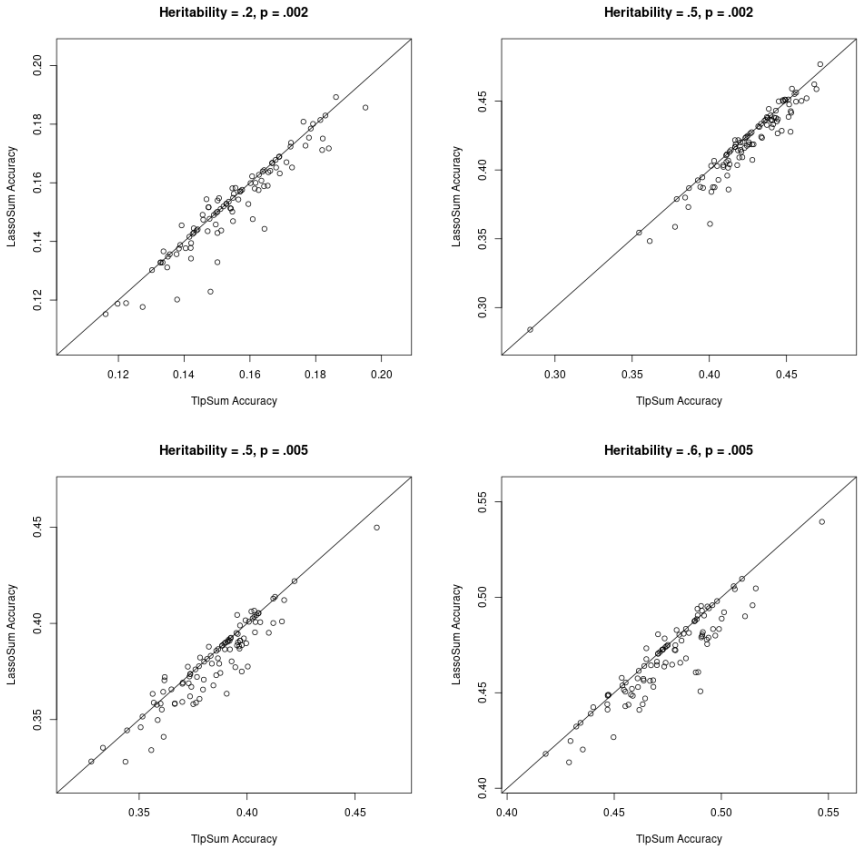

Figure 3.4: Predictive r^2 on out-of-sample data for TlpSum and LassoSum for each of the 100 replications at each of the four simulation settings with allelic heterogeneity. Lines are at a 45 degree angle through the origin, and not a line of best fit. Points below the line indicate better performance of TlpSum.

may produce a smaller number of nonzero effect size estimates. This has been demonstrated in previous literature [17]. Thus, the TLP may be more parsimonious when the truth is sparse. We investigate the number of nonzero parameter estimates for sparse situations in simulations 1, 2, and 3. We find that the TLP produces sparser estimates than the LASSO and the elastic net in the case where $p = .001$ for simulation 1, $p = .0005$ for simulation 2, and $p = .0005$ for simulation 3. The results are not as clear for the cases where the fraction of causal SNPs p is larger. We suspect this is because all of the selected values for the tuning parameter s are nonzero in the case

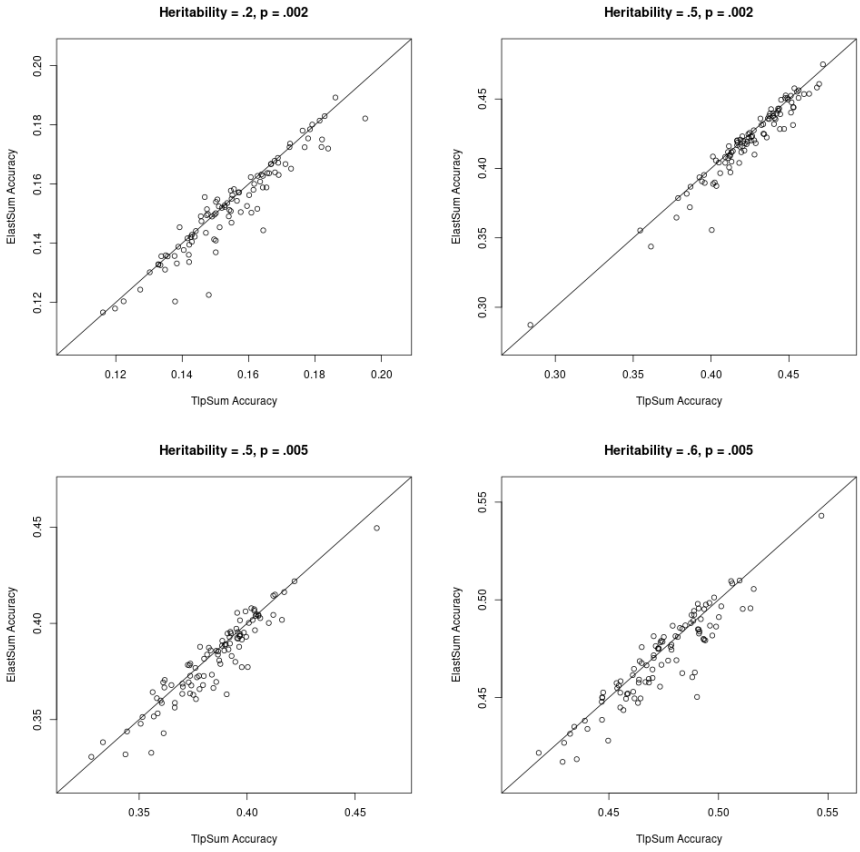

Figure 3.5: Predictive r^2 on out-of-sample data for TlpSum and ElastSum for each of the 100 replications at each of the four simulation settings with allelic heterogeneity. Lines are at a 45 degree angle through the origin, and not a line of best fit. Points below the line indicate better performance of TlpSum.

where $p \in [.01, .1]$, and some of the time when $p = .001$ in simulations 2 and 3. This means that we do not have a "true" TLP, as is described in **(2.1)** and illustrated in equation (3.5). In the case where $p = .001$ in simulation 1 and $p = .0005$ in simulations 2 and 3, the optimal value of s is zero for all models or nearly all models, giving us a "true" TLP. We present the results for the three sparse simulation settings in figure 3.6. Given that the models all achieve similar predictive performance on out of sample data as illustrated in figures 1, 2, and 3, we see that the TLP can generate the same amount of predictive power with sparser models.

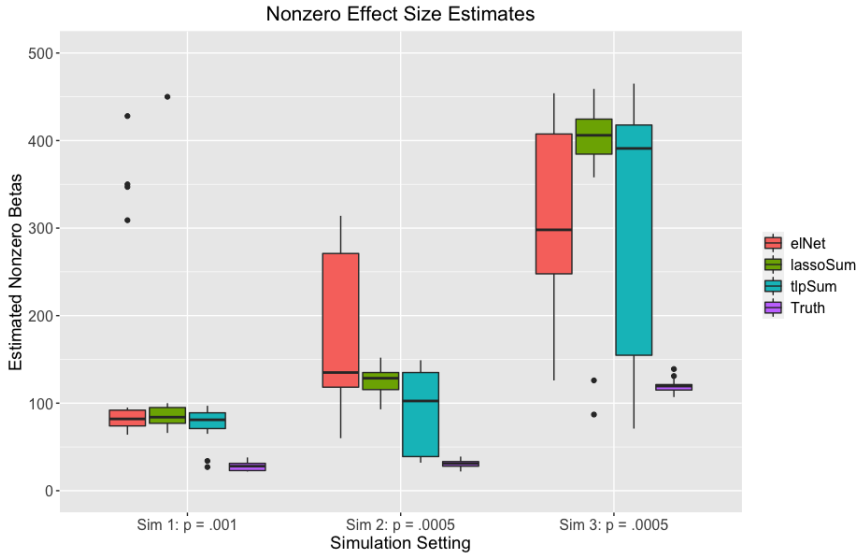

Figure 3.6: Number of nonzero effect sizes estimated by the three penalized regression methods as compared to the true number of nonzero effects, for the three sparse simulation settings.

Also of interest is the number of true nonzero effects that are estimated to be nonzero by the penalized regression models. We can think of this as a binary prediction problem, where we are trying to predict which effects are nonzero. This information is presented in figure 3.7 for the three sparse simulation settings. We see that the TLP has nearly the same number of true positives as the elastic net and LASSO, while having fewer total nonzero estimated effects, as displayed in figure 3.6. This corresponds to a higher precision, as displayed in figure 3.8. Note that precision corresponds to $\frac{TP}{TP+FP}$, where TP is the number of true positives, and FP is the number of false positives.

These simulation results provide evidence that, when the truth is sparse, the TLP may produce sparser effect size estimates and reduce the number of false positives, while capturing nearly the same number of true positives. This indicates that TLP models maintain predictive accuracy while being closest to the true structure of effects, thus facilitating the estimation of parsimonious models.

3.3.2 Simulation study for model selection

Using the simulation structure described in **(3.3.1)**, we assessed the comparative accuracy of model selection methods. We compared the three model selection methods we proposed, namely the pseudo AIC, pseudo BIC, and quasi-correlation, to existing model selection method pseudovalidation. Note that these four model selection methods do not require the existence of individual level tuning or training data, and are thus more widely applicable, especially in the framework of summary statistics and reference panels. As a point of comparison, we also include the performance of AIC and BIC for the model as fit on the training data (the so-called 'true AIC' and 'true BIC'). The true AIC and true BIC assume that we have access to individual level genotype data for the training dataset, which is not generally the case. They also directly use the true residual variance $\tilde{\sigma}^2$, which must be estimated in practice. We also compare the performance of selecting the model with maximum r^2 on the tuning data, which is a widely applied model selection criteria. This assumes that we have individual level phenotype data for the tuning dataset, which may not be the case.

We split the WTCCC data into four disjoint datasets as described below. This allowed us to simulate a setting where our proposed pseudo AIC / BIC and quasi-correlation could be estimated in a realistic setting. As described in section **(3.2.4)**, it is important not to reuse the same reference panel for the penalized regression methods and the model fitting methods. This explains the presence of two 'tuning' datasets. This practice, where we essentially split the reference panel in half and use one half for the penalized regression methods and the other half for the model fitting methods, is used in our real data applications as well. The four datasets are as follows:

- The training data $\mathbf{X}_{tr}$, which we used to estimate univariate summary statistics for each SNP. $\mathbf{X}_{tr}$ had sample size 6240.
- The tuning-1 data $\mathbf{X}_{tu1}$, which was used as a refernece panel for the model estimation methods, namely TlpSum and LassoSum. $\mathbf{X}_{tu1}$ had sample size 3119.
- The tuning-2 data $\mathbf{X}_{tu2}$, which was used as a reference panel for the model selection metrics that required a reference panel: namely pseudo AIC, pseudo BIC, pseudovalidation, and quasi-correlation. $\mathbf{X}_{tu2}$ had sample size 1560.

- The testing data $\mathbf{X}_{te}$, which was used to evaluate the performance of the polygenic risk scores. $\mathbf{X}_{te}$ had sample size 1560.

For this simulation study, we used simulation setting 1: that, is, we used SNPs from chromosomes 1 and 4 from the WTCCC study, and simulated phenotypes from the point-normal model, varying the fraction of causal SNPs p. We used the same filtering steps as described in section **(3.3.1)**. We estimated univariate summary statistics from the training data ($N = 6240$). We then used the tuning-1 data ($N = 3119$) as a reference panel to estimate polygenic risk scores using TlpSum and LassoSum. For TlpSum, we used a three dimensional matrix of tuning parameters λ, s, τ to generate a set of candidate polygenic risk scores. For LassoSum, we used a two dimensional matrix of tuning parameters λ, s to generate a set of candidate polygenic risk scores. The results from applying the model selection metrics to LassoSum are presented here. The results of applying the model selection metrics to the TlpSum models are similar (results not shown).

For the estimation of the pseudo AIC, pseudo BIC, quasi-correlation, and pseudovalidation, we used the tuning-2 dataset ($N = 1560$) as a reference panel. Although pseudovaldiation does not require the tuning data to be split in half as pseudo AIC and pseudo BIC do, we note that using the split tuning data versus the full tuning data made no difference in practice for pseudovalidation. For the quasi-correlation criteria we used summary statistics estimated from the tuning-2 data. Using the seven model fitting criteria that we described, we selected a best model in accordance with each of the criteria. We then measured the predictive r^2 of that model applied to the testing data. This was repeated for each of the 20 simulations, across three different values of p, the fraction of causal SNPs.

In addition to considering quasi-correlation as a model selection metric, we have also proposed using quasi-correlation as a measure of model fit; that is, as a way to compare the performance of different models. In the case where we do not have individual level testing data, we will not be able to use many common measures of predictive performance. If we have access to summary statistics from the testing data, we will be able to use quasi-correlation. We want the relative performance of the different model selection methods as measured by predictive r^2 on the testing data to be the same as the relative performance as measured by quasi-correlation. Note that we have two

different applications of quasi-correlation here; we are using it for model selection, and to quantify model performance. Quasi-correlation for model selection is estimated using summary statistics from the tuning-2 data; this corresponds to the 'Qcor' bar group in the bar chart. Quasi-correlation for quantifying model performance is estimated using summary statistics from the testing data; this corresponds to the red bars in the bar chart. The results are displayed in figure 3.9.

These results show that quasi-correlation performs well as a model selection method, outperforming all other metrics except for tuning r^2, which it performs equivalently to. The pseudo AIC and pseudo BIC perform relatively similarly to the true AIC and true BIC, although the true AIC and BIC do perform equivalently or better in all cases, which is to be expected. Additionally, we see that the pseudo AIC outperforms pseudovalidation in the case where $p = .01$. The methods perform similarly when $p = .001$ and $p = .1$. In this simulation, pseudovalidation, pseudo AIC, and pseudo BIC all appear to be reasonable methods for model selection when validation data is not available. A more thorough analysis of the accuracy of the different components of the estimation of pseudo AIC / BIC and quasi-correlation is located in Appendix **B.6**.

We also assess the usefulness of quasi-correlation as a measure of model fit. Figure 3.9 shows that the relative performance of the model selection methods as measured by quasi-correlation squared and testing r^2 are generally equivalent. Quasi-correlation appears to slightly overestimate the testing r^2 a majority of the time, and the standard deviation across the twenty replications is a somewhat larger. Nevertheless, we can conclude that quasi-correlation does a good job approximating the testing r^2 given the high degree of similarity between the testing r^2 and squared quasi-correlation estimates.

These results demonstrate that quasi-correlation approximates the predictive performance of selected models well on average. Also of interest is how well quasi-correlation performs within a single replication, i.e. whether quasi-correlation can generally differentiate between the predictive performance on out-of-sample data for a set of candidate models. This particularly concerns the performance of a set candidate models on a single out-of-sample dataset, rather than the average across twenty replications as shown in figure 9. Results described in Appendix **B.4** indicate that the quasi-correlation generally does this well.

3.3.2.1 Investigation of selected models

We examine the model selection performance of pseudo AIC and pseudo BIC as applied to penalized regression models in the summary statistic framework, and generally conclude that they demonstrate good performance.

Via simulation, we show that pseudo AIC and pseudo BIC select sparser models than pseudovalidation, and that pseudo AIC and pseudo BIC generally reproduce the true model more accurately. In particular, we consider simulation setting 1 from **(3.3.1)**. For each of the 20 replications at each of the three fractions of causal SNPs p, we compare the number of nonzero effect sizes for each of the three model selection methods to demonstrate the tendency of the methods to select models of differing sparsity. We describe the precision, recall, and F1 score for the model selected by each the three model selection methods to demonstrate the degree to which selected models recapture the true model. These measures of accuracy are considered in the following context. A 'true positive' occurs when a SNP with a nonzero effect size has an estimated nonzero effect in the corresponding model. Likewise, a false positive occurs when a model estimates a SNP effect to be nonzero and the true SNP effect is zero. If we define TP as the number of true positives captured by a model, FP as the number of false positives, and FN as the number of false negatives, we can define precision as $\frac{TP}{TP+FP}$ and recall as $\frac{TP}{TP+FN}$. The F1 score is defined as $F_1 = 2\frac{precision \times recall}{precision+recall}$, which is the harmonic mean of precision and recall. We consider the application of model selection to a set of candidate LassoSum models; we believe that the performance would be similar for TlpSum. We expect that the pseudo AIC and pseudo BIC may select sparser models than pseudovalidation, and that the selected models may perform better as measured by precision and F1 score. We generally expect the models selected by pseudovalidation to display better recall, given that they have more estimated nonzero effects. The results are displayed in figures 3.10 and 3.11.

We see that the three model selection methods perform equivalently when $p = .1$, but there is discrepancy when $p < .1$. Pseudo BIC selects substantially sparser models than either pseudo AIC or pseudovalidation, while pseudo AIC selects somewhat sparser models than pseudovalidation. Figure 11 shows that pseudo BIC substantially outperforms pseudo AIC and pseudovalidation according to the precision and F1 score

metrics, although pseudo BIC performs less well as measured by recall. Pseudo AIC outperforms pseudovalidation as measured by precision and F1 score as well. Given that F1 score can be considered an overall measure of binary classification performance that considers precision and recall, it is reasonable to state that pseudo BIC substantially outperforms the other two model selection methods, indicating that it best reproduces the true model.

Pseudo AIC and pseudo BIC impose model sparsity according to established theory, while pseudovalidation selects a model by minimizing training error under an ad hoc condition that imposes some sparsity. Section **(3.3.2)** demonstrates that these three selection methods perform reasonably similarly as measured by predictive accuracy on out-of-sample data, but there is moderate discrepancy among the three methods in ability to reproduce the true model. In applications where it is important to select only those variants that are truly associated, such as the selection of valid instruments for a TWAS-type analysis [8], it may be preferable to use pseudo AIC or pseudo BIC.

3.3.3 Application to lipids

We leverage our methodology to perform model estimation and model selection for GWAS analyses of lipid data. We estimate models based on summary statistics from the Teslovich et al. study [80]. We assess model accuracy via quasi-correlation, using summary statistics from the UK BioBank as our out-of-sample data [19]. For partial validation and the estimation of quasi-correlation for model selection on a third dataset, we use summary statistics from the Global Lipids Genetics Consortium, or GLGC [27]. We consider three different phenotypes in this analysis: high-density lipoprotien (HDL), low-density lipoprotein (LDL), and triglycerides (TG). We use the 1000G data as a reference panel [50], and limit the reference panel to only those individuals of European ancestry.

For each study, we did quality control as follows. We removed SNPs with $MAF < .01$ in the reference panel or in the Teslovich data. We then determined the subset of SNPs that was present in all four datasets: the Teslovich data, the GLGC data, the BioBank data, and the 1000G data. We excluded all SNPs not in the intersection of these datasets. We did LD clumping using the 1000G data as a reference panel and univariate p-values from the Teslovich data, ensuring that no two SNPs were in LD

$R^2 > .9$. This was done to ensure convergence of our penalized regression methods, and shouldn't substantively affect the results, given that we don't expect many informative SNPs to be pruned away. We removed all ambiguous SNPs (i.e. those SNPs with alleles A/T or C/G), and all SNPs with allele coding irreconcilably different between the datasets. After quality control, we had 640,675 SNPs for TG, 639,754 SNPs for LDL, and 642,675 SNPs for HDL. The Teslovich and GLGC studies are meta-analyses, so the sample size varies by SNP. The BioBank study has equal sample size for all SNPs. We present the median sample size for each study and phenotype in table 3.1.

Table 3.1: Median sample size for each study in the lipid analysis.

	Teslovich	GLGC	BioBank
TG	95,877	90,976	343,992
LDL	94,769	89,855	343,621
HDL	99,179	94,277	315,133

Using these sets of SNPs, we estimated a set of 48 candidate polygenic risk scores for each lipid phenotype. We estimated polygenic risk scores via TlpSum using 48 unique sets of tuning parameters τ, s and λ, with the summary statistics from the Teslovich study as our training data. We split the 1000G data into two groups of equal sample size as described in **(3.2.4)**, and used one half of this data as the reference panel for estimating the TlpSum models. The other half was used to estimate the model fitting criteria. We then estimated model fitting criteria pseudo AIC, pseudo BIC, and pseudovalidation. We also estimated quasi-correlation for model selection by using the GLGC study as our out-of-sample data. There is substantial overlap between the samples used in the GLGC study and the Teslovich study; however, given that the study populations are not identical, we believe it is reasonable to apply the quasi-correlation here.

We present the accuracy of each method, as measured by quasi-correlation on the BioBank data, in table 3.2. In this case, none of the model selection methods perform particularly well, given that all methods select a model that performs worse than the best performing model. The accuracy of the best performing model is quantified in the 'Maximum' column, and represents the maximum quasi-correlation attained by any of the 48 candidate models predicted into the BioBank data. We see that the pseudo AIC,

pseudo BIC, and quasi-correlation all outperform pseudovalidation for all three lipid phenotypes. Given the relatively small amount of heritability captured, even by the best performing models, and the smaller sample size of the Teslovich study, this is likely a scenario where it is important to impose model sparsity during model selection. Because the pseudo AIC and pseudo BIC impose more model sparsity than pseudovalidation, and tend to be more parsimonious in recapturing the true model, the performance of the models selected by pseudo AIC and pseudo BIC are superior to the model selected by pseudovalidation in this application. On balance, the best performing model selection method is quasi-correlation, given that it selects the model with the best performance for two of the three lipid phenotypes. This is reasonable, given that quasi-correlation for model selection leverages information from a third dataset.

In this application, we demonstrate that pseudo AIC and pseudo BIC select models with superior predictive accuracy on out-of-sample data as compared to pseudovalidation for all three lipid phenotypes. We demonstrate the usefulness of quasi-corrleation for model selection given a third dataset by showing that it selects models with good predictive accuracy on out-of-sample data. Likewise, we use quasi-correlation to assess predictive performance on out-of-sample data. Without quasi-correlation, it would not be possible to leverage summary statistic data as out-of-sample data for this purpose.

Table 3.2: Model performance, as measured by quasi-correlation of the model predicted into the BioBank data, for each model selection method. Models were estimated via TlpSum on the Teslovich data.

	Quasi-cor	Pseudo AIC	Pseudo BIC	Pseudoval	Maximum
TG	.14	.13	.11	.10	.22
LDL	.12	.16	.14	.11	.21
HDL	.20	.18	.18	.17	.30

3.3.4 Application to lung cancer

We apply our penalized regression methods and pseudo AIC / BIC model selection to summary statistics from a large lung cancer meta-analysis [81]. The published summary statistics contain information on 10,633 unique SNPs, all with marginal $p < 10^{-6}$. The summary statistics are drawn from meta-analyses where the sample size can vary by

SNP. The distribution of the sample sizes is left-tailed, with greater than half of the summary statistics corresponding to the maximum sample size of 85,716. The smallest sample size is $\sim 10,000$.

We apply the polygenic risk scores estimated from the McKay meta-analysis to the EAGLE study. The EAGLE study, downloaded from dbGap [52], is a case-control study conducted in northern Italy, with sample size $N = 3936$. There are 1945 cases and 1991 controls. The EAGLE study was genotyped on a set of 561,466 SNPs on the Illumina HapMap550v3-B array. The data was imputed to the 1000G Phase 3 V5 reference panel using the Michigan Imputation Server [51]. After imputation, we removed all SNPs with imputation quality score $R^2 < .8$, Hardy-Weinberg p-value $< 10^{-9}$, call rate $< 90\%$, and minor allele frequency $< .01$. We were left with around 7 million SNPs.

As a baseline, we describe the performance of some polygenic risk scores that do not explicitly model linkage disequilibrium. We consider a polygenic score consisting of all marginal effect sizes from the lung cancer meta-analysis [81], which we call the full polygenic risk score. Only about half of the SNPs present in the meta-analysis achieve genome-wide significance with a marginal p-value $< 5 \times 10^{-8}$. We consider a polygenic risk score including only those SNPs that achieve genome-wide significance, the so-called genome-wide polygenic risk score. We also consider the clumped polygenic risk score. We performed clumping on the summary statistics from the McKay paper using the EAGLE data as a reference panel to calculate the correlation matrix of the SNPs. We performed clumping with PLINK [53] to prune correlated SNPs such that no two remaining SNPs are in LD $R^2 > .5$. This yielded a set of 633 SNPs. The accuracy of these methods, as compared to the penalized regression methods, is displayed in figure 3.12. We see that the penalized regression methods outperform the simple polygenic risk score methods. We also see that the genome-wide PRS achieves worse performance than both the simple PRS and the full PRS. This illustrates that including SNPs that are not genome-wide significant improves prediction, and that pruning based on LD helps reduce noise and improve prediction.

We also used penalized regression methods to construct polygenic risk scores. We conducted LD clumping in PLINK with the densely imputed EAGLE data as a reference panel. This clumping ensured that no two SNPs had linkage disequilibrium $R^2 > .9$. Note that this clumping is significantly less stringent than the clumping used to generate

the clumped polygenic risk score. After this step, there were 1524 remaining SNPs. We split the EAGLE study into three datasets; the so called tuning-1 and tuning-2 datasets, each with $N = 990$, and the test dataset with N = 1980. The tuning-1 dataset was used as a reference panel to estimate polygenic risk scores with the penalized regression methods. We estimated models without LD blocks due to the small number of candidate SNPs. The tuning-2 dataset was used as a reference panel to estimate the model fitting criteria. The testing dataset was used to evaluate predictive accuracy.

We used the observed phenotypes from the tuning-1 data to calculate the AUC for each candidate model. We then used tuning-1 AUC as a model selection criteria, selecting the model that maximized the tuning-1 AUC. The results obtained by using tuning-1 AUC for model selection, as compared to simple polygenic score methods that don't account for LD, are displayed in figure 3.12. We compare the performance of tuning-1 AUC to pseudo AIC, pseudo BIC, and pseudovalidation, which do not require the existence of individual level tuning data. When calculating the pseudo AIC and pseudo BIC, we regularized the estimated covariance matrix as in section **(3.2.4)**. To facilitate the calculation of pseudo AIC / BIC for data with a binary response, we performed the method outlined in **(3.2.4)** to convert the univariate logistic regression estimates to linear regression estimates. The performance of these methods, as applied to sets of candidate models generated via TlpSum and LassoSum, are displayed in figure 3.13.

We note, somewhat unexpectedly, that pseudovalidation outperforms the pseudo AIC and pseudo BIC. In this application, we are only considering SNPs with a small marginal p-value, and have done pruning to ensure that no two SNPs are in LD $R^2 > .9$. Thus, it might be that most SNPs under consideration are truly associated with the phenotype. Pseudovalidation estimates predictive r^2 on the training data, which will generally select a model with many SNPs. In this scenario, where the majority of SNPs under consideration are truly associated, there may be little possibility of overfitting, and adding many SNPs to the model may be desirable. Thus, the penalty on model size imposed by the pseudo AIC and pseudo BIC likely degrades their performance. For the application to TlpSum, pseudovalidation selects a model with 1310 nonzero parameters, which is 86.0% of the total variables considered. Pseudo AIC selects a model with 608 nonzero paramteres (39.9%), while pseudo BIC selects a model with 8 nonzero

parameters (0.5%). This is further evidence that pseudo AIC and pseudo BIC impose substantial model sparsity as compared to pseudovalidation, which is shown via simulation in **(3.3.2.1)**. The increased sparsity is evidently not useful in this application.

Even given the issues with applying pseudo AIC and pseudo BIC to binary data, we view this application to lung cancer as evidence that the estimation of polygenic risk scores via penalized regression on summary statistics is useful, even when we have only a small subset of summary statistics. The penalized regression methods outperform the naive methods for PRS estimation that don't estimate SNP effects under LD, demonstrating the usefulness of accounting for linkage disequilibrium in effect size estimation even when applied to a small subset of highly marginally significant SNPs. Pseudovalidation performs fairly well for model selection, and the pseudo AIC performs decently as well, although the performance of the pseudo BIC is quite poor for the TlpSum models. We note that lung cancer is not a strongly heritable disease, and the best models only achieve an AUC of .61. Thus, it may be difficult for pseudo AIC and pseudo BIC to select the best model, given that the difference in predictive accuracy between the candidate models is fairly small and the signal is not very strong.

3.4 Discussion

In this chapter, we propose applying the Truncated Lasso penalty and the elastic net penalty to calculate polygenic risk scores using summary statistic data and linkage disequilibrium information. We demonstrate via simulation that the TlpSum produces sparser models when the underlying genetic architecture is sparse, and does a good job recovering truly nonzero effect sizes while limiting false positives. Additionally, we demonstrate that the TlpSum improves predictive accuracy as compared to other penalized regression models when applied to data simulated under widespread allelic heterogeneity. We propose methods for estimating model fit statistics AIC and BIC for polygenic risk scores in the case where we have only summary statistic data and linkage disequilibrium information. This facilitates model selection in the case where we do not have access to validation data. This complements existing method pseudovalidation, which may tend to select overfit models. We also propose the so-called quasi-correlation, which allows us to quantify the predictive accuracy of a polygenic

risk score on out-of-sample data for which we have only summary statistic information. Quasi-correlation can also be used to leverage information from a third 'tuning' dataset of summary statistics for model selection. These methods in totality broaden the scope of the application of polygenic risk scores. Using only summary statistics and publicly available reference panels, we can estimate polygenic risk scores, perform model selection given a candidate set of polygenic risk scores, and quantify the predictive accuracy of these polygenic risk scores on out-of-sample summary statistic data. This facilitates the construction of validated polygenic risk scores ready for use on new data. Additionally, it facilitates the application of polygenic risk scores to large summary statistic data, generating robust models based on large studies. These models can be used infer the genetic architecture of complex phenotypes.

We demonstrate via simulation that penalized regression with the TLP penalty performs well as compared to existing methods, improving predictive performance in the context of allelic heterogeneity and inducing sparsity when the true model is sparse. We investigate the comparative performance of the pseudo AIC, pseudo BIC, pseudovalidation, and quasi-correlation for model selection via simulation, demonstrating that quasi-correlation performs well in all cases, and that pseudo AIC and pseudo BIC outperform pseudovalidation in some cases. Pseudo AIC and pseudo BIC demonstrate some desirable model selection properties in simulation, with pseudo BIC in particular tending to recover the true model better than pseudovalidation. We also show via simulation that quasi-correlation approximates the actual predictive r^2 well, indicating that it is an appropriate and robust measure of model fit. We demonstrate the usefulness of pseudo AIC and BIC and quasi-correlation for model selection by demonstrating their superior performance to pseudovalidation in an application to a large GWAS of lipid data. We demonstrate an application of penalized regression and model fitting methods to a large lung cancer meta-analysis, demonstrating that penalized regression methods improve accuracy as compared to simple polygenic risk score methods. We additionally demonstrate the application of pseudo AIC and BIC methods to a GWAS analysis with a binary phenotype. In Appendix **B.2**, we apply penalized regression and model fit methods to large summary statistic data of the height phenotype, which allows us to assess the performance of our penalized regression methodology and model selection methods on a large GWAS for a highly heritable phenotype.

3.5 Data Availability

The genotype data used in the simulation studies was downloaded from the Wellcome Trust Case Control Consortium. The EAGLE study was downloaded from dbGaP, with accession number phs000093.v2.p2. The lung cancer meta-analysis data corresponds to supplementary table S3 in the meta-analysis paper [81]. The 1000 Genomes reference data was downloaded from the 1000 Genomes project. The two summary statistic datasets on height were downloaded from UK Biobank and GIANT as described in the respective articles [19, 4]. The BioBank data on lipids was downloaded as described in the BioBank article [19]. The 'Teslovich' and 'GLGC' data on lipids were downloaded from `http://csg.sph.umich.edu/abecasis/public/lipids2010` and `http://csg.sph.umich.edu/abecasis/public/lipids2013/`, respectively. The R package for our methods is publicly available on Github: `https://github.com/jpattee/penRegSum`.

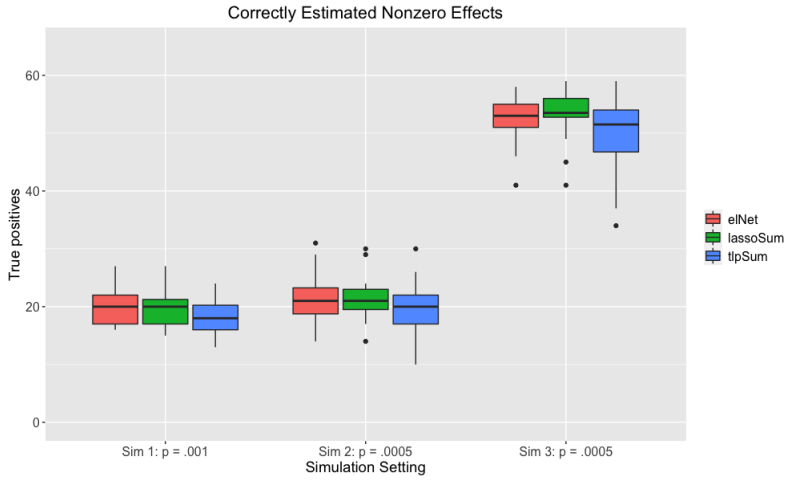

Figure 3.7: Number of true positives for the three penalized regression methods in the three sparse simulation settings.

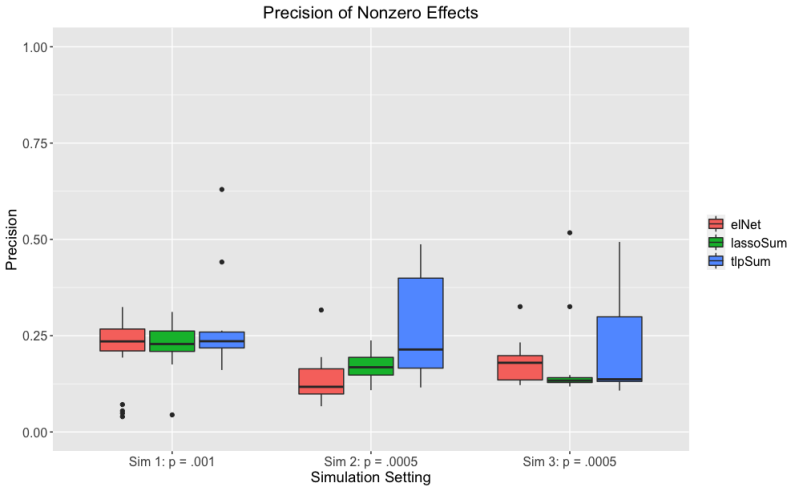

Figure 3.8: Precision of estimated nonzero effect sizes for the penalized regression methods applied to the three sparse simulation settings.

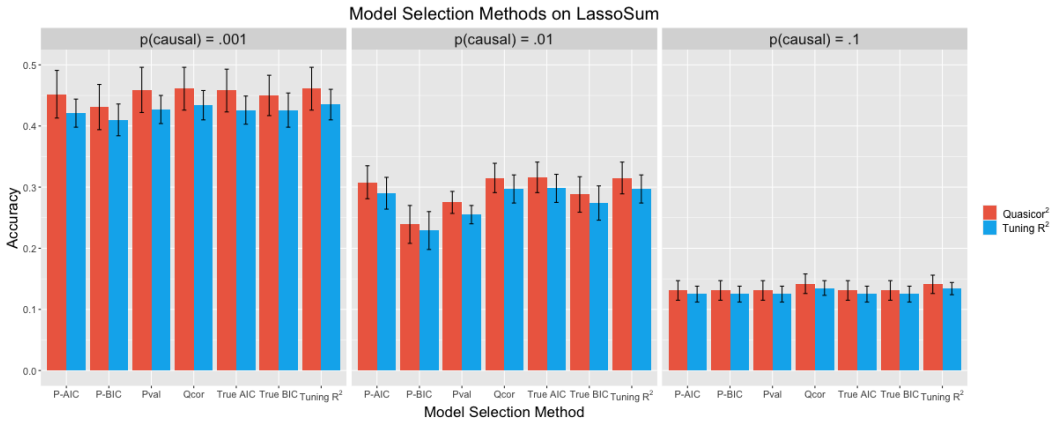

Figure 3.9: Performance of the seven different model selection methods applied to a set of candidate LassoSum models. Performance is measured by r^2 on the testing data (the right bar in each group), and by squared quasi-correlation on the testing data (the left bar in each group). Error bars represent the standard deviation across 20 replications.

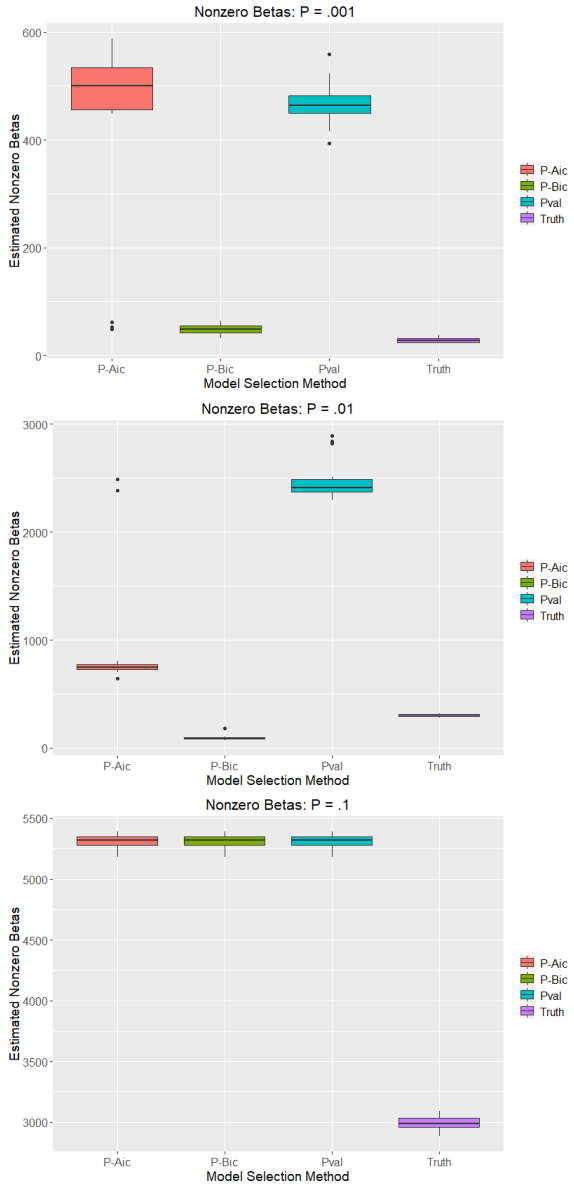

Figure 3.10: Number of estimated nonzero effects for each model selection method across each of the simulation settings in simulation 1. Models were selected from a set of candidate LassoSum models.

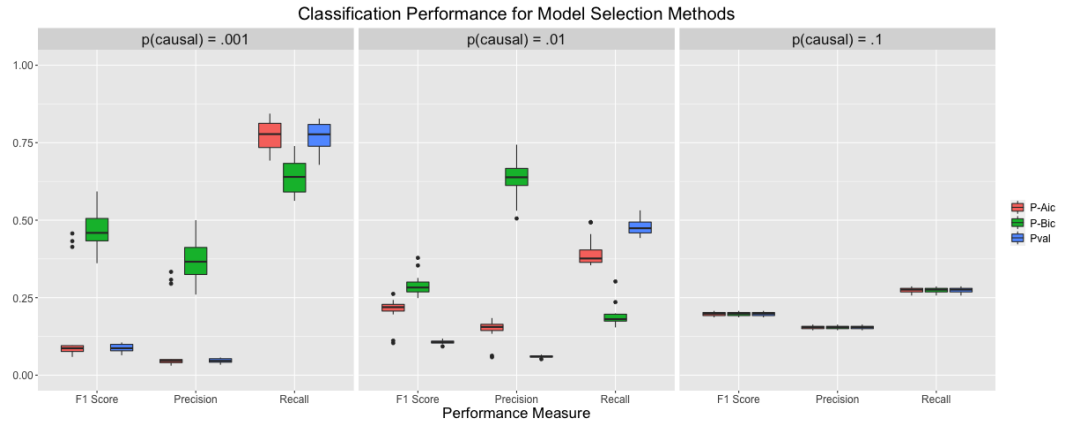

Figure 3.11: Performance of the selected models for each of the model selection methods across the different simulation settings of simulation 1, as measured by precision, recall, and F1 score. The leftmost box in each grouping of three corresponds to pseudo AIC, the center corresponds to pseudo BIC, and the rightmost corresponds to pseudovalidation. Models were selected from a set of candidate LassoSum models.

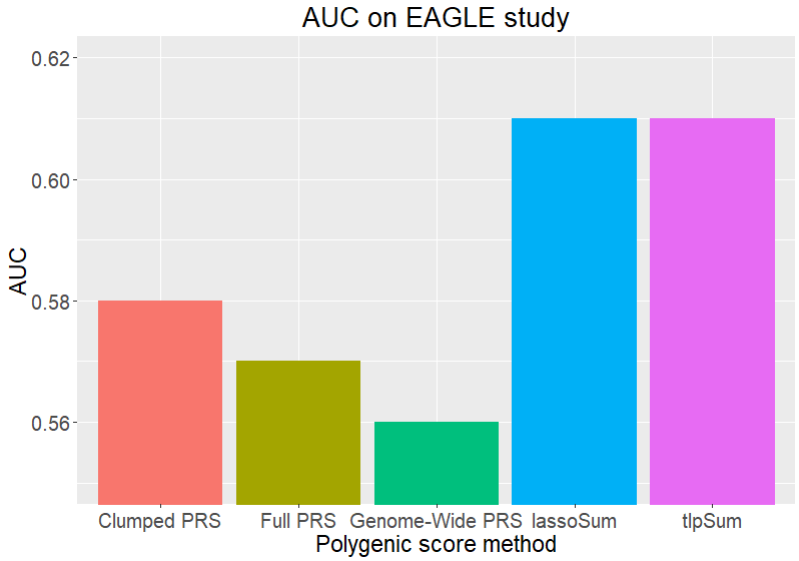

Figure 3.12: AUC on the EAGLE study for different methods of estimating polygenic risk scores.

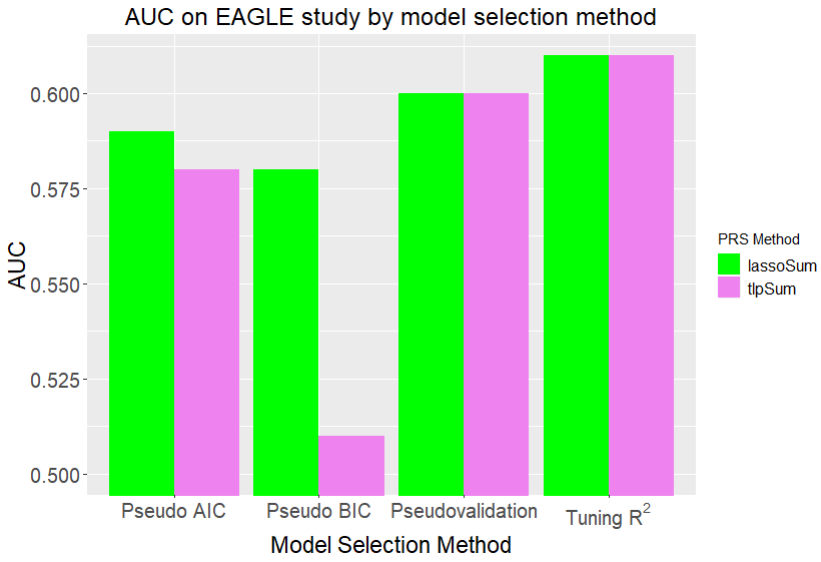

Figure 3.13: Performance of different model selection methods applied to candidate sets of TlpSum (the left bar in each group) and LassoSum models (the right bar in each group), as measured by AUC on the testing EAGLE data.

Chapter 4

Leveraging summary statistic data from the UK Biobank to identify endophenotypes associated with Alzheimer's disease

4.1 Introduction

Alzheimer's disease (AD) is a complex and multifactorial disease associated with substantial morbidity and mortality that is currently without a cure [82]. The risk of AD increases substantially with age, and thus the disease burden is expected to increase worldwide as the population ages [83, 84]. Alzheimer's disease is a neurodegenerative disease characterized by progressive changes in brain structure and function. AD is characterized by a long preclinical phase in which abnormal changes in brain structure and function accumulate asymptomatically [85]. As affected individuals age, these progressive changes begin to cause symptoms and lead to the genesis of Alzheimer's, typically after the age of 65 [86]. These progressive changes that characterize the preclinical state have been associated with genetic risk factors, in particular mutations

within the APOE gene [85]. Variations in APOE have also been associated with increased Alzheimer's risk [87]. In addition to APOE, rare, fully penetrant mutations in *Amyloid Precursor protien, Presenilin 1*, and *Presenilin 2* have been associated with an autosomally dominant form of Alzheimer's disease [86]. This evidence clearly describes a genetic basis for Alzheimer's risk. However, mutations in these genes do not explain the totality of the heritability of Alzheimer's disease, which is estimated to range from 60% to 80% [86]. In addition to the risk conferred by APOE, many other genetic loci have been associated with AD via genome-wide association studies (GWAS) [20]. Analyses of complex, multifactorial traits such as educational status, high cholesterol, diabetes, and high blood pressure have been associated with risk of AD as well [88, 89, 90]. This provides evidence that AD is a complex disease with a nuanced architecture of genetic risk.

Given the multifactorial nature of AD and the lack of a cure, identifying potentially modifiable risk factors is an area of great current interest [88, 91, 92]. Previous observational analyses have associated certain risk factors with increased Alzheimer's risk; however, these observational studies have difficulty characterizing whether an observed association is truly causal. Given this, it is of interest to find robust causal associations between modifiable endophenotypes and Alzheimer's risk. One framework which facilitates causal claims is so-called 'Mendelian randomizaiton', which is an instrumental variable approach that uses inherited single nucleotide polymorphisms as instruments. This approach has been used to associate AD risk with educational attainment [93]. Other studies have taken a more unstructured approach, instead analyzing many potential endophenotypes simultaneously to find associations with AD risk. Given the presumed multifactorial and complex genetic architecture of AD risk, many potentially associated endophenotypes may not be currently known. Given this, such an approach may be useful for generating avenues of further investigation.

One particular paper of interest, the so-called BADGERS method [22], scans 1,738 heritable traits from the UK Biobank and analyzes them in conjunction with GWAS data from the International Genomics of Alzheimer's Project (IGAP) to find novel associations between certain endophenotypes and AD risk. We take a similar approach in this paper. However, we extend upon the BADGERS approach by using a more powerful approach to estimating polygenic risk scores on summary statistics. Whereas the

BADGERS approach uses pruning and thresholding to estimate polygenic risk scores for the Biobank data, we use a penalized regression approach to estimate polygenic risk scores. There are two benefits of this approach. Firstly, estimating polygenic risk sores on summary statistics via penalized regression has been shown to increase predictive accuracy as compared to pruning and thresholding [16]. A more predictive polygenic risk score may facilitate the identification of additional associated endophenotypes. Secondly, penalized regression methods coupled with applications of the AIC and BIC for summary statistic data allow us to rigorously impose sparsity onto our polygenic risk score models. This may lessen the influence of pleiotropy. For the first point, we show that our approach identifies new putative associations between endophenotypes in the UK Biobank and AD, while picking up much of the same signal as the BADGERS approach. For the second, we provide evidence that our approach decreases the effect of pleiotropy on endophenotype-AD associations, thus potentially making the putative associations more robust.

4.2 Methods

4.2.1 Two-stage least squares

We contextualize our approach within the framework of two-sample two stage least squares (TS-2SLS). We use the language of 'two stage least squares' to denote the use of instrumental variables and call upon the framework to make causal claims, while noting that we do not apply least squares for model estimation. Ours is a two-sample approach, given that the first stage is estimated using data from the UK Biobank, and the second is estimated using data from IGAP. Motivated by the transcriptome wide association study framework (TWAS), 2SLS analyses have become popular in recent years for identifying endophenotypes that modulate the effect of inherited SNPs on a complex phenotype of interest [8]. While the TWAS approach specifically identifies associated genes by using gene expression as the endophenotype of interest, this approach can be extended to any endophenotype. This is the approach we take in our analysis.

We illustrate the true causal model via a directed acyclic graph in figure 4.1. We consider $\mathbf{Y}$ and $\mathbf{Z}$ to be some phenotypes, and we are interested in testing for the association of $\mathbf{Y}$ with $\mathbf{Z}$, corresponding to the dotted line in figure 1. In our application, $\mathbf{Z}$

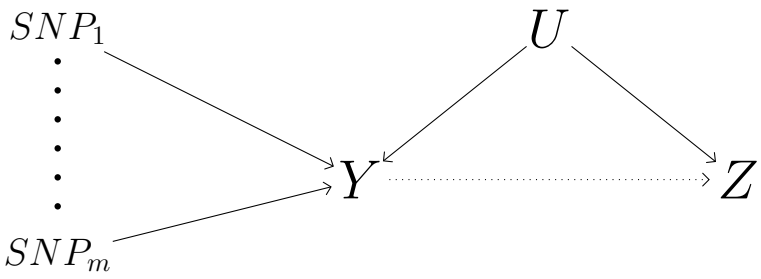

Figure 4.1: Directed acyclic graph demonstrating true causal model for TS-2SLS applications.

is AD status, and $\mathbf{Y}$ is some endophenotype from the UK Biobank data. $\mathbf{U}$ represents possible unobserved confounders. The SNPs represent inherited single nucleotide polymorphisms and are used as instrumental variables. The two-stage least squares analysis makes the following assumptions: (1) SNPs are associated with $\mathbf{Y}$ (valid instruments assumption), (2) SNPs are not associated with $\mathbf{U}$, and (3) conditional on $\mathbf{Y}$, SNPs are not associated with $\mathbf{Z}$ (no pleiotropy assumption).

The model can be formulated as follows. Consider the design matrix $\mathbf{X}_{n\times m}$ to be a genotype matrix of SNPs in additive coding. We assume the following true model:

$$\mathbf{Y} = \mathbf{X}\mathbf{W} + \phi_1\mathbf{U} + \boldsymbol{\epsilon}_1 \tag{4.1}$$

$$\mathbf{Z} = \alpha\mathbf{Y} + \phi_2\mathbf{U} + \boldsymbol{\epsilon}_2 \tag{4.2}$$

Note again that the confounders $\mathbf{U}$ are unobserved, and thus we may have endogeneity. If we had access to individual level data (that is, $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$), we could use the following working model for stage 1:

$$\mathbf{Y} = \mathbf{X}\mathbf{W} + \boldsymbol{\epsilon}_1 \tag{4.3}$$

From this stage 1 model we can calculate estimated SNP effects $\widehat{\mathbf{W}}$ and thus $\hat{\mathbf{Y}}$. We then have the following working model for stage 2:

$$\mathbf{Z} = \alpha\hat{\mathbf{Y}} + \boldsymbol{\epsilon}_2 \tag{4.4}$$

We would then perform inference on α to test for association of $\mathbf{Y}$ with $\mathbf{Z}$ as mediated by SNP effects. However, in our application to summary statistics from the UK Biobank data and the IGAP data, we do not have access to individual level data.

4.2.2 Considerations for summary statistics

We now consider the case where we do not have access to individual level data: that is, we do not have genotype matrix $\mathbf{X}$ or phenotype vectors $\mathbf{Y}, \mathbf{Z}$. Instead, we have univariate summary statistics quantifying the association between the SNPs and $\mathbf{Y}$, and likewise summary statistics quantifying the univariate association between the SNPs and $\mathbf{Z}$. Additionally, we consider the use of two-sample two-stage least squares (TS-2SLS); that is, the SNPs are measured on different individuals for the two sets of summary statistics. Formally, we have a set of summary statistics $\mathbf{S}_1$ quantifying the univariate association between $\mathbf{X}_1$ and $\mathbf{Y}$, and a second set of summary statistics $\mathbf{S}_2$ quantifying the univariate association between $\mathbf{X}_2$ and $\mathbf{Z}$. $\mathbf{X}_1$ and $\mathbf{X}_2$ correspond to non-overlapping samples, which reflects the two-sample nature of our analysis. We define $\mathbf{S}_i$ as containing the following information: marginal regression estimates generated via single linear regression $\hat{\boldsymbol{\beta}}_i^M$ and associated standard errors $SE(\hat{\boldsymbol{\beta}}_i^M)$. From this information, we also have univariate p-values, and univariate Z statistics $\boldsymbol{\gamma}_i$.

In a typical 2SLS approach, we would estimate SNP effects $\mathbf{W}$ with some estimator $\widehat{\mathbf{W}}$ and proceed to estimate $\hat{\mathbf{Y}}$, and test for association between $\hat{\mathbf{Y}}$ and $\mathbf{Z}$. This is not straightforward to do in the context of summary statistics. Existing method BADGERS proposes some summary statistic based approximations for doing so. Briefly, the Z-statistic corresponding to a significance test for parameter α is represented thusly:

$$Z_B = \widehat{\mathbf{W}}^T \boldsymbol{\Gamma} \boldsymbol{\gamma}_2 \tag{4.5}$$

Let us define $\boldsymbol{\Gamma}$ as a diagonal matrix with diagonal entries $\Gamma_{jj} = \sqrt{\frac{Var(\mathbf{X}_j)}{Var(\hat{\mathbf{Y}})}}$. Here, $Var(\mathbf{X}_j)$ corresponds to the variance of SNP j, and $Var(\hat{\mathbf{Y}})$ corresponds to the variance the predicted phenotype $\hat{\mathbf{Y}}$ in the target data $\mathbf{X}_2$. Note that we assume no individual level genotype data, making the above estimations intractable. We instead assume some reference panel $\tilde{\mathbf{X}}$ with approximately the same covariance structure as $\mathbf{X}_2$, and use this to estimate the variance of the SNPs and the variance of $\hat{\mathbf{Y}}$. $\boldsymbol{\gamma}_2$ is a vector of univariate Z statistics quantifying the assocation between the columns of $\mathbf{X}_2$ and $\mathbf{Z}$.

4.2.3 Contrasting BADGERS with new penalized regression approach

In their application, the authors of BADGERS use univariate regression estimates $\hat{\boldsymbol{\beta}}_1^M$ in conjunction with pruning and thresholding as estimates of the SNP effects $\mathbf{W}$. In particular, BADGERS uses a p-value cutoff of .01 and a LD cutoff of .1. This is an appealing approach in terms of ease of application, given that univariate regression estimates are readily available from reference data. However, we believe this approach may be suboptimal. Generally, we expect SNP effects for any phenotype to be sparse, meaning that the majority of the entries in $\mathbf{W}$ are zero. We make the valid IV assumption in the 2SLS approach, which is violated when we include SNPs with nonzero effect. We would like some way of imposing sparsity in our estimation of $\mathbf{W}$. Some sparsity is imposed by the pruning and thresholding approach, although this approach is highly dependent on cutoffs for LD and especially p-value. Selection of these cutoffs is not straightforward in the absence of validation data, which we assume is the case here. We would like to impose sparsity in a more rigorous and principled way, and ideally one that is data-driven and can vary by endophenotype. An additional concern is that the use of univariate regression estimates does not consider the joint structure of the true effects. To address these issues, we estimate models using the truncated LASSO penalty for summary statistic data, i.e. the TlpSum method detailed in chapter 3. Selection of tuning parameters for the TlpSum is difficult in the absence of validation data. To address this problem, we implement model selection methods pseudovalidation [16], pseudo AIC, and pseudo BIC. These approaches for model selection substantially influence ensuing inference, and in this chapter we compare their properties and relate them to the three assumptions of 2SLS.

A secondary concern is that the association test defined in (4.5) may be suboptimal as well. The derivation of Z_B assumes that the variance of $\mathbf{Z}$ explained by $\mathbf{Y}$ is negligible; this causes the test statistic to be conservative. Additionally, this testing structure assumes that all SNPs with nonzero weights have equal and nonzero effect size. We instead consider the use of weighted adaptive sum of powered score tests (or weighted aSPU tests), which robustly identify associations given a variety of underlying genetic architectures. The use of weighted aSPU tests for the identification of associated endophenotypes is well documented in literature [30]. Specifically, we consider the SPU(1)

('Sum test'), SPU(2) ('SSU test'), and aSPU tests. We briefly describe the weighted SPU framework here. Consider that, instead of score statistics, we use Z statistics $\boldsymbol{\gamma}_2$. For each of the m SNPs, we have estimated effects $\widehat{\mathbf{W}}$. We define weighted Z statistics as:

$$\tilde{\boldsymbol{\gamma}}_2 = (\boldsymbol{\gamma}_{21}\hat{w}_1, \ldots, \boldsymbol{\gamma}_{2m}\hat{w}_m)'$$

We define the weighted SPU(γ) test statistic as follows:

$$T_{SPU(\tau)} = \sum_{i=1}^{m} \tilde{\boldsymbol{\gamma}}_{2i}^{\tau}$$

We note that the weighted sum test confers specific a specific causal interpretation that does not hold for other choices of τ. Additionally, while the weighted SSU and aSPu tests demonstrate good power and robust identification of associations when the no pleiotropy assumption holds, they incur substantial type I error under pleiotropy. Thus, we consider only the weighted sum test in application to real data.

4.2.4 Comparison to other methods

We compare our penalized regression approach described in **(4.2.3)** and the BADGERS method discussed in **(4.2.2)** to other methods for instrumental variable analysis on genetic data, namely two-sample Mendelian randomization and single and multivariable tests of association based on $\hat{\boldsymbol{\beta}}_i^M$. We briefly describe these methods here.

Mendelian randomization (MR) uses SNPs as instrumental variables to assess the causal relationship between endophenotype $\mathbf{Y}$ and complex trait $\mathbf{Z}$. In contrast to our approach and the BADGERS approach, which leverage polygenic risk scores in stage 1 to jointly model the genetic component of endophenotype $\mathbf{Y}$, MR analyses use univariate association statistics directly. For some SNP i with associated SNP-endophenotype marginal regression estimate $\hat{\beta}_{1i}^M$ and SNP-trait marginal regression estimate $\hat{\beta}_{2i}^M$, the estimated causal effect of the endophenotype on the trait as mediated by SNP i is $\hat{\beta}_i^C = \hat{\beta}_{2i}^M / \hat{\beta}_{1i}^M$, with associated standard error $SE(\hat{\beta}_i^C) = SE(\hat{\beta}_{2i}^M)/\hat{\beta}_{1i}^M$. A test for significance can then be performed on $\hat{\beta}_i^C$ to test for a causal association of endophenotype $\mathbf{Y}$ and trait $\mathbf{Z}$ mediated by SNP i.

There are a couple of limitations of MR as phrased above. Firstly is that this approach considers only a single SNP at a time, and thus cannot leverage multiple

causal SNPs to increase power. Several methods have been developed to ameliorate this issue and consider multiple SNPs at once; they include simple mode, weighted mode, weighted median, inverse variance weighting, and Egger regression. These methods informatively aggregate multiple instruments (say q) to increase the power to detect a causal association. These are implemented in our application via the MR-Base R package [94]; we give particular consideration to Egger regression and inverse variance weighting. A pertinent issue is how to select the q relevant SNPs. Note that selecting invalid instruments, i.e. SNPs that are not associated with endophenotype $\mathbf{Y}$, will violate assumption 1 of the 2SLS analysis. A practical approach to this is to only use SNPs that are associated with endophenotype $\mathbf{Y}$ below some p-value cutoff. A second limitation is that MR does not estimate SNP effects under linkage disequilibrium, instead using marginal effects. Thus, it can be said that the MR approach does not model the true effect structure of the SNPs. This could lead to loss of power if the true structure of SNP effects is not well modeled by the marginal effects, i.e. when multiple causal SNPs are in high linkage disequilibrium.

Single tests of association based on $\hat{\boldsymbol{\beta}}_i^M$ are similar to the weighted Sum test, which is the $\tau = 1$ case of the weighted aSPU test strucutre we describe in **(4.2.3)**. The distinction between the two approaches is elucidated in previous literature [95]. Expanding on this approach are multivariable tests of association based on $\hat{\boldsymbol{\beta}}_i^M$. The multivariable approach facilitates the modeling of some imputed endophenotype $\mathbf{Y}_k$ conditional on some set of other imputed endophenotypes. Thus, given a set of endophenotypes that are marginally associated with our trait of interest $\mathbf{Z}$, the multivariable approach allows us to parse which endophenotypes from this set are robustly associated with $\mathbf{Z}$. We describe the modeling considerations below. Credit for this idea is given to Knutson et al [96].

Say we are investigating r endophenotypes for causal association with our trait of interest $\mathbf{Z}$. We use the structure of TS-2SLS. Let us denote the genotype matrix for endophenotype i as $\mathbf{X}^{(i)}$, the phenotype vector as $\mathbf{Y}_i$, and the SNP effects as $\mathbf{W}^{(i)}$. Note that $\mathbf{W}^{(i)}$ is a $p \times 1$ vector, where p is the number of SNPs. If a given trait is modeled by p_i SNPs with $p_i < p$, let the remaining $p - p_i$ elements be set to zero. We

have working model for stage 1 as in equation 1.

$$\begin{cases} \mathbf{Y}_1 = \mathbf{X}^{(1)}\mathbf{W}^{(1)} + \boldsymbol{\epsilon}^{(1)} \\ \vdots \\ \mathbf{Y}_r = \mathbf{X}^{(r)}\mathbf{W}^{(r)} + \boldsymbol{\epsilon}^{(r)} \end{cases} \tag{4.6}$$

We use equation 1 to estimate $\hat{\mathbf{Y}}_i$ given some $\widehat{\mathbf{W}}_i$ for each of the r endophenotypes. Let us now define matrix $\mathbf{W} = \left(\mathbf{W}^{(1)}, \ldots, \mathbf{W}^{(r)}\right)$. In stage 2, we want to test association between the r endophenotypes and our trait of interest $\mathbf{Z}$. To do this, we use the working model described in equation 4.7.

$$\mathbf{Z} = \hat{\mathbf{Y}}_1\beta_1 + \cdots + \hat{\mathbf{Y}}_r\beta_r \tag{4.7}$$

We wish to estimate conditional effects $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_r)'$ and their variances. In the case where we have individual level data, this could be done as in equations 4.8 and 4.9.

$$\hat{\boldsymbol{\beta}} = (\mathbf{W}'\mathbf{X}'\mathbf{X}\mathbf{W})^{-1}\mathbf{W}'\mathbf{X}'\mathbf{Z} \tag{4.8}$$

$$var(\hat{\boldsymbol{\beta}}) = (\mathbf{W}'\mathbf{X}'\mathbf{X}\mathbf{W})^{-1}\frac{\mathbf{Z}'\mathbf{Z} - \hat{\boldsymbol{\beta}}'\mathbf{W}'\mathbf{X}'\mathbf{Z}}{n-r} \tag{4.9}$$

Estimating the marginal effects $\hat{\boldsymbol{\beta}}^{\zeta} = (\hat{\beta}_1^{\zeta}, \ldots, \hat{\beta}_r^{\zeta})'$ is done as follows:

$$\hat{\boldsymbol{\beta}}^{\zeta} = (diag(\mathbf{W}'\mathbf{X}'\mathbf{X}\mathbf{W}))^{-1}\mathbf{W}'\mathbf{X}'\mathbf{Z}, \tag{4.10}$$

With associated standard errors

$$Var(\hat{\beta}_i^{\zeta}) = (\mathbf{W}_i'\mathbf{X}_i'\mathbf{X}_i\mathbf{W}_i)^{-1}\frac{\mathbf{Z}'\mathbf{Z} - \hat{\beta}_i^{\zeta}\mathbf{W}_i'\mathbf{X}_i'\mathbf{Z}}{n-1}. \tag{4.11}$$

Given these estimates, we can perform significance tests for the conditional and the marginal estimates. Note that the formulation above assumes individual level data. If we do not have access to individual level data, we must make some approximations based on summary statistics. In particular, we need approximations of $\mathbf{X}'\mathbf{X}$, $\mathbf{Z}'\mathbf{Z}$, $\mathbf{X}'\mathbf{Z}$. These approximations for summary statistics are well characterized chapter 3 i.e. in equations (3.6), (3.7), and (3.8). An additional consideration is that the summary statistics for the IGAP data ($\hat{\boldsymbol{\beta}}_2^M$ in our notation) are for logistic regression. We convert those estimates to linear regression estimates as in **(3.2.4)**, and then proceed with the estimation of these $\mathbf{X}'\mathbf{X}$, $\mathbf{Z}'\mathbf{Z}$, $\mathbf{X}'\mathbf{Z}$ quantities via summary statistics.

4.2.5 Discussion of modeling assumptions

We comment here on the application of our methodology with respect to the three assumptions for TS-2SLS, namely: (1) SNPs are associated with $\mathbf{Y}$ (valid instruments assumption), (2) SNPs are not associated with $\mathbf{U}$, and (3) conditional on $\mathbf{Y}$, SNPs are not associated with $\mathbf{Z}$ (no pleiotropy assumption). We compare three different methods for stage 1 model selection in this paper, namely pseudovalidation, pseudo AIC, and pseudo BIC. As characterized in chapter 3, pseudovalidation will select the densest models, pseudo BIC the sparsest, and pseudo AIC in between. Likewise, pseudo BIC has the best performance recovering the true model as measured by precision (i.e. the proportion of estimated nonzero effects that are truly nonzero), followed by pseudo AIC, and lastly pseudovalidation. This has relevance to modeling assumptions (1) and (3). Straightforwardly, a model selection method that tends to have lower precision will violate the valid instruments assumption, in that it will select a larger proportion of SNPs that are not associated with $\mathbf{Y}$. Assumption (3) is relevant especially in the context of complex, multifactorial endophenotypes where a large number of SNPs may be causal genome-wide for a given endophenotype. If a model selection method selects a model for some endophenotype $\mathbf{Y}_i$ with a large proportion of nonzero weights, the probability increases that a substantial proportion of those upweighted SNPs are also associated with another causal endophenotype $\mathbf{Y}_k$. Thus, a significant association between $\mathbf{Y}_i$ and $\mathbf{Z}$ may be confounded by the association of $\mathbf{Y}_k$ with $\mathbf{Z}$.

We comment on the relevance of these assumptions to our methodology throughout the paper. We demonstrate in real data application that the BADGERS method and the pseudovalidation approach for model selection behave similarly, while pseudo AIC and pseudo BIC behave substantially differently. We demonstrate some results illustrating that associated endophenotypes identified by pseudovalidation may be confounded by pleiotropic effects, whereas this effect may be less pronounced in those models selected by pseudo AIC and pseudo BIC. We in particular seek to distinguish pseudo AIC and pseudo BIC, which implement sparsity according to well defined theory, from pseudovalidation, which imposes a penalty on model size in a more ad hoc manner. This makes the case that an approach to Biobank-wide TS-2SLS inference may be best facilitated by the application of methods for model estimation and selection that impose sparsity and

parsimony in a principled way; namely, the utilization of TlpSum for model estimation in conjunction with pseudo AIC and pseudo BIC for model selection.

4.3 Results

4.3.1 Simulation study

To demonstrate the usefulness of our approach, we conduct a simulation study to examine the power and type I error of methods for TS-2SLS. Specifially, we investigate power and type I error in the case where there is and is not pleiotropy, respectively. This simulation study is an extension of simulation 1 from section **(3.3.1)**, which I will briefly recap here.

We simulated quantitative phenotypes using genotype data from the Wellcome Trust Case Control Consortium, or WTCCC [76]. The simulation used genotype data from chromosomes 1 and 4 ($\sim 30,000$ SNPs). The data was comprised of 10,919 individuals. The data were split into three sets: training (corresponding to $\mathbf{X}_1$), which consisted of 6240 individuals, tuning (corresponding to $\tilde{\mathbf{X}}$, which consisted of 3119 individuals, and testing (corresponding to $\mathbf{X}_2$), which consisted of 1560 individuals. We pruned SNPs such that no two included SNPs were in linkage disequilibrium higher than 0.9. We additionally removed all ambiguous SNPs (A/T, C/G), and all SNPs with $MAF < .01$.

We simulated SNP effect sizes $\mathbf{W}$ from the point normal model:

$$W_j \sim_{iid} \begin{cases} N(0, \frac{h^2}{Mp}), \text{ with probability p} \\ 0, \text{ with probability 1-p} \end{cases}$$

Where h^2 is the SNP-based heritability of the disease (0.5 in our simulation), M is the number of SNPs, and p is the fraction of causal SNPs. We considered values for $p \in [.01, .001]$. Using the genotype data from the WTCCC and SNP effects $\mathbf{W}$ as simulated above, we simulated $\mathbf{Y}$ as in equation 4.12. We conduct 20 replications; that is, 20 different vectors $\mathbf{W}$. From training data (corresponding to $\mathbf{X}_1$) with $N = 6240$, we estimated summary statistics quantifying the association between $\mathbf{X_1}$ and $\mathbf{Y}$. We then estimated TlpSum models using tuning data $\tilde{\mathbf{X}}$ as a reference panel. In the testing data $\mathbf{X}_2$ with simulated phenotype $\mathbf{Y}$, we simulated $\mathbf{Z}$. In the case where we are simulating

pleiotropy, we also simulate $\mathbf{Y}_2$ which is in pleiotropy with $\mathbf{Y}$, and simulated $\mathbf{Z}$ based on $\mathbf{Y}$ and $\mathbf{Y}_2$. We used the following model to simulate $\mathbf{Y}$ and $\mathbf{Z}$ under no pleiotropy:

$$\mathbf{Y} = \mathbf{X}_2\mathbf{W} + \boldsymbol{\epsilon}_1 \tag{4.12}$$

$$\mathbf{Z} = \alpha\mathbf{Y} + \boldsymbol{\epsilon}_2 \tag{4.13}$$

We used the following model to simulate under pleiotropy:

$$\mathbf{Y} = \mathbf{X}_2\mathbf{W} + \boldsymbol{\epsilon}_1 \tag{4.14}$$

$$\mathbf{Y}_2 = \mathbf{X}_2\mathbf{W}_2 + \boldsymbol{\epsilon}_1 \tag{4.15}$$

$$\mathbf{Z} = \alpha_1\mathbf{Y} + \alpha_2\mathbf{Y}_2 + \boldsymbol{\epsilon}_2 \tag{4.16}$$

Note that we did not simulate endogeneity here, i.e. $\phi_1, \phi_2 = 0$ as per equation (4.1). For the no pleiotropy simulations, the $\mathbf{Y}$-based heritability of $\mathbf{Z}$ was varied as follows: $h_z^2 \in [0, .003, .005, .008, .01, .025]$. For all nonzero values of h_z^2, we simulated two $\mathbf{Z}$ phenotypes per each of the 20 simulation settings, for a total of 40 $\mathbf{Z}$ phenotypes per nonzero h_z^2 value. We use these to establish the power of the different approaches to 2SLS association testing. For each of the 20 simulation settings, we simulated five $\mathbf{Z}$ phenotypes with $h_z^2 = 0$ to assess the type I error of the testing procedures. This entails a total of 100 $\mathbf{Z}$ phenotypes with $h_z^2 = 0$.

Pleiotropy was imposed via simulation as follows. For each replication, we simulated $\mathbf{W}_2$ such that 75% of the SNPs with nonzero effect size estimates in $\mathbf{W}$ also had nonzero effect size estimates in $\mathbf{W}_2$. This is so-called 75% pleiotropy. This is a straightforward violation of the no plieotropy assumption, i.e. assumption 3 outlined in **(4.2.1)**. Heritability of $\mathbf{Z}$ in terms of $\mathbf{Y}$ and $\mathbf{Y}_2$ was simulated according to table 4.1. Note the term 'heritability' is somewhat loose here; it is difficult to define an exact heritability when simulating under equation (4.16) due to the correlation between $\mathbf{Y}$ and $\mathbf{Y}_2$. Thus, $\mathbf{Z}$ was simulated such that the heritability for $\mathbf{Y}$, $\mathbf{Y}_2$ are those in table 4.1 assuming no correlation between $\mathbf{Y}$, $\mathbf{Y}_2$. For each scenario with $\alpha_1 = 0$, we conducted 60 replications. For each scenario with $\alpha_1 \neq 0$, we conducted 40 replications.

We used estimated TlpSum models as the weights in the weighted SPU framework described in **(4.2.3)**. We considered three different methods for selecting TlpSum models: pseudo AIC, pseudo BIC, and pseudovalidation. We also conducted association testing using the BADGERS framework; that is, estimating $\mathbf{W}$ using P+T with p-value

α_1	α_2
0	.01
0	.05
0	.08
0	.15
0	.2
0	.3
.003	.01
.003	.05
.005	.01
.005	.05

Table 4.1: Simulated heritabilities under pleiotropy.

cutoff .01 and LD cutoff .1. For $\widehat{\mathbf{W}}$ in equation (4.3) we also considered using the full vector of marginal SNP effects $\hat{\boldsymbol{\beta}}$ from the training data. Lastly, we conducted inference using Mendelian randomization, in particular the methods inverse variance weighting (IVW) and MR-Egger. Instruments were selected to be those SNPs with univariate $p < 5 \times 10^{-8}$ for the association between $\mathbf{X}_2$ and $\mathbf{Y}$.

Results comparing the performance of various methods under no pleiotropy are in figure 4.2. Results comparing power and type I error of the various methods under pleiotropy are in figures 4.3, 4.4. Results for weighted SSU and aSPU tests are excluded in the above figures. Despite the reasonable power to detect associations of weighted SSU and aSPU tests, they show very poor type I error control under pleiotropy, as detailed in figure 4.5. This motivates the exclusion of weighed SSU and aSPU tests from our application to the IGAP data.

Figure 4.2 demonstrates that our approach can control type 1 error and achieve better power than the BADGERS approach (denoted PRS P+T) when the simulation involves no pleiotropy. The three model selection methods are relatively indistinguishable in this simulation in terms of respective power. All model selection methods outperform BADGERS. Generally, all methods demonstrate better power when the proportion of causal SNPs $p = .001$; this coheres with our simulation results in **(3.3.1)** showing increased predictive accuracy when $p = .001$. We note the relatively good performance of Mendelian randomizaiton with inverse variance weighting when the proportion of causal SNPs $p = .001$, but the perfomance decays somewhat when the fraction of causal

SNPs increases to $p = .01$. Given that the MR methods selected instrument SNPs with a 5×10^{-8} univariate p-value cutoff, it's likely that these methods did a better job of selecting useful instruments when the causal proportion of SNPs is smaller and thus individual effect sizes is larger. MR-Egger is substantially underpowered in both simulations.

In figure 4.3, we see that many methods for TS-2SLS inference have some trouble controlling type I error when there is substantial pleiotropy. The most type I error is incurred by our methods, i.e. TlpSum weights with tuning parameters selected via pseudo AIC, pseudo BIC, and pseudovalidation. The PRS P+T and simple PRS methods also incur substantial type I error. The Mendelian randomization methods incur notably less type I error. In figure 4.4, we see that generally our methods have superior power when $\mathbf{Z}$ is simulated under pleiotropy.

In figure 4.5, we see that the weighted SSU and aSPU tests incur massive amounts of type I error under pleiotropy. This motivates the exclusion of these methods when we apply our methodology to real data. Given that weighted SSU and aSPU tests will maintain substantial power even if many upweighted SNPs are not truly causally associated, the upweighting of even a few SNPs that are causally associated with $\mathbf{Z}$ via pleiotropy will likely generate a false positive. Thus, despite the improved power of these methods, they may not be appropriate analysis tools when there may be substantial pleiotropy, such as when complex, multifactorial traits are used as endophenotypes.

We note here that we fail to see the improvement of pseudo AIC and pseudo BIC in controlling type I error under pleiotropy as compared to pseudovalidation and PRS P+T (i.e. BADGERS), which we suggest is the case in **(4.2.5)**. Given the results in real data application that do indicate that the pseudovalidation approach has more problems with pleiotropy, we conclude here that more simulations with varied settings are needed. In particular, it may be beneficial to simulate $\mathbf{Y}$ with a SNP-based heritability less than .5. In practice, most endophenotypes will have substantially lower SNP-based heritability, and thus penalties on model size may become more relevant. In the case where $\mathbf{Y}$ is well modeled by SNP effects, the model selection methods may select fairly similar models, and thus it is more difficult to differentiate among their performance. Nevertheless, this simulation study establishes the following: the improved power of our methodology as compared to BADGERS and especially MR, the issue of type I error under pleiotropy,

and the relative infeasibility of applying weighted SSU and aSPU tests in situations where we suspect substantial pleiotropy among endophenotypes.

4.3.2 UK Biobank and IGAP data

We apply the weighted sum testing framework described in **(4.2.3)** to the UK Biobank [19] and International Genomics of Alzheimer's Project (IGAP) data [20]. In this application, we consider 1738 heritable traits in the UK Biobank as our stage 1 phenotypes $\mathbf{Y}$ and Alzheimer's status as our case control phenotype $\mathbf{Z}$ in stage 2. These 1738 traits were identified as nominally heritable in the BADGERS application [22]; that is, an LD score regression estimate of heritability attained $p < .05$. The analysis proceeds as follows. For each of the 1,738 heritable traits in the UK Biobank, we use summary statistics to estimate penalized regression models using TlpSum, and then do model selection using pseudo AIC, pseudo BIC, and pseudovalidation. As a reference panel for the TlpSum models, we use the 1000G data for 503 individuals of European descent [50]. As a reference panel for the model fitting criteria and the weighted sum test, we use densely sequenced ADSP data for 612 controls of European descent. We use summary statistics from the IGAP data weighted by the coefficients from the TlpSum models to conduct weighted sum testing.

Summary statistic files (version 3) for the 1,738 heritable traits were downloaded from the UK Biobank. Summary statistic files contain information on univariate linear regression estimates for each SNP, as well as associated standard errors. Linear regression was used to estimate SNP effects for all phenotypes, including binary and ordinal phenotypes. Sex and ten principal components were included as covariates [97]. Sample size for the UK Biobank data varied by trait. The IGAP data is a meta-analysis of four GWAS studies of AD, with 17,008 cases and 37,154 controls. We limited the data to only those SNPs present in all four datasets; that is, the two summary statistic datasets UK Biobank and IGAP, and the two reference panels 1000G and ADSP. We then performed LD pruning using the 1000G data as a reference such that no two remaining SNPs are in LD $R^2 > .9$. This left us with a set of roughly 750,000 SNPs. These 750,000 were further reduced to a set of 100,000 by sure independence screening [98] before TlpSum model estimation. Sure independence screening was performed uniquely for each of the 1738 heritable endophenotypes. We estimated candidate TlpSum models using a three

dimensional grid search over λ, τ, and s, and used the model selection methods to select from among the candidate models. We note that, for pseudovalidation, local FDR rates were estimated after initial LD pruning, but before the application of SIS. Thus, local FDR rates were estimated on the set of $\sim 750,000$ SNPs. After models were fit and selected, the chosen TlpSum model was used as weights for a weighted sum test. The weighted sum test was estimated using summary statistics from the IGAP data, and the ADSP data as a reference panel.

4.3.3 Weighted sum testing results

Using weighted sum tests weighted by effect size estimates from TlpSum models, we identify numerous endophenotypes from the UK Biobank potentially associated with AD risk. Using a multiple testing correction of 1738 to account for endophenotype multiplicity and thus a significance cutoff of .05/1738, pseudovalidation identifies 68 associated endophenotypes, while pseudo AIC identifies 38 and pseudo BIC identifies 25. P-values delineated by endophenotype category for the three model selection methods are displayed in figure 4.6. As a point of comparison, p-values for the BADGERS analysis are reproduced in figure 4.7. Note that figure 4.7 directly uses data from the BADGERS supplementary file. P-values smaller than 1×10^{-15} were truncated in order to facilitate plotting.

We notice some ostensible similarities between these results. Both BADGERS and pseudovalidation identify a substantial proportion of cognitive and educational traits as associated with AD risk. This coheres with previous analyses identifying educational attainment as potentially causal for AD [93]. There are notable discrepancies in the results from these analysis; for example, pseudo BIC identifies a comparatively larger number of medication related endophenotypes than the other approaches, while finding only one associated cognitive / educational endophenotype.

We further illustrate the similarities and differences between the four approaches here. To compare the p-values for the four different analyses, we present scatter plots of $-log_{10}(p)$ for the four analyses in figure 4.8, and a table of the correlation of $-log_{10}(p)$ values in table 4.2. For both the calculation of correlations and the scatterplotting, P-values were truncated at 1×10^{-15}. Additionally, p-values corresponding to traits with null models selected for stage 1 were set equal to one. From these results, we see that

there is a reasonable degree of similarity between the four analyses. The most similar are pseudovalidation and BADGERS; this makes intuitive sense, because these methods may have a less stringent penalty on model size, and may tend to overfit the training data. Despite the similarity between BADGERS and pseudovalidation, we reiterate that pseudovalidation finds 68 significant traits as opposed to 50 for BADGERS. This may be because the true genetic effects are better estimated with the TlpSum than they are with pruning and thresholding; this coheres with our previous work comparing the predictive accuracy of polygenic risk score models in chapter 3. We note the relatively low correlation between the p-values from BADGERS, and the p-values from the pseudo AIC and pseudo BIC analyses. Additionally, pseudo AIC and pseudo BIC find a fairly disjoint set of significant traits as compared to BADGERS. A Venn diagram describing the overlapping results of the four analyses is displayed in figure 4.9.

	Pseudo BIC	Pseudoval	BADGERS
Pseudo AIC	.61	.56	.48
Pseudo BIC		.46	.38
Pseudoval			.81

Table 4.2: Correlation of $-log10(p)$ for the four analyses.

Nine endophenotypes were found to be associated with AD risk by all four approaches. Those traits, and associated p-values, are listed in table 4.3. Note that p-values $< 1 \times 10^{-50}$ are truncated. There is some redundancy in the endophenotypes described in this table; given the richly phenotyped nature of the UK Biobank and our agnostic approach that considers all endophenotypes, this is reasonable.

We see some trends in the significantly associated endophenotypes. All methods repeatedly identify strong associations related to diagnoses of dementia in the subject, the subject's mother, or the subject's father. Multiple medications and traits relating to high cholesterol are also identified, which is reasonable given that high cholesterol is a known risk factor for AD [89]. Some methods, in particular pseudo BIC, identify multiple medications related to asthma or other lung issues. This is potentially meaningful, given that some studies indicate asthma may be associated with increased risk of AD [99]. Risk of diabetes and high blood pressure in family members are also identified as significant. This coheres with literature on Alzheimer's risk factors [88]. Some other

Endophenotype	Pseudo AIC	Pseudo BIC	Pseudoval	BADGERS
Mother still alive	1.00E-50	1.00E-50	3.36E-12	5.10E-08
High cholesterol	1.00E-50	1.00E-50	2.26E-22	6.73E-16
Taking Simvastatin	5.54E-39	1.00E-50	3.56E-09	1.82E-05
Taking Atorvastatin	1.13E-11	1.00E-50	1.98E-48	1.90E-08
Father AD	1.00E-50	1.00E-50	1.00E-50	7.85E-20
Mother AD	1.00E-50	1.00E-50	1.00E-50	2.50E-61
Taking some cholesterol med	1.00E-50	1.00E-50	3.70E-26	8.04E-08
Taking some cholesterol med	1.00E-50	1.00E-50	4.95E-14	3.32E-07
Any dementia	2.19E-23	2.19E-23	1.00E-50	2.04E-09

Table 4.3: Nine phenotypes identified as significant by all four approaches, and corresponding p-values.

identified endophenotypes are unusual, bordering on nonsensical. For example, pseudovalidation and BADGERS both find a strong association of AD with cheese intake. It seems fairly unlikely that there is a significant effect of cheese intake on AD as mediated by SNP effects. Results such as these are likely false positives due to pleiotropy.

It's clear that the four modeling approaches each pick up substantial signal in that several significantly associated endophenotypes are identified by each approach. There is some overlap in the results of the four approaches, although it's clear that each approach picks up a unique signal as the results are quite different. A deeper discussion of these results and the consequences of the different modeling techniques for the 2SLS modeling assumptions is in section **(4.4)**.

4.3.4 Single and multivariable association tests based on marginal effects

To further investigate the significant associations identified by the weighted sum test in **(4.3.3)**, we apply single and multivariable tests of association based on marginal SNP effect estimates as described in **(4.2.4)**. Note that single tests of association based on $\hat{\boldsymbol{\beta}}_M$ are very similar, but not identical, to weighted sum tests. A more in depth exploration of the sum test versus tests based on $\hat{\boldsymbol{\beta}}_M$ can be found in Pan et al [95].

We estimated multivariable tests of association as follows. We considered three different multivariable models: one of the 68 endophenotypes identified as significant via pseudovalidation (and those corresponding pseudovalidation weights $\widehat{\mathbf{W}}$), one of the 38

endophenotypes identified by pseudo AIC (and selected pseudo AIC weights), and one of the 25 identified as significant by pseudo BIC (and selected pseudo BIC weights). A given set of endophenotypes as described above corresponds to the $\mathbf{Y}_1, \ldots, \mathbf{Y}_r$ in equation (4.6). Before modeling, we imputed endophenotypes into the ADSP data for each of the three models, and iteratively pruned away the endophenotype with the highest variance inflation factor until no remaining endophenotype had a variance inflation factor > 10. This reduced the number of endophenotypes to 59 for pseudovalidation, 18 for pseudo BIC, and 36 for pseudo AIC.

For the pseudovalidation analysis, 54 of the 59 traits are found to be significant with the univariable analysis, whereas only 32 traits are found to be significant with the multivariable approach. For the pseudo AIC analysis, 32 out of the 36 traits are found to be significantly associated with the univariable approach, and 27 were found significant with the multivariable approach. For the pseudo BIC analysis, 16 traits were significantly associated with the univariable approach, while 15 traits were significnatly associated with the multivariable approach. This is evidence that a conditional analysis substantially reduces the number of associated traits for pseudovalidation, while the same does not occur for the pseudo AIC or the pseudo BIC. This is evidence that pleiotropy may be a substantial driver of significant associations identified with pseudovalidation, while not so for the pseudo AIC and pseudo BIC.

4.3.5 Mendelian randomization results

We conducted a Mendelian randomization analysis via summary statistics using the 1738 heritabile phenotypes from UK Biobank and summary statistics for Alzheimer's from IGAP. The analysis was done in R using the 'TwoSampleMR' package. We considered two Mendelian randomization approaches, namely MR-Egger and inverse variance weighting. We selected instruments to be those SNPs with univariate p-value $< 5 \times 10^{-8}$ that remained after LD clumping with an LD cutoff $R^2 < .001$. After a Bonferroni correction for the 1738 traits, MR-Egger found 14 significant associations, while inverse variance weighting found 17 significant associations. These results are displayed in figure 4.10.

The Mendelian randomization approach is underpowered to detect effects as compared to the PRS-based TS-2SLS approaches. It could be that the true structure of

the SNP effects is non-sparse and substantially influenced by linkage disequilibrium, meaning that PRS approaches are much more powerful. Alternatively, it may be that the PRS-based approaches used in our analysis and in BADGERS are more susceptible to pleiotropy. We also note that the significant traits identified by Mendelian randomization methods are fairly sensitive to the p-value cutoff used to determine which SNPs are included as instrumental variables. One benefit of our pseudo AIC and pseudo BIC methodologies is that they impose sparsity on stage 1 models to select instrumental variables according to established theory, while the p-value cutoff chosen for Mendelian randomization analyses is somewhat ad hoc.

4.4 Characterizing stage 1 models

4.4.1 Nonzero parameters selected

Here, we present some characteristics for the models selected via the three model selection methods. In particular, we describe the number of nonzero parameters for the selected models. Evidence from chapter 3 indicates that pseudovalidation will select less sparse models than pseudo AIC, which will in turn select less sparse models than pseudo BIC. We find that to generally be the case in this application. Pseudo BIC selects the null model for 390 endophenotypes, whereas pseudo AIC selects the null model for 94 endophenotypes. Pseudovalidation never selects the null model. This can be considered an additional control for heritability. It is possible that some of the traits, which were selected using LD score regression to assess heritability from summary statistics, are not well modeled by additive polygenic risk scores. It is not possible to select the null model using pseudovalidation. Given the feasibility that the null model may be appropriate for a given endophenotype, the ability to select the null model is potential advantage of the pseudo AIC and pseudo BIC in this framework. The BADGERS analysis induced some sparsity by first removing those SNPs with univariate $p > .01$ and clumping with an LD cutoff of $r^2 = .1$ and a radius of 1Mb. The information on the number nonzero SNP weights in the BADGERS analysis is not immediately available, and so is not included here. The number of nonzero parameters for each of the model selection methods is presented in figure 4.11. We see that the pseudo BIC selects substantially sparser models than the pseudo AIC, which in turn selects substantially sparser models

than pseudovalidation.

It is clear that pseudo AIC and pseudo BIC induce substantially more sparsity than pseudovalidation. Given the likelihood of pseudovalidation to select dense models, it is more likely that those models leveraged in a TS-2SLS framework will lead to more frequent and severe violations of the valid instruments assumption and the no pleiotropy assumption. This is additional evidence that models selected via the pseudo AIC and the pseudo BIC may address some of the issues with model violations that arise in this analysis.

4.4.2 Recurrent SNPs

To further investigate pleiotropy we characterize the number of so-called 'recurrent SNPs' for the three model selection methods. The number of times that a SNP recurs is defined as follows. For a given SNP i, it's recurrence is defined as the number of endophenotypes for which a nonzero weight is assigned to that SNP. Recurrence is considered in the context of all 1738 traits to get a general sense of possible pleiotropy, and in the context of only significantly associated traits to get a more granular perception of how pleiotropy may be driving significant associations. These plots are displayed in figures 4.12 and 4.13. The set of candidate SNPs is those $\sim 750,000$ SNPs remaining after QC as described in **(4.3.2)**. Note that sure independence screening is performed before model estimation, and the procedure is unique to each endophenotype. So not all of these 750,000 SNPs are candidates in a given TlpSum model; only 100,000 are. This may explain why there is an identical set of SNPs with no nozero weights for each of the model selection methods in figure 4.12; these may be those SNPs that never make it past the SIS step.

Figure 4.12 indicates a modest degree of shared SNPs between the 1738 traits for all of the model selection methods. To clarify the meaning of these plots, consider the green bar in the category labelled '11-50' in figure 4.12. This bar indicates that roughly 6×10^5 SNPs recur between 11 and 50 times for pseudo BIC models; that is, they have nonzero weights in models for 11 - 50 different endophenotypes. As expected, pseudo BIC models correspond to a lower degree of recurrence than pseudo AIC and pseudovalidation models. Despite the general tendency of pseudo AIC to estimate sparser models than pseudovalidation as demonstrated in figure 4.11, there doesn't appear to

be a substantive difference in recurrence between pseudo AIC and pseudovalidation. Despite the general tendency of pseudo AIC and pseudo BIC to select sparser models, there are indeed cases where both method select highly dense models. This may explain the relative lack of SNPs with recurrence in < 10 traits.

Figure 4.13 displays SNP recurrence in traits that are significantly associated with Alzheimer's status. We note the substantially larger number of traits with recurrence ≥ 3 for pseudovalidation. This is in part because pseudovalidation selects less sparse models, and in part because pseudovalidation finds more significant traits. We note the relatively low amount of recurrence in pseudo BIC models in figure 4.13. This may indicate a lower degree of pleiotropy driving significant associations.

These plots further demonstrate the potential pleiotropy between imputed endophenotypes. Figure 4.13 is especially informative because it is limited to only those significant traits. This is potentially more evidence that pleiotropy is driving the pseudovalidation results.

4.4.3 Heatmaps and clustering for imputed endophenotypes

In order to investigate pleiotropy, we impute endophenotypes into reference data and perform correlation and clustering analyses. We use the stage 1 models estimated from the UK Biobank summary statistics to impute endophenotypes for 612 individuals of European descent from the ADSP study. In the interest of investigating how pleiotropy is driving significant associations with AD, only those traits achieving significance via the weighted sum test after a Bonferroni correction of $\alpha = .05/1738$ were included in these analyses. Correlation results are presented in the form of heatmaps in figures 4.14 and 4.15. The three heatmaps correspond to the three model selection methods.

The heatmaps demonstrate some substantial correlation between the imputed endophenotypes. There is strikingly high correlation between some pseudo BIC endophenotypes. This is demonstrated by the perfect correlation between the four traits in the upper left area of the pseudo BIC heatmap in figure 4.14. In this case, pseudo BIC selects a model with the same single nonzero SNP (located in the major histocompatibility complex on chromosome 6) for all four traits. There are other endophenotypes with substantial correlation in the pseudo BIC heatmap as well. This is potentially more likely with pseudo BIC than the other model selection methods; given that it tends

to select sparse models, if there are several pleiotropic traits that share a set of causal SNPs, pseudo BIC may be more likely to select only those SNPs and not pick up any additional noise. We see from the pseudo AIC and pseudovalidation heatmaps that there are some highly correlated endophenotypes, but generally there are not large groups of endophenotypes that are very highly correlated. Nevertheless, as demonstrated in **(4.3.4)**, we see substantial changes in significant associations when a multivariable approach is applied to the pseudo AIC and especially the pseudovalidation models. It may be that some meaningful correlation structure is not fully described by pairwise correlations.

Dendrograms were drawn based on results from hierarchical agglomerative clustering using complete linkage, where Euclidean distance was used to calculate similarity. The UK Biobank data defines eight general categories for endophenotypes. We label each leaf of the dendrogram with the corresponding category. This is to demonstrate whether phenotypes tend to cluster according to these categories. The results are presented in figure 4.16.

We see from the dendrograms that traits tend to cluster according to their category (i.e. 1-8 in the figure legends), although this behavior is by no means absolute. This is expected, given the similar genetic architecture of related traits. It appears that the tendency of endophenotypes to cluster by category is strongest in the pseudo BIC models, followed by pseudo AIC, and lastly pesudovalidation. This may mean that the highly parameterized pseudovalidation models may be picking up additional noise, which distorts the tendency of traits to cluster by type and may cause problematic pleiotropy.

4.5 Discussion

In this work, we apply a two sample two-stage least squares framework to scan the UK Biobank for endophenotypes possibly associated with Alzheimer's risk in the IGAP cohort. This approach extends upon the recent BADGERS analysis of UK Biobank and IGAP data. Whereas BADGERS uses pruning and thresholding to estimate polygenic risk scores for stage 1 modeling, we use TlpSum for that purpose. TlpSum has well characterized improvements over pruning and thresholding in terms of predictive

accuracy. However, TlpSum requires the specification of tuning parameters, and it is not straightforward to select tuning parameters in the absence of validation data. For this purpose, we experiment with three methods for model selection in stage 1, namely pseudo AIC, pseudo BIC, and pseudovalidation. Choice of model selection method has a substantial effect on subsequent inference, and we explore the different properties and contrast them with BADGERS. In general, pseudovalidation performs similarly to BADGERS. Notably, 42 of the 50 traits identified as significant by the BADGERS application to IGAP are also identified as significant by the pseudovalidation approach. Although the pseudo AIC and pseudo BIC approaches find fewer marginal associations than BADGERS, there is some evidence that model selection via the pseudo AIC and pseudo BIC leads to less serious violations of the modeling assumptions of 2SLS, namely the valid instruments assumption and the assumption of no pleiotropy. All polygenic score based methods, i.e. our three approaches and BADGERS, identify more associated endophenotypes than Mendelian randomization approaches.

To expand upon the marginal analysis used to identify possibly associated endophenotypes, we conduct a multivariable analysis based on marginal SNP effect estimates. While most endophenotypes significantly associated by pseudo AIC and pseudo BIC via the weighted sum test are also found to be significant under the multivariable approach, close to half of significant traits identified by pseudovalidation and the weighted sum test are not found significant under the multivariable approach. This is evidence that pleiotropy among endophenotypes may be driving a substantial proportion of the associations identified by pseudovalidation. We present additional evidence that model selection via the pseudo AIC and pseudo BIC reduces pleiotropy and in general violates fewer of the assumptions of 2SLS. Namely, we show that pseudo AIC and pseudo BIC select sparser models.

In conclusion, the application of penalized regression models and summary statistic based model selection methods provide new insights and identify new endophenotypes that are possibly causally associated with AD. A comparison to the BADGERS approach and a careful examination of how modeling choices modulate adherence to the modeling assumptions of two-stage least squares indicate the need for powerful, sparse models in 2SLS and close examination of the confounding effect of pleiotropy.

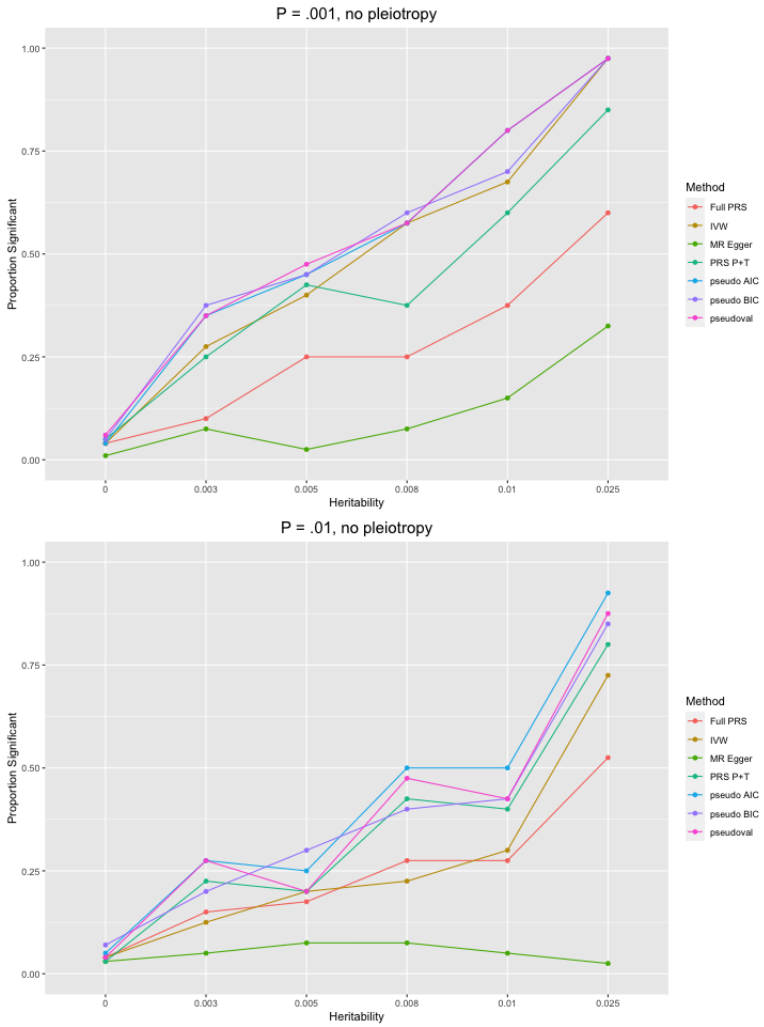

Figure 4.2: Significant tests (at $p < .05$) for different methods of TS-2SLS analysis. Proportion of causal SNPs varied from $p = .001$ and $p = .01$, no pleiotropy.

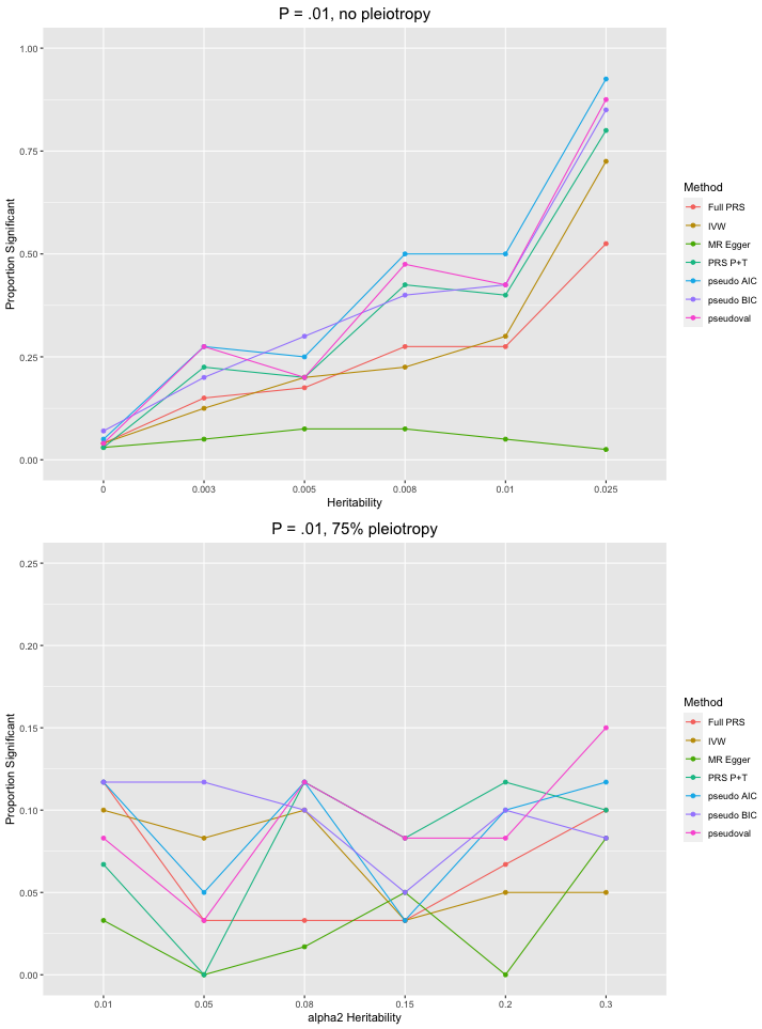

Figure 4.3: Type I error (at $p < .05$) for different methods of TS-2SLS analysis. Proportion of causal SNPs varied from $p = .001$ and $p = .01$, pleiotropy 75%. X-axis values are heritability for $\mathbf{Y}_2$; heritability of $\mathbf{Y}$ not mediated by $\mathbf{Y}_2$ is zero.

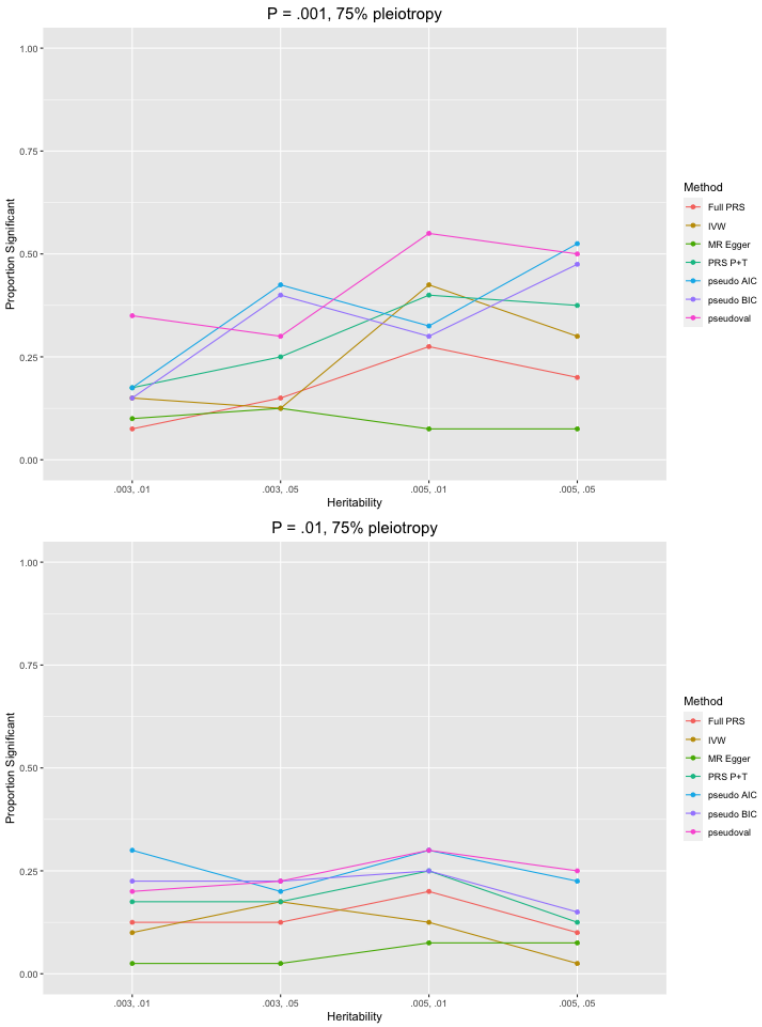

Figure 4.4: Significant tests (at $p < .05$) for different methods of TS-2SLS analysis. Proportion of causal SNPs varied from $p = .001$ and $p = .01$, pleiotropy 75%. X-axis values are ordered pairs for α_1, α_2 terms in equation (4.16).

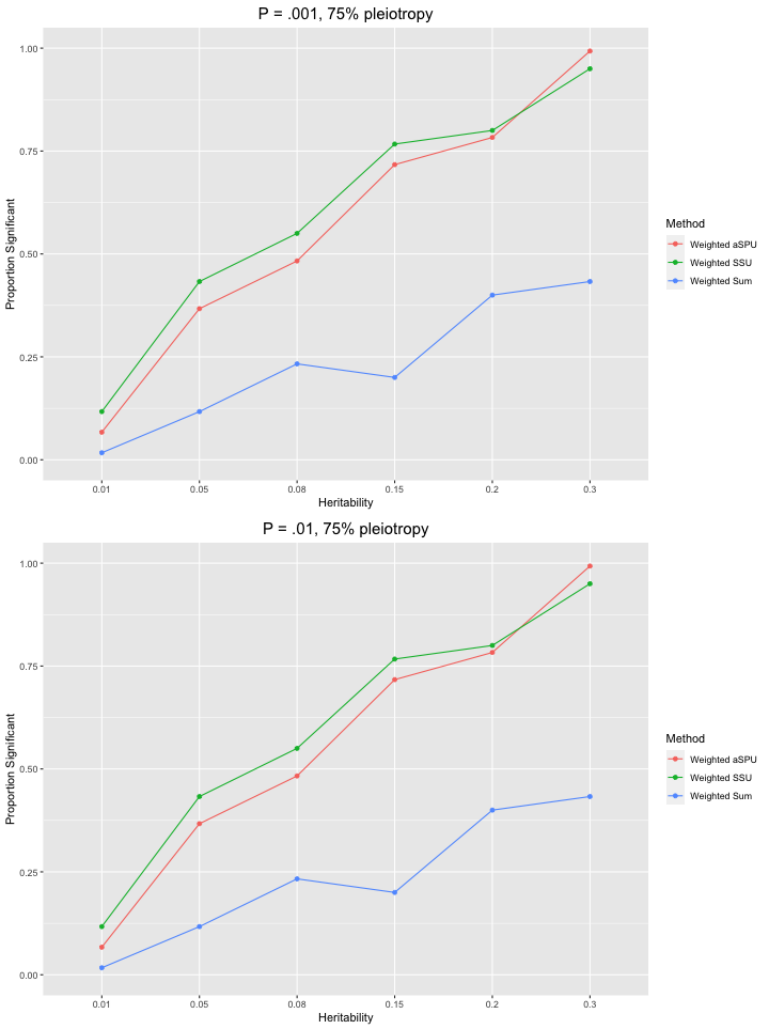

Figure 4.5: Type I error (at $p < .05$) for weighted SPU tests with different values of τ. Proportion of causal SNPs varied from $p = .001$ and $p = .01$, pleiotropy 75%. X-axis values are heritability for $\mathbf{Y}_2$; heritability of $\mathbf{Y}$ not mediated by $\mathbf{Y}_2$ is zero.

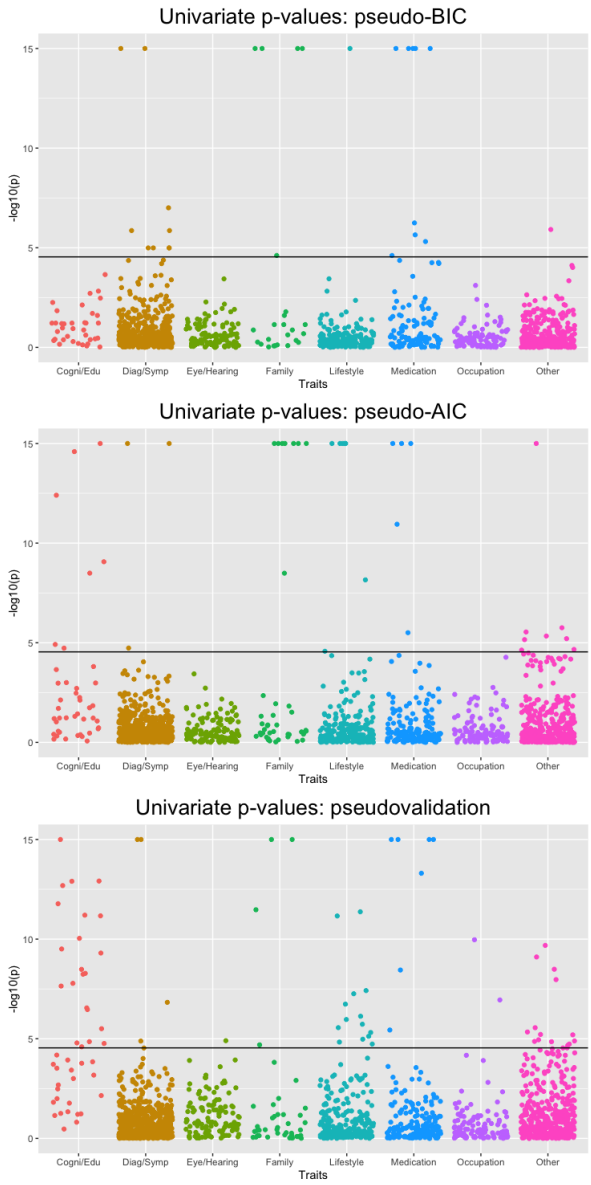

Figure 4.6: Univariate p-values for 1,738 heritable traits, using IGAP as the stage 2 data. P-values truncated at 1×10^{-15}. Horizontal line marks Bonferroni cutoff at .05 / 1738. Stage 1 weights are coefficients from TlpSum models estimated on the UK Biobank data, with the three model selection methods compared.

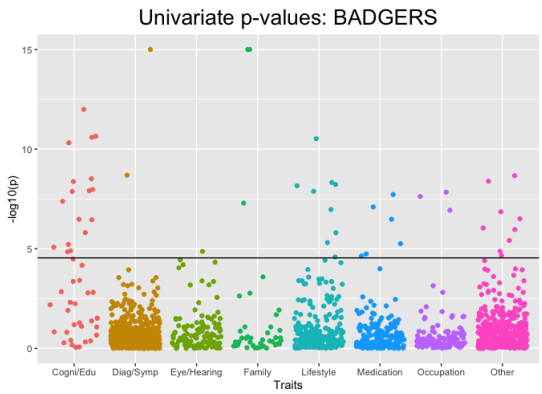

Figure 4.7: Univariate p-values for the BADGERS analysis of 1,738 heritable traits, using IGAP as the stage 2 data. P-values truncated at 1×10^{-15}. Horizontal line marks Bonferroni cutoff at .05 / 1738.

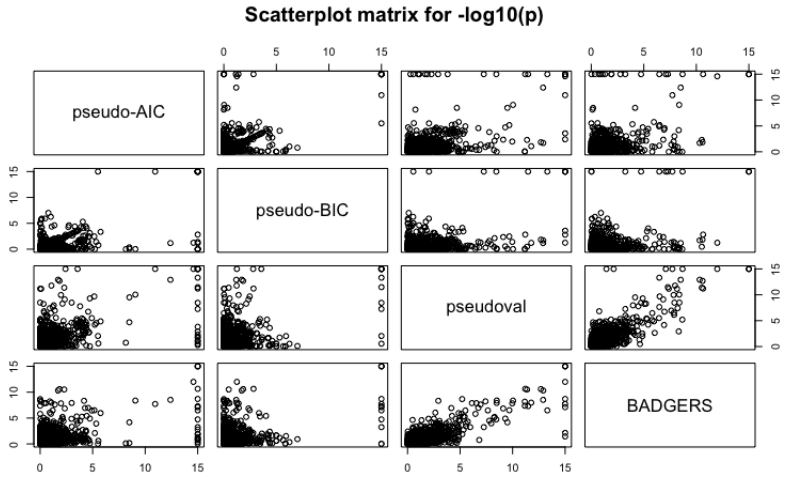

Figure 4.8: Scatterplots of $-log10(p)$ for each of the four analyses. Each plotted point corresponds to one of the 1,738 traits.

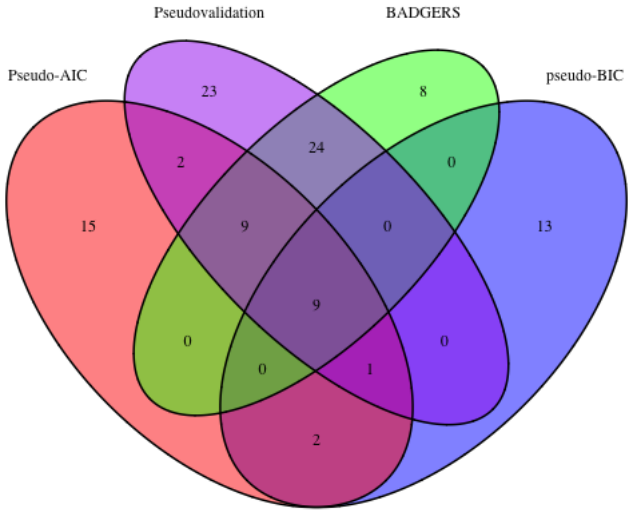

Figure 4.9: Venn diagram depicting the overlap of traits with a significant association at $\alpha = .05/1738$ for the four analyses.

Figure 4.10: P-values for Mendelian randomization analysis of the IGAP data, using instruments from 1,738 heritable UK Biobank traits. P-values truncated at 1×10^{-15}. Horizontal line marks Bonferroni cutoff at .05 / 1738.

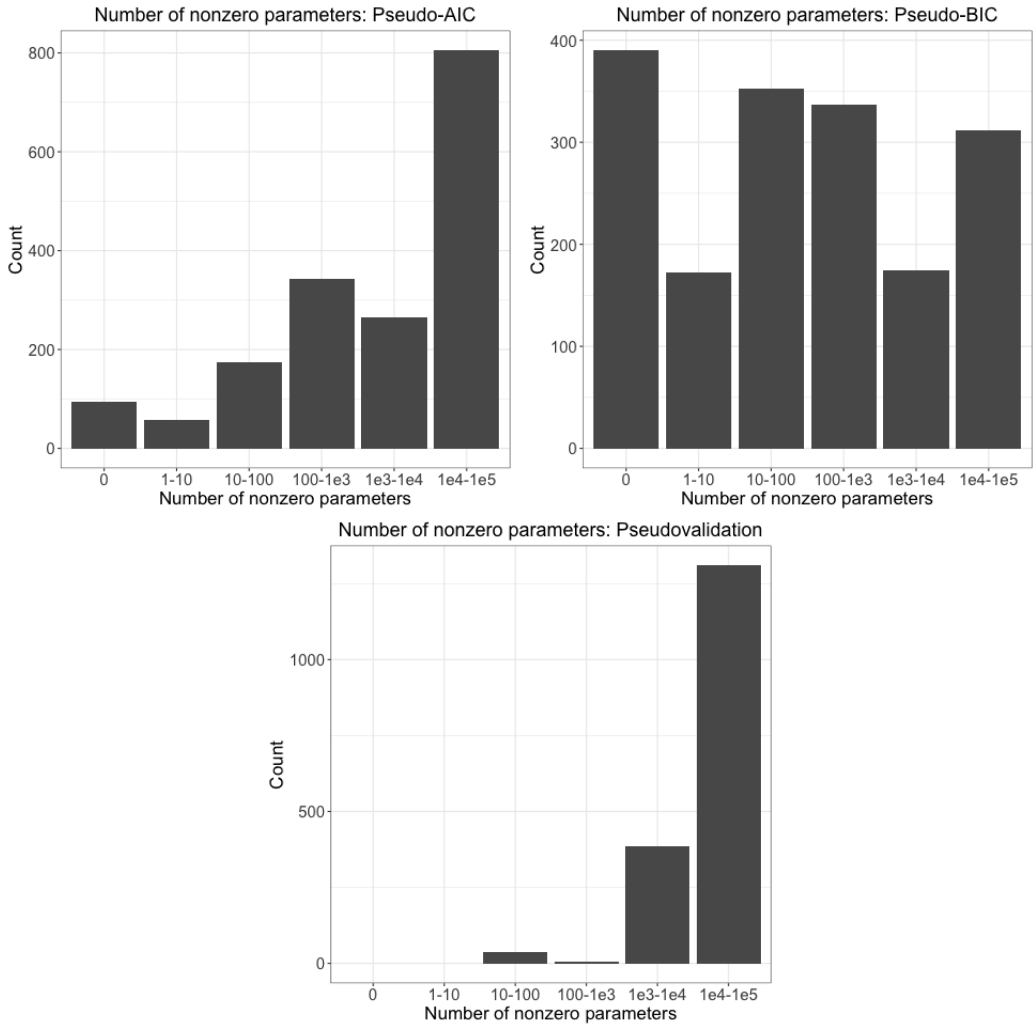

Figure 4.11: Number of nonzero parameters in the stage 1 model selected by each of the three model selection methods. Corresponds with the number of SNPs with nonzero weights in the stage 2 analysis.

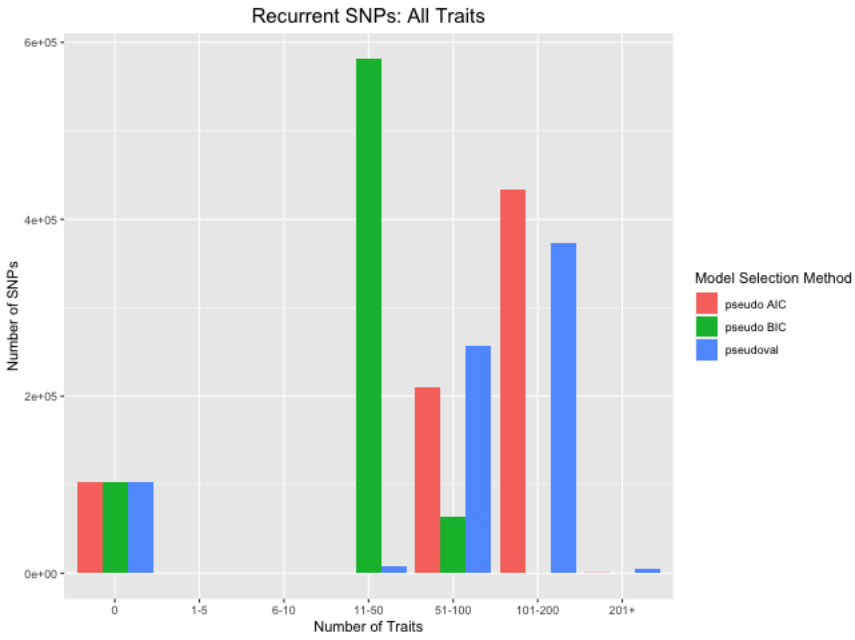

Figure 4.12: Recurrent SNPs for all 1738 traits. Each x-axis group is the number of traits which share a nonzero SNP weight, each bar within a group represents one of the three model selection methods. Bar height is the number of SNPs.

Figure 4.13: Recurrent SNPs for significant traits. Each model seleciton method corresponds to a different set of significant traits, so results may not be exactly comparable. Each x-axis group is the number of traits which share a nonzero SNP weight, each bar within a group represents one of the three model selection methods. Bar height is the number of SNPs.

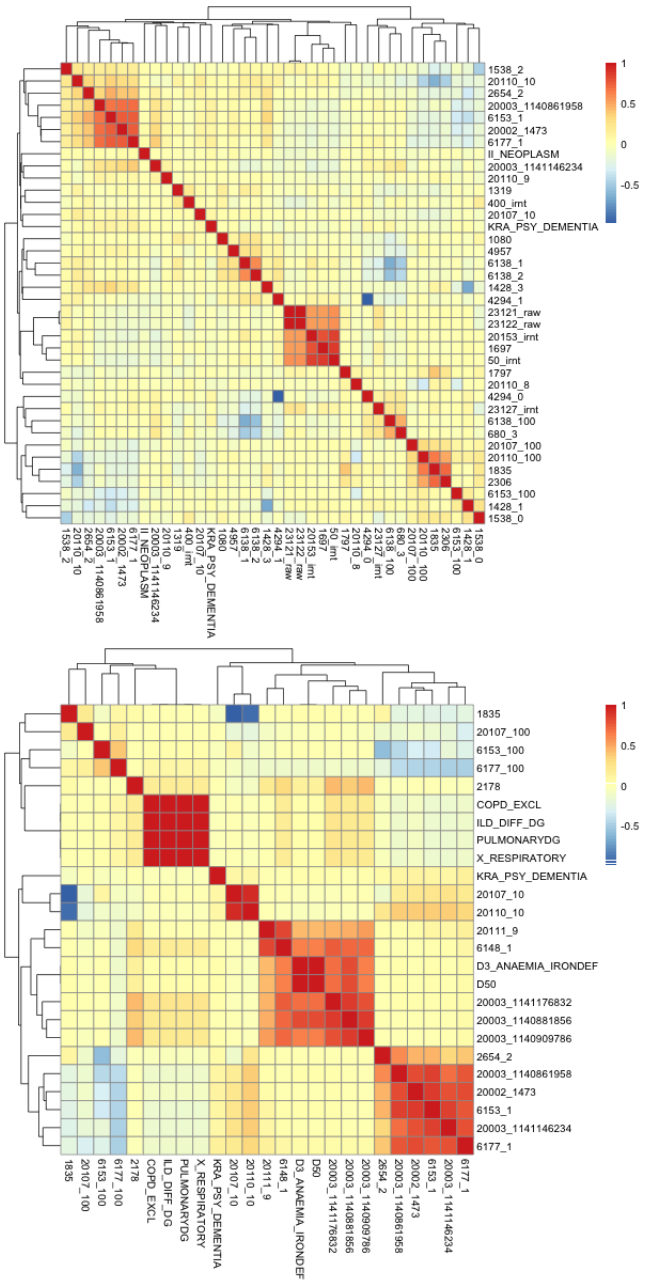

Figure 4.14: Heatmaps depicting correlation between endophenotypes imputed into the ADSP data. The top heatmap is for pseudo AIC, the bottom is for pseudo BIC.

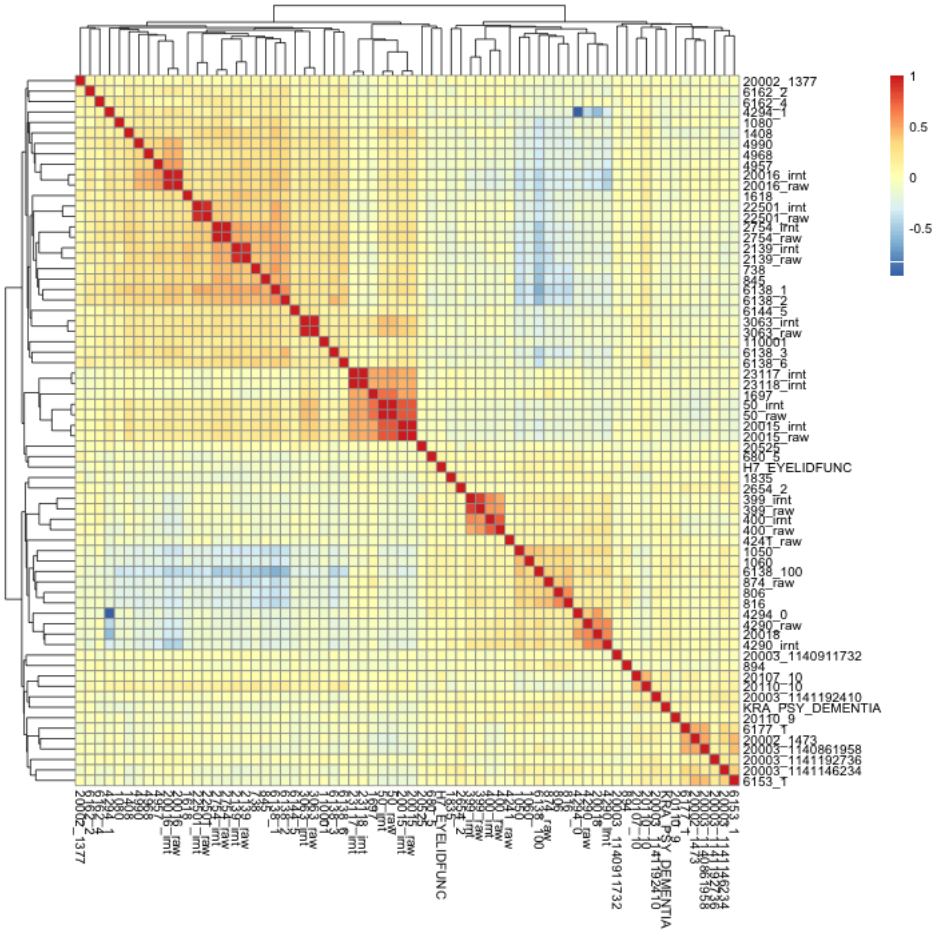

Figure 4.15: Heatmap depicting correlation between endophenotypes imputed into the ADSP data. This heatmap is for pseudovalidation.

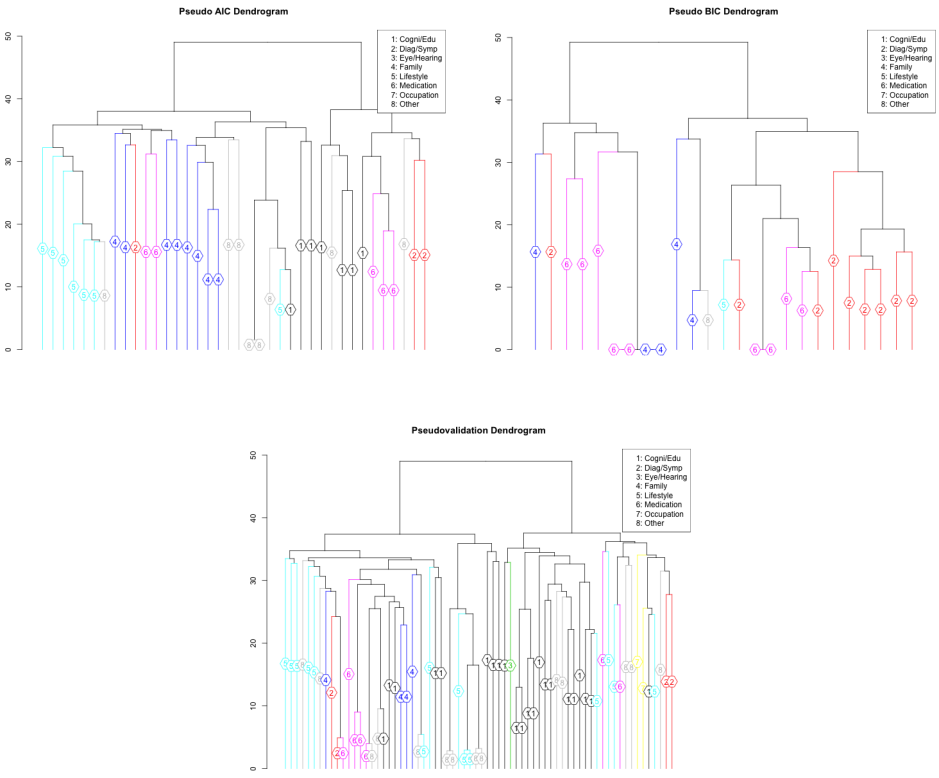

Figure 4.16: Dendrograms for hierarchical agglomerative clustering for the three model selection metrics. From left to right, top to bottom: pseudo AIC, pseudo BIC, pseudovalidation.

Chapter 5

Discussion and future work

The central aim of this dissertation is to develop approaches for the analysis of genetic and genomic data that ameliorate some of the issues with standard GWAS analyses. In this work, we develop methods that improve power, interpretability, and predictive performance, with the additional benefit that these methods are generally applicable in the summary statistic framework and can robustly incorporate multiple sources of data. Furthermore we apply these methods to genetic and genomic data on lung cancer, blood lipid levels, height, and Alzheimer's disease to develop biological insight and uncover the genetic etiology of these complex traits.

In chapter 2, we demonstrate that a novel approach to estimating functional weights for a TWAS-type analysis of lung cancer improves power as compared to existing methods. Furthermore this TWAS-type approach allows us to identify putatively causal genes, which is a more interpretable result than a single associated SNP. In particular, our method effectively incorporates both somatic and germline genetic data in the estimation of functional weights. We demonstrate that this approach has superior power as compared to a naive approach that does not incorporate somatic information. This integrative approach is an effective way of leveraging data from multiple sources to improve the power to detect germline variants associated with cancer. This approach could be extended to any cancer phenotype, under the condition that tumor samples are matched with normal tissue samples from the same individuals. Generally data from The Cancer Genome Atlas fulfills this requirement, and thus is a potentially rich source of data from which to build functional weights.

In chapter 3 we develop methods for the estimation and selection of polygenic risk scores on summary statistics. In particular, we extend the truncated LASSO penalty and the elastic net penalty to the summary statistic framework, and demonstrate the good performance of these methods in extensive simulation and application to real data. We demonstrate via simulation and application to real data that our novel methods for model selection, the pseudo AIC and the pseudo BIC, facilitate the selection of powerful and sparse models. These methods address the issue of model selection in the absence of validation data, which is a substantial concern in the application of penalized regression methodology to summary statistics. Lastly, we develop the quasi-correlation metric, which facilitates the use of summary statistic data as out-of-sample data for the assessment of model performance. We show via simulation that quasi-correlation approximates the predictive r^2 well for out-of-sample data. A useful extension of this work would be an application to complex traits that display widespread allelic heterogeneity, to investigate whether the improved performance of TlpSum demonstrated in simulation extends to real data.

In chapter 4 we implement a new approach for inference in the context of a two-stage least squares analysis of summary statistic data. We consider an application of two sample two-stage least squares where data from both stages are comprised of summary statistics. We estimate stage 1 models from GWAS summary statistics for 1,738 heritable phenotypes in the UK Biobank. We leverage these stage 1 models to identify associations between imputed endophenotypes and Alzheimer's risk in the IGAP study. This hypothesis free, phenome-wide scan of a large set of endophenotypes is facilitated by our methodology from chapter 3, which implements a wide variety of functionality for polygenic score estimation on summary statistics. Our approach reaffirms some endophenotypes that are known to be causal for Alzheimer's, while also finding novel putatively associated endophenotypes. The primary remaining issue to address is the question of pleiotropy. A potentially satisfactory approach is a careful assessment of possible evidence of pleiotropy and the ensuing consequences for inference, as we have done in chapter four. Even better would be an approach to two-sample two stage least squares inference using polygenic risk scores that is robust to model violations of the no pleiotropy assumption.

References

[1] P.M. Visscher, N.R. Wray, Q. Zhang, P. Sklar, M.I. McCarthy, M.A. Brown, and J. Yang. 10 years of gwas discovery: Biology, function, and translation. *American Journal of Human Genetics*, 101(1):5 – 22, 2017.

[2] T Manolio, F Collins, N Cox, D Goldstein, L Hindorff, et al. Finding the missing heritability of complex diseases. *Nature*, 461:747–753, 2009.

[3] AE Locke, B Kahali, S Berndt, A Justice, TH Pers, et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518:197–206, 2015.

[4] AR Wood, T Esko, J Yang, S Vedantam, TH Pers, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics*, 46(11):1173–86, 2014.

[5] P Visscher, MA Brown, MI McCarthy, and J Yang. Five years of GWAS discovery. *American Journal of Human Genetics*, 90(1):7–24, 2012.

[6] Z. Wei, K. Wang, H.Q. Qu, H. Zhang, J. Bradfield, C. Kim, et al. From disease association to risk assessment: an optimistic view from genome-wide studies on type 1 diabetes. *PLoS Genetics*, 5(10):e1000678, 2009.

[7] P.D. Pharaoh, A.C. Antoniou, D.F. Easton, and B.A. Ponder. Polygenes, risk prediction, and targeted prevention of breast cancer. *New England Journal of Medicine*, 358(26):2796 – 803, 2008.

[8] A Gusev, A Ko, H Shi, G Bhatia, W Chung, et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics*, 48(3):245–252, 2016.

[9] E Gamazon, H Wheeler, K Shah, et al. A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics*, 47(9):1091–1098, 2015.

[10] M Wainberg, N Sinnott-Armstrong, N Mancuso, AN Barbeira, DA Knowles, D Golan, R Ermel, A Ruusalepp, et al. Opportunities and challenges for transcriptome-wide association studies. *Nature Genetics*, 51(4):592–599, 2019.

[11] L Ding, MH Bailey, E Porta-Pardo, V Thorsson, A Colaprico, D Bertrand, et al. Perspective on oncogenic processes at the end of the beginning of cancer genomics. *Cell*, 173(2):305–320, 2018.

[12] H Carter, R Marty, M Hofree, AM Gross, J Jensen, KM Fisch, X Wu, et al. Interaction landscape of inherited polymorphisms with somatic events in cancer. *Cancer Discovery*, 7(4):410–423, 2017.

[13] H. Lango Allen, K. Estrada, G. Lettre, S.I. Berndt, M.N. Weedon, F. Rivadeneira, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, 467(7317):832–838, 2010.

[14] N.R. Wray, S.H. Lee, D. Mehta, A.A. Vinkhuyzen, F. Dudbridge, and Middeldorp C.M. Research review: Polygenic methods and their application to psychiatric traits. *Journal of Child Psychology and Psychiatry*, 55(10):1068–1087, 2014.

[15] B. Vilhjálmsson, J. Yang, H.K. Finucane, A. Gusev, S. Lindström, S. Ripke, et al. Modeling linkage disequilibirum increases accuracy of polygenic risk scores. *American Journal of Human Genetics*, 97:576–592, 2015.

[16] T.S.H. Shin, R.M Porsch, S.W. Choi, X. Zhou, and P.K. Sham. Polygenic scores via penalized regression on summary statistics. *Genetic Epidemiology*, 41(6):469–480, 2017.

[17] X. Shen, W. Pan, and Y. Zhu. Likelihood-based selection and sharp parameter estimation. *J Am Stat Assoc*, 107(497):223–232, 2012.

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

[19] C Sudlow, J Gallacher, N Allen, V Beral, et al. Uk biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Medicine*, 12(3):e1001779, 2015.

[20] J Lambert, C Ibrahim-Verbaas, D Harold, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for alzheimer's disease. *Nature Genetics*, 45:1452 – 1458, 2013.

[21] DM Evans and G Davey Smith. Randomization: New applications in the coming age of hypothesis-free causality. *Annu Rev Genomics Hum Genet.*, 16:327–350, 2015.

[22] D Yan, Bowen Hu, BF Darst, et al. Biobank-wide association scan identifies risk factors for late-onset alzheimer's disease and endophenotypes. *BiorXiv doi: https://doi.org/10.1101/468306*, 2018.

[23] J Pattee, X Zhan, G Xiao, and W Pan. Integrating germline and somatic genetics to identify genes associated with lung cancer. *Genetic Epidemiology*, 44(3):233–247, 2020.

[24] S Morgenthaler and WG 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*, 615(1-2):28–56, 2007.

[25] BE Madsen and SR Browning. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genetics*, 5(2):e1000384, 2009.

[26] W Pan, F Han, and X Shen. Test selection with application to detecting disease association with multiple snps. *Human Heredity*, 69(2):120–130, 2010.

[27] C.J. Willer, E.M. Schmidt, S. Sengupta, G.M. Peloso, S. Gustafsson, et al. Discovery and refinement of loci associated with lipid levels. *Nature Genetics*, 45:1274 – 1283, 2013.

[28] X He, CK Fuller, Y Song, Q Meng, B Zhang, X Yang, and H Li. Sherlock: detecting gene-disease associations by matching patterns of expression QTL and GWAS. *American Journal of Human Genetics*, 92(5):667–680, 2013.

[29] C Grisanzio, L Werner, D Takeda, BC Awoyemi, MM Pomerantz, H Yamada, et al. Genetic and functional analyses implicate the NUDT11, HNF1B, and SLC22A3 genes in prostate cancer pathogenesis. *Proceedings of the National Academy of Sciences of the United States of America*, 109(28):11252–7, 2012.

[30] Z Xu, C Wu, P Wei, and W Pan. A powerful framework for integrating eQTL and GWAS summary data. *Genetics*, 207(3):893–902, 2017.

[31] YR Su, C Di, S Bien, L Huang, X Dong, G Abecasis, S Berndt, et al. A mixed-effects model for powerful association tests in integrative functional genomics. *Am J Human Genetics*, 102:904–919, 2018.

[32] Z Xu, C Wu, and W; Alzheimer's Disease Neuroimaging Initiative Pan. Imaging-wide association study: Integrating imaging endophenotypes in GWAS. *Neuroimage*, 10.1016:159–169, 2017.

[33] Marylyn Ritchie. Large-scale analysis of genetic and clinical patient data. *Annual Review of Biomedical Data Science*, 1:263–274, 2018.

[34] The Cancer Genome Atlas, Accessed 2018. "The results shown here are in whole or part based upon data generated by the TCGA Research Network: `http://cancergenome.nih.gov/`".

[35] C Curtis, S Shah, SF Chin, G Turashvili, OM Rueda, MJ Dunning, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486:346–352, 2012.

[36] A Portela and M Esteller. Epigenetic modifications and human disease. *Nature Biotechnology*, 28(10):1057–1068, 2010.

[37] D Aran, M Sirota, and AJ Butte. Systematic pan-cancer analysis of tumour purity. *Nature Communications*, 6:8971, 2015.

[38] Q Li, JH Seo, B Stranger, A McKenna, I Pe'er, T LaFramboise, et al. Integrative eQTL-based analyses reveal the biology of breast cancer loci. *Cell*, 152(3):633–641, 2013.

[39] X Zhou, P Carbonetto, and M Stephens. Polygenic modeling with Bayesian sparse linear mixed models. *PLoS Genetics*, 9:e1003264, 2013.

[40] W. Pan, J Kim, Y Zhang, X Shen, and P Wei. A powerful and adaptive association test for rare variants. *Genetics*, 197(4):1081–1095, 2014.

[41] A Gusev, K Lawrenson, F Segato, MAS Fonseca, S Kar, et al. A transcriptome-wide association study of high-grade serous epithelial ovarian cancer identifies new susceptibility genes and splice variants. *Nature Genetics*, 51(5):815–823, 2019.

[42] Y Hu, L Mo, Q Lu, H Weng, J Wang, SM Zekavat, et al. A statistical framework for cross-tissue transcriptome-wide association analysis. *Nature Genetics*, 51:568–576, 2019.

[43] P Geeleher, A Nath, F Wang, Z Zhang, AN Barbeira, et al. Cancer expression quantitative trait loci (eQTLs) can be determined from heterogeneous tumor gene expression data by modeling variation in tumor purity. *Genome Biology*, 19(1):130, 2018.

[44] I Sur and J Taipale. The role of enhancers in cancer. *Nature Reviews Cancer*, 16:483–493, 2016.

[45] MP Menden, FP Casale, J Stephan, G Bignell, F Iorio, U McDermott, et al. The germline genetic component of drug sensitivity in cancer cell lines. *Nature Communications*, 9(1):1–8, 2018.

[46] A Tutt, H Tovey, MCU Cheang, S Kernaghan, L Kilburn, P Gazinska, J Owen, et al. Carboplatin in BRCA1/2-mutated and triple-negative breast cancer BRCAness subgroups: the TNT trial. *Nature Medicine*, 24(5):628–637, 2018.

[47] O Stegle, L Parts, R Durbin, and J Winn. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eqtl studies. *PLoS Computational Biology*, 6(5):1–11, 2010.

[48] O Stegle, L Parts, M Piipari, J Winn, and R Durbin. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nature Protocols*, 7(3):500–507, 2012.

[49] CH Mermel, SE Schumacher, B Hill, ML Meyerson, R Beroukhim, and G Getz. Gistic2.0 facilitates sensitive and confident localization of the targets of focal copy-number alteration in human cancers. *Genome Biology*, 12(4):R41, 2011.

[50] The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526:68–74, 2015.

[51] S. Das, L Forer, S Schonherr, C Sidore, AE Locke, A Kwong, et al. Next-generation genotype imputation service and methods. *Nature Genetics*, 48(10):1284–1287, 2016.

[52] dbGaP, Accessed 2017. The data/analyses presented in the current publication are based on the use of study data downloaded from the dbGap website, under phs000093.v2.p2 `https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?id=phs000093` and phs001273.v2.p2 `https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001273.v1.p1`.

[53] Shaun Purcell. PLINK Version 1.9, 2018. `http://pngu.mgh.harvard.edu/purcell/plink/`.

[54] S Purcell, B Neale, K Todd-Brown, L Thomas, MAR Ferreira, et al. PLINK: a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics*, 81(3):559–75, 2007.

[55] B. Pasaniuc, N Zaitlen, H Shi, G Bhatia, A Gusev, et al. Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics*, 30(20):2906–2914, 2014.

[56] C Willer, Y Li, and GR Abecasis. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, 26(17):2190–2191, 2010.

[57] Y Bosse and CI Amos. A decade of GWAS results in lung cancer. *Cancer Epidemiology, Biomarkers and Prevention*, 27(4):363–379, 2018.

[58] C Yu, S Chen, Y Guo, and C Sun. Oncogenic TRIM31 confers gemcitabine resistance in pancreatic cancer via activating the NF-xB signaling pathway. *Theranostics*, 8(12):3224 – 3236, 2018.

[59] H Li, Y Zhang, Y Zhang, X Bai, Y Peng, and He P. TRIM31 is downregulated in non-small cell lung cancer and serves as a potential tumor suppressor. *Tumour Biology*, 35(6):5747–52, 2014.

[60] J Wang, Q Liu, S Yuan, W Xie, Y Liu, Y Xiang, N Wu, L Wu, X Ma, T Cai, Y Zhang, Z Sun, and Y Li. Genetic predisposition to lung cancer: comprehensive literature integration, meta-analysis, and multiple evidence assessment of candidate-gene association studies. *Scientific Reports*, 7(1):8371, 2017.

[61] I Favorskaya, Y Kainov, G Chemeris, A Komelkov, I Zborovskaya, and E Tchevkina. Expression and clinical significance of CRABP1 and CRABP2 in non-small cell lung cancer. *Tumour Biology*, 35(10):10295–300, 2014.

[62] L Zhang, D Zhou, W Guan, W Ren, W Sun, J Shi, Q Lin, J Zhang, et al. Pyridoxine 5'-phosphate oxidase is a novel therapeutic target and regulated by the TGF-B signalling pathway in epithelial ovarian cancer. *Cell Death and Disease*, 8(12):3124, 2017.

[63] JT Bell, AA Pai, JK Pickrell, DJ Gaffney, R Pique-Regi, JF Degner, et al. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biology*, 12(1):R10, 2011.

[64] S. Purcell, N.R. Wray, J.L. Stone, P.M. Visscher, M.C. O'Donovan, P.F. Sullivan, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, 460:748–752, 2009.

[65] S. Ripke, C. O'Dushlaine, K. Chambert, J.L. Moran, A.K. Kähler, S. Akterin, S.E. Bergen, et al. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nature Genetics*, 45(10):1150–1159, 2013.

[66] D.M. Ruderfer, A.H. Fanous, S. Ripke, A. McQuillin, R.L. Amdur, et al. Polygenic dissection of diagnosis and clinical dimensions of bipolar disorder and schizophrenia. *Molecular Psychiatry*, 19(9):1017–24, 2014.

[67] PJ Newcombe, CP Nelson, NJ Samani, and F Dudbridge. A flexible and parallelizable approach to genome-wide polygenic risk scores. *Genetic Epidemiology*, 43(7):730 – 741, 2019.

[68] T Ge, C Chen, Y Ni, et al. Polygenic prediction via bayesian regression and continuous shrinkage priors. *Nature Communications*, 10:1776, 2019.

[69] LR Lloyd-Jones, J Zeng, J Sidorenko, et al. Improved polygenic prediction by bayesian multiple regression on summary statistics. *Nature Communications*, 10:5086, 2019.

[70] S Song, W Jiang, L Hou, and H Zhao. Leveraging effect size distributions to improve polygenic risk scores derived from summary statistics of genome-wide association studies. *PLos Comput Biol*, 16(2): e1007565:https://doi.org/10.1371/journal.pcbi.1007565, 2020.

[71] Y Zhu, X Shen, and W Pan. High-dimensional constrained maximum likelihood inference. *Journal of the American Statistical Association*, 115(529):217–230, 2019.

[72] T. Berisa and J.K. Pickrell. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics*, 32(2):283–5, 2016.

[73] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1:302–332, 2007.

[74] H. Zou, T. Hastie, and R. Tibshirani. On the degrees of freedom of the lasso. *The Annals of Statistics*, 35 (5):2173–2192, 2007.

[75] L. Song, A. Liu, and J. Shi. Summaryauc: a tool for evaluating the performance of polygenic risk prediction models in validation datasets with only summary level statistics. *Bioinformatics*, 2019.

[76] Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–78, 2007.

[77] F Hormozdiari, A Zhu, G Kichaev, CJ Ju, et al. Widespread allelic heterogeneity in complex traits. *American Journal of Human Genetics*, 100(5):789 – 802, 2017.

[78] W Chen, BR Larrabee, IG Ovsyannikova, RB Kennedy, IH Haralambieva, GA Poland, and DJ Schaid. Fine mapping causal variants with an approximate bayesian method using marginal test statistics. *Genetics*, 200(3):719–36, 2015.

[79] C Benner, CC Spencer, AS Havulinna, V Salomaa, S Ripatti, and M Pirinen. Finemap: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics*, 32(10):1493–501, 2016.

[80] T.M. Teslovich, K. Musunuru, A.V. Smith, A.C. Edmondson, I.M. Stylianou, et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, 466:707–713, 2010.

[81] J.D. McKay, R.J. Hung, X. Zong, R. Carreras-Torres, D.C. Christiani, N.E. Caporaso, M. Johansson, X. Xiao, et al. Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. *Nature Genetics*, 49(7):1126–1132, 2017.

[82] O Lindberg, M Walterfang, JCL Looi, N Malykhin, P Ostberg, et al. Hippocampal shape analysis in alzheimer's disease and frontotemporal lobar degeneration subtypes. *Journal of Alzheimer's Disease*, 30(2):355 – 365, 2012.

[83] M Prince, R Bryce, A Albanese, E adn Wimo, et al. The global prevalence of dementia: a systematic review and metaanalysis. *Alzheimers Dement.*, 9(1):63–75, 2013.

[84] C Reitz and R Mayeux. Alzheimer disease: epidemiology, diagnostic criteria, risk factors and biomarkers. *Biochemical Pharmacology*, 88(4):640–651, 2014.

[85] RJ Caselli and EM Reiman. Characterizing the preclinical stages of alzheimer's disease and the prospect of presymptomatic intervention. *J Alzheimers Dis.*, 33 Suppl 1(0 1):S405-S416., 2013.

[86] C Van Cauwenberghe, C Van Broeckhoven, and K Sleegers. The genetic landscape of alzheimer disease: clinical implications and perspectives. *Genetics in Medicine*, 18(5):421–430, 2016.

[87] CC Liu, CC Liu, T Kanekiyo, H Xu, and G Bu. Apolipoprotein e and alzheimer disease: risk, mechanisms and therapy. *Nat Rev Neurol.*, 9(2):106–118, 2013.

[88] S Norton, FE Matthews, DE Barnes, K Yaffe, and C Brayne. Potential for primary prevention of alzheimer's disease: an analysis of population-based data. *Lancet Neurology*, pages 788 – 794, 2014.

[89] M Sjogren and K Blennow. The link between cholesterol and alzheimer's disease. *World J Biol Psychiatry*, 6(2):85–97, 2005.

[90] ES Sharp and M Gatz. Relationship between education and dementia: an updated systematic review. *Assoc Disord.*, 25(4):289–304, 2011.

[91] SD Østergaard, S Mukherjee, SJ Sharp, P Proitsi, et al. Associations between potentially modifiable risk factors and alzheimer disease: a mendelian randomization study. *PLOS Medicine*, 12(6):e1001841, 2015.

[92] SC Larsson, M Traylor, R Malik, M Dichgans, S Burgess, et al. Modifiable pathways in alzheimer's disease: Mendelian randomisation analysis. *BMJ*, 359:j5375, 2017.

[93] NS Raghavan, B Vardarajan, and R Mayeux. Genomic variation in educational attainment modifies alzheimer disease risk. *Neurol Genet*, 5(2) e310, 2019.

[94] G Hemani et al. The mr-base platform supports systematic causal inference across the human phenome. *ELife*, 7, 2018.

[95] Wei Pan. Asymptotic tests of association with multiple snps in linkage disequilibrium. *Genetic Epidemiology*, 33(6):497 – 507, 2009.

[96] KA Knutson, Y Deng, W Pan, and ADNI. Implicating causal brain imaging endophenotypes in alzheimer's disease using multivariable iwas and gwas summary data. *Preprint*, 2020.

[97] Description of uk biobank qc. `http://www.nealelab.is/blog/2017/9/11/details-and-considerations-of-the-uk-biobank-gwas`. Accessed: 2020-06-11.

[98] J Fan and LV Jinchi. Sure independence screening for ultra high-dimensional feature space. *J. Roy. Statist. Soc. Ser. B*, 70:849–911, 2008.

[99] SY Kim, C Min, DJ Oh, and HG Choi. Risk of neurodegenerative dementia in asthma patients: a nested case-control study using a national sample cohort. *BMJ Open*, 9(10):e030227, 2019.

Appendix A

Supplementary material for Chapter 2

A.1 Quantile-quantile plots for gene based tests

We present in figures A.1 - A.6 quantile-quantile plots for each of the gene based tests to demonstrate the control of type I error. The majority of the QQ plots demonstrate adequate type I error control. Note that plots with 'Zoom' in the title have excluded outliers, so the general behavior of the test is more apparent. Because the p-values of the GWAS analysis are somewhat deflated due to the ImpG methodology used to do imputation for the OncoArray study, the p-values for the gene-based tests may be slightly deflated, and the genomic inflation factor λ for some of the gene-based tests is < 1. Genomic inflation factors for each test are displayed in table A.1. Because weighted gene-based tests can be thought of as an association test between gene expression and phenotype, and gene effects tend to be less sparse than SNP effects, we can't compare the inflation factors seen here to the inflation factors from a GWAS. We don't believe that the few tests with $\lambda > 1.1$ are evidence of inflation, and believe that type I error is well controlled.

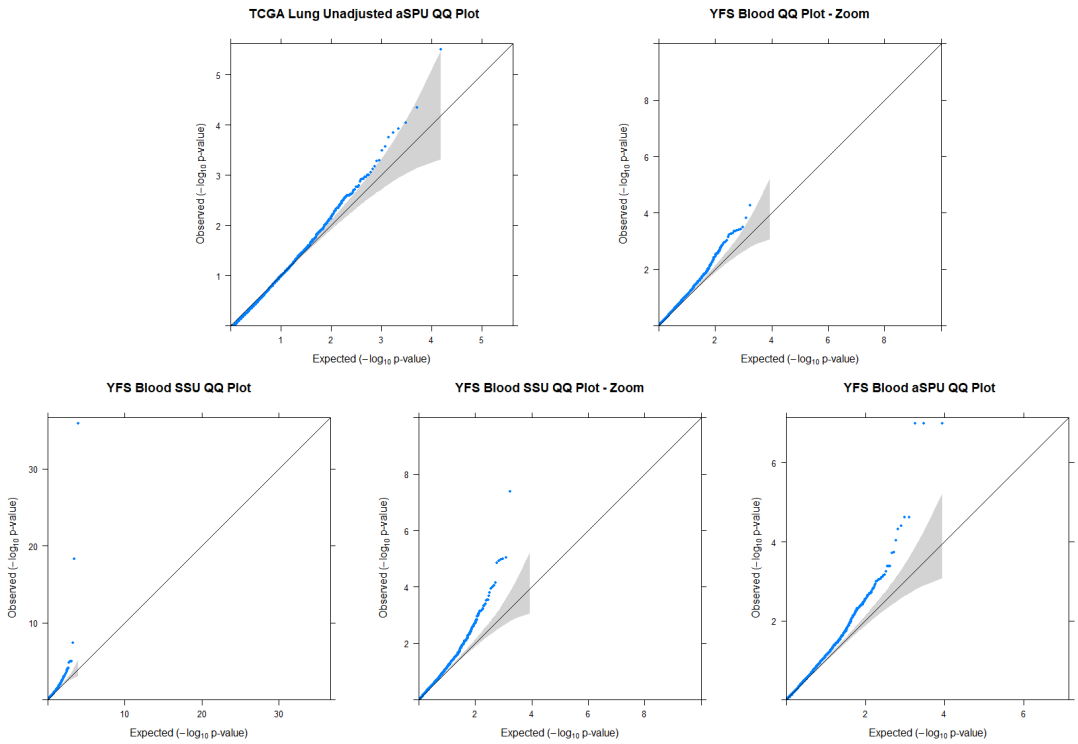

Figure A.1: Quantile-Quantile plots for the weighted tests generated with the YFS blood weights.

A.2 Comparison of weighted gene-based tests

As an extension upon section **(2.3.8)**, we display the correlation of the log of the p-values corresponding to the weighted Sum, SSU and aSPU tests for each set of weights. This is intended to demonstrate the degree of similarity in the behavior of different weighted gene-based tests. The logarithm was taken to emphasize the effect of smaller p-values, since larger p-values are likely to correspond to non-associated genes and thus be uninformative. The results are displayed in tables A.2, A.3, and A.4. For the weighted Sum test, the performance of each set of weights is highly driven by the specific weights and less so by the underlying GWAS, as demonstrated by the low correlation between each set of weights and the unweighted Sum test. For the SSU and aSPU tests, the performance of each set of weights is driven more by the underlying GWAS, as demonstrated by the higher correlation with the unweighted tests. For the

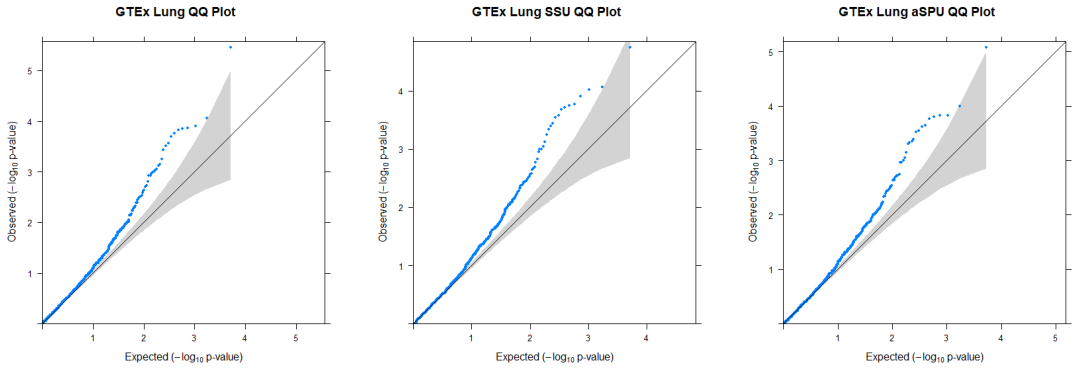

Figure A.2: Quantile-Quantile plots for the weighted tests generated with the GTEx lung weights.

Weights	Sum Test	SSU Test	aSPU Test
YFS Blood	1.07	1.14	1.07
GTEx Lung	1.03	1.01	1.03
GTEx Blood	1.21	1.17	1.15
TCGA Adjusted	1.01	1.02	.98
TCGA Unadjusted	.99	.98	.99
Unweighted	1.07	.92	1.00

Table A.1: Genomic inflation factor λ for each gene-based test.

Sum, SSU, and aSPU tests, the p-values corresponding to the somatic-adjusted TCGA weights are more correlated with the weights derived from normal tissue than are the p-values corresponding to the unadjusted TCGA weights. This is further evidence that adjusting for somatic features causes the weighted test statistics from tumors to behave more similarly to the weighted test statistics corresponding to normal tissue weights, which demonstrates the utility of adjusting for somatic features and tumor purity.

A.3 Imputation of GTEx expression into TCGA data

In order to validate the analysis and to illustrate the similarity between the GTEx weights and the TCGA weights, we impute gene expression into the TCGA data using the GTEx lung and GTEx blood weights. The GTEx weights were estimated using normal (i.e. non-tumor) tissue. If our method of adjusting for tumor purity and somatic mutations is useful, we expect that the GTEx imputed expression level is more closely correlated with the somatic-adjusted expression than the unadjusted expression. By

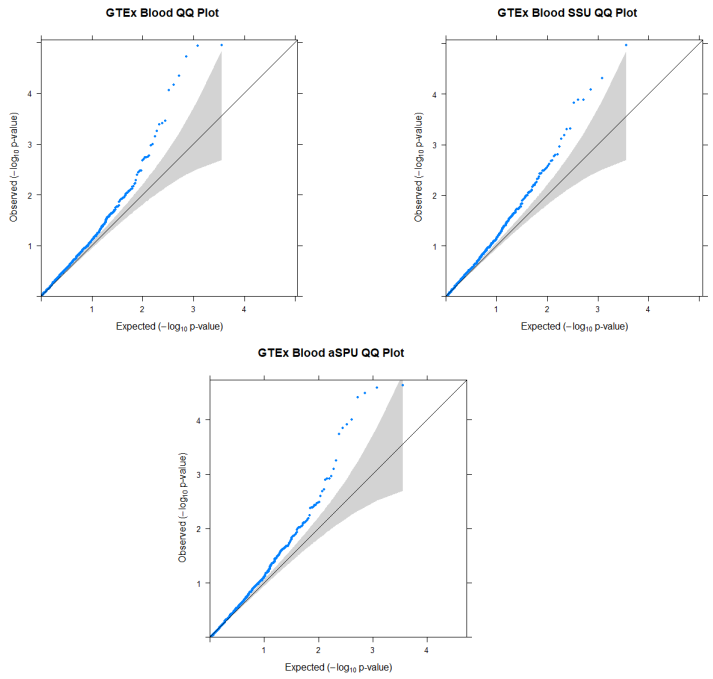

Figure A.3: Quantile-Quantile plots for the weighted tests generated with the GTEx lung weights.

"somatic-adjusted expression", we are referring to the residual expression as described in section **(2.2.2)**. By "unadjusted expression", we are referring to the expression adjusted for PEER factors as described in section **(2.3.7)**. We find that that the GTEx imputed expression levels are more closely correlated with the somatic-adjusted expression, for both sets of GTEx weights.

For each set of GTEx genes, we consider a subset of genes with cross-validated accuracy $R^2 > .1$ in the GTEx data for which we also have TCGA gene expression data. For the GTEx lung weights, there were 909 such genes. Of these 909 genes, the expression level imputed into the TCGA data was better correlated (as measured by Pearson's r) with the somatic-adjusted expression than the unadjusted expression for 590 genes, or 65.0% of genes. This information is displayed in figure A.7. Using a difference of proportions test to test if this proportion is significantly different than .5, we get a highly significant p-value $< 2.2 \times 10^{-16}$. This is evidence that the imputed expression from the GTEx lung weights is better correlated with the somatic-adjusted

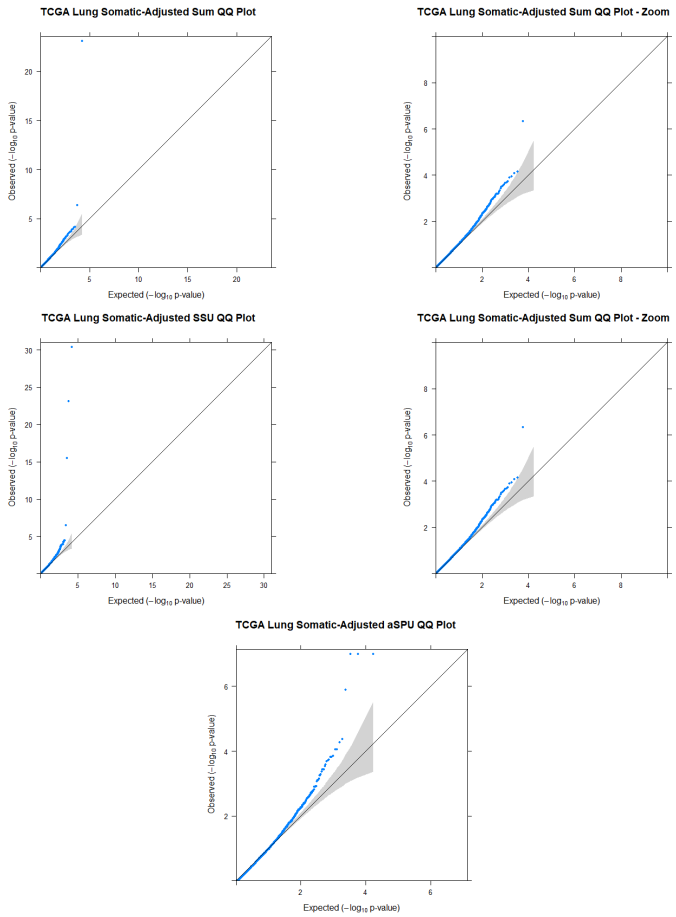

Figure A.4: Quantile-Quantile Plots for TCGA tumor weights, adjusted for copy number variance, methylation status, tumor purity, and twenty PEER factors.

TCGA gene expression than the unadjusted TCGA gene expression, indicating that gene expression adjusted for somatic features and tumor purity better resembles gene expression in normal tissue. The mean correlation of the imputed GTEx Lung expression with the unadjusted TCGA gene expression is .17, with standard deviation .15. The mean correlation of the imputed GTEx Lung expression with the somatic-adjusted TCGA gene expression is .18, with standard deviation .15.

For the GTEx blood weights, there are 433 genes with cross-validated accuracy $R^2 > .1$ in the GTEx data for which we also have TCGA gene expression data. Of these 433 genes, 290 of them were better correlated (as measured by Pearson's r) with the

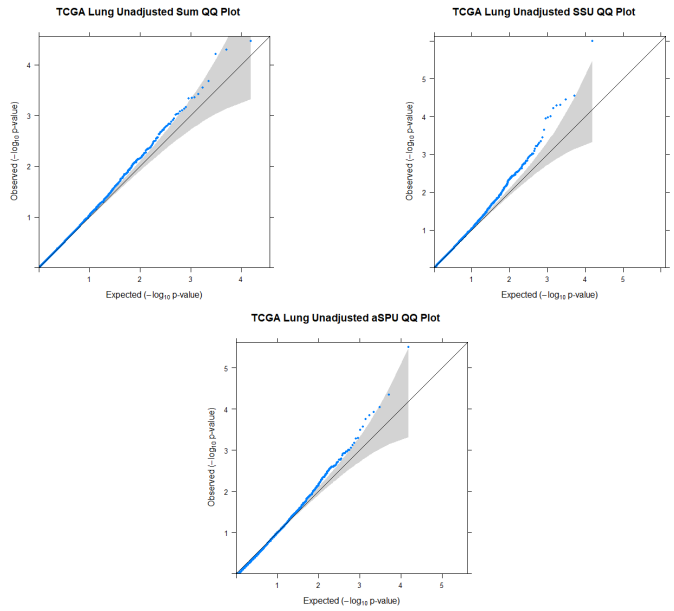

Figure A.5: QQ Plots for TCGA tumor weights adjusted for 20 PEER factors, but no somatic features or tumor purity.

somatic-adjusted expression than the unadjusted expression. This corresponds 67.0% of genes. This information is displayed in figure A.8. The difference of proportions test yields a p-value of 2.28×10^{-11}. This is evidence that the imputed expression from the GTEx blood weights is better correlated with the somatic-adjusted TCGA expression than the unadjusted TCGA expression. The mean correlation of the imputed GTEx Blood expression with the unadjusted TCGA gene expression is .19, with standard deviation .17. The mean correlation of the imputed GTEx Lung expression with the somatic-adjusted TCGA gene expression is .20, with standard deviation .17.

We note that, for both the GTEx Lung and the GTEx Blood weights, there are some genes for which the imputed gene expression is substantially better correlated with the unadjusted expression than the somatic-adjusted expression. For both sets of weights, there are four genes for which $r_{unadjusted} - r_{adjusted} > .2$. For each of these genes, we find that methylation status controls a substantial proportion of the heritability. In these cases, it is likely that methylation is under genetic control via mQTLs. By regressing out the effect of methylation, we are regressing out some of the SNP effects, and the resulting model has less predictive power.

	GTEx Lung	GTEx Blood	TCGA Adjusted	TCGA Unadjusted	Unweighted
YFS Blood	.584 (994)	.667 (821)	.237 (2372)	.187 (2123)	.105 (4346)
GTEx Lung		.782 (949)	.492 (1278)	.408 (1160)	.209 (1801)
GTEx Blood			.462 (869)	.353 (760)	.245 (1257)
TCGA Adjusted				.683 (5640)	.159 (8558)
TCGA Unadjusted					.177 (7776)

Table A.2: Table displaying the correlation of the logarithm of the p-values for weighted Sum tests with different sets of functional weights. The first number is the Pearson correlation coefficient r. The number in parentheses is the number of genes in common among the two sets of weights.

	GTEx Lung	GTEx Blood	TCGA Adjusted	TCGA Unadjusted	Unweighted
YFS Blood	.681 (994)	.716 (821)	.333 (2372)	.297 (2123)	.554 (4346)
GTEx Lung		.818 (949)	.591 (1278)	.568 (1160)	.386 (1801)
GTEx Blood			.526 (869)	.442 (760)	.296 (1257)
TCGA Adjusted				.751 (5640)	.492 (8558)
TCGA Unadjusted					.315 (7776)

Table A.3: Table displaying the correlation of the logarithm of the p-values for weighted SSU tests with different sets of functional weights. The first number is the Pearson correlation coefficient r. The number in parentheses is the number of genes in common among the two sets of weights.

	GTEx Lung	GTEx Blood	TCGA Adjusted	TCGA Unadjusted	Unweighted
YFS Blood	.634 (994)	.690 (821)	.340 (2372)	.273 (2123)	.425 (4346)
GTEx Lung		.754 (949)	.552 (1278)	.511 (1160)	.420 (1801)
GTEx Blood			.474 (869)	.375 (760)	.322 (1257)
TCGA Adjusted				.648 (5640)	.355 (8558)
TCGA Unadjusted					.305 (7776)

Table A.4: Table displaying the correlation of the logarithm of the p-values for weighted aSPU tests with different sets of functional weights. The first number is the Pearson correlation coefficient r. The number in parentheses is the number of genes in common among the two sets of weights.

This analysis validates our approach of adjusting for somatic features and tumor purity to reduce the noise in gene expression in tumor cells. It is also a further confirmation that our analysis is sound and we are producing meaningful weights from the tumor cells.

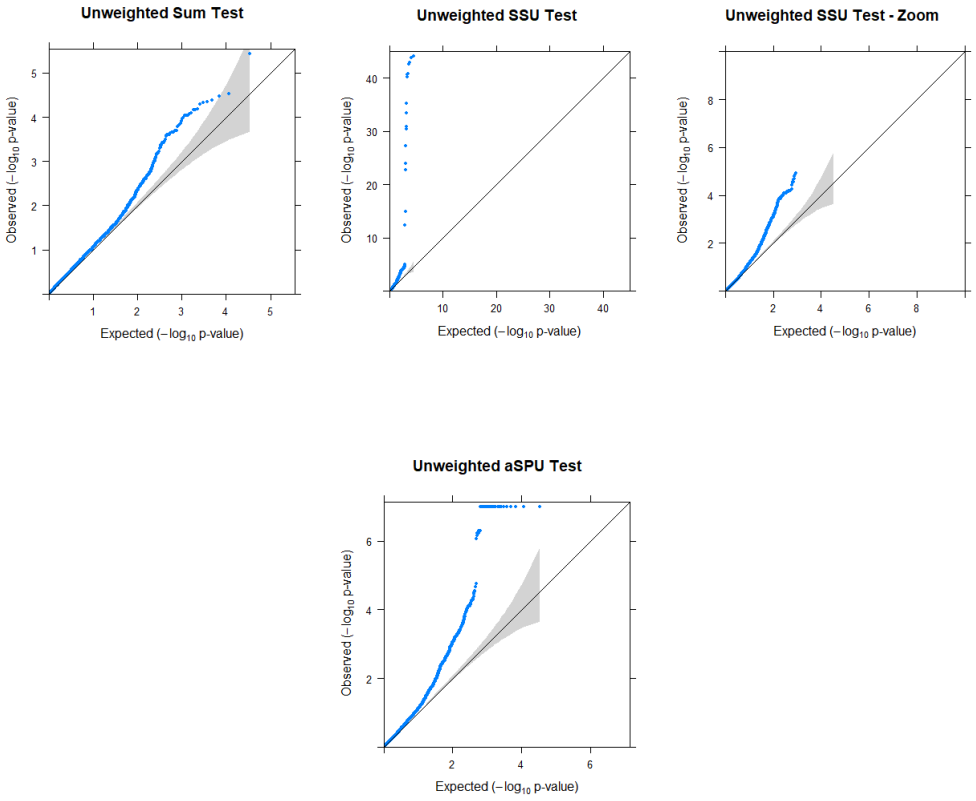

Figure A.6: QQ Plots for unweighted gene-based tests. The SSU and aSPU tests show a moderate degree of inflation. This is due to the highly correlated nature of the tests, given that many genes overlap significantly. The majority of genes located near the highly significant GWAS peak on chromosome 15 have a significant or nearly significant p-value.

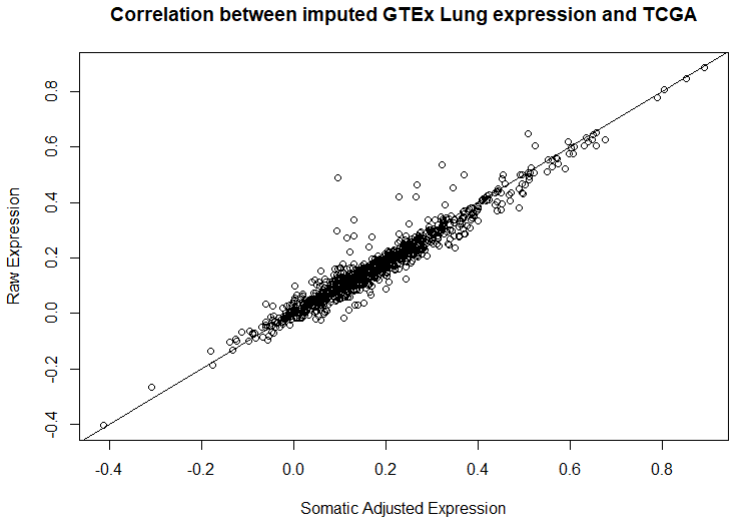

Figure A.7: Pearson's r values for the GTEx Lung expression imputed into the TCGA data, as compared to the unadjusted and the somatic-adjusted expression level.

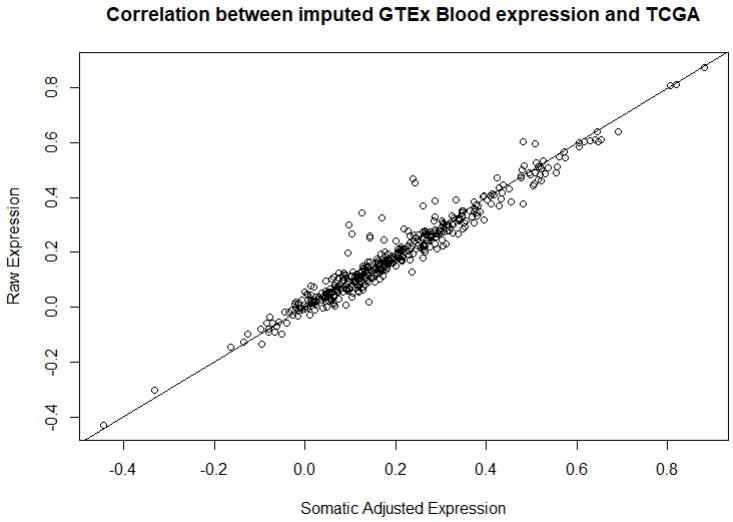

Figure A.8: Pearson's r values for the GTEx Blood expression imputed into the TCGA data, as compared to the unadjusted and the somatic-adjusted expression level.

Appendix B

Supplementary material for chapter 3

B.1 Coordinate Descent Algorithms

We provide a software to perform the estimation of the penalized regression models described above. The software solves the objective function (3.5) via coordinate descent.

First, we describe the coordinate descent algorithm for the LASSO penalty. We have the penalized regression objective function, standardized $n \times p$ design matrix $\mathbf{X}$, and standardized response vector $\mathbf{y}$. Let our vector of temporary effect size estimates, which is updated elementwise by the coordinate descent algorithm, be denoted $\tilde{\boldsymbol{\beta}}$. We loop over $\tilde{\boldsymbol{\beta}}$ to update each element $\tilde{\beta}_j$ in sequence. We iterate this process until convergence, which is defined as two successive iterations that produce no elementwise change above some (small) threshold. The updating formula is as follows [73]:

$$\tilde{\beta}_j \leftarrow S(\sum_{i=1}^{n} x_{ij}(y_i - \tilde{y}_i^{(j)}), \lambda) \tag{B.1}$$

where S is the soft-thresholding operator, and $\tilde{y}_i^{(j)} = \sum_{k \neq j} x_{ik}\tilde{\beta}_k$. We can also express the updating formula as:

$$\tilde{\beta}_j \leftarrow S(\sum_{i=1}^{n} x_{ij}y_i - x_{ij}\sum_{k \neq j} x_{ik}\tilde{\beta}_k, \lambda) \tag{B.2}$$

We can make substitutions equivalent to equation (4) to derive an updating formula that can be used in our framework of summary statistics and reference data. We also define the following quantity: $\tilde{\boldsymbol{\beta}}_{j=0}$ is equal to $\tilde{\boldsymbol{\beta}}$ with the jth element equal to zero. Given these, we can represent equation (B.2) as:

$$\tilde{\beta}_j(\lambda) \leftarrow S([\mathbf{r} - \mathbf{R_s}\tilde{\boldsymbol{\beta}}_{j=0}(\lambda)]_j, \lambda) \tag{B.3}$$

where $[\mathbf{r} - \mathbf{R}\tilde{\boldsymbol{\beta}}_{j=0}(\lambda)]_j$ denotes the jth element of the vector. The above coordinate descent algorithm for summary statistic data is implemented by Shin et al [16] in their LassoSum package.

A similar process follows for the TLP and the elastic net. In the elastic net, we have the following updating formula [73]:

$$\tilde{\beta}_j = \frac{S(\sum_{i=1}^{n} x_{ij}(y_i - \tilde{y}_i^{(j)}), \lambda\alpha)}{1 + (1-\alpha)\lambda} \tag{B.4}$$

We make the substitutions specified above to get the following update formula which allows us to estimate the model in the summary statistic framework:

$$\tilde{\beta}_j = \frac{S([\mathbf{r} - \mathbf{R_s}\tilde{\boldsymbol{\beta}}_{j=0}]_j, \lambda\alpha)}{1 + (1-\alpha)\lambda} \tag{B.5}$$

We now present the updating formula for the TLP [17]. We introduce the following notation: consider that estimated effect size $\tilde{\beta}_j^{(m)}$ is from the mth iteration of the coordinate descent algorithm. Given this, we define the updating formula as follows:

$$\tilde{\beta}_j^{(m)} \leftarrow \begin{cases} \sum_{i=1}^{n} x_{ij}(y_i - \tilde{y}_i^{(j)}), & \text{if } \tilde{\beta}_j^{(m-1)} > \tau \\ S(\sum_{i=1}^{n} x_{ij}(y_i - \tilde{y}_i^{(j)}), \lambda), & \text{if } \tilde{\beta}_j^{(m-1)} < \tau \end{cases} \tag{B.6}$$

Which gives us the following update formula after substituting:

$$\tilde{\beta}_j^{(m)}(\lambda) \leftarrow \begin{cases} [\mathbf{r} - \mathbf{R_s}\tilde{\beta}_{j=0}(\lambda)]_j, & \text{if } \tilde{\beta}_j^{(m-1)} > \tau \\ S([\mathbf{r} - \mathbf{R_s}\tilde{\beta}_{j=0}(\lambda)]_j, \lambda), & \text{if } \tilde{\beta}_j^{(m-1)} < \tau \end{cases} \tag{B.7}$$

B.2 Application to Height

We leverage our quasi-correlation and model selection methodologies to assess the fit of penalized regression models on large summary statistic data for height. We estimate

polygenic risk scores on the UK BioBank data for the height, then assess the accuracy of these polygenic risk scores on the GIANT consortium data.

In this analysis, we used the UK BioBank height data [19] as our training dataset ($\sim$ 14 million SNPs, $N \sim 360,000$), 1000G data [50] as our reference panel ($\sim$ 9.5 million SNPs, $N \sim 503$) and the GIANT data [4] as our testing dataset ($N \sim 130,000$, $\sim$ 2.5 million SNPs). We limited the 1000G data to only those individuals of European ancestry. We limited the data to include only the subset of SNPs contained in all three studies. From this subset, we excluded SNPs with minor allele frequency $< .01$ in the 1000G data or the UK BioBank data. Then, we performed LD clumping using the 1000G data for LD information and the p-values from the UK BioBank study. The clumping was not especially stringent, only ensuring that no two SNPs with $R^2 > .9$ were included. From these, we also pruned out ambiguous SNPs (A/T, C/G). This left us with a set of $\sim 715,000$ SNPs.

With these SNPs, we constructed a set of candidate polygenic risk scores using TlpSum and LassoSum. We then performed model selection using our pseudo AIC and pseudo BIC methods, and compared these results to the existing pseudovalidation method for model selection [16]. We split the 1000G data into two datasets, so called 1000G-1 and 1000G-2, with $N = 252$ and $N = 251$, respectively. 1000G-1 was used as a reference panel to estimate polygenic risk scores via TlpSum and LassoSum. 1000G-2 was used as a reference panel to estimate the model fitting criteria. For the calculation of $\hat{\tilde{\sigma}}^2$ in the pseudo AIC / BIC, we used stringent clumping such that only those SNPs with marginal $p < 10^{-40}$ were leading a clump, and no two SNPs in LD $R^2 > .5$ were included. We regularized the estimated covariance matrices as described in section **(3.2.4)**. We present in figure B.1 the accuracy of the polygenic risk scores estimated via penalized regression as applied to the GIANT data.

In this application, TlpSum and LassoSum perform more or less equivalently. We see that adding more parameters to the model generally increases the predictive accuracy of the model as applied to the GIANT data. Because height is highly heritable phenotype, our training data has large sample size, and we have performed LD clumping to remove highly correlated SNPs, it is possible that many of the candidate SNPs are truly associated with the height phenotype. This means that adding more parameters

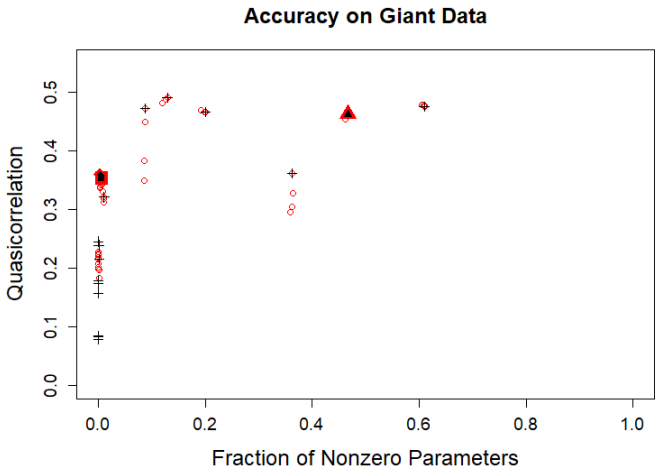

Figure B.1: Quasi-correlation versus fraction of nonzero parameters. The black crosses represent LassoSum models, while the red circles represent TlpSum models. Each data point represents a unique set of tuning parameters. The triangular points represent the model selected by pseudovalidation. The square points represent the model selected by pseudo AIC, while the diamond points represent the models chosen by pseudo BIC.

to the model is generally better, which may account for the better performance of pseudovalidation as compared to pseudo AIC / BIC. The pseudo AIC / BIC select models that perform relatively well as measured by quasi-correlation, and have a small proportion of active parameters. As the proportion of active parameters increases, the models achieve modest but noticeable performance gains. However, the AIC and BIC prefer the models that perform decently while having a small proportion of active parameters. This mirrors the behavior we saw in simulation and described in **(3.3.2.1)**.

B.3 Derivation of Standard Error for Linear Regression Estimates

We justify the expression for the standard error of linear regression estimates in section **(3.2.4)** as follows. Let us denote the univariate logistic regression estimates as $(\hat{b}_0, \hat{b}_1)'$, and the univariate linear regression estimates as $(\hat{\beta}_0, \hat{\beta}_1)'$. Using the law of total variance

and conditioning on $\hat{b}_0$, we get the expression

$$Var(\hat{\beta}_1) = Var\Big(\frac{e^{-\hat{b}_0}}{(1+e^{-\hat{b}_0})^2}\hat{b}_1\Big) = var(\hat{b}_1)E\Big(\frac{e^{-\hat{b}_0}}{(1+e^{-\hat{b}_0})^2}\Big)^2 + var\Big(\frac{e^{-\hat{b}_0}}{(1+e^{-\hat{b}_0})^2}\Big)E(\hat{b}_1)^2.$$

Plugging in $e^{-b_0} = \frac{p(Y=0)}{p(Y=1)}$ for $e^{-\hat{b}_0}$ in the first term gives us our expression derived in **(3.2.4)**, namely

$$var(\hat{\beta}_1) = \Big(\frac{e^{-\hat{b}_0}}{(1+e^{-\hat{b}_0})^2}\Big)^2 var(\hat{b}_1).$$

Now, if we can show that the $var\Big(\frac{e^{-\hat{b}_0}}{(1+e^{-\hat{b}_0})^2}\Big)E(\hat{b}_1)^2$ term is negligible, our expression will be justified. To do this, consider the following. We know that $e^{-\hat{b}_0} = \frac{\hat{p}_0}{1-\hat{p}_0}$, where $\hat{p}_0 = N_{control}/N$. Thus, we can define $\frac{e^{-\hat{b}_0}}{(1+e^{-\hat{b}_0})^2} = \hat{p}_0(1-\hat{p}_0)$. We now need an expression for the variance of $\hat{p}_0(1-\hat{p}_0)$. We apply the delta method for that purpose, with the function $g(p_0) = p_0*(1-p_0)$ and the distributional assumption that $\hat{p}_0 \sim N(p_0, \frac{p_0(1-p_0)}{N})$. Thus, we have the following expression:

$$\hat{p}_0(1-\hat{p}_0) \sim N(p_0(1-p_0), \frac{p_0(1-p_0)(1-2p_0)^2}{N})$$

Given this, we can state the following:

$$var\Big(\frac{e^{-\hat{b}_0}}{(1+e^{-\hat{b}_0})^2}\Big)E(\hat{b}_1)^2 = \frac{\hat{p}_0(1-\hat{p}_0)(1-2\hat{p}_0)^2}{N}\hat{b}_1^2$$

This term is negligible in GWAS applications. Given the small effect size of individual SNPs (and thus small $\hat{b}_1$) and the factor of N in the denominator which will be large in GWAS applications, this is reasonable. Thus, we exclude this term in our expression in **(3.2.4)**.

As an additional note, we only require the use of $Var(\hat{\beta}_1)$ for the estimation of $\frac{1}{N}\widehat{\mathbf{Y}'\mathbf{Y}}$ in our equation (7). An intuitive approach, the so-called 'marginal approach' in that it is derived from the marginal distribution of $\mathbf{Y}$, is detailed as follows. Note that if $\mathbf{Y}$ is a centered binary phenotype, $\frac{1}{N}\widehat{\mathbf{Y}'\mathbf{Y}}$ is simply the variance of $\mathbf{Y}$. This is simply the variance of Bernoulli variable. Given observed $\hat{p} = N_{case}/N$ (i.e. the proportion of cases in the summary statistic data), we have that $\frac{1}{N}\widehat{\mathbf{Y}'\mathbf{Y}} = \hat{p}*(1-\hat{p})$. Assuming that we have the quantities N_{case}, N in our summary statistic data, this should be straightforward to calculate. There is, however, an issue when summary statistics are

taken from a GWAS that included covariates. A detailed description of this issue is in Appendix **B.4**. For this reason, we recommend in general using the approach based on $Var(\hat{\beta}_1)$ we have detailed in **(3.2.4)**, and not using the marginal approach to estimation of $\frac{1}{N}\widehat{\mathbf{Y}'\mathbf{Y}}$

B.4 Considerations for Summary Statistics Estimated with Covariates

It is often the case that summary statistics from published GWAS are estimated via multiple regression with non-SNP covariates such as age, PCs, etc. also included. Our expressions (3.6), (3.7), and (3.8) in **(3.2.3)** have been derived assuming single linear regression. These expressions are still valid assuming multiple regression with some very mild assumptions and some changes in interpretation, detailed here.

Say we have design matrix $\mathbf{Z}_i = (\mathbf{X}_i, \mathbf{Z})$, where $\mathbf{X}_i$ is one of p SNPs in the GWAS and $\mathbf{Z}$ is a set of other covariates (i.e. age, gender, PCs). Let us denote the genotype matrix of SNPs as $\mathbf{X}$, comprised of p columns $\mathbf{X}_1, \ldots, \mathbf{X}_p$. Say our summary statistics for the SNP effects consist of estimates from multiple linear regression:

$$\mathbf{Y} = \beta_i \mathbf{X}_i + \boldsymbol{\alpha}\mathbf{Z} + \boldsymbol{\epsilon} \tag{B.8}$$

Note that $\boldsymbol{\alpha}$ may be a vector. Model (B.8) would be estimated p times for each of the p SNPs in the GWAS, giving us a set of summary statistic estimates $(\hat{\beta}_1, \ldots, \hat{\beta}_p)'$. $\hat{\beta}_i$ from model (B.8) does not have straightforward application to our equation (3.8), given that equation (3.8) is derived assuming single linear regression. Consider the following models, again for some SNP i:

$$\mathbf{Y} = \boldsymbol{\gamma}\mathbf{Z} + \boldsymbol{\omega}_Y \tag{B.9}$$

$$\mathbf{X}_i = \boldsymbol{\lambda}_i\mathbf{Z} + \boldsymbol{\omega}_X \tag{B.10}$$

We define $\mathbf{Y}_e$ and $\mathbf{X}_{ie}$, representing the residuals from models (B.9) and (B.10), as $\mathbf{Y}_e = \mathbf{Y} - \hat{\boldsymbol{\gamma}}\mathbf{Z}$ and $\mathbf{X}_{ie} = \mathbf{X}_i - \hat{\boldsymbol{\lambda}}_i\mathbf{Z}$. We can then define the following model:

$$\mathbf{Y}_e = \tau_i \mathbf{X}_{ie} + \boldsymbol{\omega}_{ie} \tag{B.11}$$

It is the case that $\hat{\tau}_i = \hat{\beta}_i$. Thus, if we replace $\mathbf{Y}$ in our equations (3.7) and (3.8) with $\mathbf{Y}_e$, i.e. the phenotype data with the effect of non-SNP covariates regressed out, the expression holds. Consider that $\mathbf{X}_e = (\mathbf{X}_{1e}, \ldots, \mathbf{X}_{pe})$. If $\mathbf{X}_e \neq \mathbf{X}$, it will be difficult or impossible to estimate the covariance matrix of $\mathbf{X}_e$ from a reference panel.

It is a widely held implicit assumption of all summary statistic based estimation of polygenic risk scores on GWAS data with covariates that $\mathbf{X}_e \neq \mathbf{X}$, and we justify that assumption here. If we examine expression (B.10), we note that it is unlikely that a substantial proportion of the variance of a single SNP $\mathbf{X}_i$ is explained by covariates $\mathbf{Z}$, which tend to be multifactorial features such as principal components, or features that are uncorrelated with the SNP such as age. Thus, we make the implicit assumption that $\boldsymbol{\lambda}_i \approx 0, \forall i$. Thus, we assume that $\mathbf{X}_e = \mathbf{X}$. Existing methods, such as LassoSum, LDPred, and pseudovalidation, are widely applied to summary statistic data that was estimated using multivariable regression, i.e. with non-SNP covariates. All these methods make the implicit assumption that $\boldsymbol{\lambda}_i \approx 0, \forall i$. Thus, we do not need to replace $\mathbf{X}$ with $\mathbf{X}_e$ in our equations (3.6) and (3.8).

This has interesting application to our equation (3.7). As referenced in **B.3**, we do not recommend the use of the marginal approach to estimation of $\mathbf{Y}'\mathbf{Y}$ for binary $\mathbf{Y}$ when summary statistics are taken from a GWAS that includes non-SNP covariates. The reasoning behind this is as follows. If the covariates $\mathbf{Z}$ control a substantial proportion of the variation in $\mathbf{Y}$, then $\mathbf{Y}_e$ will have substantially different variance than $\mathbf{Y}$. We implicitly replace $\mathbf{Y}$ with $\mathbf{Y}_e$ when summary statistics are estimated with covariates, which the marginal approach to estimation of $\mathbf{Y}'\mathbf{Y}$ does not account for. Thus, estimation of $\mathbf{Y}'\mathbf{Y}$ for binary data is best done via the methodology outlined in **(3.2.3)** and **(3.2.4)**.

B.5 Simulating Effect Sizes Under Allelic Heterogeneity

The algorithm for simulating effect sizes under allelic heterogeneity is as follows. Note that this simulation process assumes an ordering of the SNPs where nearby SNPS are in high linkage disequilibrium. We define h^2 as the SNP-based heritability of the disease (0.5 in our simulation), M as the number of SNPs, p as the fraction of causal SNPs, and q as the number of SNPs in the simulation. We considered values of $p = .005, p = .002$

and $h^2 \in [.2, .5, .6]$.
$i = 0$
while $i \leq q$:
$d \sim Bernoulli(p/5)$
if$(d == 0)$:
    $\beta_i = 0$
    $i = i + 1$
if$(d == 1)$:
    $k \sim round(unif(1, 7))$
    for $j \in [0, \ldots, k]$ $\beta_{i+j} \sim N(0, \frac{h^2}{Mp})$
    $i = i + k$

B.6 Accuracy of Summary Statistic Approximations

In this section, we present results demonstrating the accuracy of the summary statistic based approximations that comprise the estimation of the pseudo AIC, pseudo BIC, and quasi-correlation. Given that these approximations are not perfect, we are interested in determining where the sources of error come from. This section serves to demonstrate the accuracy of the various summary statistic approximations, and how that accuracy changes depending on the simulation setting. In particular, we consider simulation 1 from **(3.3.1)**. We conclude generally that the approximations are appropriate.

B.6.1 Estimating Phenotypic Out-of-Sample Variance

Estimating the quasi-correlation requires estimating the phenotypic variance of the out-of-sample testing data. We show in figure B.2 that, for the three simulation settings with different proportions of causal SNPs as described in this section, our methodology for estimating the variance of the out-of-sample phenotype as described in section **(3.2.5)** is highly accurate. Note that each SNP generates an estimate of out of sample variance; we used the median of the p estimates.

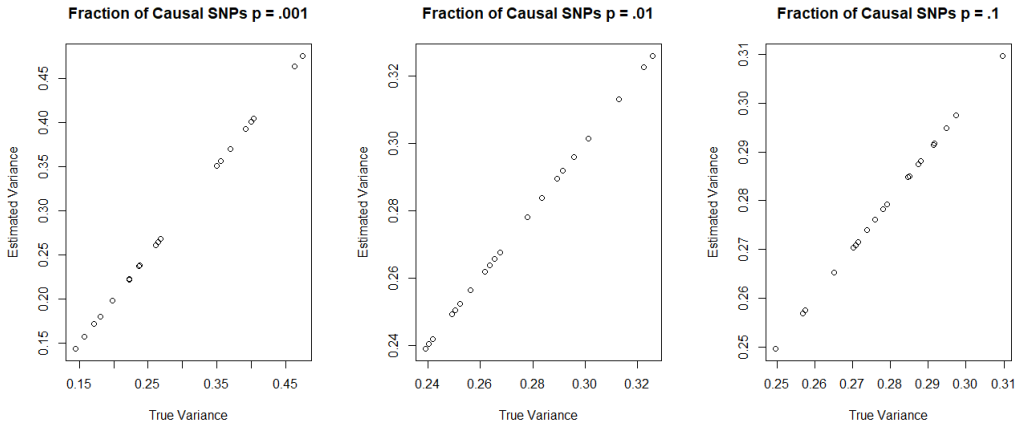

Figure B.2: Plot of the estimated variance versus the true variance for each of the three simulation settings. Each data point represents one of the twenty different simulations. The Pearson correlation coefficient $r = 1$ for each of these three plots.

B.6.2 Accuracy of Residual Variance Estimation

An important component of our pseudo AIC / BIC methods is the estimation of the residual variance $\tilde{\sigma}^2$. We present here the accuracy of the residual variance estimation for the three simulation scenarios described in this section Note that, when estimating $\tilde{\sigma}^2$, we regularize the estimated covariance matrix as described in section **(3.2.4)**. We select the set of SNPs used to estimate the residual variance as follows for the three scenarios. For the scenario where the probability of a SNP being causal is $p = .1$, we do clumping and pruning such that only SNPs with marginal p-value $< .01$ are included, and no two SNPs are in LD $r^2 > .2$. When the proportion of causal SNPs is $p = .01$, we do clumping and pruning with a marginal p-value cutoff of $< .001$ and an LD cutoff of $r^2 > .2$. Likewise, when the proportion of causal SNPs is $p = .001$, we have a marginal p-value cutoff of $< 1 \times 10^{-4}$ and an LD cutoff of $r^2 > .2$. Note that our method estimates the proportion of residual variance that is not explained by ordinary least squares linear regression. This is an application where OLSE will fail to capture much of the heritable variance, because of the sparse signal and nuanced correlation structure of the data. This means that our estimates $\hat{\tilde{\sigma}}^2$ are biased upwards. Nevertheless, we show that the estimates $\hat{\tilde{\sigma}}^2$ are well correlated with the true residual variance, which is known in this simulation setting. This information is displayed in figure B.3.

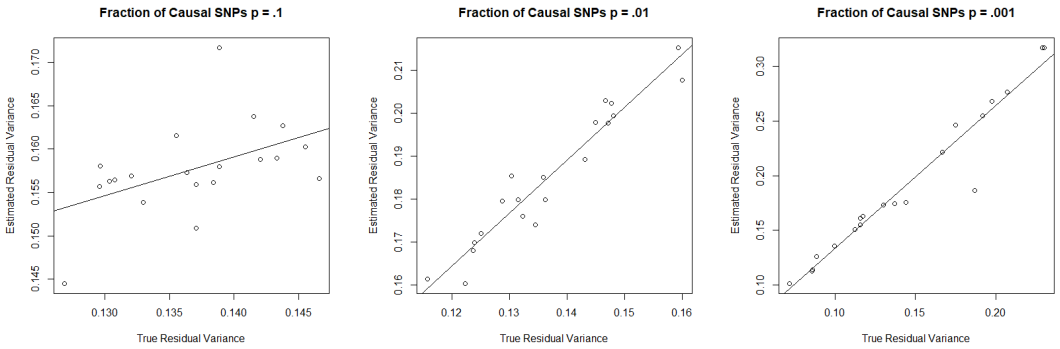

Figure B.3: Plot of the estimated residual variance versus the true residual variance for 20 replications at each of the three simulation settings. Note that the residual variance is overestimated in all three simulation settings.

We see that the residual variance is significantly easier to model when the fraction of causal SNPs is smaller. When the proportion of causal SNPs is $p = .001$, we have a correlation between true and estimated residual variances of $r = .97$. Likewise, when the proportion of causal SNPs is $p = .01$, we have a correlation of $r = .96$. When the proportion of causal SNPs is $p = .1$, we have a correlation of $r = .49$. To see why, consider that estimation of $\hat{\hat{\sigma}}^2$ can be thought of as prediction, where we are using ordinary least squares estimates to predict the phenotype. As demonstrated in figures 1, 2 and 3, prediction is more difficult as the proportion of causal SNPs p increases. Likewise, estimation of $\hat{\hat{\sigma}}^2$ is more difficult as p decreases.

B.6.3 Accuracy of SSE estimation

Another component of our pseudo AIC / BIC methods is the estimation of model SSE on the training data, as described in equation (3.10). For this method, we regularize the estimated covariance matrix as described in section **(3.2.4)**. The estimation of the residual variance involves three approximations: the approximation of $\frac{1}{n}\widehat{\mathbf{X}^T\mathbf{X}}$ (3.6), the approximation of $\frac{1}{N}\widehat{\mathbf{Y}'\mathbf{Y}}$ (3.7), and the approximation of $\frac{1}{N}\widehat{\mathbf{X}'\mathbf{Y}}$ (3.8). We find that the approximation described by equation (3.7) is nearly always accurate, as demonstrated in appendix **B.6.1**. Likewise, the approximation described in equation (3.8) is very accurate. The issue arises in the approximation described in equation (6). $\frac{1}{n}\widehat{\mathbf{X}^T\mathbf{X}}$ is used to estimate the variance of the predicted phenotype in the unseen training data.

It is difficult to estimate this variance, because there may be overfitting effects that are difficult to account for by using reference panel data to estimate covariance, especially for under-penalized models with a large proportion of active parameters. We offset this somewhat with the penalty described in **(3.2.4)**, but is difficult to account for completely.

The phenomenon described above explains the difference between the estimated and true SSE values. We present three different plots for each different proportion of causal SNPs in figures B.4, B.5, and B.6. Each plot represents one of the twenty replications. Each point on the plot represents a unique TLP model, estimated via our TlpSum method.

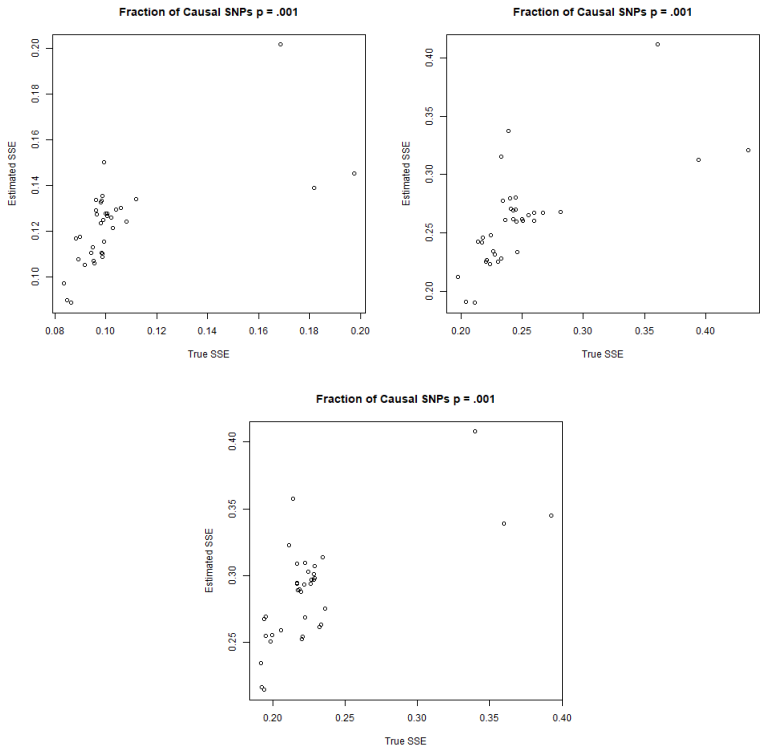

Figure B.4: True versus estimated SSE for TlpSum models applied to the simulation setting with the fraction of causal SNPs $p = .001$. These plots describe the relationship between the true and estimated SSE for three randomly chosen simulation settings among the twenty we conducted. Each plotted point represents one of 36 candidate models. The estimated Pearson correlation coefficients for the plots were, from left to right: $r = .65$, $r = .69$, $r = .67$

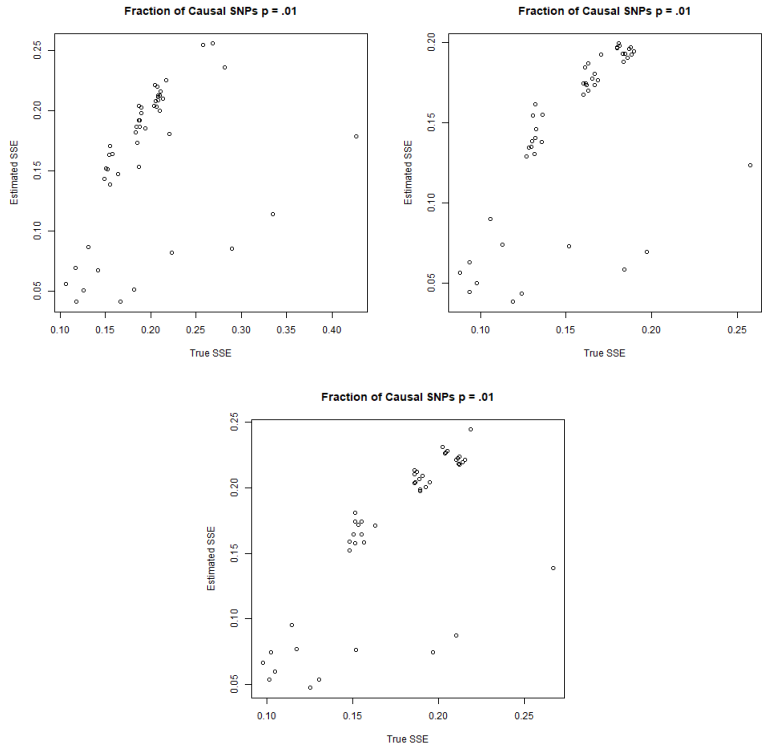

Figure B.5: True versus estimated SSE for TlpSum models applied to the simulation setting with the fraction of causal SNPs $p = .01$. These plots describe the relationship between the true and estimated SSE for three randomly chosen simulation settings among the twenty we conducted. Each plotted point represents one of 48 candidate models. The estimated Pearson correlation coefficients for the plots were, from left to right: $r = .41$, $r = .61$, $r = .73$

We see, contrary to the behavior of the estimated residual variance described in appendix **B.6.2**, that the SSE is easier to estimate as the fraction of causal SNPs p increases. To demonstrate why, consider that SSE/σ^2 is distributed χ^2_{n-k}, where k is the number of active parameters in the model. As the fraction of causal SNPs p decreases, the number of active parameters in a given model k will decrease, meaning that the variance of $SSE/\sigma^2 \sim \chi^2_{n-k}$ will increase, making estimation more difficult. Most important is the behavior in the bottom left corner of these plots: we want the models with small true SSE also have small estimated SSE. By and large, we find that this is the case. We do see some systematic underestimation of the SSE. This is

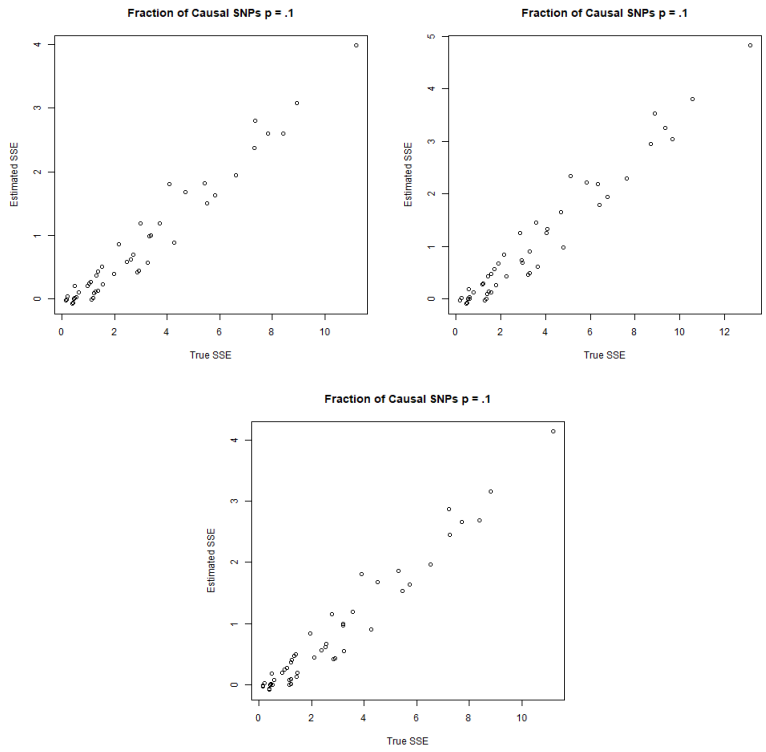

Figure B.6: True versus estimated SSE for TlpSum models applied to the simulation setting with the fraction of causal SNPs $p = .1$. These plots describe the relationship between the true and estimated SSE for three randomly chosen simulation settings among the twenty we conducted. Each plotted point represents one of 48 candidate models. The estimated Pearson correlation coefficients for the plots were, from left to right: $r = .98$, $r = .98$, $r = .98$

especially apparent by the plots for $p = .1$, where some of the estimated SSE values are in fact negative. We stress that the estimation of these SSE values is of use for model comparison especially in the context of AIC and BIC, and shouldn't necessarily be used as a reliable estimate of the magnitude of the SSE. This behavior is due to the systematic underestimation of the $\boldsymbol{\beta}^{'}\mathbf{X}^{'}\mathbf{X}\boldsymbol{\beta}$ term that occurs when a reference panel is used to approximate the $\mathbf{X}^{'}\mathbf{X}$ matrix. This effect is more pronounced as the number of active parameters in the model grows.

B.6.4 Accuracy of Quasi-Correlation

Section **(3.3.2)** demonstrates that quasi-correlation does a good job approximating the true predictive r^2 on out-of-sample data for selected models. Here we expand on those results, showing that quasi-correlation generally does a good job approximating the predictive r^2 for all candidate models.

We investiage the performance of quasi-correlation as follows. For each of the twenty simulation settings, we have a set of candidate models. For each candidate model, we calculate the predictive accuracy on the testing data using predictive r^2, which requires individual level data, and squared quasi-correlation, which requires only summary statistics. We then calculate the correlation between the predictive r^2 values and the squared quasi-correlation values. Thus, for each combination of simulation setting (defined by the fraction of causal SNPs) and model estimation method (i.e. TlpSum or LassoSum), we have a vector of twenty correlations. Entry i of the vector corresponds to the correlation between the predictive r^2 values and the squared quasi-correlation values for the candidate models in replication i. This information is displayed in figure B.7.

The reasoning behind this investigation is as follows. More of interest than whether the quasi-correlation can precisely estimate the magnitude of the predictive r^2 is whether the quasi-correlation can reliably differentiate among the out-of-sample predictive performance for a set of candidate models. If such a differentiation can be made, we can draw conclusions about the comparative quality of different models. Generally, figure B.7 indicates that predictive r^2 and squared quasi-correlation are well correlated, indicating that we can sufficiently differentiate between candidate models.

Figure B.7 demonstrates that quasi-correlation approximates true correlation well in the majority of simulations. In the simulation setting with the proportion of causal SNPs $p = .001$, there are a small number of replications where the squared quasi-correlation does a somewhat poor job of approximating the predictive r^2 for the LassoSum. Nevertheless, the median value is well above .9 when considering all simulations. Candidate sets of TlpSum models are generally well differentiated by squared quasi-correlation. Generally, the better performance of quasi-correlation in differentiating TlpSum models as opposed to LassoSum models can be explained as follows. Candidate sets of TlpSum models generally contain models with a wider spread of predictive r^2 values on out

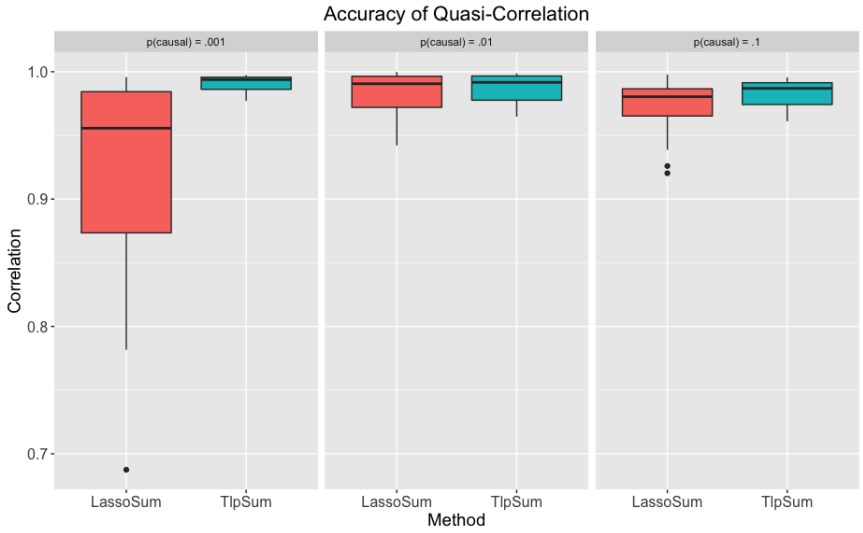

Figure B.7: Accuracy of quasi-correlation approximations in simulation 1. Boxplots represent distribution across twenty replications at each simulation setting of the correlation between predictive r^2 and squared quasi-correlation on out-of-sample testing data.

of sample data, due to the three-dimensional grid search that contains some infeasible values of τ. Quasi-correlation can differentiate quite well between models with substantially different predictive r^2 performance on out of sample data. For several replications, candidate sets of LassoSum models contain models that perform reasonably similarly on out-of-sample data. Quasi-correlation has more difficulty differentiating between these similar models, thus the correlation is lower. However, the satisfactory performance of quasi-correlation for model selection and and for assessment of model performance that we see in figure 3.9 indicates that application of quasi-correlation to LassoSum models is feasible. From the results displayed here and in **(3.3.2)**, we can generally conclude that quasi-correlation is an appropriate and robust measure of predictive performance on out-of-sample data.

B.7 Additional simulation results

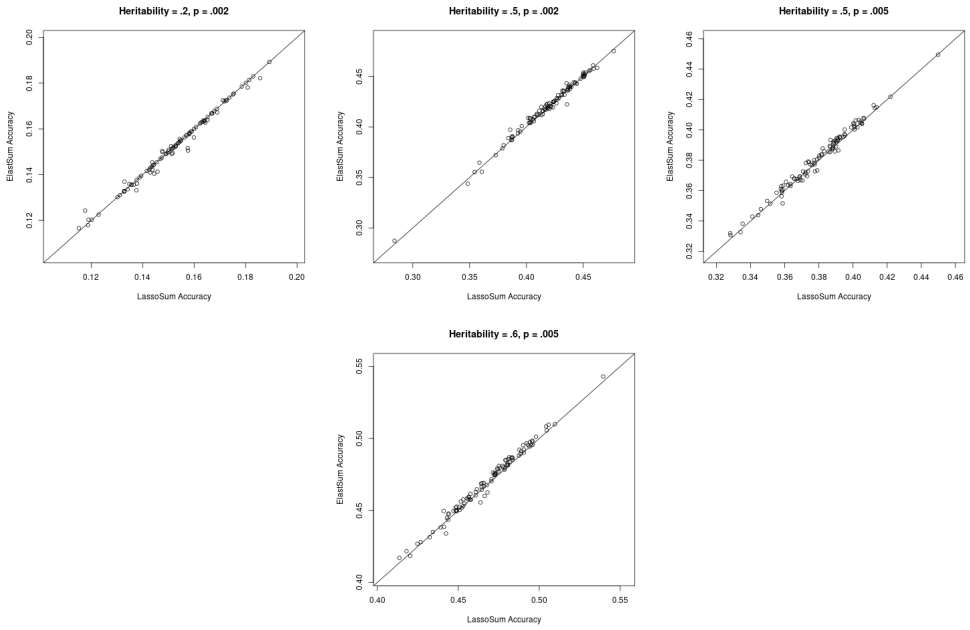

Figure B.8: Predictive r^2 for LassoSum and ElastSum for each of the 100 replications at each of the four simulation settings with allelic heterogeneity. Lines are at a 45 degree angle through the origin, and not a line of best fit. Points below the line indicate better performance of LassoSum.

	ElastSum	TlpSum
LassoSum	2.37E-10 (.001, .002)	3.05E-10 (.004, .008)
ElastSum		1.17E-5 (.002, .006)

Table B.1: P-values and 95% CIs for paired t-test applied to predictive r^2 on out of sample data for the allelic heterogeneity simulation in **(3.1.1)**. Ranges are for the method in the column label less the method in the row label. Simulation setting with $h^2 = .6, p = .005$.

	ElastSum	TlpSum
LassoSum	1.18E-6 (.001, .002)	2.69E-8 (.003, .006)
ElastSum		4.32E-5 (.002, .005)

Table B.2: P-values and 95% CIs for paired t-test applied to predictive r^2 on out of sample data for the allelic heterogeneity simulation in **(3.1.1)**. Ranges are for the method in the column label less the method in the row label. Simulation setting with $h^2 = .5, p = .005$.

	ElastSum	TlpSum
LassoSum	2.30E-5 (.001, .002)	1.61E-10 (.004, .007)
ElastSum		5.57E-8 (.003, .006)

Table B.3: P-values and 95% CIs for paired t-test applied to predictive r^2 on out of sample data for the allelic heterogeneity simulation in **(3.1.1)**. Ranges are for the method in the column label less the method in the row label. Simulation setting with $h^2 = .5, p = .002$.

	ElastSum	TlpSum
LassoSum	.075 (-6.37E-4, 3.16E-5)	6.59E-5 (.001, .003)
ElastSum		5.98E-6 (.001, .004)

Table B.4: P-values and 95% CIs for paired t-test applied to predictive r^2 on out of sample data for the allelic heterogeneity simulation in **(3.1.1)**. Ranges are for the method in the column label less the method in the row label. Simulation setting with $h^2 = .2, p = .002$.