THE UNIVERSITY OF CHICAGO

METHODS FOR INTEGRATIVE MULTI-OMICS ASSOCIATION ANALYSIS USING SUMMARY STATISTICS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF PUBLIC HEALTH SCIENCES

BY
KEVIN J. GLEASON

CHICAGO, ILLINOIS
AUGUST 2020

For my wife, Tracey, whose constant support and numerous sacrifices made this dissertation possible.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

There are many people whose support and assistance made it possible to complete this dissertation. First, I want to thank my advisor, Dr. Lin Chen, for her guidance, insights, and dedication to my training and these research projects. I would also like to thank my dissertation committee: Drs. Brandon Pierce (chair) and Yuan Ji, and especially Dr. Fan Yang, for significant input and effort. Thank you to multiple co-authors and collaborators, especially Drs. Xin He, Jiebiao Wang, Jubao Duan, Hae Kyung Im and Eric Gamazon, for helpful discussions in the developments of these works.

I would like to thank the study participants and research staff of the various consortia whose research generated data and summary statistics used in this work, as well as several individuals involved in providing and/or preprocessing data: Drs. Francois Aguet and Kristin Ardlie, Mr. Alvaro Barbeira, and Mr. Rodrigo Bonazzola.

Finally, thank you to the Department of Public Health Sciences students, faculty, and staff (especially Ms. Michele Thompson), members of the University of Chicago community, and many friends and family members for your unwavering support. While there are too many of you to thank individually, your collective efforts enabled the completion of this work.

Grant support

This research was supported by funding from the National Cancer Institute and the Susan G. Komen Foundation under grant numbers F31CA239557 and GTDR16376189, respectively. I would like to thank Drs. Lin Chen and Habib Ahsan for co-sponsoring my F31 application, and Drs. Suzanne Conzen, Eileen Dolan, and Kathleen Goss and Ms. Michelle Domecki for their involvement with the Susan G. Komen Graduate Training in Disparities Research program. I would also like to thank Drs. Lin Chen and Kavi Bhalla for training support during research assistantships.

ABSTRACT

In the post-genome wide association study era, an important objective is developing a more comprehensive understanding of the biological mechanisms through which genetic variants affect complex traits such as disease susceptibility. Integrative multi-omics association analyses have the potential to elucidate these underlying molecular mechanisms, and the increasing availability of summary-level data makes integrative methods that use only summary statistics as input particularly valuable. But, there are several challenges in performing integrative association analyses using summary statistics. In this dissertation, we develop novel statistical methods and computational tools to address existing challenges and limitations in the joint analyses of multi-omics data from multiple perspectives. In addition to the development of general integrative analysis methods, we make tailored developments to address specific questions in identifying molecular associations of complex trait-associated genetic variants, to integrate statistics from mediation analyses, and to identify genes that are consistent with a causal model in which their expression levels affect variation of a complex trait (such as disease susceptibility). The proposed methods and tools have been applied to study multiple diseases and traits, and they can also be broadly used in many other areas to infer multi-study joint associations, conditional associations, mediations and potential causal associations with only summary statistics as input.

CHAPTER 1

INTRODUCTION

1.1 Overview

Genetic and genomic variation explain varying and often substantial proportions of variations in complex traits and diseases [Muñoz et al., 2016; Polderman et al., 2015; Cannataro et al., 2018]. Genome-wide association studies (GWAS) have identified a large number of these associations between genetic variants and complex traits [Buniello et al., 2019]. In the post-GWAS era, there is significant interest in elucidating the biological mechanisms through which genetic variants affect complex traits. For example, understanding the molecular mechanisms through which genetic variation influences disease processes is critical for identifying treatment targets and prognostic indicators as well as for developing an improved understanding of how various molecular features interact to contribute to disease etiology and progression. To identify these important molecular mechanisms, integrative association analysis methods that utilize summary statistics have become a popular approach for analyzing multi-omics data because of the rapidly increasing availability of summary-level data [Pasaniuc and Price, 2017; Aguet et al., 2019] coupled with the potential for improved power and/or a more comprehensive understanding of underlying biology that can arise from jointly analyzing multiple data types. Yet integrative analyses pose several challenges, and novel statistical methods and computational tools need development to address these challenges. In this work, we develop novel integrative analysis methods and computational tools to address existing challenges and limitations in the joint analyses of multi-omics data from multiple perspectives. In addition to the development of general integrative analysis methods, we make tailored developments to address specific questions in identifying molecular associations of complex trait-associated genetic variants, to integrate statistics from mediation analyses, and to identify genes that are consistent with a causal model in which their expression levels affect variation of a complex trait (such as disease susceptibility). The pro-

posed methods and tools have been applied to study multiple diseases and traits, including cancer phenotypes (breast, ovarian and colorectal), anthropomorphic traits, inflammatory bowel diseases, lipid traits, and schizophrenia, and they can also be broadly used in many other areas to infer multi-study joint associations, conditional associations, mediations and potential causal associations with only summary statistics as input.

1.2 Background

1.2.1 The heritability of disease and traits

For more than a century, scientists have known that human traits often cluster in families [Galton, 1886; Fisher, 1918] The familial aggregation of complex diseases [Edwards, 1969] coupled with the higher correlation of such diseases in monozygotic twins compared to dizygotic twins [Siemens, 1924; Boomsma et al., 2002; Polderman et al., 2015] spurred the search for sources of disease/trait heritability – the proportion of the total variation in a disease/trait that can be attributed to genetic variation [Visscher et al., 2008; Falconer, 1967]. Thus, studying genetic variation and its effects on human complex traits such as disease (and disease susceptibility) became a major research focus with implications for human health.

1.2.2 Germline genetic variation

Genetic risk factors for disease are often inherited through the germline. Before the advent of high-throughput single-nucleotide polymorphism (SNP) genotyping technology, researchers used pedigree information in conjunction with sparse genetic markers such as microsatellites [Hearne et al., 1992] or restriction fragment length polymorphisms (RFLP) [Botstein et al., 1980] to identify potential genetic sources of complex diseases. Given the sparsity of genotyping, pedigrees were necessary for researchers to determine which genetic markers were co-segregating with disease phenotypes in families (e.g. in analyzing genetic linkage) [Morton, 1955; Morgan, 1911]. But in recent decades, the advancement and decreasing costs of first SNP microarray [LaFramboise, 2009] and then Next Generation Sequencing (NGS) [Shendure and Ji, 2008; Baker, 2012] technologies have allowed for analyses of entire genomes from large samples of unrelated individuals [Carlson et al., 2004]. By evaluating the associations between complex diseases and loci where DNA sequences vary in the population at a single position (e.g. SNPs), such genome-wide association studies (GWAS) have the

potential to identify multiple sites in the genome with modest effects on common diseases [Hirschhorn and Daly, 2005; Manolio, 2010]. Indeed, tens of thousands of unique associations between germline genetic variants and human complex traits have been reported in published GWAS, and the list is still expanding [Buniello et al., 2019].

1.2.3 Somatic mutations

In addition to inherited germline variation, genetic variation can also affect complex traits through somatic mutations. Rather than being inherited and present in all of an individual's cells, somatic mutations arise during cell replication and may be present in only a subset of cells in an individual. There are several forms of somatic mutations, such as point mutations, chromosomal translocations, and copy number alterations (CNAs). Figure 1.1 illustrates the phenomenon of copy number alteration, where a large genomic region of a chromosome is either deleted (left) or amplified (right) during replication. Acquired somatic mutations are a hallmark of cancer, affecting all cancer types [Watson et al., 2013]. However, somatic mutations also affect other traits such as some neurological disorders [Poduri et al., 2013] and are being increasingly identified as associated with diseases other than cancer.

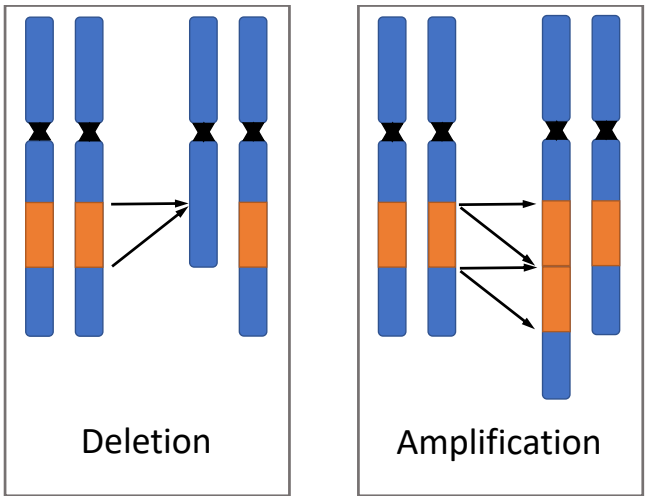


Figure 1.1: **Illustration of copy number alteration (CNA).** A large genomic region (shown in orange) is either deleted (left) or amplified (right).

1.2.4 Molecular (omics) traits

Whether genetic variation is inherited through the germline or acquired through somatic mutation, it is important to understand the biological mechanisms through which these genetic variants act to influence disease processes. Such understanding is necessary to determine the biological and clinical relevance of genetic associations, such as identifying potential treatment targets. However, these biological mechanisms are not always known. For example, the majority of trait-associated SNPs reported by GWAS reside in non-coding regions of the genome [Hindorff et al., 2009], obscuring their functional connections to disease. Because genetic variation also affects the human transcriptome [Gilad et al., 2008], DNA methylome [Smith et al., 2014], histone modification [McVicker et al., 2013; Grubert et al., 2015], proteome [Johansson et al., 2013] and other intermediate traits, studying the effects of genetic variation on intermediate, molecular phenotypes may help further understanding of the underlying mechanisms of complex diseases.

As illustrated in Figure 1.2, regions of DNA are transcribed into mRNA, which is subsequently translated into protein. Proteins carry out various functions throughout the human body, including disease-related functions such as immune [Schroeder and Cavacini, 2010] or inflammatory response [Inohara and Nuñez, 2003; Zhang and An, 2007]. Therefore, disruption to normal activity in one of these steps, or to processes that regulate them, has the potential to cause or mediate disease processes. Studying effects of genetic variation on the transcriptome and proteome, as well as other molecular omics traits, may reveal the biological mechanisms through which genetic variants affect disease.

Figure 1.2: **Illustration of transcription and translation.** Regions of DNA are transcribed into mRNA, which is then translated into protein.

Molecular quantitative trait loci (QTLs)

In order to systematically study how genetic variation affects intermediate phenotypes, many studies of molecular quantitative trait loci (QTLs) have been conducted. Molecular QTLs are sections of the genome where genetic variants are associated with variation in molecular phenotypes. These associations may occur in "cis" (i.e. with local or proximal genes/phenotypes) or in "trans" (i.e. with distal genes/phenotypes). At a QTL, one allele is associated with increased levels of the quantitative trait while the another allele is associated with decreased levels, as illustrated in Figure 1.3. In the example, an adenine nucleotide ("A") at the locus is associated with increased transcription of a nearby (a.k.a. *cis*) gene whereas a thymine nucleotide ("T") is associated with decreased transcription of the gene. Since the SNP is associated with gene expression levels, it is called an expression QTL (eQTL).

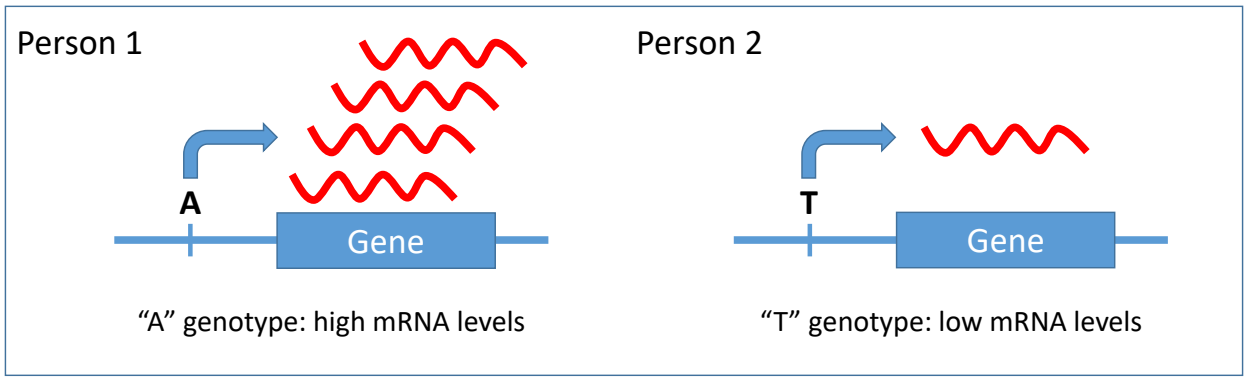


Figure 1.3: **Illustration of an eQTL.** Individuals with the "A" genotype tend to have higher levels of gene expression than individuals with the "T" genotype.

QTL studies may be valuable in elucidating the biological mechanisms through which genetic variants affect complex traits. For example, trait-associated SNPs are enriched for eQTLs [Nicolae et al., 2010]. Thus, many trait-associated genetic variants likely affect complex traits through their effects on molecular traits such as gene expression. Performing integrative analyses of disease GWAS with studies of effects of genetic variation on omics traits, such as QTL studies, may help identify the mechanisms underlying disease processes.

1.3 Integrative omics analysis

In order to gain a complete understanding of the dynamic system of the mechanisms and the interplay among different omics data types underlying complex diseases and traits, methods have been proposed to simultaneously consider and analyze multiple omics data types. Since multiple molecular phenomena interact to contribute to phenotypic traits, the joint study of different layers of molecular structures will develop a more comprehensive understanding of how molecular mechanisms affect susceptibility and other disease-related traits [Kristensen et al., 2014]. Because no single "omics" trait fully encapsulates molecular contributions to disease processes, such "multi-omics" approaches have significant potential to more fully elucidate the molecular mechanisms underlying disease etiology [Rotroff and Motsinger-Reif, 2016]. Furthermore, whereas associations with one data type identify correlations, integration of multi-omics data may identify potentially causative associations that can be further tested in follow-up experiments [Hasin et al., 2017]. Since associations may also differ by tissue type [The GTEx Consortium, 2017], cell type [van der Wijst et al., 2018], or other contexts [Yao et al., 2014; Zhernakova et al., 2017], integrative analyses are necessary for a more complete portrait of the regulatory landscape through which genetic variants act to influence disease processes.

A challenge in jointly analyzing multi-omics data is that often not all omics traits of interest will be measured in the same sample. And even when all traits of interest are measured in the same study, there may be imperfect overlap between subjects (i.e. some subjects are missing measurements for some traits or in some contexts/conditions like cell/tissue type). Restricting analyses to subjects with all traits measured may limit power for the traits (e.g. gene expression) that have a larger sample size than the traits that are more challenging or expensive to ascertain (e.g. protein abundance).

Thus, integrative analyses methods requiring only summary statistics as input offer significant value in elucidating the molecular mechanisms through which genetic variants affect complex traits such as disease susceptibility. Because summary statistics of the effects of

genetic variants on both complex and molecular traits are being made increasingly publicly available [Pasaniuc and Price, 2017; Aguet et al., 2019; Bonder et al., 2017], integrative analysis methods using summary statistics also provide researchers an opportunity to integrate data from traits not measured in their individual studies or to reanalyze existing data to derive new insights. But integrative analysis using summary statistics presents several challenges. There may be heterogeneity between studies or traits being analyzed, such as differences in effect size distributions. If the data comes from overlapping samples, the presence of correlation among traits within the same individual makes it challenging to differentiate between biological correlation, which is the phenomenon of interest, and sample correlation, which is a nuisance that needs to be accounted for. Correlations may also be present between the genetic variants (e.g. due to linkage disequilibrium, or "LD"), making it challenging to differentiate between mere correlations and causal associations. Researchers may also be interested not only in "omnibus" testing, but in detecting specifically the subset(s) of traits with which a genetic variant demonstrates association. As the number of traits/studies/conditions being considered grows, not only will computational challenges increase, but it will also become more likely to detect joint associations by chance, necessitating proper multiple testing adjustment.

In the next three subsections, we review the basic tenets of three types of integrative analysis methods based on summary statistics for studying the effects of genetic variants on complex traits and/or molecular phenotypes: meta-analysis methods, colocalization methods and two-sample MR methods.

1.3.1 Meta-analysis Methods

Meta-analysis methods combine the results of multiple studies of the same phenomenon to derive a pooled estimate for inference. Traditional meta-analysis approaches, such as Fisher's method for combining P-values [Fisher, 1932] and Stouffer's method for combining z-scores [Stouffer et al., 1949], are designed to evaluate studies of the same association measured in

multiple distinct samples. The measurement from each study is considered as one sample estimate of a common underlying parameter in a population. Because the methods assume there is a common true effect size underlying each statistic, traditional meta-analysis methods may not be well-suited to perform integrative analyses of different data types since each data type may have a different underlying effect size (and effects between data types may even be in opposite directions). Traditional meta-analysis methods also assume each statistic comes from an independent sample, an assumption that may be violated in the case where integrative analyses of the different data types are conducted in overlapping samples (though some meta-analysis methods have been developed to relax the independence assumption; see e.g. Brown, 1975, and Kost and McDermott, 2002).

1.3.2 Colocalization Methods

Colocalization methods have become a popular approach for performing integrative analyses of the effects of genetic variants on multiple traits. These methods attempt to distinguish cases of colocalization – when two or more traits share a common underlying causal variant – from cases where apparent associations in a region come from distinct causal variants. Here we review several commonly used colocalization methods: coloc, moloc, enloc, and eCAVIAR. These methods have identified known and novel candidate genes underlying psychiatric disorders [Giambartolomei et al., 2018], diabetes traits [Hormozdiari et al., 2016], obesity-related traits [Giambartolomei et al., 2014; Wen et al., 2017], and others. A primary strength of the methods is moving beyond the observed correlation between traits towards identifying potentially causal associations. Common limitations include the inability to account for potential sample correlations and restrictions of the number of traits that can be integrated. All methods allow for presence of linkage disequilibrium (LD), but only eCAVIAR incorporates LD correlation estimates directly.

coloc [Giambartolomei et al., 2014] is a Bayesian statistical test for colocalization that analyzes exactly two traits in a single genomic region to quantify the probability that the

two traits share a causal variant in the region. The authors note that a major focus is interpreting patterns of (inferred) LD in the locus. The method assumes that traits are measured in unrelated individuals and that effect sizes for the two traits are independent. coloc also assumes that there is no more than one causal variant for each trait in the region, an assumption that is violated in regions harboring allelic heterogeneity (AH) as discussed in Hormozdiari et al. (2016). It requires specification of prior probabilities for being associated with only the first trait, only the second trait and both traits. As discussed and demonstrated in Wen et al. (2017), analysis results may be sensitive to this prior specification.

moloc [Giambartolomei et al., 2018] extends the methodology of coloc to more than 2 traits. Like coloc, moloc assumes that traits are measured in unrelated individuals, that effect sizes for the traits are independent, and that there is no more than one causal variant for each trait in the region. The method requires prior specification for the probabilities of each causal association pattern (referred to as "configurations"). For J traits, this requires the specification of 2^J priors. The R package implementation of moloc makes the simplifying assumption that each n-way configuration has the same prior probability. That is, the prior probability of being associated with only trait X is the same as the prior probability of being associated with only trait Y; the prior probability of being associated with both X and Y is the same as the prior probability being associated with both Y and Z; etc.

enloc [Wen et al., 2017] uses a Bayesian hierarchical model that treats the latent association status of molecular QTLs as variant-level annotations for candidate SNPs of complex traits. It is implemented to study exactly one complex trait GWAS with exactly one QTL study. In recognition of the importance of LD, the method studies one LD block at a time, but it does not directly incorporate LD correlation estimates in its analysis. enloc does not adjust for potential sample correlation between the two traits being analyzed.

eCAVIAR [Hormozdiari et al., 2016] is a colocalization method that performs Bayesian fine-mapping within each LD block to obtain estimates of the posterior inclusion probability (PIP) for each variant of being causal in a GWAS study and an eQTL study. By assuming

that the prior probability of the same variant being causal in both a GWAS and eQTL study is independent, the colocalization posterior probability (CLPP) for each variant can be obtained as the product of the posterior probabilities of causality in each trait. In its analysis, eCAVIAR incorporates LD correlation estimates, which can be obtained from reference panels if individual level data is not available. The method does not adjust for potential sample correlation between the two traits being analyzed.

Thus, common limitations to colocalization methods include an inability to account for potential sample correlations and restrictions on the number of traits that can be integrated. Other limitations may include assumptions about the (maximum) number of causal variants within a genomic region and specification of a potentially large number of priors with sensitivity to that specification.

1.3.3 Two-sample Mendelian Randomization Methods and recent developments

Another integrative association analysis method that utilizes summary statistics is two-sample Mendelian Randomization. Mendelian Randomization (MR) is an instrumental variables (IV) approach that utilizes germline genetic variants as instruments to assess the causal association between an exposure and an outcome [Katan, 1986; Smith and Ebrahim, 2003; Thomas and Conti, 2004]. MR treats genetic inheritance as a natural experiment in which inheritance of genetic variants for an offspring from each of its parents during meiosis follows a pseudo-random process (hence "Mendelian Randomization"). The genotype serves as an instrument for an exposure of interest to assess the causal effect of that exposure on a disease (or other complex trait). When subject-level genotypes and measurements of exposure and disease status are all available within the same sample, traditional IV approaches such as two-stage least squares regression may be used for MR analysis [Burgess et al., 2017]. However, often these three data are not measured in the same large sample, creating a need for MR methods that can be performed using summary statistics from analyses conducted

in more than one sample.

In two-sample MR, summary statistics from a sample assessing the marginal effects of the genetic variants (IVs) on the exposure are integrated with summary statistics from another sample assessing the marginal effects of the IVs on the outcome (e.g. disease susceptibility). The two sets of summary statistics are evaluated for consistency with a model in which the exposure is causally associated with the outcome. The rationale for two-sample MR is depicted in Figure 1.4. When measurements of genotypes, exposure and outcome (e.g. disease susceptibility) are not all available in the same sample, the ratio of the marginal effect of a SNP on the outcome (β_Y) to the marginal effect of the SNP on the exposure (β_X) is a commonly used estimand for the effect of exposure on disease (γ).

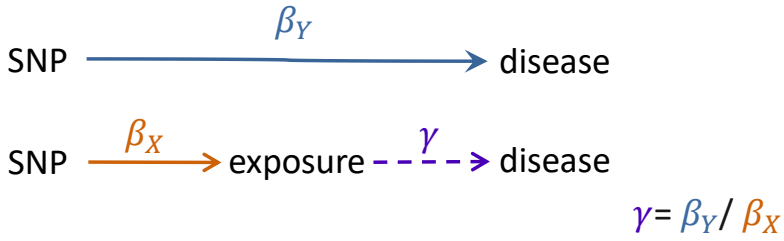


Figure 1.4: **Illustration of the concept of two-sample Mendelian Randomization (MR).** A genetic variant (SNP) is used as an instrumental variable (IV) to assess a potential causal relationship between an exposure and outcome (e.g. disease susceptibility). Marginal summary statistics assessing the effect of the SNP on the exposure (β_X) from one study are integrated with marginal summary statistics from another study assessing the effects of the SNP on an outcome (β_Y) such as disease susceptibility to infer the effect of the exposure on the outcome (γ). The ratio β_Y/β_X forms one commonly used estimand for γ.

The individual ratio estimates of the effect of the exposure on the outcome for each SNP i, $\hat{\gamma}_i = \frac{\hat{\beta}_{Y_i}}{\hat{\beta}_{X_i}}$, can be combined into a single estimate, as is achieved in the inverse-variance weighted (IVW) method [Burgess et al., 2013]. The IVW method yields equivalent estimates to the two-stage least squares method often used when individual data (rather than only summary statistics) are available [Burgess et al., 2016]. A challenge in MR methods not addressed by the IVW method, however, is that the "exclusion restriction" criteria of IV analyses (in which the only pathway through which the IV may affect the outcome is through the exposure) may often be violated in MR due to pervasive pleiotropy throughout

the genome [Lawlor et al., 2008; Morrison et al., 2020]. Genetic variants that affect the outcome through other pathways besides the exposure are invalid IVs. To address the issue of invalid IVs and other challenges in two-sample MR, several robust two-sample MR methods have been proposed.

Some robust two-sample MR methods detect outlying ratio estimates among the genetic variants, and subsequently remove or downweight the outliers. MR-PRESSO detects outliers through an iterative leave-one-out process that identifies outliers when substantial decreases occur in the residual sums of squares (RSS) in regressions of the IV-outcome statistics ($\hat{\beta}_{Y_i}$) onto the IV-exposure statistics ($\hat{\beta}_{X_i}$) [Verbanck et al., 2018]. Outliers are removed, and IVW is performed on the remaining genetic variants. Penalized regression methods to remove outliers in the MR setting add an intercept for each variant to the IVW regression model and then apply a penalty (e.g. lasso) to the intercepts [Windmeijer et al., 2019]. The penalty tuning parameter is iteratively increased until reaching a heterogeneity criterion. Genetic variants with non-zero intercept terms are removed from the analysis. Other approaches use robust regression methods (such as MM-estimation with or without Tukey's bisquare objective function) or penalized weights to lessen but not exclude the effects of outliers on the causal estimate [Rees et al., 2019].

Other robust two-sample MR methods model the distributions of estimates from invalid IVs. MR-Egger incorporates an intercept term in a regression model that represents the average pleiotropic effect of the genetic variants [Bowden et al., 2015]. Under the Instrument Strength Independent of Direct Effect (InSIDE) assumption that the pleiotropic effects are uncorrelated with the IV-exposure association, MR-Egger gives consistent estimates of the causal effect [Bowden et al., 2015]. MRMix attempts to divide variants into four categories based on their exposure and outcome association status, and uses a spike-detection algorithm under a normal-mixture model for the underlying effect-size distributions [Qi and Chatterjee, 2019]. MR-RAPS uses random intercepts to model pleiotropy, assuming that pleiotropic effects follow a normal distribution centered on zero with unknown variance, with or without

contamination (where most effects follow a normal distribution but some may be much larger) [Zhao et al., 2019]. A profile-likelihood function is used to estimate the causal effect and the variance of the pleiotropic effects, with Tukey's biweight or Huber's loss functions used for robustness against outliers.

Many of the aforementioned robust two-sample MR methods were motivated by settings where the exposure is an intermediate complex trait (e.g. blood pressure, body mass index or lipid measurements) that may be affected by many SNPs throughout the genome. To properly assess the causal effect between exposure and outcome, these methods require a large number of IVs and/or IVs that are relatively independent. Thus, they may not be well-suited for settings where the exposure is a molecular omics trait (e.g. gene expression) since omics traits are often most strongly regulated in cis, with a limited number of potentially correlated genetic variants available as candidate IVs. There is a need to develop novel two-sample MR methods motivated by settings where a limited number of correlated genetic variants can be used as IVs.

1.4 Summary

In this work, motivated by the need to elucidate the molecular mechanisms through which genetic variants affect disease susceptibility and to further identify the putative causal molecular risk factors for complex diseases/traits, we develop integrative multi-omics analysis methods and computational tools that take summary statistics as input and jointly analyze the effects of genetic variants and/or molecular traits on multiple complex and/or omics traits.

The rest of the dissertation is organized as follows. In Chapter 2, we develop a general integrative association analysis method for multi-omics data and apply the method to study the cis- and trans-acting effects of copy number alterations (CNAs) on omics traits in multiple cancer types. In Chapter 3, we extend the proposed framework to integrate product of coefficients test statistics to identify cis-mediated trans-associations that are shared across

conditions, and we use the method to identify shared cis-mediated trans-protein associations of CNAs in breast and ovary tumors. In Chapter 4, we tailor the integrative analysis method to assess the molecular mechanisms through which GWAS SNPs might affect complex traits by accounting for linkage disequilibrium (LD) when analyzing associations of germline variants. The method is applied to study multi-omics associations of known breast cancer-risk SNPs as well as to identify potential pleiotropic associations shared between complex traits while elucidating their associations with gene expression in disease-relevant tissues. In Chapter 5, we develop a two-sample Mendelian Randomization (MR) method that leverages the rich variation in multi-tissue eQTL reference datasets to detect genes associated with a complex trait, and we apply the method to identify genes with potentially causal associations with schizophrenia risk. Finally, in Chapter 6, we summarize the presented works and suggest possible directions for future research.

CHAPTER 2

AN INTEGRATIVE ASSOCIATION ANALYSIS METHOD FOR MULTI-OMICS DATA

2.1 Introduction

Genome-wide associations studies (GWAS) of germline variation effects [Buniello et al., 2019; Tam et al., 2019] and analyses of the effects of somatic mutations [Watson et al., 2013; Poduri et al., 2013; Erickson, 2003] have established an important role of DNA variation in disease etiology and other complex traits. It is also known that DNA variation affects gene expression [Aguet et al., 2019; Shao et al., 2019; Jia and Zhao, 2017], protein abundance [Suhre et al., 2017; Sun et al., 2018; Jia and Zhao, 2017], and other molecular (i.e. omics) traits [McClay et al., 2015; Aguet et al., 2019; McVicker et al., 2013; Li et al., 2017b]. It is likely that DNA variation affects complex traits through its effects on molecular traits [Albert and Kruglyak, 2015; Civelek and Lusis, 2014]. To more fully understand the mechanisms through which DNA variation affects complex traits, it may be necessary to jointly analyze the effects of DNA variation on multiple molecular traits. While all molecular traits of interest may not be measured in the same sample, summary statistics from high-throughput analyses of omics traits are becoming increasingly available. Thus, statistical methods that perform integrative multi-omics association analysis using summary statistics have the potential to elucidate the molecular mechanisms through which DNA variation affects disease phenotypes.

Integrative multi-omics analysis may help identify molecular mechanisms underlying cancer phenotypes [Kristensen et al., 2014]. It is well known that cancer initiation and progression is affected by alterations of DNA sequences [Futreal et al., 2004]. These DNA variations include inherited germline polymorphisms, which affect disease susceptibility, and somatic mutations that arise during cell replication [Futreal et al., 2004]. A form of DNA alteration common in tumors is copy number alteration (CNA) [Beroukhim et al., 2010; Vogelstein et al., 2013] – the amplification or deletion of a large genomic region. In particular, the

amplification of oncogene regions or deletion of tumor suppressor regions may play an important role in tumorigenesis [Wee et al., 2018; Zhao and Zhao, 2016; Vogelstein et al., 2013]. However, the ubiquity of CNAs in tumors makes it challenging to differentiate causal cancer driver mutations from random passenger mutations [Beroukhim et al., 2007; Taylor et al., 2008; Stratton et al., 2009; Vogelstein et al., 2013]. Because CNAs often affect multiple local and distal omics traits, such as gene expression and protein abundance, analyzing the effects of CNAs on omics traits may improve understanding of how CNAs influence tumorigenesis and may help distinguish driver events from passenger events [Akavia et al., 2010].

A joint analysis of the effects of CNAs on multiple molecular traits may further aid in identifying driver mutations. In addition to producing a more comprehensive understanding of how different molecular features interact, the joint analysis of multiple traits may increase confidence in research findings by confirming associations in multiple lines of evidence [Kristensen et al., 2014]. Whereas an association in one data type establishes correlation, integration of different omics types may be used to identify potential causative changes that can be further tested in follow-up experiments [Hasin et al., 2017]. In particular, investigations of CNA effects on mRNA and protein abundances within a given cancer type have illustrated the importance of integrative analyses [Mertins et al., 2016; Zhang et al., 2016a, 2014]. For example, in an analysis of CNA effects in breast tumors, Mertins et al. [2016] found that "CNA events with a tumour-promoting outcome more likely lead to cis-regulatory effects on both the protein and mRNA level, whereas CNA events with no documented role in tumorigenesis are more likely to be neutralized on the protein level". And Zhang et al. [2014] proposed that integrating proteomics data may enable prioritization of candidate driver genes since the authors found many fewer CNAs with strong effects on protein abundance compared to mRNA abundance in colorectal tumors. Thus, integrative analyses of multi-omics data within a cancer type may aid in distinguishing important driver mutations. However, when jointly analyzing multiple omics traits, the number of samples with all omics traits being measured is often limited. Therefore, we propose to conduct marginal analyses

of effects within each omics trait, using all samples having that trait measured, and then to perform integrative analysis using the sets of summary statistics from the marginal analyses.

Integrative analyses may also be used to jointly analyze associations in multiple related diseases, such as distinct cancer types. Such joint analysis approaches may highlight shared treatment targets or prognostic indicators as well as boost confidence in the findings that are shared across cancer types. Indeed, a "pan-cancer" analysis of CNAs found that recurrent alterations across multiple tumor types show enrichment of cancer-related genes [Kim et al., 2013]. Another analysis found that amplification regions shared across several cancer types frequently contained functionally validated oncogenes while deletion regions frequently contained functionally validated tumor suppressors [Beroukhim et al., 2010]. Thus, integrating multi-omics data across cancer types has the potential to further highlight genes of important significance (such as oncogenes) across multiple cancer types.

There are many benefits to integrative analyses of summary statistics. They provide a more comprehensive picture of how features interact than do analyses of individual features, increase confidence in shared associations, and may boost power for discovery by leveraging shared correlations across traits. However, integrative analyses also present challenges. When integrating studies of different data types, the effect size distributions may vary. In other words, even when there is a non-zero association for each of two data types, the effect size may not be the same (and may even be in opposite directions). Since large consortia often collect multiple data types on the same or overlapping sets of individuals, it is necessary to be able to account for sample correlation to distinguish between the apparent joint associations that arise simply due to this sample correlation from joint associations due to biological correlation. Also, as the number of traits being studied grows, so too does the computational burden and data complexity, making integrative analyses of multiple data types challenging.

To address existing challenges in conducting integrative analyses, in this work we develop a statistical method and computational tool to integrate summary statistics of multi-omics

data – the Package in R for Integrative Multi-Omics association analysis (Primo). Primo takes as input summary statistics from multiple studies or of multiple traits and performs integrative analysis to estimate probabilities of joint associations. Primo is flexible in many aspects: it allows unknown and arbitrary study heterogeneity and can detect coordinated effects from multiple studies while not requiring the effect sizes to be the same; it allows the summary statistics to be calculated from studies with independent or overlapping samples with unknown sample correlations; and it is not an omnibus test for association, but rather can be used to calculate the probability of each observation (e.g. genetic variant or gene) belonging to each type (or groups) of interpretable association patterns (e.g. the probability of copy number alteration also being associated with at least one/two cis omics-traits).

We used simulations to test the performance of Primo in estimating key parameters and in analyzing sets of summary statistics from correlated samples. We applied Primo to identify joint associations of the effects of DNA copy number alterations (CNAs) on gene expression and protein abundance in three different cancer types (i.e. breast, ovarian and colorectal) using data from The Cancer Genome Atlas (TCGA) [Koboldt et al., 2012; Bell et al., 2011; Muzny et al., 2012] and the Clinical Proteomic Tumor Analysis Consortium (CPTAC) [Mertins et al., 2016; Zhang et al., 2016a, 2014]. The simulations and data applications demonstrate the utility of Primo in performing genome-wide integrative analysis of multi-omics data from different data types or sets of samples.

2.2 Methods

2.2.1 Primo as a general framework for assessing joint associations across data types

Here we first introduce the general Primo association framework. As a general integrative association method, Primo takes as input multiple sets of association summary statistics from different studies of different data types. The multiple sets of summary statistics could

be gene-level associations of DNA copy number alterations to multiple molecular traits in multiple tumor types (Chapter 2), or one or more sets of complex trait-GWAS statistics and multiple sets of omics/multi-omics QTL statistics (see Chapter 4), or could even be from studies beyond the complex and/or omics trait-associations of somatic or germline variation.

Consider an $m \times J$ matrix of association statistics, $\mathbf{T}$, consisting of the summary statistics for the associations of m observations with J types of traits from J studies with independent or correlated samples. The observations could be genes (if studying gene-level associations), or SNPs (if studying germline variant-level associations), or some other unit being analyzed. Note that here a "study" refers to a study of associations to a particular trait in a particular condition/cell-type/tissue-type. For each observation (here a row in the matrix $\mathbf{T}$), the underlying association status to the j-th ($j = 1, \ldots, J$) trait is binary. Considering all observations in the genome, there are a total of $K = 2^J$ possible association patterns to J traits. We use a $K \times J$ binary matrix, $\mathbf{Q}$, to denote all of the possible association patterns. And $q_{kj} = 1$ implies *the presence of association* with the j-th trait in the k-th association pattern, and $q_{kj} = 0$ implies *no association*. Figure 2.1 presents an illustrative example of the $\mathbf{Q}$ matrix when studying the effects of DNA copy number alteration (CNA) on gene expression in three cancer types.

For each observation i, there must be one and only one true underlying association pattern. Primo calculates the probability of a given observation being in each of the K mutually exclusive association patterns by borrowing information across observation in the genome and across J traits. More specifically, let a_i denote the true association pattern for observation i. Then the probability that observation i belongs to association pattern k is given by:

$$P(a_i = k|T_i, \pi_k) = \frac{\pi_k D_k(T_i)}{\sum_{b=1}^{K} \pi_b D_b(T_i)}, \tag{2.1}$$

where T_i is a vector of J association statistics and is also the i-th row in the $\mathbf{T}$ matrix, π_k represents the overall proportion of observations in the genome belonging to the k-th association pattern ($k = 1, \ldots, K$), and $D_k(\cdot)$ is the multivariate density function of J sets

k	Q matrix $K = 2^J$			Interpretation
	Breast	Ovarian	Colorectal	CNA associated with expression in:
1	0	0	0	No assoc. in any cancer type
2	1	0	0	Breast only
3	0	1	0	Ovarian only
4	0	0	1	Colorectal only
5	1	1	0	Breast+Ovarian ; not Colorectal
6	1	0	1	Breast+Colorectal ; not Ovarian
7	0	1	1	Ovarian+Colorectal ; not Breast
8	1	1	1	All 3 cancer types

Figure 2.1: **Example of the binary matrix, Q.** The example also provides interpretations of association patterns for an analysis of the effects of DNA copy number alteration (CNA) on gene expression in Breast, Ovarian and Colorectal cancers for $j = 1, 2$ and 3, respectively. The red box shows how association patterns can be collapsed into groups of interest (here, summing probabilities across the patterns in the red box would yield the probability of association between CNA and gene expression in *at least two* cancer types).

of statistics, conditioning on the k-th association pattern. Here π_k captures the biological co-occurrence frequency of the k-th association pattern in the genome, with $\sum_k \pi_k = 1$. For example, in Figure 2.1, π_8 is the proportion of genes in the genome where DNA copy number alteration (CNA) is associated with gene expression in all three cancer types.

In estimating a mixture distribution of K components, the performance of estimation and subsequent inference depend on how well different mixing components separate from each other. When K is moderate to large, it is challenging to simultaneously estimate the distributions of mixing components (D_k's) and the mixing proportions (π_k's). Different from previous work [Wei et al., 2015], Primo first estimates the pattern-specific multivariate density function D_k for each of the association pattern by borrowing information across observations and traits. (See section 2.2.3 for detailed estimation procedures when J sets of association statistics were calculated from independent or correlated samples, as well as sections 2.2.2 and 2.2.5 for discussion of two versions of the method for integrating t-statistics or P-values, respectively.) Then Primo estimates π_k's via the Expectation-Maximization algorithm [Dempster et al., 1977]. When D_k's are reasonably-estimated, the one-step estimates of π_k's can well capture the overall proportions of different association patterns and there is no need to re-iterate and re-estimate D_k's and π_k's. Based on (2.1), we can obtain the posterior probabilities of SNP i being in each of the K possible association patterns. See Figure 2.2 for an overview of the Primo algorithm.

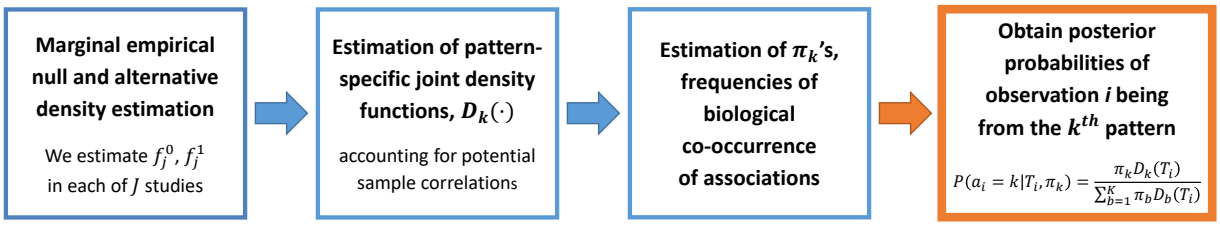


Figure 2.2: **Illustrative overview of Primo.** The main steps of the Primo algorithm for assessing joint associations.

An advantage of Primo is that one may collapse many association patterns based on biological interpretations and obtain the posterior probabilities of groups of patterns of interest

by summing over the probabilities of those mutually exclusive patterns. As illustrated in Figure 2.1, when $J = 3$, there are 8 possible association patterns. We may collapse the association patterns into interpretable groups. For example, we may be interested in the genes where DNA copy number alteration (CNA) is associated with expression in at least two cancer types. And we can obtain the probability estimate by summing over the posterior probabilities of patterns 5-8. When J is large, some specific patterns might not be present in the genome. With collapsed patterns, this would not be an issue, and both interpretability and robustness of the results are enhanced.

2.2.2 Estimating empirical null and alternative marginal density functions for each of the J studies using the limma method

Here we describe the Primo algorithm when the test-statistics for integrative analysis are approximately normal (e.g. z-scores or t-statistics). For each of the J studies, we first adopt the limma method [Smyth, 2004; Ritchie et al., 2015] to calculate a set of *moderated* t-statistics by replacing the error variance estimates in the classical t-statistic calculation with the posterior variances. The new variance shrinks the observed sample variance towards a prior that is estimated across all observations in the data, and stabilizes the variance estimation across the genome. It also penalizes the observations with large t-statistics but small variances.

Next, for each study j, we estimate the empirical null and alternative marginal density functions, $\hat{f}_j^0(\cdot)$ and $\hat{f}_j^1(\cdot)$, respectively, based on all the moderated t-statistics in the genome for the study. Here one needs to specify a key parameter for each study, the proportion of study-specific non-null statistics (i.e. with associations), θ_j^1. We then adopt the limma method to estimate $\hat{f}_j^0(\cdot)$ and $\hat{f}_j^1(\cdot)$ (illustrated in Figure 2.3). Under the null hypothesis, the moderated t-statistic follows a t-distribution with a mean of zero and moderated degrees of freedom d_j in the j-th study, allowing for an empirical null distribution slightly deviating from the parametric t-distribution. Under the alternative, the moderated t-statistic follows a

scaled t-distribution, still with degrees of freedom d_j and a mean of zero allowing for different directions of effects in different studies, and an observation-specific scaling factor v_{ij} ($v_{ij} \geq 1$) estimated from the data. The scaling factor is calculated as $v_{ij} = (1 + v_{0j}/w_{ij})^{1/2}$, where v_{0j} is the variance hyperparameter for the prior placed on non-zero effect size coefficients and w_{ij} is an observation-specific weight for observation i. In the analyses presented in this chapter, we set $w_{ij} = 1$ for each gene i (see Section 4.2.2 for discussion of SNP-specific weights for the variant-level analyses of Chapter 4). The degrees of freedom d_j is estimated from the data as $d_j = d_{0j} + d_{1j}$, where d_{1j} is the (original) degrees of freedom of the summary statistics in study j and d_{0j} is the degrees of freedom hyperparameter for the prior on the unknown variances of effect sizes. Estimation is performed using an empirical Bayes approach as described in Smyth (2004) and implemented in the limma package in R [Ritchie et al., 2015].

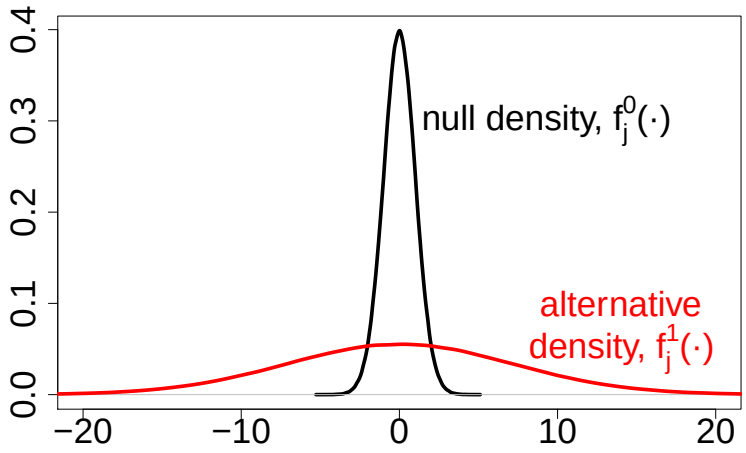


Figure 2.3: **Illustration of marginal null and alternative densities for moderated t-statistics.** Under the null hypothesis, the moderated t-statistics follow a t-distribution with moderated degrees of freedom (with density function $f_j^0(\cdot)$). Under the alternative hypothesis, the moderated t-statistics follow a *scaled* t-distribution with moderated degrees of freedom (with density function $f_j^1(\cdot)$).

With the estimated marginal null and alternative density functions from each study, the joint density functions for all K association patterns can be calculated as described in Section 2.2.3.

2.2.3 Estimating pattern-specific multivariate density functions when input summary statistics are calculated from independent or overlapping samples

With J independent studies, the pattern-specific multivariate density function D_k for the k-th association pattern is given by

$$D_k(T_i) = \prod_{j=1}^{J} f_j^0(t_{ij})^{1-q_{kj}} f_j^1(t_{ij})^{q_{kj}}. \tag{2.2}$$

where q_{kj} is the association status of the k-th pattern in study j. For example, given the association status being $q_k = (1, 1, 0, 0)$, the joint density D_k is modeled as the product of the alternative marginal density functions from the first two studies and the null marginal density functions from the other two studies, $D_k = f_1^1 \cdot f_2^1 \cdot f_3^0 \cdot f_4^0$.

In estimating a pattern-specific multivariate density function D_k from J correlated studies, we obtain the empirical null and alternative marginal distributions as non-scaled and scaled t-distributions, respectively, in each of the J studies. Then we further approximate them with normal distributions with zero means and variances being $\sigma_{ikj}^2 = v_{ij}^{2\times q_{kj}} \cdot \frac{d_j}{d_j-2}$, where v_{ij} is the scaling factor under the alternative. When the association status indicator $q_{kj} = 0$ for the j-th study under pattern k, i.e., no association, $\sigma_{ikj}^2 = \frac{d_j}{d_j-2}$. Since J studies are correlated due to possible sample overlap with an unknown correlation matrix of $\mathbf{\Gamma}$, similar to Urbut et al. (2019) we pool all the statistics likely to be from the null pattern to estimate their correlation matrix as the estimate for $\mathbf{\Gamma}$. Under certain assumptions, the correlation matrix of test statistics approximates the sample correlation matrix and the sample correlation under the null represents the correlation due to sample overlap. Here we estimate the $J \times J$ correlation matrix using observations with absolute statistics less than 5 in all J studies. Then, we approximate the pattern-specific multivariate density function D_k as $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_k^{1/2}\mathbf{\Gamma}\mathbf{\Sigma}_k^{1/2})$, where $\mathbf{\Sigma}_k$ is a diagonal matrix with diagonal elements of σ_{ikj}^2's.

Note that here the normal approximations of multivariate density functions enjoy computational efficiency, and moreover (as described in Section 4.2.3), they facilitate the estimation of conditional density functions. Also note that Primo separates sample correlations $\mathbf{\Gamma}$ from biological correlations/co-occurrences captured by π_k's in the subsequent estimation and inference.

2.2.4 Estimating the false discovery rate (FDR) to account for multiple testing

The genome-wide integration of summary statistics allows for testing of multiple hypotheses (e.g. testing each gene in the genome for association with copy number alteration). However, it is necessary to account for these multiple tests in order to control the Type I error. One way to account for multiple hypothesis testing is to control the false discovery rate (FDR) – the expected proportion of "discoveries" (i.e. rejections of the null hypothesis) that are false (i.e. the null hypothesis is true) [Benjamini and Hochberg, 1995]. For a pattern of interest, Primo calculates the estimated FDR [Storey and Tibshirani, 2003] for multiple testing adjustment as follows:

$$\text{estFDR}(\lambda) = \frac{\sum_i (1 - \hat{P}_i) 1(\hat{P}_i \geq \lambda)}{\#\{\hat{P}_i \geq \lambda\}}, \tag{2.3}$$

where λ is the probability threshold and $\hat{P}_i$ is the estimated probability of SNP i being in the (collapsed) pattern of interest. Users may determine the FDR for a pre-selected posterior probability threshold λ, or pass a grid of possible λ's to determine the threshold which controls the FDR at a desired level.

2.2.5 Primo for integrating P-values from multiple studies

In addition to integrating t-statistics or effect sizes and variance estimates, Primo can also jointly analyze J sets of P-values, chi-squared statistics, or other second-order association

statistics. We model the pattern-specific multivariate density functions and still use equations (2.1) in obtaining the posterior probabilities for each observation being in each pattern.

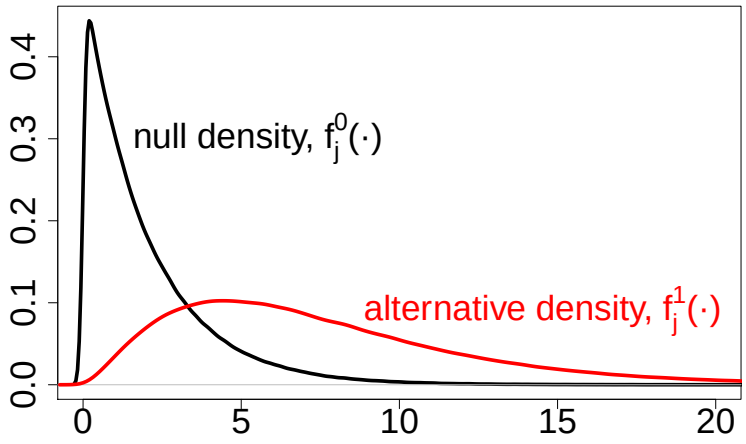


Figure 2.4: **Illustration of marginal null and alternative densities for $-2\log(P)$.** Under the null hypothesis, $-2\log(p_{ij})$ follows a χ^2_2 distribution (with density function $f_j^0(\cdot)$). Under the alternative hypothesis, $-2\log(p_{ij})$ follows a mixture of non-central χ^2 distributions, which is approximated by a scaled chi-squared distribution with certain degrees of freedom, $A_j\chi^2_{d'_j}$.

In estimating the marginal null and alternative density functions for each study j, f_j^0 and f_j^1 (as illustrated in Figure 2.4), we make the following modification. We first take negative two times the log of P-values as our test statistics, $\mathbf{T}$. Under the null hypothesis, $t_{ij} = -2\log(p_{ij})$ follows a χ^2_2 distribution. Under the alternative, the P-value distributions may vary (e.g. locus by locus). In the genome, the alternative distribution of $-2\log(p_{ij})$ $(i = 1, \ldots, m)$ follows a mixture of non-central chi-squared distributions, which can be approximated by a scaled chi-squared distribution with certain degrees of freedom, $A\chi^2_d$ [Satterthwaite, 1946; Solomon and Stephens, 1977]. Note that we do not assume P-values under the alternative follow the same distribution, rather we approximate the mixture of chi-squared distributions using a scaled chi-squared distribution. To estimate a study-specific scaling factor $A_j > 0$ and degree of freedom d'_j that best approximate the tail of the alternative distribution in study j, we use a numerical optimization algorithm to find values which minimize the differences between the P-values of T_j under a mixture of $A_j\chi^2_{d'_j}$ and

χ^2_2 distributions given the mixing proportion θ^1_j for the study, and their nominal P-values based on their ranks.

More specifically, let $t_{ij} = -2\log(p_{ij})$ for SNP i in study j. Then the cumulative distribution function of t_{ij} is given by

$$F(t_{ij}; A_j, d'_j, \theta) = (1-\theta^1_j)G(t_{ij}; 2) + \theta^1_j G(\frac{1}{A_j}t_{ij}; d'_j)$$

where $G(\cdot;\nu)$ is the cumulative distribution function of a χ^2_ν variable. Let r_{ij} be the rank of SNP i in study j when the t_{ij} are sorted in descending order. To estimate A_j and d'_j, we use the optimization algorithms implemented in the R nloptr package [Johnson, 2018] to minimize the following objective function:

$$\sum_{i:\, r_{ij} \le \max\{20, \frac{m}{2}\theta^1_j\}} \left| 1 - F(t_{ij}; A_j, d'_j, \theta^1_j) - \frac{r_{ij} - 0.5}{m} \right|.$$

Since associations can be sparse (i.e., θ^1_j being close to zero) in the genome, it is more important to well approximate the tail of the alternative distribution than the first two moments (mean and variance). As such, we sum over the most extreme tail statistics or at least the 20 most extreme statistics. In Section 2.3.2, we assess the performance of the approximation via simulation studies, especially when associations are sparse. When the J studies are independent, the multivariate density function is modeled as the product of the individual density functions, as in Equation (2.2). When the J studies are correlated, we proceed in a similar manner as when t-statistics are used as input, except that the multivariate normal distribution is replaced by a multivariate gamma distribution with joint densities estimated using copulas.

Note that the t-statistic-based Primo – Primo(t) – and the P-value-based Primo – Primo(P) – may produce slightly different results due to different estimation algorithms of the D_k's. Primo(t) requires both effect sizes and standard errors as input. When those

statistics are not available, or when F-tests or other second-order tests are used in association analysis, or when one-sided tests are preferred if a same direction of association effects are expected for biological reasons, then users may instead use Primo(P).

2.2.6 Extensions of Primo when J is large

When jointly analyzing a large number of sets of association summary statistics, the number of possible joint association patterns $K = 2^J$ increases exponentially with the number of sets of statistics, J. When $J = 15$, there are 32,768 possible association patterns and the calculation for all K patterns can be computationally expensive. One may reduce the number of patterns under consideration to only the major and interpretable patterns [Urbut et al., 2019]. However, the selection of major and interpretable patterns is still a challenge. Additional work is still needed in future research. When analyzing a large number of sets of association statistics of similar types, one possible strategy is to group sets of statistics into major and relatively independent groups $g = 1, \ldots, G$, each with $J_g < 10$ sets of statistics. Then one can apply Primo to calculate the posterior probabilities within each group and take the products of the probabilities between groups to obtain the overall probabilities for all groups in the association patterns of interest. For example, the posterior probability of a SNP being associated with at least 1 (omics) trait in G groups of studies is given by

$$P = 1 - \prod_{g=1}^{G} \Pr(\text{the SNP is not associated with any trait in group } g),$$

where the probability of the SNP being not associated with any trait in group g can be calculated by separately applying Primo to the low-dimensional J_g set of statistics within the g-th group.

When jointly analyzing unbalanced numbers of summary statistics of different data types (e.g., 10 sets of statistics with gene expression as the outcome and 1 set where the outcome is protein abundance), caution should be taken as the joint association results can be dominated

by one data type (here, gene expression), which is not ideal. One may first collapse those J sets of statistics by data types, and apply Primo in a hierarchical fashion to the (converted) summary statistics from multiple data types. This direction will be explored in future work.

2.3 Simulations

We evaluated the performance of Primo in a variety of simulated scenarios. In each scenario, we simulated the test statistics for associations of observations with J traits. Test statistics under the null hypothesis of no association were simulated from a standard normal distribution; test statistics under the alternative were simulated from a normal distribution with mean 0 and standard deviation of 10 (allowing effect sizes to be positive or negative). For each simulated dataset, we ran two versions of the Primo algorithm using t-statistics and P-values as input, denoted as Primo (t) and Primo (P), respectively. We repeated each simulation 100 times and compared the performance of the two versions of Primo.

2.3.1 Accurate estimation of proportions (π) even for very sparse joint associations

It is known to be challenging to estimate π_k's when associations are sparse, i.e., π_k's being very close to zero for patterns with associations. In Scenarios 1a and 1b, we showed that in analyzing independent and correlated sets of summary statistics, respectively, Primo can well-estimate the π_k's despite very sparse associations. In each scenario, we simulated test statistics for $J = 3$ traits for 10 million observations, first under independence and then with pairwise (Pearson) correlation of 0.3 between each set of statistics. In Scenario 1a, we simulated true $\pi_k = (7 \times 10^{-4},\ 2 \times 10^{-4},\ 1 \times 10^{-4})$ for observations being associated with only one, exactly two, and all three traits, respectively. Scenario 1b simulated even sparser associations for the third trait, with $\pi_k = (7 \times 10^{-6},\ 2 \times 10^{-6},\ 1 \times 10^{-6})$ for observations being associated with only the third, the third and first or second, and all

three traits, respectively. Table 2.1(A) shows true π_k's, and the average estimates for π_k's by Primo based on t-statistics or P-values. In Table 2.1(B), we also show the performance of estimation of π_k's when the marginal alternative proportions θ_j^1's are mis-specified. As shown, Primo estimates the π_k's with reasonable accuracy even when the associations are very sparse and when the marginal alternative proportions θ_j^1's are under-specified.

Table 2.1: **Performance of Primo in estimating proportions of association patterns, $\hat{\pi}$.** Average estimates of $\hat{\pi}$ shown over 100 simulations. Scenario 1a simulates sparse associations for $J = 3$ traits. Scenario 1b simulates even sparser associations for the third trait. (A) When θ_j^1 are correctly specified. (B) When θ_j^1 are under-specified ($\theta_j^1/10$) or over-specified ($\theta_j^1 \times 10$).

(A)

Scenario	Method	π_k(%)							
		q_k=(0 0 0)	(1 0 0)	(0 1 0)	(0 0 1)	(1 1 0)	(1 0 1)	(0 1 1)	(1 1 1)
1a		Independent							
	True	99.720	0.070	0.070	0.070	0.020	0.020	0.020	0.010
	Primo (t)	99.714	0.075	0.075	0.075	0.017	0.017	0.017	0.009
	Primo (P)	99.742	0.069	0.069	0.069	0.015	0.015	0.015	0.007
		Correlated							
	True	99.720	0.070	0.070	0.070	0.020	0.020	0.020	0.010
	Primo (t)	99.718	0.073	0.073	0.073	0.018	0.018	0.018	0.009
	Primo (P)	99.746	0.067	0.067	0.067	0.016	0.016	0.016	0.007
1b		Independent							
	True	99.840	0.070	0.070	0.0007	0.020	0.0002	0.0002	0.0001
	Primo (t)	99.830	0.073	0.073	0.0067	0.017	0.0003	0.0003	0.0001
	Primo (P)	99.850	0.066	0.066	0.0043	0.014	0.0002	0.0002	0.0001
		Correlated							
	True	99.840	0.070	0.070	0.0007	0.020	0.0002	0.0002	0.0001
	Primo (t)	99.834	0.072	0.072	0.0054	0.017	0.0003	0.0003	0.0001
	Primo (P)	99.853	0.064	0.064	0.0039	0.015	0.0003	0.0003	0.0001

(B)

Scenario	Specific.	Method	π_k(%)							
			q_k=(0 0 0)	(1 0 0)	(0 1 0)	(0 0 1)	(1 1 0)	(1 0 1)	(0 1 1)	(1 1 1)
1a			Independent							
		True	99.720	0.070	0.070	0.070	0.020	0.020	0.020	0.010
	Under	Primo (t)	99.767	0.061	0.061	0.061	0.015	0.015	0.015	0.006
		Primo (P)	99.822	0.049	0.049	0.049	0.009	0.009	0.009	0.002
	Over	Primo (t)	99.648	0.091	0.091	0.091	0.022	0.022	0.022	0.014
		Primo (P)	99.424	0.158	0.158	0.157	0.028	0.028	0.028	0.018
			Correlated							
		True	99.720	0.070	0.070	0.070	0.020	0.020	0.020	0.010
	Under	Primo (t)	99.766	0.061	0.061	0.061	0.015	0.015	0.015	0.007
		Primo (P)	99.823	0.048	0.048	0.048	0.010	0.010	0.010	0.003
	Over	Primo (t)	99.648	0.091	0.091	0.091	0.022	0.022	0.022	0.014
		Primo (P)	99.436	0.147	0.147	0.146	0.035	0.035	0.035	0.021
1b			Independent							
		True	99.840	0.070	0.070	0.0007	0.020	0.0002	0.0002	0.0001
	Under	Primo (t)	99.868	0.058	0.058	0.0024	0.014	0.0001	0.0002	0.0001
		Primo (P)	99.903	0.044	0.044	0.0004	0.008	0.0001	0.0001	0.0001
	Over	Primo (t)	99.785	0.091	0.091	0.0091	0.023	0.0004	0.0004	0.0002
		Primo (P)	99.634	0.160	0.160	0.0147	0.030	0.0005	0.0005	0.0002
			Correlated							
		True	99.840	0.070	0.070	0.0007	0.020	0.0002	0.0002	0.0001
	Under	Primo (t)	99.867	0.058	0.058	0.0021	0.014	0.0002	0.0002	0.0001
		Primo (P)	99.903	0.044	0.044	0.0004	0.008	0.0001	0.0001	0.0001
	Over	Primo (t)	99.779	0.093	0.093	0.0114	0.022	0.0005	0.0005	0.0002
		Primo (P)	99.640	0.153	0.153	0.0133	0.034	0.0019	0.0019	0.0020

2.3.2 Numerical optimization simulation for alternative distributions of $-2\log(P)$-values

We evaluated the performance of Primo in estimating the scaling factor (A_j) and degrees of freedom (d'_j) parameters of the alternative distribution for $t_{ij} = -2\log(p_{ij})$ $(1, \ldots, m)$. Under different specifications of the proportion of statistics coming from the alternative distribution (θ_j^1), we simulated 10 million test statistics. Test statistics under the null hypothesis of no association were simulated from a χ_2^2 distribution; test statistics under the alternative were simulated from a $A_j\chi^2_{d'_j}$ distribution. Figure 2.5 compares the density curves of the true alternative density to the alternative densities estimated by Primo over 1000 simulations for $A_j = 4.5$ and $d'_j = 7$. As shown, the density curves estimated by Primo reasonably approximate the true density curve even when the proportion of statistics coming from the alternative distribution is sparse ($\theta_j^1 = 1 \times 10^{-4}$; panel 2.5A) or very sparse ($\theta_j^1 = 1 \times 10^{-5}$; panel 2.5B).

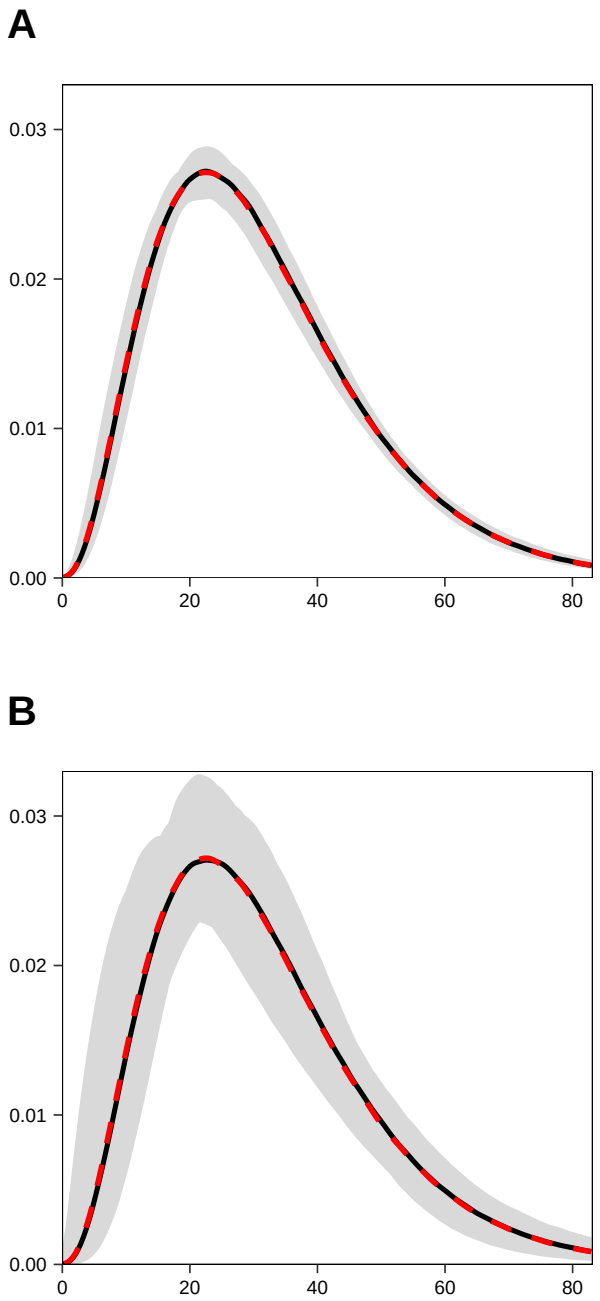


Figure 2.5: **Performance of Primo in estimating parameters for alternative distribution of $-2\log(P)$ or χ^2 statistics.** For $A_j = 4.5$ and $d'_j = 7$, the true density curve is shown by the dotted red curve in A and B. For $\theta_j^1 = 1 \times 10^{-4}$ (A) and $\theta_j^1 = 1 \times 10^{-5}$ (B), the density curves estimated by the median estimates of the parameters over 1000 simulations are given by the black curve. The shaded gray area shows the curves of the parameters between the $5th$ and $95th$ percentiles. Primo reasonably estimates the scaling factor (A_j) and degrees of freedom (d'_j) for the alternative distribution, even when associations are sparse.

2.3.3 The performance of Primo in jointly analyzing associations to multiple traits

In Scenario 2, we simulated correlated test statistics with pairwise correlations of 0.3 among $J = 3$ traits for 1 million observations. $\pi_k = 1 \times 10^{-3}$, 5×10^{-4}, 5×10^{-4} for the patterns where observations are associated with only one, exactly two, and all of the three traits, respectively. Here we compared the observed and estimated false discovery rates (FDRs) and power to detect associations to all three traits and to at least one trait, based on Primo versus Fisher's method [Fisher, 1918]. The results with correctly specified, under-specified and over-specified marginal non-null proportions (θ_j^1's) are shown in Table 2.2. When θ_j^1's are well-specified (Scenario 2a in Table 2.2), Primo nicely controlled the FDR even in the presence of study/sample correlations – highlighting one advantage of Primo in integrating potentially correlated multi-omics data. As shown in Table 2.2, the estimated FDR (estFDR) is very close to the observed FDR for Primo. Fisher's method, as a combination method for testing omnibus hypotheses, can only be used to detect observations with associations to at least one trait, and is not applicable to detect associations to all traits. The FDR [Storey and Tibshirani, 2003; Storey et al., 2015] for Fisher's method calculated from the nominal P-values are not well controlled due to correlations among test statistics, as expected. At similar power levels, the observed FDRs of Fisher's method are also much higher than those of Primo.

In this simulation, the true θ_j^1's are 2.5×10^{-3}. In Scenario 2b, we under-specified θ_j^1 to be 2.5×10^{-4}. As shown in Table 2.2, although power might decrease to some extent, the FDRs are reasonably controlled. In Scenario 2c, when θ_j^1's are over-specified as 2.5×10^{-2}, we observed slightly inflated FDRs. As such, we suggest to obtain reasonable estimates for θ_j^1's based on the current data and the literature, or under-specify θ_j^1's to be more conservative.

Table 2.2: **Simulation results evaluating the performance of Primo.** When $J = 3$ with correlated samples, we compared Primo versus Fisher's method in detecting associations to at least 1 trait and assessed Primo's performance in detecting associations to all traits. We also assess the performance of Primo when parameters are correctly, under- and over-specified. PP := posterior probability; estFDR := estimated FDR.

Scenario	Method	Association to at least one trait						Association to all three traits					
		PP $\geq$ 0.90			PP $\geq$ 0.80			PP $\geq$ 0.90			PP $\geq$ 0.80		
		true FDR (%)	estFDR (%)	Power (%)	true FDR (%)	estFDR (%)	Power (%)	true FDR (%)	estFDR (%)	Power (%)	true FDR (%)	estFDR (%)	Power (%)
2a	Primo (t)	0.1	0.2	75.0	0.3	0.4	75.9	0.8	0.8	46.0	2.0	2.0	49.6
	Primo (P)	0.2	0.3	73.8	0.4	0.6	74.8	0.7	0.8	42.8	1.5	1.8	45.9
		5% estFDR			10% estFDR								
	Fisher's	23.5	-	77.0	36.0	-	78.4	-	-	-	-	-	-
2b	Primo (t)	0.1	0.2	73.9	0.1	0.3	74.8	0.5	0.8	43.2	1.0	1.7	46.2
	Primo (P)	0.1	0.1	65.9	0.1	0.1	66.6	0.1	0.1	28.8	0.1	0.1	29.9
2c	Primo (t)	0.6	0.2	76.6	1.6	0.6	77.8	6.1	1.3	47.8	13.2	3.7	53.3
	Primo (P)	1.6	0.8	74.9	5.3	3.2	76.9	0.4	4.6	40.4	2.7	7.4	47.1

In Scenario 3, we simulated correlated test statistics for associations to five traits for 1 million observations with pairwise study-study correlations of 0.3. $\pi_k = 5 \times 10^{-4}$, 2×10^{-4}, 1×10^{-4} for the patterns where observations are associated with one to two, three to four, and all of the five traits, respectively. Results are presented in Table 2.3.

Table 2.3: **Simulation results evaluating the performance of Primo for integrating summary statistics from 5 studies/traits.** When $J = 5$ with correlated samples, we evaluated the performance of Primo. PP := posterior probability; estFDR := estimated FDR.

Scenario 3	Association to at least one trait						Association to at least three traits						Association to all five traits					
	PP $\geq$ 0.90			PP $\geq$ 0.80			PP $\geq$ 0.90			PP $\geq$ 0.80			PP $\geq$ 0.90			PP $\geq$ 0.80		
	true FDR (%)	estFDR (%)	Power (%)	true FDR (%)	estFDR (%)	Power (%)	true FDR (%)	estFDR (%)	Power (%)	true FDR (%)	estFDR (%)	Power (%)	true FDR (%)	estFDR (%)	Power (%)	true FDR (%)	estFDR (%)	Power (%)
Primo (t)	0.1	0.1	77.4	0.2	0.3	78.2	0.1	0.7	68.6	0.2	1.6	73.2	0.2	1.4	36.7	0.7	3.0	41.7
Primo (P)	0.1	0.2	75.9	0.3	0.4	76.7	0.1	0.5	62.5	0.3	1.1	66.3	0.3	1.0	30.2	0.9	2.3	33.9

Overall, Primo yields good control of FDRs and high power in detecting various patterns of joint associations, even for a moderately large number of sets of summary statistics and in the presence of study correlations.

2.4 Data applications

2.4.1 Effects of DNA copy number alterations (CNA) on cis- gene expression and protein abundance in tumors from multiple cancer types

The Cancer Genome Atlas (TCGA)

To analyze the effects of DNA copy number alterations (CNA) on cis-gene expression and global cis-protein abundance in tumors from multiple cancer types, we utilized data from The Cancer Genome Atlas (TCGA). TCGA is a multi-center project that was established to catalog genomic mutations and molecular changes that occur in cancer [Hutter and Zenklusen, 2018]. The project collected data from more than 11,000 cases and over 30 tumor types [Hutter and Zenklusen, 2018]. Data collected by TCGA includes single nucleotide polymorphism (SNP) genotyping, exome sequencing, DNA copy number variation measurements, RNA sequencing, DNA methylation, microRNA sequencing, etc.

A subset of TCGA subjects were included in The Clinical Proteomic Tumor Analysis Consortium (CPTAC) program, which applied standarized proteome analysis platforms in order to analyze protein abundance in tumor tissues [Ellis et al., 2013]. Among the stated goals of CPTAC are to study variant protein sequences corresponding to somatic mutations and evaluate relationships between mutation frequency and variant protein expression, and to determine how copy number variation translates into differences in protein expression [Ellis et al., 2013].

Data for TCGA breast, ovarian and colorectal cancer subjects was downloaded through the Genomic Data Commons (GDC) data portal [Grossman et al., 2016], and downloaded and processed using TCGA-Assembler2 [Wei et al., 2018]. RNA sequencing, used to measure gene expression levels, was performed in tumor tissues using the Illumina HiSeq 2000 RNA Sequencing platform. RNA-Seq expression levels were quantified using HTSeq-count [An-

ders et al., 2015]. Protein abundance was measured in tumor tissue samples using iTRAQ (isobaric tag for relative and absolute quantitation) mass-spectometry (MS) in experiments conducted by the Clinical Proteomic Tumor Analysis Consortium (CPTAC) [Ellis et al., 2013]. The protein abundance was measured using the Log Ratio (i.e. the log of the ratio between the spectral count of a protein in a sample versus the spectral count of the protein in the reference sample). Prior to analysis, expression and protein measurements were transformed to the quantiles of the standard normal distribution (separately for each gene or protein).

TCGA DNA copy number measurements used Affymetrix SNP 6.0 array data to infer the copy number of repeats of genomic regions using circular binary segmentation analysis [Olshen et al., 2004]. Copy number values were transformed into segment mean values as $\log_2(\text{copyNumber}/2)$. We used functions in TCGA-Assembler2 to summarize copy number values at the gene-level, centered on 0 [Wei et al., 2018].

Clinical covariates were downloaded through the GDC using the TCGAbiolinks package in R [Colaprico et al., 2016]. We generated genotype PCs using autosomal bi-allelic variants on the Affymetrix SNP 6.0 array with the following filters in PLINK 1.90 [Chang et al., 2015; Purcell and Chang, 2017]: minor allele frequency ≥ 0.05; Hardy-Weinberg Equilibrium p-value ≥ 0.0001; pairwise linkage disequilibrium $R^2 \leq 0.2$. To account for systematic variation in the outcomes (i.e. gene expression and protein abundance) that were not captured by measured covariates, we generated surrogate variables using the sva package in R [Leek and Storey, 2007; Leek et al., 2019].

Summary statistics from gene-level linear regressions

Consider a study of gene-level CNA of m genes in a single cancer type. We are interested in the effects of CNA on cis-gene expression (mRNA) levels and (global) cis-protein abundance in tumor samples. Let n_e subjects have measurements of both mRNA and CNA; and n_p subjects have measurements of both protein and CNA. We use the following regression

models to assess the effects of CNA on the molecular traits:

$$\begin{aligned} \text{cis-mRNA}_{ui} &= \alpha_{0i} + \alpha_{1i}\text{CNA}_{ui} + \alpha_{2i}^T\text{cov}_u + \epsilon_{ui}, \\ \text{cis-protein}_{vi} &= \beta_{0i} + \beta_{1i}\text{CNA}_{vi} + \beta_{2i}^T\text{cov}_v + \epsilon_{vi}, \end{aligned} \tag{2.4}$$

where $u \in \{1, ..., n_e\}$ and $v \in \{1, ..., n_p\}$ are subject indices; $i \in \{1, ..., m\}$ are gene indices; the cov's are sets of covariates adjusted for in the regression analyses; and the ϵ's are error terms. Of primary interest are the coefficients for CNA, α_{1i} and β_{1i}. By modeling each molecular trait separately in (2.4), we avoid restricting the analysis to subjects having complete measurements of all molecular traits, which may limit power for traits with much larger sample sizes (here, gene expression). Integrative analysis using these summary statistics can then leverage the potential shared correlations between the molecular traits.

In addition to jointly analyzing the effects of CNA on multiple molecular traits (i.e. gene expression and protein abundance) within one cancer type, we are interested in jointly analyzing the effects of CNA across multiple cancer types. For each gene $i \in \{1, ..., m\}$, we use the following regression models to assess the effects of CNA within molecular traits and within cancer types:

$$\begin{aligned} \text{Breast} \begin{cases} \text{cis-mRNA}_{ui} &= \alpha_{b0i} + \alpha_{b1i}\text{CNA}_{ui} + \alpha_{b2i}^T\text{cov}_u + \epsilon_{ui}, \\ \text{cis-protein}_{vi} &= \beta_{b0i} + \beta_{b1i}\text{CNA}_{vi} + \beta_{b2i}^T\text{cov}_v + \epsilon_{vi}, \end{cases} \\ \text{Ovary} \begin{cases} \text{cis-mRNA}_{wi} &= \alpha_{o0i} + \alpha_{o1i}\text{CNA}_{wi} + \alpha_{o2i}^T\text{cov}_w + \epsilon_{wi}, \\ \text{cis-protein}_{zi} &= \beta_{o0i} + \beta_{o1i}\text{CNA}_{zi} + \beta_{o2i}^T\text{cov}_z + \epsilon_{zi}, \end{cases} \\ \text{Colorectal} \begin{cases} \text{cis-mRNA}_{si} &= \alpha_{c0i} + \alpha_{c1i}\text{CNA}_{si} + \alpha_{c2i}^T\text{cov}_s + \epsilon_{si}, \\ \text{cis-protein}_{ti} &= \beta_{c0i} + \beta_{c1i}\text{CNA}_{ti} + \beta_{c2i}^T\text{cov}_t + \epsilon_{ti}, \end{cases} \end{aligned} \tag{2.5}$$

where u, v, w, z, s and t are subject indices; and other terms have similar interpretations as in the regressions described in Models (2.4). Of primary interest are the coefficients of

the effects of CNA on the molecular traits (i.e. the $\alpha_{.1i}$'s and $\beta_{.1i}$'s). We repeat the set of regressions for each gene $i \in \{1, ..., m\}$ to obtain genome-wide sets of summary statistics.

In the current analyses, all regressions were adjusted for genotype principal components (PCs) to account for population stratification (the first 5 PCs were used for all analyses except for protein abundance in colorectal tumors, which used the first 3 PCs because of the small sample size). All regressions also included the following covariates: age, tumor purity [Aran et al., 2015], cancer stage (dichotomized as: stage III/IV or lower than stage III), and up to 50 surrogate variables. Analyses of breast tumors were restricted to female subjects and additionally adjusted for histological subtype (infiltrating ductal, infiltrating lobular, mucinous, metaplastic, mixed histology or other), and estrogen receptor (ER) and progesterone receptor (PR) status (+/-). Analyses of ovary tumors were also adjusted for tumor grade (dichotomized as: grade 3/4 or lower than grade 3). Analyses of colorectal tumors were also adjusted for gender and primary tumor site (colon or rectum). The number of surrogate variables (SVs) used in the regressions depended on sample size: 50 SVs for sample size ≥ 300; 10 SVs for sample size ≥ 50 and < 300; 5 for sample size < 50.

After performing regressions for each molecular trait in each cancer type for each gene, we conducted integrative analyses to detect joint associations of the effects of CNA.

Integrative analysis

We are interested in detecting joint associations of the effects of CNA of m genes on J types of molecular traits. More specifically, we are interested in assessing the effects of gene-level CNA on cis-gene expression and cis-protein abundance in breast tumors, on cis-gene expression and cis-protein abundance in ovary tumors, and on cis-gene expression and cis-protein abundance in colorectal tumors (thus, $J = 6$). We analyzed $m = 5,167$ genes having CNA, gene expression and protein abundance measurements for for all three cancer types. Let T denote the $m \times J$ matrix of t-statistics of CNA effects obtained by the regression equations in (2.5). These regressions used the tumor samples having CNA and all covariates measured

in the respective models: 1002 and 74 samples for gene expression and protein abundance, respectively, in breast; 301 and 121 for gene expression and protein abundance, respectively, in ovary; and 357 and 39 for gene expression and protein abundance, respectively, in colorectal. To obtain the probability for each association pattern for each gene, we apply Primo to T as described in Sections 2.2.2 and 2.2.3. We used $\theta_j^1 = 0.1$ as the alternative proportion specified for each trait j, which should be a conservative (i.e. under) estimate given the ubiquity of cis-effects of CNA on both expression levels and protein abundance in tumors [Zhang et al., 2016a; Mertins et al., 2016].

After applying Primo, we identified genes having a posterior probability $> 75\%$ and estimated false discovery rate $< 10\%$ in collapsed patterns of interest. We are primarily interested in genes whose CNAs are associated with both cis-gene expression and cis-protein abundance in a given tumor type. There were 3763, 3233, and 241 genes whose CNAs were associated with both omics traits in at least 1, 2, or 3 tumor types, respectively (estimated FDR of 1.5, 4.0 and 9.9%). Thus, a high proportion of gene-level CNAs demonstrated cis-associations with multiple molecular traits (gene expression and protein abundance) in multiple tumor types.

We performed gene set enrichment analysis (GSEA) of the 241 genes whose CNAs were associated with both cis-gene expression and cis-protein abundance in all three cancer types using the hallmark gene set collection curated by the Molecular Signatures Database (MSigDB) v7.0 [Liberzon et al., 2015; Subramanian et al., 2005]. Table 2.4 shows the gene sets in which these 241 genes demonstrated enrichment (FDR $< 10\%$). The 241 genes showed enrichment in several sets of well-defined biological states or processes that are relevant to cancer, including fatty acid metabolism [Koundouros and Poulogiannis, 2019], oxidative phosphorylation [Ashton et al., 2018], genes regulated by MYC [Dang, 2012], and genes encoding cell cycle related targets of E2F transcription factors [Kent and Leone, 2019].

The gene whose CNAs have the highest probability of associations with both cis-gene expression and cis-protein abundance in all three tumor types (posterior probability $> 99.99\%$)

Table 2.4: **Gene set enrichment analysis (GSEA) results of genes whose CNAs were associated with all cis-omics traits in tumors.** The table present gene sets showing enrichment at false discovery rate (FDR) < 10% for 241 genes whose CNAs were associated with all 6 cis-omics traits analyzed (gene expression and protein abundance in breast, ovary, and colorectal tumors).

Gene Set Name	Description	(# Genes) / (#Genes in Set)	p-value	FDR q-value
FATTY ACID METABOLISM	Genes encoding proteins involved in metabolism of fatty acids	13 / 158	3.1×10^{-11}	1.1×10^{-9}
OXIDATIVE PHOSPHORYLATION	Genes encoding proteins involved in oxidative phosphorylation	14 / 200	4.7×10^{-11}	1.1×10^{-9}
MYC TARGETS V2	A subgroup of genes regulated by MYC - version 2 (v2)	7 / 58	8.5×10^{-8}	1.4×10^{-6}
PEROXISOME	Genes encoding components of peroxisome	8 / 104	3.5×10^{-7}	4.3×10^{-6}
PROTEIN SECRETION	Genes involved in protein secretion pathway	7 / 96	2.7×10^{-6}	2.7×10^{-5}
E2F TARGETS	Genes encoding cell cycle related targets of E2F transcription factors	9 / 200	5.8×10^{-6}	4.1×10^{-5}
MYC TARGETS V1	A subgroup of genes regulated by MYC - version 1 (v1)	9 / 200	5.8×10^{-6}	4.1×10^{-5}
ESTROGEN RESPONSE EARLY	Genes defining early response to estrogen	8 / 200	4.4×10^{-5}	2.7×10^{-4}
BILE ACID METABOLISM	Genes involve in metabolism of bile acids and salts	6 / 112	8.2×10^{-5}	4.6×10^{-4}
CHOLESTEROL HOMEOSTASIS	Genes involved in cholesterol homeostasis	5 / 74	1.1×10^{-4}	5.5×10^{-4}

is *LARP4B*. Evidence suggests that *LARP4B* serves as a tumor-suppressor in glioma [Koso et al., 2016], is downregulated in prostate cancer [Yin et al., 2019], and is correlated with survival status in liver cancer [Li et al., 2019]. And the La-related protein (LARP) family has recently been suggested as targets for cancer therapy [Stavraka and Blagden, 2015]. Analysis of DNA copy number alterations of TCGA glioblastoma samples revealed *LARP4B* to be one of only four genes with heterozygous deletion in over 80% of glioblastoma samples [Koso et al., 2016]. The present integrative analysis provides evidence of the cis-effects of CNAs of this gene in three additional cancer types, demonstrating the importance of integrative analysis of related diseases.

The integrative analysis of effects of CNAs on cis-omics traits in multiple tumor types demonstrated that key driver mutations are frequently associated with multiple cis-omics traits (i.e. gene expression and protein abundance) in multiple tumor types (i.e. breast, ovarian, and/or colorectal). These findings are consistent with previous research evaluating CNA effects on gene expression and protein abundance within a single cancer type [Zhang et al., 2016a; Mertins et al., 2016; Zhang et al., 2014] and pan-cancer analysis of CNA effects on gene expression [Shao et al., 2019].

Possibly due to limited sample size (39) in the analysis of colorectal protein abundance, we observed a significant decrease from the number of genes associated with both cis-omics traits in at least 2 tumor types (3233) to the number of genes associated with both cis-omics

traits in all three tumor types (241), even though the correlation between gene expression and protein abundance in colorectal tumors was often high (with a mean Spearman correlation of 0.47 [Zhang et al., 2014]). However, this observation highlights the potential for integrative analysis methods, such as Primo, to boost power by borrowing information across studies. A univariate analysis of the effects of CNA on protein abundance in colorectal tumors would have identified 3 genes at the Bonferroni threshold of $p < (0.05/5167)$, 100 genes at FDR q-value < 0.05, and 215 genes at q-value < 0.10 [Storey and Tibshirani, 2003]. Thus, the 241 genes identified as being associated with both cis-omics traits in all three tumor types actually exceeds the number of genes that would have been identified by univariate analysis of CNA effects on colorectal protein abundance after multiple testing adjustment. Since the 241 genes associated with both cis-omics traits in all three tumor types include 52 genes not included in the 215 genes with q-value < 0.10 in the colorectal protein abundance analysis, we also observe that jointly analyzing multiple traits in integrative analysis can reveal additional insights beyond just taking the intersection of identified results from univariate analyses.

2.4.2 Trans-omics effects of DNA copy number alterations (CNAs)

As noted in the introduction, a challenge in understanding effects of CNAs in the context of cancer is distinguishing driver events, which may be causally involved in tumorigenesis, from passenger mutations [Zack et al., 2013]. In the previous section, we performed integrative analyses of *cis*-omics effects of CNAs across multiple molecular traits and multiple cancer types to detect joint associations. In addition, it is well known that CNAs affect distal gene expression levels and protein abundances [Srihari et al., 2016; Mertins et al., 2016; Zhang et al., 2014, 2016a]. And CNAs with strong trans effects are "more likely to elicit a molecular phenotype and confer selective advantages" for cancerous cells in tumors [Zhang et al., 2016a]. Thus, identifying CNA "hubs" that affect multiple trans-genes helps to identify driver mutations. Analyzing trans effects may also be important for studying cancer related phenotypes besides tumor initiation. For example, an analysis of ovarian cancer found that

the four CNAs that affected the largest number of trans proteins were all highly associated with survival [Zhang et al., 2016a]. Integrating studies of CNA trans-effects across multiple cancer types may further highlight important genes with pleiotropic effects on related diseases.

Summary statistics from linear regressions of trans-gene cis-CNA pairs

For the analyses of trans-omics effects of DNA copy number alterations (CNAs), we use the same TCGA/CPTAC data described in Section 2.4.1. We define "trans" as the genes on a different chromosome from the cis-gene whose CNA is being evaluated. For each gene $i \in \{1, ..., m\}$, we wish to assess the association between CNA of gene i, and gene expression and protein abundance of each of its m_i' trans-genes (in multiple cancer types). For each gene $i \in \{1, ..., m\}$, we use the following regression models to assess the effects of CNA within trans molecular traits and within cancer types:

$$\begin{aligned}
\text{Breast} \begin{cases} \text{trans-mRNA}_{ui'} &= \alpha_{b0i'} + \alpha_{b1i'}\text{CNA}_{ui} + \alpha_{b2i'}^T\text{cov}_u + \epsilon_{ui'}, \\ \text{trans-protein}_{vi'} &= \beta_{b0i'} + \beta_{b1i'}\text{CNA}_{vi} + \beta_{b2i'}^T\text{cov}_v + \epsilon_{vi'}, \end{cases} \\
\text{Ovary} \begin{cases} \text{trans-mRNA}_{wi'} &= \alpha_{o0i'} + \alpha_{o1i'}\text{CNA}_{wi} + \alpha_{o2i'}^T\text{cov}_w + \epsilon_{wi'}, \\ \text{trans-protein}_{zi'} &= \beta_{o0i'} + \beta_{o1i'}\text{CNA}_{zi} + \beta_{o2i'}^T\text{cov}_z + \epsilon_{zi'}, \end{cases} \\
\text{Colorectal} \begin{cases} \text{trans-mRNA}_{si'} &= \alpha_{c0i'} + \alpha_{c1i'}\text{CNA}_{si} + \alpha_{c2i'}^T\text{cov}_s + \epsilon_{si'}, \\ \text{trans-protein}_{ti'} &= \beta_{c0i'} + \beta_{c1i'}\text{CNA}_{ti} + \beta_{c2i'}^T\text{cov}_t + \epsilon_{ti'}, \end{cases}
\end{aligned} \tag{2.6}$$

where u, v, w, z, s and t are subject indices; $i' \in \{1, ..., m_i'\}$ is the trans gene index; the cov's are sets of covariates adjusted for in the regression analyses; and the ϵ's are error terms.. Of primary interest are the coefficients of the effects of CNA on the trans molecular traits (i.e. the $\alpha_{\cdot 1i'}$'s and $\beta_{\cdot 1i'}$'s). For a given gene i, we repeat the set of regressions for each of its trans genes $i' \in \{1, ..., m_i'\}$. Then, we repeat the set of regressions for each gene $i \in \{1, ..., m\}$ to

obtain genome-wide sets of summary statistics.

Note that the models in 2.6 are similar to the models in 2.5. The primary difference is that for the CNAs for each gene i, the models in 2.5 assess only the expression of cis-gene i as an outcome whereas the models in 2.6 assess the effect of i's CNAs on *each* trans-gene for i. Thus, there are many more associations tested in 2.6 (25,277,071 sets of the six models in the presented analysis, since each of the 5,167 genes has around five thousand trans-genes) than are tested in 2.5 (5,167 sets of the six models in the presented analysis) since 2.5 has a one-to-one mapping of CNAs to cis-gene expression or cis-protein abundance.

Integrative analysis

We are interested in detecting joint associations of the trans-effects of gene-level CNAs of $m = 5,167$ genes on $J = 6$ types of molecular traits (i.e. trans-gene expression and trans-protein abundance in each of breast, ovary and colorectal tumors). The total number of cis-CNA trans-gene pairs that was analyzed was $h = \sum_{i=1}^{m} m_i' = 25,277,071$. Let T denote the $h \times J$ matrix of t-statistics of CNA effects obtained by the regression equations in (2.6). To obtain the probability for each association pattern for each gene, we apply Primo to T as described in Sections 2.2.2 and 2.2.3. We used $\theta_j^1 = 10^{-4}$ as the alternative proportion specified for each trait j.

Of primary interest is identifying CNAs having trans-omics effects in more than one cancer type. Because the proteomic datasets have low sample sizes for trans analyses, we focus on identifying joint associations of CNAs with trans-gene expression. It is known that breast cancer and ovarian cancer share several commonalities in risk factors and etiology. These include hereditary mutations (e.g. mutations in *BRCA1/BRCA2* [Petrucelli et al., 2010] or the susceptibility locus at 19p13 [Lawrenson et al., 2016]) as well as acquired mutations (e.g. in the *PIK3CA* gene region [Campbell et al., 2004]). For each cis-CNA trans-gene pair in the Primo analysis, we obtained the probability of being "at least associated with gene expression in both breast and ovary tumors" by summing the probabilities over the

patterns where the summary statistics for breast gene expression and ovary gene expression both come from their alternative distributions. That is, we sum the probabilities from each pattern k such that $q_{kj_{be}} = 1$ for the column j_{be} corresponding to gene expression in breast tumors and $q_{kj_{oe}} = 1$ for the column j_{oe} corresponding to gene expression in ovary tumors.

At the posterior probability threshold of 78.1% (FDR $<$ 10%), there were 2101 cis-CNA trans-gene pairs showing association with gene expression in both breast and ovary tumors. Figure 2.6 connects each cis-CNA and trans-gene pair by chromosome and position for the 2101 identified pairs. The figure shows that trans-effects of CNAs on expression occur throughout the genome and reveals that the CNAs of genes on one chromosome may affect expression of genes on many other chromosomes. The CNAs of 428 genes were associated with expression of multiple trans-genes, including the CNAs of 27 cancer-associated genes reported in the Catalogue of Somatic Mutations in Cancer (COSMIC) database [Forbes et al., 2016]. The gene whose CNAs were associated with the most trans associations (7) is *CTCF*, a tumor suppressor gene that represses binding to promoters of *MYC* [Gombert and Krumm, 2009]. *CTCF* is one of the most frequently mutated genes in endometrial, breast, and head and neck cancers [Lawrence et al., 2014].

Next, we sought to identify CNAs having joint associations with trans-gene expression levels in breast, ovarian and colorectal tumors. By summing over probabilities of the patterns where the summary statistics for breast gene expression, ovary gene expression and colorectal gene expression all come from their alternative distributions, for each cis-CNA trans-gene pair we obtained the probability of being "at least associated with gene expression in breast, ovarian, and colorectal tumors". At the posterior probability threshold of 75% (FDR of 8.8%), there were 368 cis-CNA trans-gene pairs showing association with gene expression in breast, ovarian, and colorectal tumors. The CNAs of 15 genes were associated with expression of multiple trans-genes, including the CNAs of 2 cancer-associated genes reported in COSMIC [Forbes et al., 2016]: *CTCF* and *CDH1*. *CDH1* is known to be involved in tumorigenesis [Semb and Christofori, 1998] as well as tumor progression, invasion

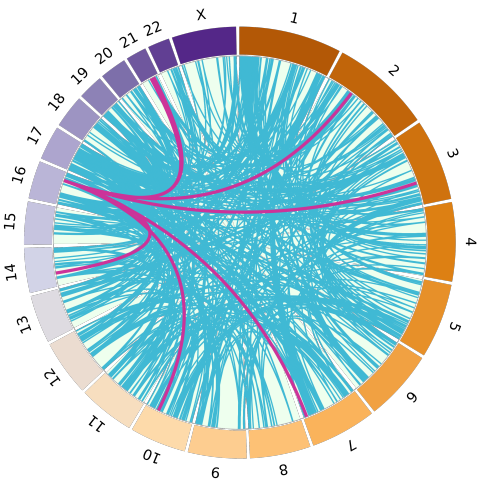


Figure 2.6: **Tumor cis-CNA trans-gene pairs in breast and ovary tumors by chromosome and position.** Each colored line connects a cis-CNA and trans-gene with whose expression it is associated in both breast and ovary tumors as identified by Primo analysis. The tumor suppressor gene *CTCF* was the cis-hub with the most trans-associations (pink lines). CNAs of *CTCF* were associated with expression of 7 trans-genes on 6 chromosomes (two on chr 21).

and metastasis [Hu et al., 2016]. It is downregulated through hypermethylation in ovarian cancer [Wu et al., 2014] and ductal breast cancer [Shargh et al., 2014], is frequently mutated in bladder cancer [Al-Ahmadie et al., 2016], and methylation of its promoter region is associated with overall survival in endometrial cancer [Yi et al., 2011]. Furthermore, germline mutations in *CDH1* pose a significant risk factor for gastric and breast cancers [Hansford et al., 2015].

The results of the trans analyses demonstrate the potential for integrative analyses to highlight important genes whose mutations are relevant to multiple cancer types. And we will further examine this question in Chapter 3 by integrating product of association/mediation test statistics to further map putative causal mediator genes for distal trans-targets.

2.5 Discussion

In this work, we developed a general integrative association approach and computational tool – Primo – for assessing joint associations across studies and/or data types. Primo allows for

unknown study heterogeneity and sample correlation and uses only summary statistics as input. It detects joint associations while allowing different effect sizes (and different directions of effects) on different types of traits and considers multiple testing adjustment by guiding selection of a posterior probability threshold that controls the false discovery rate (FDR). Primo is computationally efficient and capable of integrating the t-statistics, P-values or other second order statistics of a moderate to large number of traits. In simulation studies, we demonstrated that Primo performs well in estimating association pattern proportions, even when associations are sparse, and maintains reasonable power and controls the false discovery rate (FDR) when the proportion of observations coming from the alternative hypothesis is correctly or under-specified.

We applied Primo to jointly analyze the effects of copy number alterations (CNAs) on gene expression and protein abundance in three distinct cancer types (breast, ovarian and colorectal). Our analyses revealed that CNAs are frequently associated with cis-gene expression and cis-protein abundance, and that these associations are frequently shared across cancer types. We also identified a number of associations between CNAs and expression of trans-genes on distal chromosomes. The gene whose CNAs were associated with the most trans-genes was a known cancer gene frequently mutated in multiple cancer types, highlighting the potential for integrative analyses to identify important hub genes. One limitation of the presented trans-association analysis is that it was not always clear which cis-gene/cis-protein in a particular genomic region was the one through which the CNAs were affecting the trans-gene/trans-protein expression/abundance since the gene-level CNAs of genes in a given genomic region are correlated. In Chapter 3, we propose an extension of the Primo framework to integrate mediation test statistics to help identify the putative cis-protein through which CNAs of a genomic region are affecting trans-protein abundance.

There are some additional points to consider for the current work. First, Primo requires the specification of a key parameter for each study, the study-specific alternative proportion (θ_j^1), and the results may suffer from slightly inflated FDR when those parameters are highly

over-specified (e.g. by an order of magnitude). Conversely, when the parameters are under-specified to some extent (e.g. by an order of magnitude), there may not be a significant loss of power. Thus, we suggest the use of stringent (i.e. conservative) estimates of θ_j^1 in analyses. Second, Primo accounts for possible correlation of the test statistics due to sample overlap (i.e. sample overlap causing correlation between test statistics within a row of the matrix of test statistics). However, the present work does not account for possible correlation *between* rows of test statistics (i.e. correlation within a column, as may occur due to linkage disequilibrium if the units of observation are genetic variants, for example). This is a major focus of and motivation for the work presented in Chapter 4. Third, the number of possible association patterns grows exponentially with the number of studies being integrated (i.e. 2^J for J studies). When jointly analyzing a large number of sets of summary statistics, the computation time of Primo to assess all possible association patterns can increase substantially. The current work proposed a quick extension by applying Primo to groups of sets of summary statistics. An alternative strategy would be to resample a subset of observations during each iteration of the EM algorithm that estimates the π vector. This will be explored in future work.

In this work we applied a novel, general integrative analysis method and computational tool – Primo – to detect multi-omics effects of copy number alterations (CNAs) on molecular phenotypes in tumor tissues from multiple tumor types. However, the Primo framework may applied to many other settings and can be extended to a broader class of association, conditional association and mediation analyses. In Chapter 3, we develop a version of Primo to integrate product of coefficient test statistics to identify robust cis-mediated trans-associations that are replicated across multiple studies or conditions. In Chapter 4, we make tailored developments to assess the molecular mechanisms through which GWAS SNPs might affect complex traits by accounting for linkage disequilibrium (LD) when analyzing associations of germline variants. In these and other settings, the integrative analysis performed by Primo help in providing a more robust, comprehensive and precise assessment

of multi-omics associations.

The R package Primo is freely available at: https://github.com/kjgleason/Primo. A tutorial for using the Primo R package is included in Appendix A.

CHAPTER 3

AN INTEGRATIVE PRODUCT OF COEFFICIENTS ASSOCIATION ANALYSIS METHOD TO IDENTIFY DISTAL ASSOCIATIONS SHARED ACROSS CONDITIONS

3.1 Introduction

The effects of genetic variation on molecular phenotypes such as gene expression and protein abundance have been well-established [Aguet et al., 2019; Suhre et al., 2017; Jia and Zhao, 2017]. These include the effects of germline variants on molecular traits – as quantitative trait loci (QTLs) – as well as effects of somatic variation (e.g. copy number alterations) on molecular phenotypes [Shao et al., 2019; Jia and Zhao, 2017]. Most of the identified associations between genetic variants and molecular traits occur in cis [Aguet et al., 2019; Gong et al., 2018; Bryois et al., 2014], where the genetic variant is associated with variation in a local gene, protein or other molecular phenotype. However, a large proportion of the effects of genetic variation on molecular phenotypes likely occurs in trans [Liu et al., 2019b], where a genetic variant is associated with a more distal omics trait. Trans-associations may be of particular interest in studying the effects of copy number alterations (CNAs) on omics traits in cancer cells since CNAs with strong trans-effects may be more likely to confer selective advantages [Zhang et al., 2016a], and CNAs with preserved effects on many distal omics traits across multiple tumor types are more likely to be cancer drivers with multiple functional consequences in contributing to tumor initiation and progression. Thus, developing a better understanding of the trans-effects of both germline variants and somatic mutations is key to creating a more comprehensive portrait of the regulatory landscape and how genetic variants affect omics traits.

Many of the trans-effects of genetic variants may be mediated through the effects of those genetic variants on cis-omics traits [Pierce et al., 2014; Yang et al., 2017]. Identifying the cis-

mediators of trans-associations increases the interpretability of findings and produces a more comprehensive understanding of the mechanisms through which genetic variants affect omics traits. In the case of trans-associations of CNAs, mediation analysis may help distinguish which cis-gene/cis-protein in a genomic region may be driving the trans-association. Whereas the gene-level CNAs of genes in a cis-genomic region are correlated, making it challenging to distinguish which gene-level CNA is truly associated with the trans-omic trait following direct trans-association testing of the total cis-CNA effect, studying CNA trans-associations in a cis-mediation framework may help reveal the putative cis-gene/cis-protein mediators and provide possible interpretations of the underlying mechanism.

Performing integrative analysis of cis-mediated trans-associations to identify joint associations shared across conditions may further elucidate underlying biological mechanisms of key importance. In addition to identifying robust associations, such analyses have the potential to identify associations of high public health and biological importance given that they are shared across multiple related conditions. When studying effects of CNAs on omics traits in tumors, identifying associations shared by different cancer types may highlight therapeutic targets of potentially preserved and high impact given their joint association in multiple related diseases.

In this chapter, we extend the Primo framework to integrate product of coefficient test statistics testing indirect (mediation) effects across conditions to identify shared cis-mediated trans-associations. We evaluate performance of the method, Primo(med), through simulations. We apply the method to identify trans-protein associations of CNAs mediated by cis-protein abundance shared in breast and ovary tumors, and identify several cis-hubs (with multiple trans-associations) that have known associations to cancer phenotypes. The results demonstrate the potential for Primo(med) to use integrative mediation analysis to highlight dynamic mechanisms in the genome.

3.2 Methods

3.2.1 Mediation

In this work, we propose to integrate mediation test statistics to identify shared trans-associations mediated by cis-associations. A graphical depiction of mediation is shown in Figure 3.1. In a traditional mediation analysis, there is an observed (and presumed causal) association between an independent variable X and dependent variable Y, implying that the effect $\tau \neq 0$ in Figure 3.1A. Of interest is whether or not the association between X and Y is mediated through a variable M. If M is a mediator of the association between X and Y as shown in Figure 3.1B, then X affects M ($\alpha \neq 0$) and subsequently M affects Y ($\beta \neq 0$). In the case of full mediation, the entire effect of X on Y is mediated through M ($\tau' = 0$). In the case of partial mediation, some of the effect of X on Y occurs through mediation by M while some of the effect occurs through other pathway(s) ($\tau' \neq 0$). Typically, mediation is assessed using the three regression equations:

$$Y = \tau_0 + \tau X + \epsilon_1 \tag{3.1}$$

$$M = \alpha_0 + \alpha X + \epsilon_2 \tag{3.2}$$

$$Y = \beta_0 + \beta M + \tau' X + \epsilon_3 \tag{3.3}$$

Without loss of generality, we omitted additional covariates in the models for simplicity, but other covariates (e.g. potential confounders) may also be included in the regression equations.

In the case of trans-associations mediated by cis-associations, X and M occur in cis (local association) while Y is in trans with X and M (distal association). A cis-mediated trans-association may be present when there is: non-zero total trans-association ($\tau \neq 0$); non-zero association between the cis-independent variable X and cis-mediator M ($\alpha \neq 0$); and non-zero conditional association between cis-mediator M and trans-dependent variable

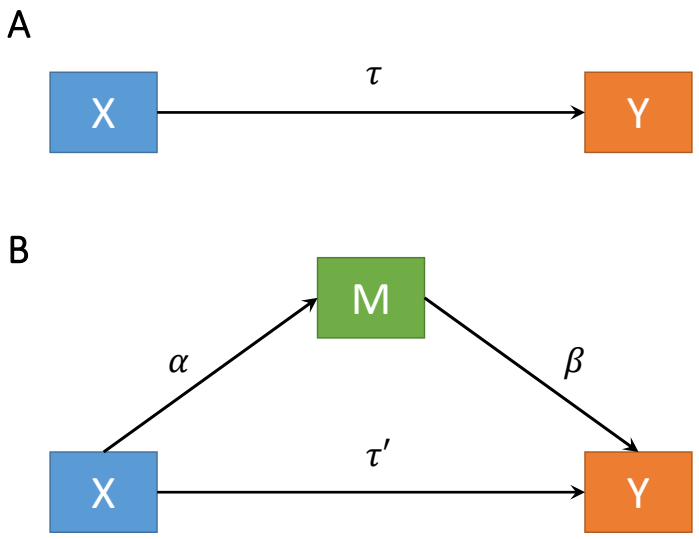


Figure 3.1: **Illustration of mediation.** The graph depicts the pathways involved in mediation and the parameters typically tested in regression models.

Y, conditioning on cis-independent variable X ($\beta \neq 0$). An additional set of assumptions, that there is no unmeasured (unadjusted) confounding in any of the relationships between cis-independent variable X, cis-mediator M, and trans-dependent variable Y is discussed in the next section.

Confounding and adjustment in high-dimensional data

For independent variable X, mediator M and dependent variable Y, mediation analyses assume that there is no unmeasured (unadjusted) confounding in the association between:

1. X and Y
2. X and M
3. M and Y

Additionally, there should be no M-Y confounder that is itself affected by X. Figure 3.2 illustrates how unadjusted confounding may lead to detection of spurious associations. In the figure, there is a true association between the independent variable X and mediator M ($\alpha \neq$

0). If there is no true association between the mediator M and dependent variable Y ($\beta = 0$), but there is a confounder H of M and Y, then failing to adjust for the confounding effect of H may result in detection of a spurious mediation effect. Similarly, spurious mediation effects may be detected when there are unadjusted confounders of X and M (e.g. $\beta \neq 0$ but $\alpha = 0$) or of X and Y (e.g. if exactly one of α or β were non-zero).

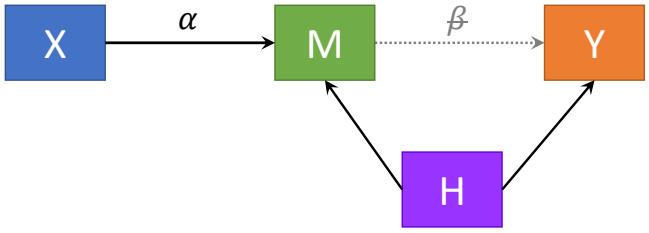


Figure 3.2: **Illustrative example of confounding in mediation analyses.** The example shows how unmeasured or unadjusted confounding of the association between mediator M and dependent variable Y by another variable H may lead to detection of a spurious mediation effect. Black lines depict true associations while gray dotted lines depict no association.

The high-dimensionality of omics data provides an opportunity to reduce spurious associations due to unmeasured (and therefore unadjusted) confounding. For example, say we are studying trans-associations on protein abundance mediated by cis-protein abundance. Let's say there is a variable, H, that confounds the relationship between cis-protein M_p and trans-protein $Y_{p'}$. If H confounds the relationship between one pair of proteins (M_p and$Y_{p'}$), meaning it is causally associated with abundance of more than one protein, then it may be associated with the abundance of (potentially many) more proteins. Data reduction techniques which create new variables that account for systematic variability in the data, such as principal components analysis, may thus be useful in reducing spurious associations due to unmeasured confounding. When known or suspected confounders are measured, approaches that account for variability in the data while "protecting" the the effects of measured covariates, such as surrogate variable analysis [Leek and Storey, 2007] or PEER factors [Stegle et al., 2012], may be used in conjunction with the measured confounders. These strategies

may aid in reducing detection of spurious associations due to confounding, as we demonstrate using simulations in Appendix B.2.

3.2.2 Product of coefficient tests of the indirect (mediation) effect

As discussed and evaluated in MacKinnon et al. (2002), there are several widely used methods to test for mediation. One set of approaches evaluates the indirect effect ($X \to M \to Y$) using the product of coefficients ($\hat{\alpha}\hat{\beta}$ from regression equations 3.2 and 3.3). Evaluating whether mediation is present tests the hypothesis that the product of α and β is zero: $H_0 : \alpha\beta = 0$. Inference for this hypothesis necessitates an estimate of the uncertainty in the product. Perhaps the most commonly used method to estimate the uncertainty of the product was proposed in Sobel (1982).

Sobel test

Based on regression equations 3.2 and 3.3, the Sobel test evaluates whether there is a non-zero indirect (i.e. mediation) effect by testing the hypothesis that the product of α and β is zero: $H_0 : \alpha\beta = 0$. Uncertainty in the product is quantified by a formula derived using the multivariate delta method based on the first order Taylor series approximation [Sobel, 1982]. Using the estimates from regression models 3.2 and 3.3, the Sobel test statistic is given by $z_{\text{Sobel}} = \frac{\hat{\alpha}\hat{\beta}}{\hat{\sigma}_{\hat{\alpha}\hat{\beta}}}$, where $\hat{\sigma}_{\hat{\alpha}\hat{\beta}} = \sqrt{\hat{\alpha}^2\hat{\sigma}^2_{\hat{\beta}} + \hat{\beta}^2\hat{\sigma}^2_{\hat{\alpha}}}$ ($\hat{\sigma}$'s denoting the standard errors of the regression coefficients). As shown in Sobel (1982), when at least one of $\alpha \neq 0$ or $\beta \neq 0$, the asymptotic distribution of z is normal. (When $\alpha = 0$ and $\beta = 0$, the conditions of the delta method used by [Sobel, 1982] are not met; in Appendix B.1, we show by simulations that in finite samples, the distribution of z is symmetric around zero with thinner tails than $N(0, 1)$ when both $\alpha = 0$ and $\beta = 0$). Inference using the Sobel test statistic z is typically performed by comparison to $N(0, 1)$, the asymptotic distribution of z when one of $\alpha = 0$ or $\beta = 0$ (making true the null hypothesis $H_0 : \alpha\beta = 0$).

Other approaches to test product of coefficients of the indirect effect

Other approaches have also been suggested to test the product of coefficients of the indirect effect. The standard error of the product $\hat{\alpha}\hat{\beta}$ based on first and second order Taylor series expansion [Aroian, 1944] is: $\hat{\sigma}_{\hat{\alpha}\hat{\beta}} = \sqrt{\hat{\alpha}^2\hat{\sigma}^2_{\hat{\beta}} + \hat{\beta}^2\hat{\sigma}^2_{\hat{\alpha}} + \hat{\sigma}^2_{\hat{\alpha}}\hat{\sigma}^2_{\hat{\beta}}}$. Because the additional summand $(\hat{\sigma}^2_{\hat{\alpha}}\hat{\sigma}^2_{\hat{\beta}})$ is strictly non-negative, the standard error based on Sobel (1982) forms a lower bound for the standard error based on Aroian (1944), making the Aroian test statistic more conservative than the Sobel test statistic. And as shown in MacKinnon (2002), the Sobel test statistic already tends to perform conservatively in finite samples, especially when sample size is small.

The standard error of the product $\hat{\alpha}\hat{\beta}$ based on the unbiased variance of the product of two normal variables [Goodman, 1960] is: $\hat{\sigma}_{\hat{\alpha}\hat{\beta}} = \sqrt{\hat{\alpha}^2\hat{\sigma}^2_{\hat{\beta}} + \hat{\beta}^2\hat{\sigma}^2_{\hat{\alpha}} - \hat{\sigma}^2_{\hat{\alpha}}\hat{\sigma}^2_{\hat{\beta}}}$. In contrast to the approach based on Aroian (1944), the product of variances is subtracted, rather than added. When the observed product of variances of the cofficients is larger than the sum $\hat{\alpha}^2\hat{\sigma}^2_{\hat{\beta}} + \hat{\beta}^2\hat{\sigma}^2_{\hat{\alpha}}$, the standard error is undefined using this method.

MacKinnon and colleagues have proposed several methods to test the mediation effect [MacKinnon et al., 2002]. These include the distribution of the product of two standard normal variables: $z_\alpha z_\beta$, where $z_\alpha = \frac{\hat{\alpha}}{\hat{\sigma}_{\hat{\alpha}}}$ and $z_\beta = \frac{\hat{\beta}}{\hat{\sigma}_{\hat{\beta}}}$ [MacKinnon et al., 1998]. This approach has slightly inflated Type I error rate in small sample sizes [MacKinnon et al., 2002]. When associations are sparse, even slight inflation in Type I error rate can lead to a significantly inflated false discovery rate (FDR). Other approaches, such as the empirical distribution [MacKinnon et al., 1998] and asymmetric confidence interval [MacKinnon and Lockwood, 2001] approaches proposed by MacKinnon and colleagues, or the standardized variables approach proposed by Bobko and Rieck (1980), were not considered in this work and are left to the reader for exploration.

Permutation methods estimating uncertainty in $\hat{\alpha} \cdot \hat{\beta}$

As discussed in MacKinnon et al. (2002), existing methods to perform inference on the null hypothesis $H_0 : \alpha\beta = 0$ testing for an indirect (mediation) effect vary in power and their ability to control the Type I error rate. Those methods that control the Type I error rate, including the Sobel test, suffer from low statistical power, particularly in small sample sizes [MacKinnon et al., 2002]. Because the distributional assumption made by many mediation methods that the sampling distribution of the product of coefficients $\hat{\alpha}\hat{\beta}$ follows a normal distribution is often violated in finite samples, permutation methods have been proposed to estimate a (possibly non-normal) distribution for $\hat{\alpha}\hat{\beta}$ [Taylor and MacKinnon, 2012; Koopman et al., 2015]. In this work, we consider a permutation method in which we permute the residuals from the original regression equations 3.2 and 3.3. The approach is similar to the permutation test of the Indirect Effect under Full Models (IEFM) described in Kroehl et al. (2020).

Specifically, in calculating permutation-based test statistics of the indirect (mediation) effect, we follow the following steps:

1. Fit the models $M = \alpha_0 + \alpha X + \mathbf{C}\xi_M + \epsilon_M$ and
   $Y = \beta_0 + \beta M + \tau' X + \mathbf{C}\xi_Y + \epsilon_Y$.
   Here, $\mathbf{C}$ is a matrix of covariate(s) (including possible confounders), ξ is a vector of effects of the covariates, and the rest of the terms are as previously described.

2. Obtain residuals from models in step 1: e_M and e_Y. For $p = 1, ..., P$ permutations, permute the residuals, labeled e^*_{Mp} and e^*_{Yp}.

3. For each permutation, calculate $M^*_p = \hat{M} + e^*_{Mp}$ and $Y^*_p = \hat{Y} + e^*_{Yp}$.

4. For each permutation, fit the regression models from step 1, replacing M with M^*_p and Y with Y^*_p. Estimate $\hat{\alpha}^*_p$ and $\hat{\beta}^*_p$.

5. Calculate $\hat{\sigma}_{\hat{\alpha}\hat{\beta}} = \sqrt{\frac{\sum_p (\hat{\alpha}^*_p \hat{\beta}^*_p - \overline{\hat{\alpha}^* \hat{\beta}^*})^2}{P-1}}$, the standard deviation of $\hat{\alpha}^* \hat{\beta}^*$.

6. Calculate $z_{\text{perm}} = \frac{\hat{\alpha}\hat{\beta}}{\hat{\sigma}_{\hat{\alpha}\hat{\beta}}}$.

z_{perm} is then the test statistic testing the indirect (mediation) effect of X on Y through mediator M.

3.2.3 Primo for integrative mediation analyses

To identify cis-mediated trans-associations that are shared across conditions (e.g. cancer/tumor type), we propose the algorithm Primo(med), described in the next few sections.

Distribution of product of coefficients of the indirect (mediation) effect

In this section, we describe the rationale for developing a new version of the Primo algorithm for integrating product of coefficient test statistics. In section 2.2.2, we described a version of the Primo algorithm – Primo(t) – to use when the test-statistics for integrative analysis are approximately normal. Given the work of Sobel in showing that the asymptotic distribution of $\hat{\alpha}\hat{\beta}$ is normal (under the condition that at least one of α or β is non-zero) [Sobel, 1982], it may seem natural to apply the Primo(t) algorithm to a matrix of Sobel or other product of coefficient test statistics. However, as discussed in MacKinnon et al. (2002), whereas the distribution of $\hat{\alpha}\hat{\beta}$ may be asymptotically normal, in finite sample sizes the distribution is not normally distributed but rather is "often asymmetric with high kurtosis". The limma method, which is used by Primo(t) to estimate the marginal alternative distribution of the test statistics, makes the assumption that the distribution of an estimator conditional on the true parameter follows a normal distribution. In the context of the product of coefficients, the assumption would be: $\hat{\alpha}\hat{\beta}|\alpha, \beta, \sigma^2 \sim N(\alpha\beta, v\sigma^2)$, which is violated in finite sample sizes.

While the distribution of the product of coefficients for a particular mediation trio may not follow a normal distribution in finite sample sizes, it may still be reasonable to model the distribution of mediation test statistics across the genome using a normal distribution. As shown in the simulations presented in Appendix B.1, when the non-zero distributions of

α and β followed normal distributions across the genome, the Sobel test statistics followed a bell-shaped distribution symmetric around zero. This observation motivates the algorithm for Primo for integrating z-scores of product of coefficient test statistics.

Primo for integrating product of coefficient z-scores

The Primo method for integrating product of coefficient z-scores, Primo(z), proceeds similarly to the general Primo integrative association analysis method described in Section 2.2.1. In estimating the marginal null and alternative density functions for each study j, f_j^0 and f_j^1, we make the following modification. Let $\mathbf{T}$ be the matrix of z-scores (product of coefficient mediation test statistics). Under the null hypothesis, the z-scores in study j follow the standard normal distribution. Under the alternative, the test statistics in study j are modeled using a normal distribution with fatter tails than the standard normal distribution. That is, under the alternative, the test statistics t_{ij} are assumed to follow a $N(0, \sigma_j^2)$ distribution with a study-specific scaling factor $\sigma_j > 1$.

To estimate a study-specific scaling factor $\sigma_j > 1$ that best approximates the tail of the alternative distribution in study j, we use a numerical optimization algorithm to find values which minimize the differences between the P-values of T_j under a mixture of $N(0, \sigma_j^2)$ and $N(0, 1)$ distributions given the mixing proportion θ_j^1 for the study, and their nominal P-values based on their ranks. More specifically, let t_{ij} be the product of coefficient z-score for mediation trio i in study j. Then the cumulative distribution function of t_{ij} is given by

$$F(t_{ij}; \sigma_j, \theta_j^1) = (1 - \theta_j^1)G(t_{ij}; 0, 1) + \theta_j^1 G(t_{ij}; 0, \sigma_j)$$

where $G(\cdot; \mu, \sigma)$ is the cumulative distribution function of a $N(\mu, \sigma^2)$ variable. Let r_{ij} be the rank of trio i in study j when the t_{ij} are sorted in descending order. To estimate σ_j, we use the optimization algorithms implemented in the R nloptr package [Johnson, 2018] to

minimize the following objective function:

$$\sum_{i:\, r_{ij} \leq \max\{20, \frac{m}{2}\theta_j^1\}} \left| 1 - F(t_{ij}; \sigma_j, \theta_j^1) - \frac{r_{ij} - 0.5}{m} \right|.$$

Once we estimate the scaling factors under the alternative distribution for each study, we proceed to estimate pattern-specific multivariate density functions.

In estimating a pattern-specific multivariate density function D_k from J (potentially correlated) studies using product of coefficient z-scores, we obtain the empirical null and alternative marginal distributions as standard normal and normal distributions with scale factor $\sigma_j > 1$, respectively, in each of the J studies. The variance of the test statistics in study j under pattern k is given by $\sigma_{kj}^2 = \sigma_j^{2 \times q_{kj}}$, where q_{kj} is the association status of the k-th pattern in study j. As in Section 2.2.3, we estimate a $J \times J$ sample correlation matrix $\mathbf{\Gamma}$ using observations with absolute statistics less than 5 in all J studies (when all studies are independent, $\mathbf{\Gamma}$ is diagonal with all diagonal elements equal to 1). Then, we approximate the pattern-specific multivariate density function D_k as $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_k^{1/2} \mathbf{\Gamma} \mathbf{\Sigma}_k^{1/2})$, where $\mathbf{\Sigma}_k$ is a diagonal matrix with diagonal elements of σ_{kj}^2's.

The Primo(med) algorithm

Finally, we propose the following Primo(med) algorithm to identify cis-mediated trans-associations shared across across J studies or conditions. For convenience (given the data application in the next section), we refer to the independent variable as "CNA", the mediator as "cis-protein", and the dependent variable as "trans-protein". The Primo(med) algorithm is:

Algorithm 1 Primo(med) for detecting cis-mediated trans-associations across studies

1. Obtain study-specific summary statistics. In each study j ($j = 1, \ldots, J$), calculate the cis-association effect $\{\hat{\alpha}_{ij}\}$ (mediator-independent variable association) for each cis-protein i ($i = 1, \ldots, m$); the cis-trans protein-protein conditional correlation effect $\{\hat{\beta}_{(ii')j}\}$ (dependent variable-mediator association) for each ordered pair of proteins (i,i') such that ($i = 1, \ldots, m$) and ($i' \in$ {proteins in trans with i}); total trans-association effect $\{\hat{\tau}_{(ii')j}\}$ (dependent variable-independent variable association) assessing the total trans-effect of CNA i on trans-protein i'; and standard errors for each estimate ($\hat{\sigma}_{\cdot}$).

2. Calculate study-specific product of coefficient test statistics. For each cis-trans protein ordered pair (i, i') in each study j, calculate the product of coefficients test statistic as $z_{(ii',j)} = \frac{\hat{\alpha}_{ij}\hat{\beta}_{(ii')j}}{\hat{\sigma}_{\hat{\alpha}_{ij}\hat{\beta}_{(ii')j}}}$. When using Sobel test statistics (recommended for large sample sizes), $\hat{\sigma}_{\hat{\alpha}_{ij}\hat{\beta}_{(ii')j}} = \sqrt{\hat{\alpha}_{ij}^2\hat{\sigma}^2_{\hat{\beta}_{(ii')j}} + \hat{\beta}^2_{(ii')j}\hat{\sigma}^2_{\hat{\alpha}_{ij}}}$. When using permutation-based test statistics (recommended for small sample sizes), $\hat{\sigma}_{\hat{\alpha}_{ij}\hat{\beta}_{(ii')j}}$ is estimated using permuations of residuals as described in Section 3.2.2.

3. Estimate cross-study trans-association probabilities using Primo(t). Using the total trans-association effect statistics $\{\hat{\tau}_{(ii')j}\}$ and their standard errors as input, estimate the posterior probabilities of cross-study trans-association using Primo(t).

4. Estimate cross-study indirect (mediation) effect probabilities using Primo(z). Using the product of coefficients test statistics $\{z_{(ii')j}\}$ as input, estimate the posterior probabilities of cross-study indirect effect using Primo(z).

5. Identify instances of cross-study cis-mediated trans-associations. Using posterior probability thresholds (λ_1,λ_2), identify which trios have *both* a posterior probability of a cross-study trans-association $> \lambda_1$ (from step 3) *and* a posterior probability of a cross-study indirect effect $> \lambda_2$ (from step 4).

3.3 Simulations

We evaluated the performance of Primo(med) in integrating product of coefficient test statistics in a variety of simulated scenarios. In each scenario, we simulated data for N_j subjects in each j of J studies of m mediation trios. Here a "study" generally represents a specific trait (e.g. gene expression or protein abundance) evaluated in a specific cell-type, tissue-type or condition. In study j, mediation trio i for subject n_j was simulated according to the models:

$$\begin{aligned} X_{n_j ij} &\sim N(0,1) \\ M_{n_j ij} &= \alpha_{ij} X_{n_j ij} + \varepsilon_{M,n_j ij} \\ Y_{n_j ij} &= \beta_{ij} M_{n_j ij} + \varepsilon_{Y,n_j ij} \end{aligned} \tag{3.4}$$

where α_{ij} is the effect of the independent variable on the mediator (i.e. the "cis effect"), β_{ij} is the effect of the mediator on the dependent variable (i.e. "cis-trans correlation"), and $\varepsilon. \overset{iid}{\sim} N(0,1)$'s are random variation. For each trio i ($i = 1, ..., m$) in study j ($j = 1, ..., J$), the key parameters in the simulations are α_{ij} (the effect of the independent variable on the mediator, or "cis effect") and β_{ij} (the effect of the mediator on the dependent variable, or "cis-trans correlation"). When a cis-mediated trans-association is present, both of these parameters are non-zero.

After simulating the data as described, trio-level test statistics in each simulation were calculated using the following regression models:

$$Y_{ij} = \tau_{0ij} + \tau_{1ij} X_{ij} + \epsilon_{\tau,ij} \tag{3.5a}$$

$$M_{ij} = \alpha_{0ij} + \alpha_{1ij} X_{ij} + \epsilon_{\alpha,ij} \tag{3.5b}$$

$$Y_{ij} = \beta_{0ij} + \beta_{1ij} M_{ij} + \tau'_{ij} X_{ij} + \epsilon_{\beta,ij} \tag{3.5c}$$

After running the regressions, we extracted the coefficient estimates and their standard errors for input into Primo(med). For input into Primo(t) testing the total trans-association, we

obtained $\hat{\tau}_{1ij}$ and $\hat{\sigma}_{\hat{\tau}_{1ij}}$. For input into Primo($z$) testing the indirect (mediated) effect, we calculated the product of coefficients test statistic as $z_{ij} = \frac{\hat{\alpha}_{1ij}\hat{\beta}_{1ij}}{\hat{\sigma}_{\hat{\alpha}_{1ij}\hat{\beta}_{1ij}}}$. For the Sobel test statistic, $\hat{\sigma}_{\hat{\alpha}_{1ij}\hat{\beta}_{1ij}} = \sqrt{\hat{\alpha}^2_{1ij}\hat{\sigma}^2_{\hat{\beta}_{1ij}} + \hat{\beta}^2_{1ij}\hat{\sigma}^2_{\hat{\alpha}_{1ij}}}$. For the permutation-based test statistic, $\hat{\sigma}_{\hat{\alpha}_{1ij}\hat{\beta}_{1ij}}$ was estimated using permuations of residuals as described in Section 3.2.2. In the simulations, we repeat the regressions for each trio i ($i = 1, ..., m$) in each study j $(1, ..., J)$ to obtain three $m \times J$ matrices of test statistics: one for Primo(t), one for Primo(z) using Sobel test statistics, and one for Primo(z) using permutation-based test statistics.

3.3.1 Evaluating the performance of Primo(med) in moderate sample sizes

In Scenario 1, we evaluated the performance of Primo(med) when sample sizes are moderate. We simulated data for $J = 3$ studies of $m = 100k$ trios for $N_j = 200$ $\forall j$ subjects. In this scenario, non-zero effects for $\alpha_{ij} \sim N(1, 0.25)$, and non-zero effects for $\beta_{ij} \sim N(0, 0.5)$.

In the simulation, the alternative hypothesis of a non-zero cis-mediated trans-association is true when there is both a non-zero "cis effect" ($\alpha_{ij} \neq 0$) and non-zero "cis-trans correlation" ($\beta_{ij} \neq 0$). We simulated $\pi_k = (5 \times 10^{-3}, 2 \times 10^{-3}, 1 \times 10^{-3})$ for patterns with cis-mediated trans-association being present in only one, exactly two, and all three studies, respectively. Thus, the true alternative proportion $\theta^1_j = 0.01$, $\forall j$. Under the null hypothesis, a trio i in study j was set to one of three categories:

1. ($\alpha_{ij} \neq 0$, $\beta_{ij} = 0$): non-zero cis effect but no cis-trans correlation (70%)
2. ($\alpha_{ij} = 0$, $\beta_{ij} \neq 0$): no cis effect but non-zero cis-trans correlation (5%)
3. ($\alpha_{ij} = 0$, $\beta_{ij} = 0$): neither cis effect nor gene-gene correlation (25%)

The setting mimics ones where cis-effects are prevalent but cis-mediated trans-associations are sparse.

Figure 3.3 shows the power (A) and false discovery rate (FDR) (B) comparison between the Primo(z)-Sobel (solid lines), Primo(z)-permuted (dashed lines) and Primo(t) for total

trans-association (dotted lines) tests over 1000 simulations. As shown in the figure, for each grouped pattern of trans-association in "at least # of studies" (line color), each method demonstrates reasonable power and control of the FDR (y-axes) across several posterior probability thresholds (x-axis), with performance of Primo(z)-Sobel and Primo(z)-permuted often indistinguishable in this setting. The simulation demonstrates reasonable performance of Primo(med) to identify cis-mediated trans-associations when sample sizes are moderate.

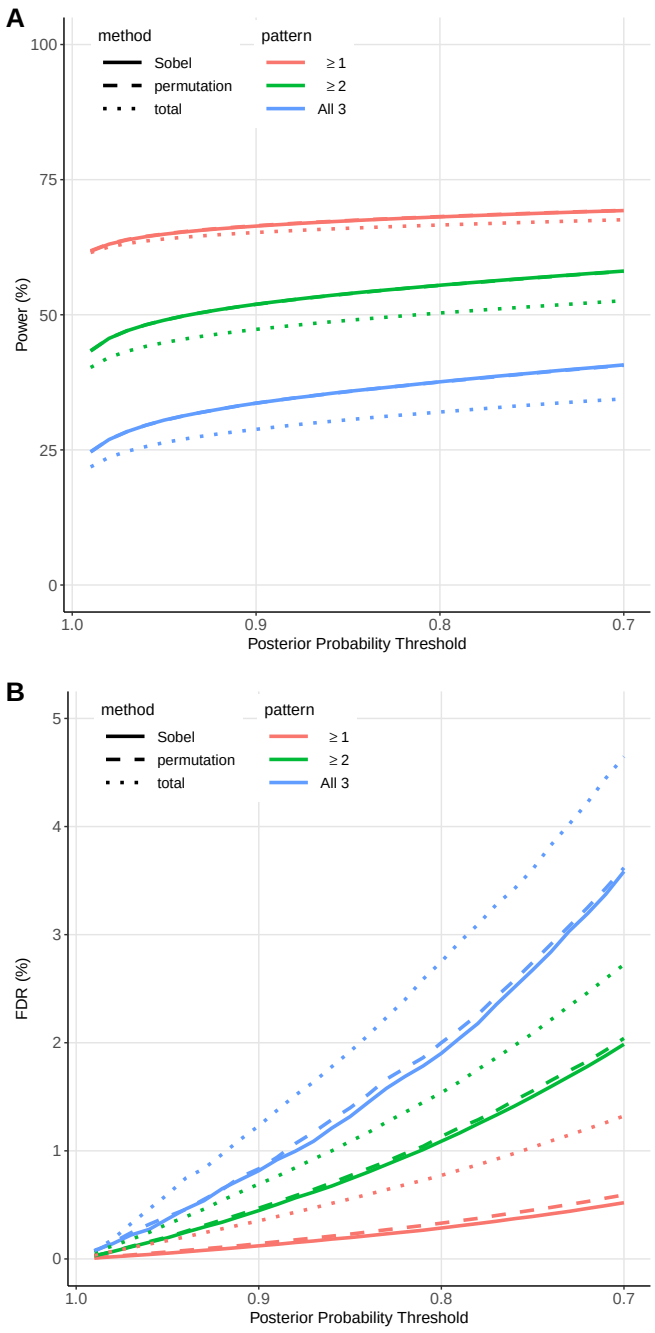


Figure 3.3: **Evaluating the performance of components of Primo(med) in assessing cis-mediated trans-associations in moderate sample sizes.** Power (A) and observed false discovery rate (FDR) (B) (y-axes) by posterior probability threshold (x-axis) shown over 1000 simulations for Primo(z)-Sobel (solid lines) and Primo(z)-permuted (dashed lines) methods for integrating product of coefficient test statistics, and Primo(t) integrating total trans-association statistics (dotted lines). Line color represents grouped association pattern ("associated with at least # studies"). In moderate sample sizes, all three methods demonstrate reasonable performance in assessing trans-associations mediated by cis. Note that the dashed lines of Primo(z)-permuted are overlayed with solid lines of Primo(z)-Sobel in (A).

3.3.2 Evaluating the performance of Primo(med) in small sample sizes

In Scenario 2, we evaluated the performance of Primo(med) in identifying cis-mediated trans-associations when sample sizes are small. We simulated data for $J = 3$ studies of $m = 100k$ trios, with sample sizes of $N_j = 100$, 50, and 20 for $j = 1$, 2, and 3, respectively. In this scenario, non-zero effects for $\alpha_{ij} \sim N(1, 0.25)$ and non-zero effects for β_{ij} are drawn from the vector $\{-0.9, -0.6, -0.3, 0.3, 0.6, 0.9\}$ with equal probabilities. π_k is simulated as in 3.3.1, with similar distributions of α_{ij} and β_{ij} under the null. Thus, the true alternative proportion $\theta_j^1 = 0.01$, $\forall j$.

Figure 3.4 shows the power (A) and false discovery rate (FDR) (B) comparison between the Primo(z)-Sobel (solid lines), Primo(z)-permuted (dashed lines) and Primo(t) for total trans-association (dotted lines) tests over 1000 simulations. As shown in the figure, Primo(z)-permuted using the permutation-based product of coefficient test statistics demonstrates considerably higher power than Primo(z)-Sobel using Sobel test statistics for the grouped association patterns of "at least 2" and "all 3" trait associations in this setting, as well as better control of FDR (which is a function of both power and the Type I error rate) for the pattern of association with "all 3" traits. The difference may be driven by the small sample sizes of the second ($N_2 = 50$) and third ($N_3 = 20$) traits in the simulation.

Based on the performance of the simulations, we suggest using the Sobel test-statistics only in larger sample sizes and permutation-based product of coefficient test statistics when sample sizes are small.

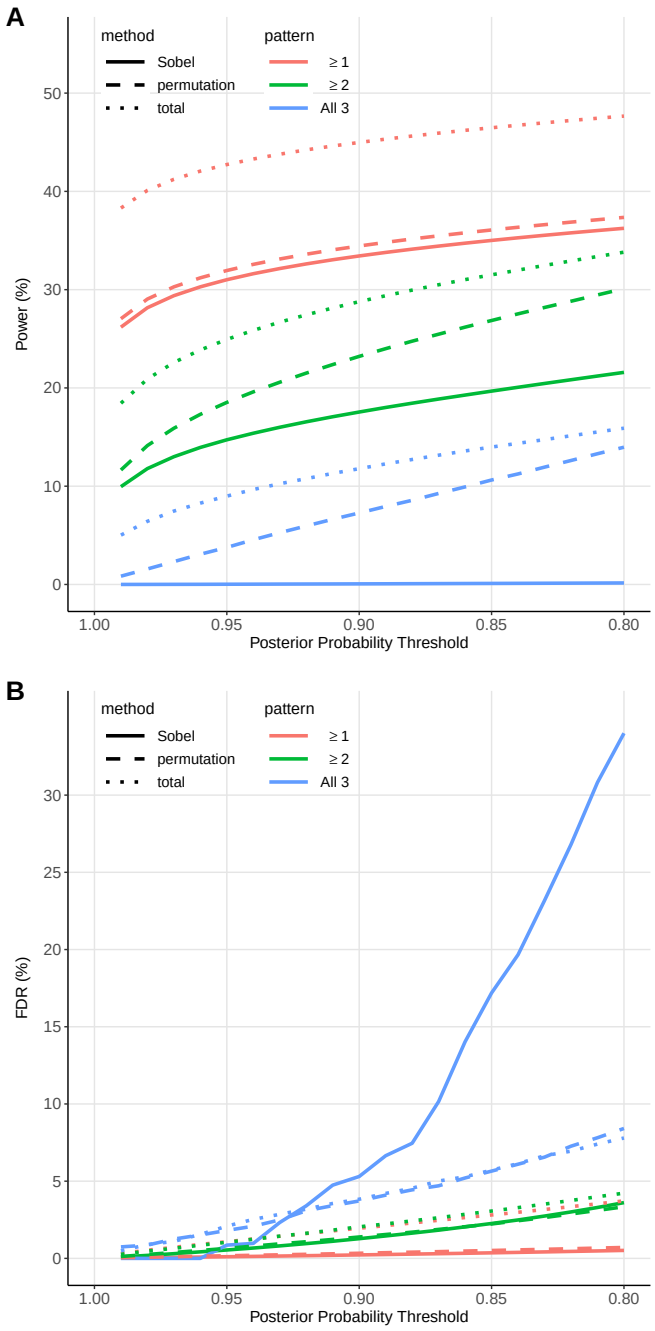


Figure 3.4: **Evaluating the performance of components of Primo(med) in assessing cis-mediated trans-associations in small sample sizes.** Power (A) and observed false discovery rate (FDR) (B) (y-axes) by posterior probability threshold (x-axis) shown over 1000 simulations for Primo(z)-Sobel (solid lines) and Primo(z)-permuted (dashed lines) methods for integrating product of coefficient test statistics, and Primo(t) integrating total trans-association statistics (dotted lines). Line color represents grouped association pattern ("associated with at least # studies"). In small sample sizes, Primo(z)-permuted using permutation based product of coefficient test statistics outperforms Primo(z)-Sobel using Sobel test statistics.

3.4 Data Application

3.4.1 Cis-mediated trans-associations of DNA copy number alterations (CNAs) on protein abundance in both breast and ovary tumors

We applied the Primo(med) method to identify cis-mediated trans-associations of DNA copy number alterations (CNAs) on protein abundance that are shared in breast and ovary tumors. In the mediation framework $X \rightarrow M \rightarrow Y$: the independent variable X is the gene-level CNA of the chromosomal region/segment of a cis-gene i; the mediator M is abundance of cis-protein i; and dependent variable Y is the abundance of a trans-protein i', where i' is in trans with i.

Summary statistics, linear regressions of trans-protein/cis-protein/CNA trios

For the analyses of trans-omics effects of DNA copy number alterations (CNAs) mediated by cis-associations, we use the same TCGA/CPTAC data described in Section 2.4.1. We define "trans" as the proteins coded by genes on a different chromosome from the cis-protein whose CNA is being evaluated. For each protein $i \in \{1, ..., m\}$, we wish to assess the association between CNA of protein i, and protein abundance of each of its m_i' trans-proteins (in multiple cancer types), mediated through the abundance of protein i (i.e. cis-mediated). For each protein $i \in \{1, ..., m\}$, we use the following regression models to assess the cis-mediated

trans-protein effects of CNA within each given cancer type (breast or ovary):

$$\begin{cases} \text{Breast} \begin{cases} \text{trans-protein}_{ui'} &= \tau_{b0i'} + \tau_{b1i'}\text{CNA}_{ui} + \tau_{b2i'}^T \text{cov}_u + \epsilon_{ui'}, \\ \text{cis-protein}_{ui} &= \alpha_{b0i} + \alpha_{b1i}\text{CNA}_{ui} + \alpha_{b2i}^T \text{cov}_u + \epsilon_{ui}, \\ \text{trans-protein}_{ui'} &= \beta_{b0i'} + \beta_{b1i'}\text{cis-protein}_{ui} + \tau'_{bi'}\text{CNA}_{ui} + \beta_{b2i'}^T \text{cov}_u + \epsilon'_{ui'}, \end{cases} \\ \text{Ovary} \begin{cases} \text{trans-protein}_{vi'} &= \tau_{o0i'} + \tau_{o1i'}\text{CNA}_{vi} + \tau_{o2i'}^T \text{cov}_v + \epsilon_{vi'}, \\ \text{cis-protein}_{vi} &= \alpha_{o0i} + \alpha_{o1i}\text{CNA}_{vi} + \alpha_{o2i}^T \text{cov}_v + \epsilon_{vi}, \\ \text{trans-protein}_{vi'} &= \beta_{o0i'} + \beta_{o1i'}\text{cis-protein}_{vi} + \tau'_{oi'}\text{CNA}_{vi} + \beta_{o2i'}^T \text{cov}_v + \epsilon_{vi'}, \end{cases} \end{cases} \tag{3.6}$$

where $\tau_{b1i'}$ and $\tau_{o1i'}$ are the (total) trans-effects of CNA i on abundance of trans-protein i' in breast and ovary tumors, respectively; α_{b1i} and α_{o1i} are the cis-effects of CNA on protein abundance for protein i in breast and ovary tumors, respectively; $\beta_{b1i'}$ and $\beta_{o1i'}$ measure the conditional correlation of proteins i and i' in breast and ovary tumors, respectively, conditioning on CNA; cov's are sets of covariates adjusted for in the regression analyses; u and v are subject indices; and ϵ's are error terms. For each protein $i \in \{1, ..., m\}$, we repeat the regression models for each of its m'_i trans-genes within each cancer type (breast and ovary).

All models were adjusted for: tumor purity, age, stage (III/IV vs. lower than III), 5 genotype PCs, and 10 surrogate variables [Leek and Storey, 2007]. Models for breast tumors were restricted to female subjects, and adjusted for hormone-receptor (HR) status and histological subtype (infiltrating ductal, infiltrating lobular, mucinous, metaplastic, mixed histology or other). Models for ovary tumors were adjusted for tumor grade (3 or 4 vs. lower than 3).

Mediation test statistics

Using the summary statistics from the regression models specified in Equations 3.6, we calculated mediation test statistics z_{Sobel} and z_{perm} for each CNA/cis-protein/trans-protein trio. Quantile-quantile (QQ) plots comparing the observed $-\log_{10}(p)$-values of the test statistics to the expected quantiles under the standard normal distribution are presented in Figure 3.5. As shown in Figure 3.5A, for the mediation analysis in breast tumors, the Sobel test statistics (black points) show deflation as demonstrated by the trend of points falling below the quantiles expected under the null (red line), possibly due to the small sample size ($N = 74$). On the other hand, the permutation-based mediation test statistics (purple points) for the analysis of breast tumors do not appear deflated in Figure 3.5A. The two sets of mediation test statistics show similar distributions in the analysis of ovary tumors (Figure 3.5B), which has a larger sample size ($N = 121$) than the breast tumor analysis. Based on the visualization of the distributions, we utilize the permutation-based mediation test statistics (z_{perm}) in the integrative analysis using Primo.

Integrative analysis

We are interested in detecting joint trans-protein associations of CNAs mediated by cis-protein abundance in breast and ovary tumors. Using summary statistics from the regression models specified in Equations 3.6, we performed integrative analysis of $\sum_i m_i' = 48,437,634$ CNA/cis-protein/trans-protein trios using Primo, with alternative proportion specified as $\theta_j^1 = 10^{-5}$ in both tumors. At a posterior probability threshold of 75%, there 2,317 CNAs with trans-protein association in both breast and ovary tumors (FDR $< 8.4\%$). Of these, 61 trios had a posterior probability $> 75\%$ of cis-mediated trans-association in both breast and ovary tumors (FDR $< 9.1\%$).

The $-\log_{10}(p)$-values of mediation effects and the mediation proportions $(\hat{\tau}_{\text{total}} - \hat{\tau}'_{\text{cis-adj}})/\hat{\tau}_{\text{total}}$ for the 61 trios with high probability of cis-mediated trans-protein association are shown in Figure 3.6. The mediation proportions ranged from 38% to $\geq 100\%$

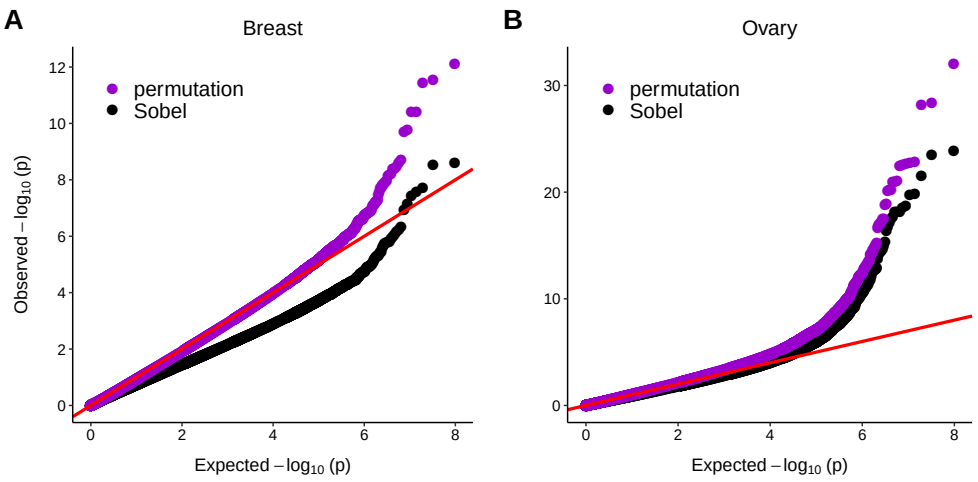


Figure 3.5: **QQ-plots comparing Sobel to permutation-based product of coefficient mediation test statistics in the integrative trans-protein CNA analysis.** Observed $-\log_{10}(p)$-values (y-axis) are plotted against the $-\log_{10}(p)$-values expected under the null hypothesis of no mediation effect (x-axis) for the Sobel test statistics (black points) and permutation-based test statistics (purple points) in the analysis of breast (A) and ovary (B) tumors. The red line (slope=1) shows the expected pattern under the null. In the analysis of breast tumors (A), the Sobel test statistics appear deflated compared to the permutation-based test statistics, possibly due to the small sample size ($N = 74$). The two methods produced similar distributions of test statistics in the analysis of ovary tumors (B), which has a larger sample size ($N = 121$) than breast.

(with medians of 78% and 92% for breast and ovary tumors, respectively), demonstrating that for the identified trios, a sizable proportion of the effect of CNAs on the abundance of the trans-protein was mediated by cis-protein abundance.

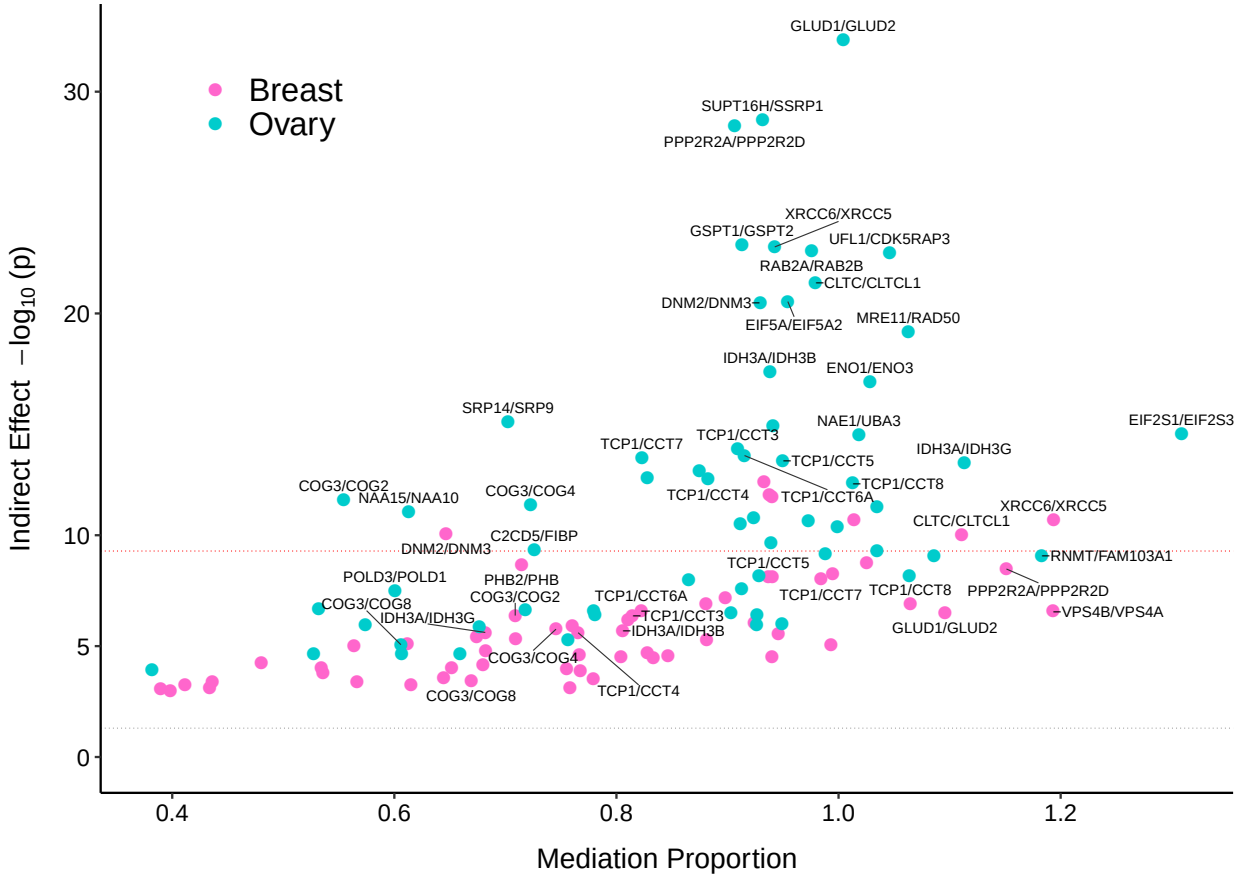


Figure 3.6: **CNA, cis-protein, trans-protein trios with high probability of mediation effects in both breast and ovary tumors.** $-\log_{10}(p)$-values of mediation (indirect) effects using permutation-based test statistics (y-axis) are plotted against the mediation proportion $(\hat{\tau}_{\text{total}} - \hat{\tau}'_{\text{cis-adj}})/\hat{\tau}_{\text{total}}$ (x-axis) for the breast (pink) and ovary (teal) analysis for the 61 trios with high probability of mediation effects in both tumors. The dotted gray line shows the nominal threshold $p = 0.05$ while the dotted red line shows the Bonferroni corrected threshold of $p = 0.05/(2 \times 48,437,634)$. cis-protein/trans-protein pairs are labeled for outlying points and the cis-hubs COG3, IDH3A and TCP1.

The CNAs of 3 cis-proteins were associated with abundance of multiple trans-proteins: COG3, IDH3A and TCP1. *COG3* frequently undergoes adenosine-to-inosine (A-to-I) RNA editing in cancers [Han et al., 2015; Peng et al., 2018], including lobular breast cancer [Shah et al., 2009]. Excessive RNA editing at the *COG3* I/V site promotes tumor proliferation and is associated with poor prognosis in glioblastoma [Silvestris et al., 2019]. Overexpression of *IDH3A* promotes tumor growth by increasing the stability and transactivation activity of HIF-1α and is associated with poor overall survival of lung and breast cancer patients [Zeng et al., 2015]. TCP1 is a member of the chaperonin containing TCP1 complex (CCT), which is required for folding of tubulin and actin proteins [Vallin and Grantham, 2019]. TCP1 was found to be necessary for growth of breast cancer cells in vitro and associated with overall survival in breast cancer patients [Guest et al., 2015]. Moreover, the members of CCT which form the trans targets of TCP1 CNAs also have known associations with cancer phenotypes. CCT3 promotes tumor cell proliferation in liver, papillary thyroid, and gastric cancers [Zhang et al., 2016b; Shi et al., 2018; Li et al., 2017a], and higher expression of *CCT3* is associated with poorer survival in liver cancer [Liu et al., 2019c]; levels of CCT4 may help predict chemoresponse in lung cancer [Epsi et al., 2019]; CCT5 has been proposed as a diagnostic biomarker for non-small cell lung cancer [Gao et al., 2017], and higher expression of *CCT5* may implicate resistance to docetaxel treatment in breast cancer [Ooe et al., 2007] ; higher levels of CCT6A were found to be associated with poorer overall survival in liver cancer and glioblastoma [Zeng et al., 2019; Hallal et al., 2019]; CCT7 has been identified as a potential biomarker for endometrial cancer [Shan et al., 2016]; and protein abundance of CCT8 is associated with the proliferation and invasion capacity of glioma cells [Qiu et al., 2015] as well as metastastis in pancreatic cancer [Liu et al., 2019a].

The results of the integrative analysis of cis-mediated trans-protein associations of CNAs demonstrate the potential for integrative mediation analysis to elucidate the mechanisms through which copy number alterations or other forms of genetic variation affect trans-omics or other distal traits. The cis-hubs (i.e. cis-proteins with multiple trans targets) identified

in such analyses may especially be of interest, given their effects on multiple trans-targets shared in multiple cancer types or other related disease phenotypes.

3.5 Discussion

In this chapter, we developed an integrative mediation analysis method – Primo(med) – by extending the Primo framework to integrate product of coefficient test statistics assessing the indirect (mediated) effect of an independent variable on a dependent variable acting through a mediator. The method can integrate Sobel test (recommended for larger sample sizes) or permutation-based (recommended for smaller sample sizes) product of coefficient test statistics. Primo(med) can be used to identify joint trans-associations that are mediated through effects on cis-omics traits. The associations detected by Primo(med) are: robust, since they are shared across multiple related conditions (such as different cancer types); interpretable, since they occur in the context of an organized mediation structure; and potentially of high public health or biological relevance since the association occurs in multiple conditions.

We applied Primo(med) to jointly analyze cis-mediated trans-protein effects of copy number alterations (CNAs) in breast and ovary tumors. We identified dozens of cis-mediated trans-associations, in each of which a substantial proportion of the total trans-association occurred by mediation through cis-protein abundance. We also identified several "cis-hubs" whose CNAs were associated with multiple trans-genes, and each cis-hub had previously reported associations to cancer-related phenotypes. The results of the analysis demonstrated the potential for Primo(med) to identify robust and interpretable trans-omics associations shared across multiple conditions.

There are some caveats to the presented work that warrant additional consideration. First, while the trans-associations identified by Primo(med) may be consistent with a causal mediation model, additional analyses and/or experiments should be conducted prior to making causal interpretations of any identified trans-associations. Second, we would like to stress the importance of accounting for possible sources of confounding in each of the regression

steps that create the summary statistics for Primo(med) since a key assumption of mediation analyses is that there is no unmeasured/unadjusted confounding in any of the relationships between independent variable (predictor), mediator and dependent variable (outcome). In addition to adjusting for measured covariates, in the presented analyses we also adjusted for surrogate variables [Leek and Storey, 2007] estimated from the matrices of protein abundance in order to account for possible unmeasured confounders that systematically affect protein abundance throughout the genome. Third, in the presented work, the pattern-specific proportions (π_k) and posterior probabilities were separately estimated for the total (direct) trans-association testing using Primo(t) and the indirect (mediation) effect association testing using Primo(z), and we used thresholding of posterior probabilities from each separately performed estimation to identify mediation trios. Neither the pattern-specific proportions nor the posterior probabilities of trios satisfying both total association and indirect effect association criteria (as required for mediation to be present) were directly estimated. Jointly estimating both criteria, to allow estimation of the probability of a mediation effect, is a direction for future research.

The Primo(med) algorithm developed in this work performs integrative mediation analysis to identify joint associations shared across conditions. We applied the method to identify cis-mediated trans-protein associations of CNAs in two tumor types, but the method could be adapted and applied to other settings. Such settings include: identifying cis-mediated trans-associations of germline variants or of somatic mutations other than CNAs; detecting instances where one omics trait (e.g. gene expression or DNA methylation) mediates the associations between genetic variants and another omics trait (e.g. protein abundance); or even integrating mediation test statistics testing associations with non-omics traits. Evaluating joint associations of (germline) genetic variants on both omics and non-omics (i.e. complex) traits is a topic further explored in Chapter 4.

CHAPTER 4

INTEGRATION OF MULTIPLE GWAS AND OMICS QTL SUMMARY STATISTICS FOR ELUCIDATION OF MOLECULAR MECHANISMS OF TRAIT-ASSOCIATED SNPS AND DETECTION OF PLEIOTROPY IN COMPLEX TRAITS

4.1 Introduction

In the post-genomic era, genome-wide association studies (GWAS) have identified tens of thousands of unique associations between germline genetic variants (e.g. single nucleotide polymorphisms, or "SNPs") and human complex traits [Buniello et al., 2019]. Most of the trait-associated SNPs have small effect sizes and many reside in non-coding regions [Edwards et al., 2013; Hindorff et al., 2009], obscuring their functional connections to complex traits. It is known that trait-associated SNPs are more likely to also be expression quantitative trait loci (eQTLs) [Nicolae et al., 2010], thus many of these SNPs likely affect complex traits through their effects on expression levels and/or other "omics" traits. Extensive evaluations of genetic effects on omics traits such as gene expression [The GTEx Consortium, 2017], protein abundance [Johansson et al., 2013], DNA methylation [Smith et al., 2014], histone modification [McVicker et al., 2013; Grubert et al., 2015], and RNA splicing [Li et al., 2016] have revealed an abundance of quantitative trait loci (QTLs) for omics traits (omics QTLs) throughout the genome. These findings suggest that integrating data from omics and multi-omics QTL studies with GWAS would help to elucidate functional mechanisms that underlie trait/disease processes. Moreover, the integrative analysis of omics and multi-omics traits would also enhance confidence in detecting true omics-associations while reducing false positive findings by observing co-occurrence of associations in multiple different data types and borrowing information across multi-omics data sources. The increasing availability of summary statistics for complex traits and omics QTL studies in many conditions and cellular

contexts [The GTEx Consortium, 2017; Pasaniuc and Price, 2017; Bonder et al., 2017; Suhre et al., 2017] provides a valuable resource to conduct integrative analyses in a variety of settings and presents an unprecedented opportunity to gain a system-level perspective of the regulatory cascade, which may highlight targets for disease prevention and/or treatment strategies.

To integrate GWAS and omics QTL summary statistics, several methods have been proposed to identify trait-associated loci that share a common casual variant with omics QTLs (often referred to as "colocalization"). The implementations of these methods allow for integration of GWAS summary statistics with 1-2 sets of QTL summary statistics [Giambartolomei et al., 2014, 2018; Wen et al., 2017; Hormozdiari et al., 2016]. There are also methods that have been proposed to directly test the molecular mechanisms through which genetic variation affects traits by integrating GWAS and eQTL summary statistics [Gusev et al., 2016; Barbeira et al., 2018]. These methods have identified known and novel candidate genes underlying psychiatric disorders [Giambartolomei et al., 2018], diabetes traits [Hormozdiari et al., 2016], obesity-related traits [Giambartolomei et al., 2014; Wen et al., 2017; Gusev et al., 2016], and other traits. By applying the integrative methods to multi-omics data, some QTL pairs such as eQTL and methylation (me)QTL pairs have also been identified with evidence of a shared causal mechanism [Giambartolomei et al., 2018; Pierce et al., 2018]. Integrating studies of multiple complex and omics traits could produce a more comprehensive picture of how cellular processes contribute to variation in complex traits.

Compared to integrating GWAS with single omics QTL statistics, studying multi-omics QTLs increases the chances of detecting the regulatory mechanisms underlying trait/disease-associated SNPs. The effect of any particular SNP may be strong for some omics traits and weak or absent for others. For example, protein (p)QTLs exist for genes lacking an apparent eQTL [Battle et al., 2015], suggesting post-transcription regulation [Chick et al., 2016]. And there could be multiple different omics QTLs in a gene region with different functions. As another example, SNPs affecting RNA splicing (splicing QTLs) may not be eQTLs in a

gene region [Li et al., 2016]. Moreover, QTL effects may vary across molecular phenotypes [Vandiedonck, 2018], tissue types [The GTEx Consortium, 2017], cell types [van der Wijst et al., 2018; Chen et al., 2016], or other contexts [Yao et al., 2014; Zhernakova et al., 2017]. For example, lead SNPs for eQTLs (eSNPs, and a "lead SNP" is the SNP with the smallest association P-value with a particular trait in the region) often vary by tissue type [The GTEx Consortium, 2017]. Jointly analyzing the omics QTL association summary statistics to more than one type of omics trait from different conditions/studies could yield a more complete portrait of the regulatory landscape. Given the increasing availability of summary statistics for omics QTLs from different studies/conditions/cell-contexts, novel methods and tools are needed to integrate GWAS with many relevant sets of omics QTL summary statistics for an improved understanding of the mechanisms of trait-associated SNPs.

Jointly analyzing more than three complex and omics traits can also be viewed as an approach for identifying shared mechanisms that underlie multiple complex traits – pleiotropic effects. Pleiotropy is ubiquitous in the genome [Sivakumaran et al., 2011; Pickrell et al., 2016]. Since pleiotropic effects often occur among related diseases and traits [Parkes et al., 2013; Cross-Disorder Group of the Psychiatric Genomics Consortium, 2013; Wu et al., 2018], shared mechanisms are likely to exist. By integrating omics QTL summary statistics from multiple trait-relevant tissue types with GWAS statistics, one can also boost power in detecting pleiotropic effects while simultaneously providing mechanistic interpretations.

Given the rich availability of omics and multi-omics QTL summary statistics and their dynamic effects in different cellular conditions, in order to provide a comprehensive mechanistic interpretation of known trait-associated SNPs, it is desirable to develop new methods that can integrate multiple sets of GWAS statistics and omics QTL statistics from different conditions/studies while accounting for study heterogeneity, potential sample correlations and linkage disequilibrium (LD). Additionally, as the number of traits/studies/conditions being considered grows, it will be more likely to detect joint associations by chance, necessitating proper multiple testing adjustment. To address those challenges, in this work we

tailor the integrative method described in Chapter 2 (Primo) to integrate summary statistics from multiple GWAS and omics QTL studies.

As described in Chapter 2, Primo is flexible in many aspects: it allows unknown and arbitrary study heterogeneity and can detect coordinated effects from multiple studies while not requiring the effect sizes to be the same; it allows the summary statistics to be calculated from studies with independent or overlapping samples with unknown sample correlations; and it is not an omnibus test for association, but rather can be used to calculate the probability of each SNP belonging to each type (or groups) of interpretable association patterns (e.g. the probability of a trait-associated SNP also being associated with at least one/two cis omics-traits). For gene regions harboring known susceptibility loci, the conditional association analysis of Primo examines the conditional associations of a known trait-associated SNP with other complex and omics traits adjusting for other lead SNPs in a gene region. It moves beyond joint association towards colocalization, and provides a thorough inspection of the effects of multiple SNPs within a region to reduce spurious associations due to LD.

We conduct extensive simulations to evaluate the performance of Primo under various scenarios in analyzing multiple sets of summary statistics from studies with correlated samples. We apply Primo to examine the omics trait association patterns for known SNPs associated with breast cancer risk by integrating multi-omics QTL summary statistics from the Genotype-Tissue Expression (GTEx) project [The GTEx Consortium, 2017] and The Cancer Genome Atlas (TCGA) [The Cancer Genome Atlas Network, 2012] with GWAS statistics from The Breast Cancer Association Consortium (BCAC) [Michailidou et al., 2017]. We also apply Primo to detect known trait-associated SNPs with pleiotropic effects to two complex traits in gene regions harboring susceptibility loci for at least one trait, while also providing mechanistic interpretations by integrating publicly available GWAS summary statistics [Liu et al., 2015; Wood et al., 2014; Locke et al., 2015; Churchhouse and Neale, 2017] with multi-tissue eQTL summary statistics from GTEx. In this work, we focus on only trait-associated SNPs and aim to provide comprehensive mechanistic interpretations of how known GWAS

SNPs affect complex traits. It should be noted that the goal of the analyses is not to identify true causal SNPs. However, the Primo algorithm is generally applicable to integrative association analysis, and when applied in other contexts, the interpretations of the results may be different.

4.2 Methods – Primo for joint association and conditional association across studies/conditions/data-types

4.2.1 Assessing joint associations of SNPs across data types

As in the general integrative association analysis version of Primo (described in Section 2.2.1), when assessing the joint association of SNPs across data types, Primo takes as input multiple sets of association summary statistics from different studies of different data types. The multiple sets of summary statistics could be one set of GWAS statistics and multiple sets of omics/multi-omics QTL statistics, or two or more sets of GWAS statistics of related traits and multiple sets of omics/multi-omics QTL statistics (e.g. different omics phenotypes or the same omics phenotype(s) measured in multiple trait-relevant tissue/cell types).

The input matrix, $\mathbf{T}_{m\times J}$, consists of the summary statistics for the associations of m SNPs with J types of traits from J studies with independent or correlated samples. Note that here a "study" refers to a study of SNPs' associations to a particular trait in a particular condition/cell-type/tissue-type. For each SNP i, there must be one and only one true underlying association pattern. Primo calculates the probability of a given SNP being in each of the $K = 2^J$ mutually exclusive association patterns by borrowing information across SNPs in the genome and across J traits. If a_i denotes the true association pattern for SNP i, then the probability that SNP i belongs to association pattern k is given by:

$$P(a_i = k|T_i, \pi_k) = \frac{\pi_k D_k(T_i)}{\sum_{b=1}^{K} \pi_b D_b(T_i)}, \qquad \text{(2.1, revisited)}$$

where T_i is a vector of J association statistics and is also the i-th row in the $\mathbf{T}$ matrix, π_k represents the overall proportion of SNPs in the genome belonging to the k-th association pattern ($k = 1, \ldots, K$), and $D_k(\cdot)$ is the multivariate density function of J sets of statistics, conditioning on the k-th association pattern.

As in the general integrative association analysis version of Primo, one may collapse association patterns based on biological interpretations and obtain the posterior probabilities of patterns of interest by summing over the probabilities of those mutually exclusive patterns. As illustrated in Figure 4.1, when $J = 4$, there are 16 possible association patterns. We may collapse the association patterns into interpretable groups. For example, here we are interested in the trait-associated SNPs that are also associated with at least 1 omics trait. And we can obtain the probability estimate by summing over the posterior probabilities of patterns 10-16.

4.2.2 Estimating empirical null and alternative marginal density functions for SNP associations

As described in Section 2.2.3, Primo estimates each pattern-specific multivariate density function D_k using estimated parameters from the null and alternative marginal density functions for each study j. When assessing the joint associations of SNPs, Primo estimates the marginal null and alternative density functions of the test statistics by adopting the limma method as described in Section 2.2.2, with the following modifications. For genetic association studies, we calculate the error variance for each SNP based on the t-statistic and the minor allele frequency (MAF) assuming that covariates are independent from genotypes. That is, the error variance for SNP i is given by $s_{ij}^2 = \text{se}^2(\hat{\beta}_{ij}) \cdot 2N_j(\text{MAF}_i)(1 - \text{MAF}_i)$, where N_j is the sample size for study j. The estimation of the empirical null and alternative marginal densities is similar to that described in the general integrative association analysis version of Primo (see Section 2.2.2), except for a modification to the calculation of the scaling factor under the alternative, v_{ij}. For genetic association studies, the scaling factor is

	Q matrix				Interpretation
k	$K = 2^J$				
	GWAS	eQTL	meQTL	pQTL	
1-8	0	0	0	0	No trait association
	0	.	.	.	
	0	1	1	1	
9	1	0	0	0	trait-association only
10	1	1	0	0	trait-assoc. and eQTL
11	1	0	1	0	trait-assoc. and meQTL
12	1	0	0	1	trait-assoc. and pQTL
13	1	1	1	0	trait-assoc. and e+meQTL
14	1	1	0	1	trait-assoc. and e+pQTL
15	1	0	1	1	trait-assoc. and me+pQTL
16	1	1	1	1	trait-assoc. and e+me+pQTL

Figure 4.1: **Example of Q matrix for mechanistic interpretations of trait-associated SNPs.** Interpretations of association patterns for an analysis of a complex trait, eQTL, meQTL and pQTL studies for $j = 1, 2, 3$ and 4, respectively. The red box shows how association patterns can be collapsed into groups of interest (here, summing probabilities across the patterns in the red box would yield the probability of association with the complex trait and *at least one* omics trait).

calculated as $v_{ij} = (1 + v_{0j}/w_{ij})^{1/2}$, where v_{0j} is the variance hyperparameter for the prior placed on nonzero effect size coefficients, and w_{ij} is a SNP-specific weight for SNP i, with $w_{ij} = 1/(2N_j \cdot \text{MAF}_i(1 - \text{MAF}_i))$.

4.2.3 Mechanistic interpretations of trait-associated SNPs via Primo conditional association analysis in gene regions harboring susceptibility loci

In order to elucidate the molecular mechanisms of known trait-associated SNPs, one may examine the omics trait associations of those SNPs by integrating GWAS and omics QTL summary statistics. However, a major challenge in such analyses is the complex LD structure among SNPs in the same gene region.

To assess whether the trait-association of a SNP i reflects an independent causal variant or is simply due to being in LD with a nearby lead SNP i', conditional association analysis is often conducted [Schaid et al., 2018]. It tests the conditional association of SNP i with the trait of interest adjusting for the genotype of the lead SNP i' and other covariates. If SNP i is no longer statistically significant after adjusting for the lead SNP, it is unlikely that the trait-association of SNP i reflects an independent causal effect.

Following this idea, to assess whether a GWAS SNP is associated with omics traits due to it being in LD with lead omics QTLs, we propose to conduct conditional association analysis with summary statistics of the GWAS SNP and lead omics QTLs as input. Here we consider a GWAS SNP i of interest and a set of lead omics SNPs I' in the gene region, where $I' = \{1', ..., L'\}$ is a set of indices. We can model the joint association statistics for SNPs i and I' in study j, i.e., $(t_{ij}, t_{1'j}, ..., t_{L'j})$, using a multivariate normal distribution, $\mathcal{N}(\mathbf{0}, \mathbf{\Lambda}_j)$, where $\mathbf{\Lambda}_j$ is the $1 + L'$ by $1 + L'$ variance-covariance matrix described as follows. The diagonal elements of $\mathbf{\Lambda}_j$ correspond to the study-specific variances of statistics of the SNPs. Specifically, the $(1, 1)$ entry of $\mathbf{\Lambda}_j$ is given by σ^2_{ij}, which is the marginal variance of

the statistic t_{ij} for SNP i in study j with $\sigma^2_{ij} = \frac{d_j}{d_j-2}$ under the null and $\sigma^2_{ij} = v^2_{ij} \cdot \frac{d_j}{d_j-2}$ under the alternative. For each lead SNP $i' \in I'$ with its most plausible association pattern $k_{i'}$, the variance of the corresponding t-statistic $t_{i'j}$ is given by $\sigma^2_{i'k_{i'}j} = v^{2q_{k_{i'}j}}_{i'j} \cdot \frac{d_j}{d_j-2}$. The off-diagonal elements of $\mathbf{\Lambda}_j$ are calculated based on the study-specific variances of the SNPs and the LD among the SNPs assuming additional covariates are independent of the SNP genotypes [Yang et al., 2012]. For instance, the covariance between t_{ij} and $t_{i'j}$ is $\sigma_{ij} \cdot \sigma_{i'k_{i'}j} \cdot \rho_{ii'}$ where $\rho_{ii'}$ is the genotype correlation coefficient of the SNPs i and $i'(\in I')$. Partitioning the variance-covariance matrix $\mathbf{\Lambda}_j$ as follows, $\Lambda_j = \begin{pmatrix} \Lambda_{j,11} & \Lambda_{j,12} \\ \Lambda_{j,21} & \Lambda_{j,22} \end{pmatrix}$ with sizes $\begin{pmatrix} 1 \times 1 & 1 \times L' \\ L' \times 1 & L' \times L' \end{pmatrix}$, we can obtain the conditional null and alternative distributions for SNP i in study j as

$$t_{ij} \mid \begin{pmatrix} t_{1'j} \\ \vdots \\ t_{L'j} \end{pmatrix} \sim \mathrm{N}(\Lambda_{j,12}\Lambda^{-1}_{j,22} \begin{pmatrix} t_{1'j} \\ \vdots \\ t_{L'j} \end{pmatrix}, \ \Lambda_{j,11} - \Lambda_{j,12}\Lambda^{-1}_{j,22}\Lambda_{j,21})$$

where $\Lambda^{-1}_{j,22}$ denotes the inverse of the matrix $\Lambda_{j,22}$. Here we approximate the conditional t-distributions with the conditional Gaussian distribution for efficient density estimation since most GWAS and omics QTL studies have sample sizes large enough for good approximation.

With the conditional null and alternative density functions for SNP i in study j adjusting for other lead omics SNPs in the region, we can proceed to obtain the pattern-specific J-variate density functions for all association patterns as outlined in the previous subsection and re-assess the probabilities of each association pattern in (2.1) using the conditional densities with the previously estimated π_k's. We propose to conduct gene-level conditional association analysis accounting for LD structures only in selected gene regions, after the SNP-level association analysis.

Figure 4.2 shows a conceptual illustration of the conditional association analysis. If the

GWAS SNP is an independent meQTL and pQTL, it remains associated with methylation and protein after adjusting for other lead SNPs in the region; and if the GWAS SNP is associated with cis-expression levels because it is in LD with the lead eSNP, it will be no longer significantly associated with expression after adjusting for the lead eSNP. With conditional association analysis, we can reduce spurious associations due to LD.

As a summary, to elucidate the molecular mechanisms of trait-associated SNPs, we first obtain the estimates of key parameters (π_k's, D_k's) by borrowing information across all SNPs and across traits/studies. Then we focus on each gene region harboring known trait-associated SNPs, and conduct a SNP-level association analysis to all traits for all SNPs in the gene region, followed by a conditional association analysis for each GWAS SNP of interest accounting for LD with other lead omics SNPs. If a GWAS SNP is no longer associated with a particular omics trait after conditioning on the lead omics SNPs, we will not consider it as being truly associated with the omics trait, i.e., the GWAS SNP is not affecting the complex trait via modulating the omics trait. Estimated FDR can be calculated similar to Equation (2.3), with the following modification.

$$\text{estFDR}(\lambda) = \frac{\sum_i (1 - \hat{P}_i) 1(\hat{P}_i \geq \lambda)}{\#\{\hat{P}_i \geq \lambda\}}, \qquad \text{(2.3, revisited)}$$

In the calculation of the numerator of the estimated FDR, for each SNP i that is no longer significant after conditional analysis, its contribution to the numerator $(1 - \hat{P}_i)1(\hat{P}_i \geq \lambda)$ in Equation (2.3) is corrected to be 1 since it is considered as an estimated false discovery.

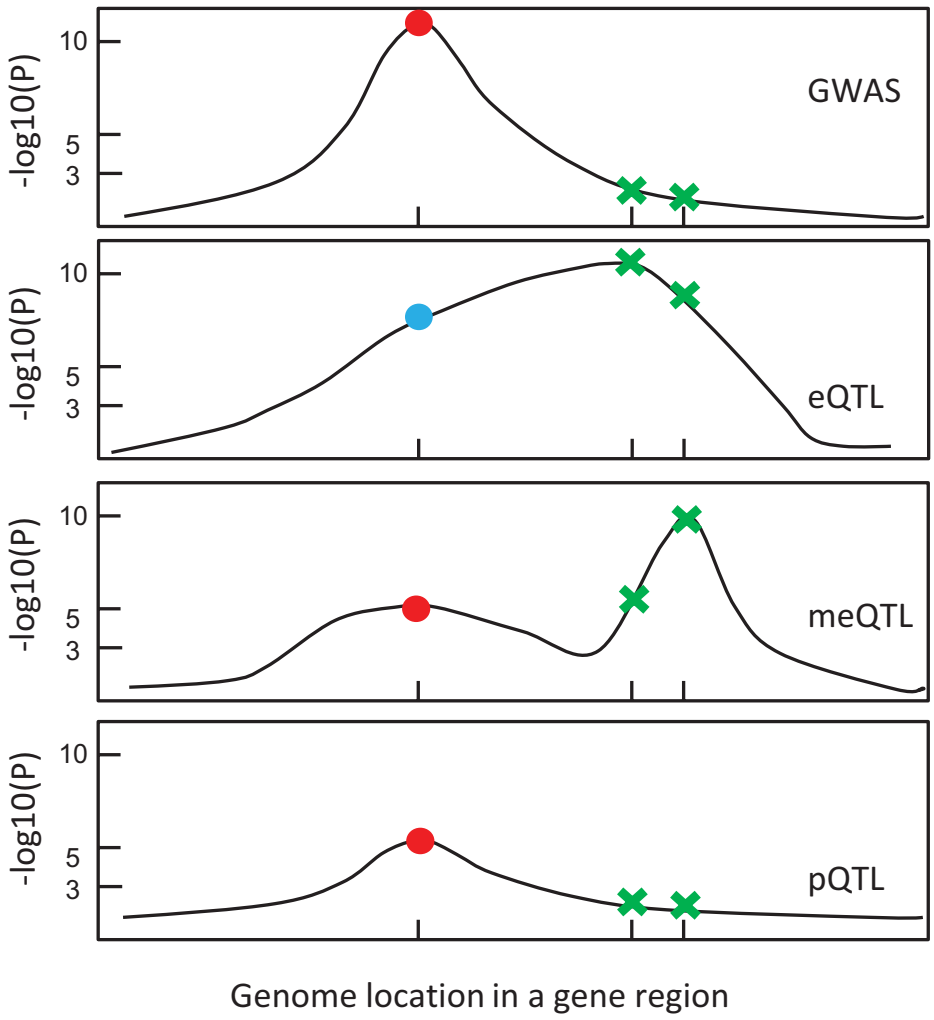


Figure 4.2: **Conceptual illustration of the conditional association analysis of Primo.** Consider a joint analysis of GWAS summary statistics and summary statistics of eQTL, meQTL and pQTL. In a gene region harboring trait-associated SNPs, there is a GWAS SNP of interest (red/blue dot) and two other confounding SNPs – the lead SNPs for eQTL and meQTL (green cross). Before conditional association analysis, the GWAS SNP is estimated to be associated with cis expression, methylation and protein levels. After adjusting for the two lead omics SNPs, the GWAS SNP is no longer associated with cis expression levels (blue dot) but is still estimated to be a me- and pQTL.

4.3 Simulations

4.3.1 Comparison with existing methods for jointly analyzing associations to three traits

In Scenario 1, we simulated genotypes and phenotypes with pairwise sample correlations of 0.2 among $J = 3$ studies for 1 million SNPs. The proportions of SNPs associated with only one, exactly two, and all of the three traits were 5×10^{-3}, 5×10^{-4}, and 5×10^{-4}, respectively. Non-zero effect sizes were simulated from a $N(0, \sigma^2)$ distribution, with $\sigma^2 = 0.25, 0.5$ and 1.0 each in one third of gene regions. We then calculated the SNP-level test statistics and P-values as input for Primo.

Here we compared the true and estimated FDRs and power to detect associations to all three traits and to at least one trait, based on Primo versus two competing methods, "moloc" [Giambartolomei et al., 2018] and Fisher's method [Fisher, 1918]. The results with correctly specified, under-specified (by 10-fold) and over-specified (by 10-fold) marginal non-null proportions (θ_j^1's) are shown in Table 4.1. When θ_j^1's are well-specified (Scenario 1a in Table 4.1), Primo nicely controlled the FDR even in the presence of unknown study/sample correlations – highlighting one advantage of Primo in integrating potentially correlated multi-omics data. Note that moloc and Primo are not directly comparable as moloc aims to assess whether three traits of interest share a causal variant in a gene region, while Primo first identifies SNPs' joint associations to multiple traits and then reduces spurious associations due to LD. Nevertheless, we show comparisons between Primo and moloc in the simulated setting. Since moloc does not output the posterior probabilities for all SNPs in every association pattern, we are only able to compare the power and FDR of Primo versus moloc in detecting associations to all three traits. We observed that Primo generally enjoys substantial power improvement, which is not surprising because the goal of moloc is more restrictive. As shown in Table 4.1, the estimated FDR (estFDR) is very close to the true FDR for Primo. Fisher's method, as a combination method for testing omnibus hypotheses, can only be used to detect

SNPs with associations to at least one trait, and is not applicable to detect associations to all traits. The estimated FDR [Storey and Tibshirani, 2003; Storey et al., 2015] for Fisher's method based on nominal P-values are not well controlled due to correlations among test statistics, as expected. At similar power levels, the FDRs observed across simulations of Fisher's method are also much higher than those of Primo.

In this simulation, the true θ_j^1's are 2.5×10^{-3}. In Scenario 1b, we under-specified θ_j^1 to be $\theta_j^1/10$. As shown in Table 4.1, although power might decrease to some extent, the FDRs are reasonably controlled. In Scenario 1c, when θ_j^1's are over-specified by an order of magnitude as $10 \times \theta_j^1$, we observed slightly inflated FDRs. As such, we suggest to obtain reasonable estimates for θ_j^1's based on the current data and the literature, or under-specify θ_j^1's to be more conservative. When θ_j^1's are correctly or under-specified in a certain range, Primo is robust to parameter specification.

Table 4.1: **Simulation results evaluating the performance of Primo in comparison to other methods.**When $J = 3$ with correlated samples, we compared Primo versus moloc and Fisher's method in detecting associations to at least 1 trait and associations to all traits and when parameters are correctly, under- and over-specified. PP := posterior probability; estFDR := estimated FDR.

Scenario	Method	Association to at least one trait						Association to all three traits					
		PP $\geq$ 0.90			PP $\geq$ 0.80			PP $\geq$ 0.90			PP $\geq$ 0.80		
		true FDR (%)	estFDR (%)	Power (%)	true FDR (%)	estFDR (%)	Power (%)	true FDR (%)	estFDR (%)	Power (%)	true FDR (%)	estFDR (%)	Power (%)
1a	Primo (t)	0.3	0.2	67.3	0.6	0.5	68.5	1.1	0.9	46.8	2.4	2.2	50.9
	Primo (P)	0.2	0.2	66.3	0.5	0.5	67.5	0.6	0.7	44.2	1.2	1.4	47.6
	moloc	-	-	-	-	-	-	0.1	1.1	13.9	0.3	1.9	14.8
		5% estFDR			10% estFDR								
	Fisher's	11.6	-	69.3	19.9	-	70.8	-	-	-	-	-	-
1b	Primo (t)	0.1	0.2	66.0	0.2	0.5	67.1	0.4	0.8	43.2	1.0	1.9	46.6
	Primo (P)	< 0.01	< 0.01	56.1	< 0.01	< 0.01	56.8	< 0.01	< 0.01	23.3	< 0.01	< 0.01	24.1
	moloc	-	-	-	-	-	-	0.2	0.8	14.9	0.4	1.7	15.9
1c	Primo (t)	1.4	0.3	69.4	3.5	1.0	71.0	4.4	1.3	49.1	9.5	3.4	55.1
	Primo (P)	1.0	0.3	67.2	3.7	1.3	69.6	0.2	1.0	39.9	1.0	2.2	46.5
	moloc	-	-	-	-	-	-	< 0.01	4.7	12.0	< 0.01	5.2	12.7

4.3.2 Evaluation of the performance of Primo conditional association analysis accounting for LD and sample correlations

In this section, we simulated association statistics for correlated SNPs in moderate to high LD and evaluated the performance of the proposed conditional association approach in the presence of LD. To simulate genotype data with a realistic LD structure, we used the sim1000G package [Dimitromanolakis et al., 2019] to simulate 1 million variants for 1000 subjects using chromosomes 8, 9 and 10 in the CEU 1,000 Genomes population [Auton et al., 2015]. We divided the genotypes into regions of 1000 consecutive SNPs in order to form gene regions. Within each region, we randomly selected one SNP with MAF > 0.1 to be the "known trait-associated SNP" and randomly selected two "confounding SNPs" in moderate to strong LD with both the trait-associated SNP and each other (pairwise $r \in [0.5, 0.8]$). Within each gene region, we then generated $J = 4$ traits. The first trait is a "complex trait" for all 1000 subjects and the three other traits are "omics traits" with sample sizes of 500, 300, and 200, respectively, resampled from the 1000 subjects. In 20% of the LD blocks the true underlying association pattern for the trait-associated SNP is (1,0,1,0) while the association patterns for the two confounding SNPs are (1,1,0,0) and (0,1,1,1), respectively. These LD blocks represent gene regions with no SNP truly associated with all traits but with multiple SNPs in LD with different association patterns. The effect sizes in these blocks ranged from $0.1 - 0.4$. We further simulated another 20% of LD blocks where the true underlying association pattern for the trait-associated SNP is (1,1,1,1) while the association patterns for the two confounding SNPs are (0,1,0,0) and (0,0,1,0), respectively. These LD blocks represent gene regions with one true causal SNP associated with all traits as well as two confounding SNPs in high LD with it. For the remaining 60% of the LD blocks, no SNPs are associated with any traits. Then we obtain the single-variant association statistics $\mathbf{T}$ for 1 million SNPs with $J = 4$ traits.

We applied Primo with $\mathbf{T}$ as input to identify SNPs associated with all traits. For each index SNP detected as significant at the probability cutoffs of 0.8 and 0.9, we further

conducted conditional association analysis, conditioning on its two confounding SNPs in moderate to high LD. The trait-associated SNPs that no longer have the highest probabilities in the pattern of $(1, 1, 1, 1)$ after conditional association analysis were not considered to be positive findings. In the calculations of the FDRs, we use the same denominators before and after conditional association analysis for fair comparison. That is, the denominators are the number of identified SNPs with associations to all traits at a given cutoff before the conditional associating analysis. After conditional analysis, the numerator (i.e. # false positive) of the true FDR is the number of SNPs that are not truly associated to all traits, yet continue to show the highest probability in the pattern of $(1, 1, 1, 1)$ after conditional association analysis. In the calculation of the numerator of the estimated FDR, for each SNP i that is no longer significant after conditional analysis, its contribution to the numerator $(1 - \hat{P}_i)1(\hat{P}_i \geq \lambda)$ in the formula (2.3) is corrected to be 1 since we considered it as an estimated false discovery.

Table 4.2 summarizes the results over 100 simulations. As shown in the table, when SNPs are in LD, we observed some slightly inflated FDRs without conditional association analysis even when θ_j^1's are correctly specified (Scenario 2a). In contrast, after accounting for LD, true FDRs are reduced and are well-controlled by the estimated FDRs. In Scenario 2b and 2c, we under-specified and over-specified θ_j^1's by 10 fold. Overall, Primo after conditional association analysis could yield nice control of FDR and maintain good power when θ_j^1's are correctly or under-specified.

Table 4.2: **Comparison of Primo results before and after conditional association analysis.** PP := posterior probability

Scenario	PP $\geq$ 0.9						PP $\geq$ 0.8					
	Before accounting for LD			After accounting for LD			Before accounting for LD			After accounting for LD		
	True FDR(%)	estFDR (%)	Power (%)	True FDR(%)	estFDR (%)	Power (%)	True FDR(%)	estFDR (%)	Power (%)	True FDR(%)	estFDR (%)	Power (%)
2a	5.1	2.5	71.3	4.1	4.7	70.3	8.7	4.6	82.8	6.4	8.6	80.8
2b	3.8	2.4	67.4	2.9	5.6	65.7	7.0	4.6	79.2	4.7	10.1	75.9
2c	7.6	2.7	76.4	6.9	3.4	76.3	13.6	5.0	88.5	11.8	7.0	88.0

4.4 Data applications

4.4.1 Description of studies and data

The Genotype-Tissue Expression (GTEx) Project

The Genotype-Tissue Expression (GTEx) project was established to study tissue-specific gene expression and regulation. The GTEx V8 dataset contains data from 948 donors and 17,382 samples from 52 tissues and two cell lines [Aguet et al., 2019]. QTL analyses were conducted using RNA-seq and genotype data from 838 donors and 15,201 samples from 49 tissue types. GTEx samples underwent whole genome sequencing (WGS) at a median depth of 32x on Illumina HiSeq 2000 or Illumina HiSeq X. Additional details about the genotyping pipeline and sample and variant quality control have been reported elsewhere [Aguet et al., 2019]. GTEx RNA sequencing was performed using the Illumina TruSeq^{TM} RNA sequencing platform. Raw sequence data were processed using the Broad's Picard pipeline [Broad Institute, 2019]. Data was aligned using STAR (v2.5.3a) [Dobin et al., 2013]. RNA-SeQC [DeLuca et al., 2012] was used for quality control and gene-level expression quantification, and TMM [Robinson and Oshlack, 2010] was used to normalize read counts. Additional details on the RNA-Sequencing pipeline and processing are reported elsewhere [Aguet et al., 2019].

GTEx *cis*-eQTL summary statistics were generated by linear regression as implemented in FastQTL [Ongen et al., 2016], adjusting for sex, genotyping platform, WGS library construction protocol (PCR-based or PCR-free), five genotype principal components (PCs), and up to 60 PEER [Stegle et al., 2012] variables.

The Cancer Genome Atlas (TCGA)

The Cancer Genome Atlas (TCGA) project is described in Section 2.4.1. Genotype data and clinical covariates for TCGA breast cancer subjects were downloaded through the Ge-

nomic Data Commons (GDC) Data Portal [Grossman et al., 2016]. TCGA subjects were genotyped on the Affymetrix Genome-wide Human SNP Array 6.0. Germline genotypes were measured in blood-derived DNA samples primarily. For subjects missing genotyping from blood samples, we used genotype measured in solid normal tissue as a surrogate. Restricting to bi-allelic variants on autosomes yielded 859,193 SNPs. After removing subjects with duplicate blood genotypes that did not match (i.e. labeling problems), there were 1094 subjects remaining. We used IMPUTE2 [Howie et al., 2009] to conduct genotype imputation using 1000 Genomes as the reference panel (phase3 v5) [Auton et al., 2015]. We performed 30 MCMC iterations, discarding the first 10 as burn-in, using 1 MB intervals for inference. SNPs with an imputation info score < 0.3 or with a minor allele frequency (MAF) < 0.01 were removed post-imputation.

Gene expression, protein abundance, and DNA methylation data for TCGA subjects were downloaded using TCGA-Assembler 2 [Wei et al., 2018]. RNA sequencing was performed in tumor tissues using the Illumina HiSeq 2000 RNA Sequencing platform. RNA-Seq expression levels were quantified using HTSeq-count [Anders et al., 2015]. DNA methylation was measured in tumor tissue samples using the Infinium HumanMethylation450 BeadChip. The level of methylation at each CpG site was measured as a β value ranging from 0 (completely unmethylated) to 1 (completely methylated). Protein abundance was measured in tumor tissue samples using iTRAQ (isobaric tag for relative and absolute quantitation) mass-spectometry (MS) in experiments conducted by the Clinical Proteomic Tumor Analysis Consortium (CPTAC) [Ellis et al., 2013]. The protein abundance was measured using the Log Ratio (i.e. the log of the ratio between the spectral count of a protein in a sample versus the spectral count of the protein in the reference sample). We restricted our analysis of protein abundance to the 77 high-quality samples as identified by Mertins et al. (2016). We further restricted the analysis to the 74 of these 77 samples from female subjects with measured tumor purity scores [Aran et al., 2015].

Prior to QTL analyses, expression, methylation and protein measurements for TCGA

samples were transformed to the quantiles of the standard normal distribution (separately for each gene, CpG site or protein). Imputed genotypes were converted to expected counts of the alternative allele. We generated *cis*-QTL summary statistics using linear regression as implemented in Matrix eQTL [Shabalin, 2012], adjusting for subject covariates. QTL analyses were restricted to female subjects. Covariates included tumor purity scores [Aran et al., 2015], cancer stage, histological subtype (infiltrating ductal, infiltrating lobular, mucinous, metaplastic, mixed histology or other), estrogen receptor (ER) and progesterone receptor (PR) status, and genotype principal components (15 PCs for expression and methylation; 3 PCs for protein). We generated genotype PCs in PLINK 1.90 [Chang et al., 2015; Purcell and Chang, 2017] using measured bi-allelic variants on autosomes with the following filters: minor allele frequency ≥ 0.05; Hardy-Weinberg Equilibrium p-value ≥ 0.0001; pairwise linkage disequilibrium $R^2 \leq 0.2$.

Cis associations were defined as: $<$ 250 kb apart from transcription start site (TSS) for expression and protein; $<$ 50 kb apart from a CpG site for methylation. For GWAS-reported SNPs that were missing based on these criteria, we selected the nearest gene (based on distance between the SNP and transcription start site) for which both gene expression and protein abundance levels were available (Application I in the main text) or for which gene expression levels were measured in both tissues (Application II in the main text). For these regions, we included associations $<$ 1 Mb apart from the transcription start site. Note that using different definitions of cis window sizes may have yielded slight differences in results. For genes in regions of GWAS-reported SNPs missing protein or methylation data, and for GWAS-reported SNPs that could not be mapped to a GTEx variant, a t-statistic of 0 was added for the missing data to allow integration of the non-missing test-statistics.

Genome-wide association studies (GWAS)

For summary statistics of genetic associations with complex traits (such as disease susceptibility), we used data from several publicly available genome-wide association studies

(GWAS). Breast cancer susceptibility GWAS summary statistics were obtained from the Breast Cancer Association Consortium (BCAC) [Michailidou et al., 2017]. Height and body mass index (BMI) GWAS summary statistics were obtained from the GIANT consortium [Wood et al., 2014; Locke et al., 2015]. GWAS summary statistics for Ulcerative Colitis and Crohn's disease were obtained from the International Inflammatory Bowel Disease (IBD) Genetics Consortium [Liu et al., 2015]. GWAS summary statistics for LDL and triglycerides were obtained from a meta-analysis of circulating metabolites quantified by nuclear magnetic resonance metabolomics [Kettunen et al., 2016] and from the Global Lipids Genetics Consortium (GLGC) [Willer et al., 2013]. For purposes of replication, summary statistics were obtained for GWAS of several traits performed in the UK Biobank [Churchhouse and Neale, 2017].

4.4.2 Application I: Understanding the mechanisms of breast cancer susceptibility loci

With over 100,000 breast cancer cases and a similar number of controls, BCAC has recently reported 174 common genetic variants associated with breast cancer risk [Michailidou et al., 2017]. In order to understand the underlying mechanisms of those susceptibility risk loci and their potential cis target genes, a recent study conducted cis-eQTL analysis using both normal and tumor breast transcriptome data and identified multiple genes likely to play important roles in breast tumorgenesis [Guo et al., 2018].

In addition to transcription, SNPs may affect cis- epigenetic features, protein abundances, and other omics traits. Functional relationships may exist among those omics traits. Therefore, we propose to jointly examine the susceptibility risk loci and their effects on multiple omics traits in tumor and normal tissues in order to better understand the mechanisms through which risk-associated SNPs act in different conditions. Moreover, this analysis will enhance our understanding of the regulatory cascade and their roles in breast tumorigenesis. The regulatory SNPs with "cascading effects" [Battle et al., 2015; Pai et al., 2015] on gene

regulation and downstream gene products are of particular interest.

In this work, we applied Primo to integrate GWAS summary statistics from BCAC with the eQTL, meQTL, and pQTL association summary statistics obtained from 1012, 762, and 74 breast tumor samples, respectively, from TCGA [The Cancer Genome Atlas Network, 2012] and CPTAC [Mertins et al., 2016], and eQTL summary statistics obtained from 396 normal breast mammary samples from GTEx [Aguet et al., 2019]. A total of 162 of the GWAS SNPs reported by Michailidou et al. (2017) reached genome-wide significance ($P < 5 \times 10^{-8}$) in the meta-analysis. And there are 158 of these SNPs with MAF $> 1\%$ in TCGA data. And the 158 breast cancer GWAS SNPs are the SNPs we examined for mechanistic interpretations, while we used genome-wide summary statistics from all SNPs to obtain estimations of key parameters. Note that one SNP could be mapped to multiple genes and multiple CpG sites. We assessed the probabilities of 32 (2^5, for GWAS and 4 omics QTLs) association patterns for each SNP-gene-CpG-protein quartet. In the conditional association analysis of gene regions harboring at least one GWAS SNP, we selected the lead SNP for each omics trait in the region and adjusted for any lead SNP outside a 5kb distance of and with LD $R^2 < 0.9$ with the GWAS-reported SNP (those with $R^2 > 0.9$ or within 5kb were considered likely to share a causal variant or too close to assess individual associations, respectively).

At the 80% probability cutoff and after conditional association analysis (estimated FDR of 4.2, 9.6, 20.2 and 13.2%), there were 52, 26, 9 and 1 GWAS SNPs out of 158 examined being associated with at least 1, 2, 3 or 4 omics traits, respectively. The three GWAS SNPs (rs11552449, rs3747479, and rs73134739) in the three genes (*DCLRE1B*, *MRPS30*, and *ATG10*, respectively) reported in Guo et al. (2018) had high probabilities of being an eQTL in both tumor and normal tissues (with probabilities of 61.1, 95.6, and >99.9%, respectively). In the *KLHDC7A* gene region, the GWAS SNP rs2992756 (indicated by red dot in Figure 4.3) is associated with the expression, methylation and global protein abundance levels of the cis-gene *KLHDC7A*. Figure 4.3 shows the plot of $-\log_{10}(P)$-values of associations to breast cancer risk and the three omics traits (with expression traits in both tumor and normal

tissue types) of *KLHDC7A* for the SNPs in the gene region. Note that the GWAS SNP is only moderately associated with the gene expression levels in the normal GTEx breast tissue with a P-value of 0.0034, highlighting the need to study omics QTLs under different conditions.

Due to limited sample sizes (74) in the pQTL analysis, only 1 out of the 158 examined breast cancer susceptibility loci was associated with cis-protein abundance levels with high confidence. However, as shown in Figure 4.4, the cis-gene expression levels and cis-protein abundances for those loci were often highly correlated, with an averaged (Pearson) correlation coefficient of r=0.396 and a median of r=0.411.

There were 16 out of 158 susceptibility loci uniquely associated with cis-methylation levels but not expression levels in either tumor or normal tissue, echoing a recent work showing both unique and shared causal mechanisms of epigenome variations and transcription [Pierce et al., 2018]. We analyzed the CpG targets of meQTLs identified by Primo for enrichment in several genomic features after performing annotation using publicly available datasets. Exons and introns were annotated using the refGene database provided through the UCSC Table Browser [Karolchik et al., 2004]. Promoter regions for genes were defined as the regions 0–1500 bases upstream of the TSS. Enhancer regions were annotated using H3K4me1 and H3K27ac histone marks in human mammary epithelial and breast myoepithelial cells using data from the Roadmap Epigenomics Project [Kundaje et al., 2015] and the Encyclopedia of DNA Elements Project (ENCODE) [Dunham et al., 2012]. We downloaded the call sets from the ENCODE portal (https://www.encodeproject.org/) [Davis et al., 2018] with the following identifiers: ENCFF001SWW, ENCFF001SWZ. To test for enrichment of features among the CpG targets, we used 10,000 bootstrapped samples of all CpG sites on the 450k array and calculated P-values using the proportion of bootstrapped samples with more extreme counts than were observed in the CpG targets identified by Primo. As shown in Figure 4.5, CpG targets of multi-omics QTLs (breast cancer susceptibility loci associated with methylation as well as gene expression and/or protein abundance) were enriched in

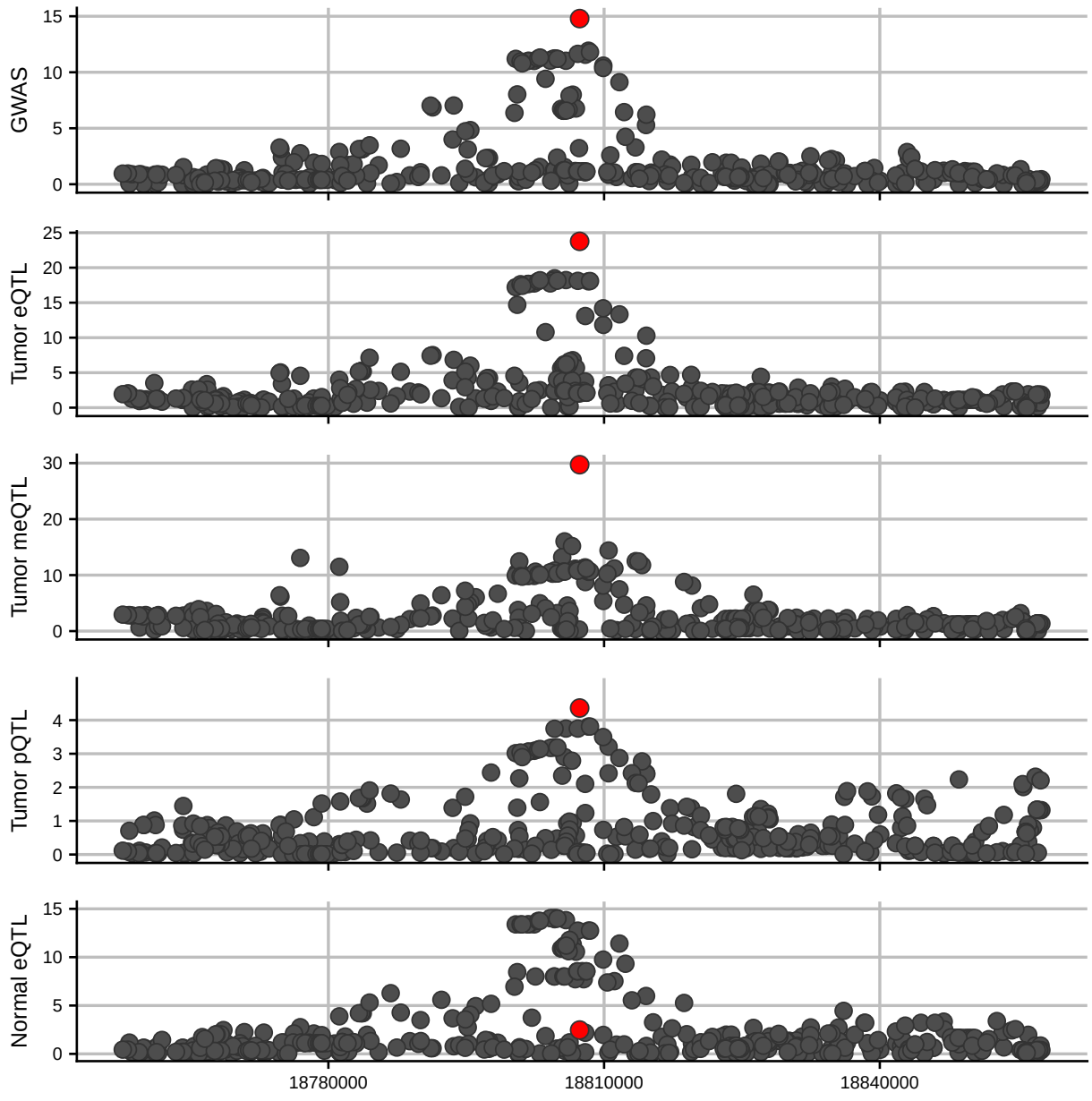


Figure 4.3: **Example of a known breast cancer susceptibility locus being associated with multi-omics traits.** At a posterior probability threshold of 80%, Primo identified SNP rs2992756 as being associated with all four omics traits for the gene *KLHDC7A*. Here shows the $-\log_{10}(P)$-values by position on Chromosome 1 in the region of the gene *KLHDC7A* for all SNPs including the breast cancer susceptibility locus (rs2992756, red dot) in GWAS (top panel) and eQTL, meQTL and pQTL analyses in tumor tissue (the next three panels, respectively) and eQTL analysis in normal tissue (bottom panel) for the gene and protein KLHDC7A and CpG site cg05040210.

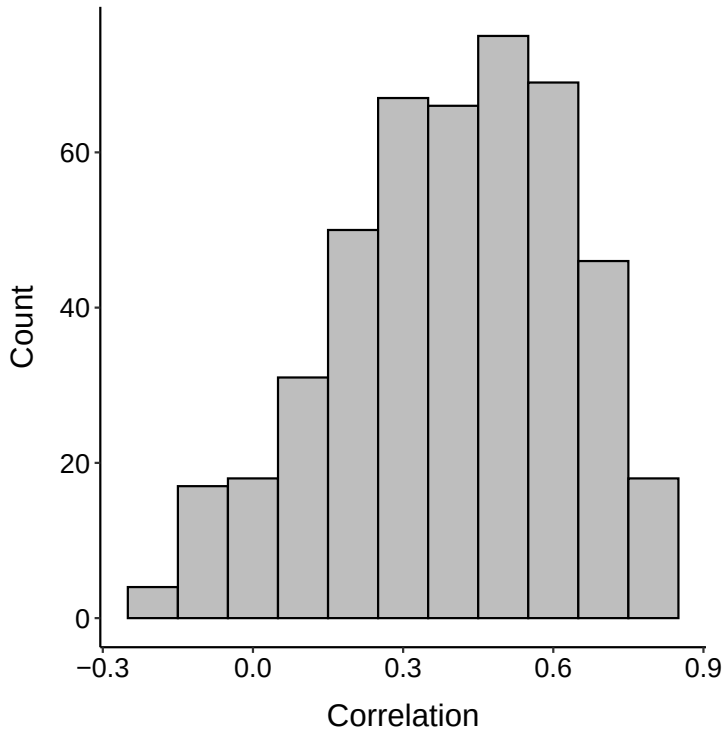


Figure 4.4: **Correlations between gene expression and protein abundance in breast cancer susceptibility loci.** As shown in the histogram, cis-gene expression levels and cis-protein abundances were often highly correlated in these loci.

CpG Island Shores ($P < 0.05$) and depleted in Open Seas ($P < 0.01$). CpG targets of multi-omics QTLs were enriched in exons ($P < 0.01$) while CpG targets of meQTL-only loci were enriched in introns ($P < 0.001$). In promoter regions, CpG targets of multi-omics QTLs were enriched ($P < 0.01$) while CpG targets of meQTLs not also associated with gene expression levels were depleted ($P < 0.001$), consistent with the involvement of promoter regions in transcription. This also shows that the integration of GWAS and multi-omics traits can provide additional insights in understanding the complex and dynamic mechanisms.

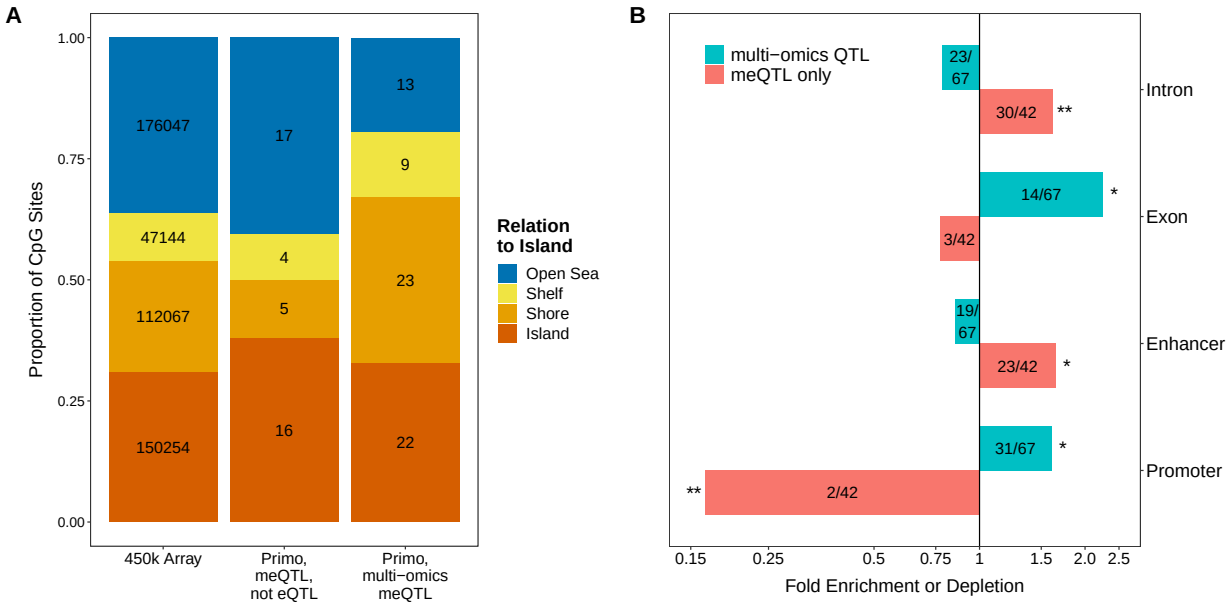


Figure 4.5: **Enrichment of CpG targets of breast cancer susceptibility loci among genomic features.** (A) Distribution with relation to islands of CpG targets of Primo-identified multi-omics QTLs and meQTL-only susceptibility loci compared with distribution of all CpGs on 450k array. Numbers represent counts of CpGs in each relationship to islands. B) Fold enrichment or depletion of genomic features among CpG targets of multi-omics QTL (cyan) and meQTL-only (pink) susceptibility loci. Feature counts out of total CpG targets are displayed within each bar. X-axis displayed on log scale. P-values were obtained by bootstrapping samples of CpGs from the full 450k array [$p < 0.01$ (*), $p < 10^{-3}$ (**)].

4.4.3 Application II: Detecting SNPs with pleiotropic effects and elucidating their mechanisms

Many genetic variants are associated with more than one complex trait [Solovieff et al., 2013; Sivakumaran et al., 2011; Pickrell et al., 2016]. Identifying such pleiotropic variants and elucidating the molecular mechanisms which underlie these multi-trait associations may enhance our understanding of the etiology of complex traits and provide additional insights into clinical treatment development [Solovieff et al., 2013]. In this section, we applied Primo to detect SNPs with pleiotropic effects to two complex traits in gene regions harboring susceptibility loci for at least one trait, and provide mechanistic interpretations by integrating pairs of publicly available complex-trait GWAS summary statistics with eQTL association summary statistics obtained from trait-relevant tissue types in the GTEx project [Aguet et al., 2019].

Height and BMI

We applied Primo to height [Wood et al., 2014] and body mass index (BMI) [Locke et al., 2015] GWAS summary statistics from the GIANT consortium (sample size $> 250,000$) with eQTL summary statistics in subcutaneous adipose ($n = 581$) and skeletal muscle ($n = 706$) tissues from GTEx for all SNPs in the genome. There are 697 height-associated SNPs reported by Wood et al. (2014) and 97 BMI-associated SNPs reported by Locke et al. (2015). Out of those SNPs reaching genome-wide significance (5×10^{-8}) for either trait, 683 were present in both sets of GWAS summary statistics and could be mapped to GTEx SNPs in cis with at least one gene measured in both tissue types. Of the 683 SNPs, 612 reached genome-wide significance for height and 78 reached genome-wide significance for BMI, with 7 reaching genome-wide significance for both. Those 683 GWAS SNPs are the SNPs of interest in our analysis of pleiotropy, while again we estimated key parameters used in Primo using genome-wide summary statistics. At the 80% probability cutoff and after conditional

association analysis accounting for LD, 32 SNPs out of 683 were detected by Primo as being associated with both complex traits (estimated FDR of 17.5%). Of these, 17 were associated with expression of at least one gene in at least 1 tissue (estimated FDR of 21.8%) and 12 were associated with expression of at least one gene in both tissues (estimated FDR of 18.4%). Furthermore, 12 of the SNPs were associated with the expression of multiple genes, highlighting the possibility that pleiotropic SNPs may affect multiple complex traits through their co-regulation of multiple genes.

To validate the 32 identified pleiotropic SNPs being associated with both height and BMI regardless of association status to cis-gene expression levels, we used GWAS summary statistics from the UK Biobank Churchhouse and Neale (2017) ($> 336k$ samples have both height and BMI measured) as a replication study. At $P < 0.0008$ (the Bonferroni threshold is calculated as $0.05/(32\times2)$, since there are two traits), 27 out of the 32 SNPs were associated with both traits in the UK Biobank, including 16 of the 17 SNPs that were also associated with gene expression. Plots of $-log_{10}(P)$-values for associations with height, BMI and expression in each tissue are presented in Appendix C for the genomic regions containing the 27 replicated SNPs.

Ulcerative Colitis and Crohn's Disease

We applied Primo to integrate GWAS summary statistics of Crohn's disease and ulcerative colitis from a study of over 20,000 samples of European Ancestry conducted by the International Inflammatory Bowel Disease (IBD) Genetics Consortium [Liu et al., 2015] with eQTL summary statistics from sigmoid colon ($n = 318$) and transverse colon ($n = 368$) tissues from GTEx [Aguet et al., 2019]. Of the 232 SNPs reported in the initial meta-analysis, 67 SNPs have reached genome-wide significance for at least one of Crohn's disease or ulcerative colitis and could be mapped to GTEx SNPs in cis with at least one gene measured in each tissue. At the 80% probability cutoff (estimated FDR of 0.8%, 5.8% and 6.4%) and after conditional association analysis accounting for LD, 37, 15 and 11 of the 67 SNPs were associated with

both complex traits, both complex traits plus gene expression in at least 1 tissue, and both complex traits plus gene expression in both tissues, respectively. We used GWAS summary statistics of self-reported Crohn's disease and self-reported ulcerative colitis from the UK Biobank [Churchhouse and Neale, 2017] to replicate our findings. At $P < 0.0007$ ($0.05/(37 \times 2)$), 4 of the 37 SNPs were associated with both traits in the UK Biobank. The relatively low replication rate might be due to the fact that in the UK Biobank data, self-reported disease status was used for both traits and the numbers of cases of Crohn's disease and ulcerative colitis are low (< 2000 and < 3000, respectively).

LDL and Triglycerides

We applied Primo to integrate GWAS summary statistics of total cholesterol in low-density lipoprotein (LDL) and trigycerides in intermediate-density lipoprotein (IDL) from a meta-analysis of 123 circulating metabolic traits (sample sizes of 13,527 for LDL and 21,559 for triglycerides, respectively) [Kettunen et al., 2016] with eQTL summary statistics from visceral omentum adipose ($n = 469$) and liver ($n = 208$) tissues. Of the 74 variants that Kettunen, et al. [2016] reported were associated with at least one circulating metabolite trait, 15 reached genome-wide significance for at least one of LDL or triglycerides in IDL and could be mapped to GTEx variants in cis with at least one gene measured in both tissues. At the 80% probability cutoff and after conditional analysis accounting for LD, all 15 variants were associated with both complex traits (estimated FDR of 1.6%). 5 variants were associated with gene expression in at least one tissue (FDR 12.0%), and 2 were associated with gene expression in both tissues (FDR 12.8%). To replicate our findings, we used GWAS summary statistics of LDL and triglycerides from the Global Lipids Genetics Consortium (GLGC) [Willer et al., 2013]. Note that this might not be an ideal replication study as there is substantial sample overlap between the discovery and replication cohorts. Nevertheless, at $P < 0.0017$ ($0.05/(15 \times 2)$), 4 of the 15 variants were associated with both traits in the GLGC study.

Summary of Pleiotropy Analyses

The analyses we conducted showed that Primo can be used to detect SNPs with pleiotropic effects on (potentially more than two) complex traits while simultaneously providing mechanistic interpretations by examining their effects on cis-gene expression levels in trait-relevant tissue types. A majority of our detected and replicated pleiotropic SNPs do not have associations reaching genome-wide thresholds for both traits. Our analyses and results underscored the value of integrating GWAS summary statistics of multiple traits with eQTLs in relevant tissue types.

4.5 Discussion

In the current work, we made a tailored development of Primo to comprehensively elucidate the molecular mechanisms of known complex-trait-associated SNPs, where we assessed the omics- or other trait-associations of known complex-trait-associated SNPs by conducting conditional association analysis in gene regions harboring known trait-associated SNPs to account for LD with other SNPs in the region. Note that in our analyses, we focused on known trait-associated SNPs reported in GWAS.

With the rapidly increasing availability of GWAS and omics QTL association summary statistics from different studies, populations, and cellular contexts, it is commonly observed that there could be multiple causal SNPs for different complex and omics traits in the same gene regions. Conducting integrative analysis of GWAS summary statistics and 1-2 sets of omics QTL statistics may provide only a partial view of the genomic activities in a region; meanwhile, if multiple omics QTL statistics are jointly analyzed, one also needs to consider the associations identified by chance and perform multiple testing adjustment. The advantage of Primo is that it can integrate a moderate to large number of sets of summary statistics from different data sources as input to provide a more comprehensive evaluation while also considering multiple testing adjustment. Additionally, Primo enjoys

other unique advantages and shows great flexibility in integrative analysis. It allows the input summary statistics to be from independent, or partially overlapped studies with unknown study correlations. It detects SNPs with coordinated effects allowing different effect sizes (and different directions of effect sizes) on different types of traits. It can also integrate one-sided P-values if the same direction of effect sizes is expected and desired. Primo can identify SNPs in different combinations of association patterns to molecular omics and complex traits. Moreover, with the conditional association analysis of Primo, we can move one step beyond association towards causation by assessing whether a GWAS SNP is also an omics QTL while adjusting for the effects of multiple lead SNPs in a gene region. The conditional association analysis can reduce spurious omics-trait associations of GWAS SNPs due to LD with the lead omics SNPs.

As described in Sections 2.2.2 and 2.2.5, we implemented two versions of Primo taking either t-statistics (or effect sizes and standard error estimates) or P-values as input. Primo is computationally very efficient and can analyze the joint associations of 30 million SNPs to five traits in dozens of minutes. We applied Primo to examine and interpret the associations to omics traits in tumor/normal tissues for known breast cancer susceptibility loci. We also applied Primo to integrate pairs of GWAS summary statistics of complex traits with eQTL summary statistics from trait-relevant tissue types from GTEx to detect pleiotropic effects and examine their mechanisms.

There are a few additional points we would like to emphasize. First, we recommend a stringent specification of the marginal study-specific alternative proportion parameters (θ_j^1's), especially when there is limited *a priori* knowledge guiding the parameter specification. Primo may suffer from slightly inflated FDR when those parameters are highly over-specified; whereas when those parameters are under-specified to an extent, there might not be much power loss. Second, the focus of the current work is to comprehensively evaluate the molecular mechanisms of known trait-associated SNPs, rather than to identify new causal SNPs for complex traits from other regions in the genome. When applying Primo in

other integrative association analyses, the interpretations of results may be different. Third, there are many existing functional annotations for SNPs that are not incorporated in the current version of Primo but have also proved to be useful. We will explore this direction in future work.

Primo was motivated by the analysis of multi-omics data, in particular the analysis of trait-associated SNPs for their molecular trait-associations. It should be noted that Primo can also be broadly applied to many other settings when data integration is needed. Primo can be used to detect associations repeatedly observed in multiple correlated or independent conditions, and those repeatedly observed associations may enhance the confidence for new discoveries, or at least provide a more comprehensive examination of how those associations may occur in different conditions.

CHAPTER 5

A ROBUST TWO-SAMPLE MENDELIAN RANDOMIZATION METHOD INTEGRATING GWAS WITH MULTI-TISSUE EQTL SUMMARY STATISTICS

5.1 Introduction

For more than a decade, genome-wide association studies (GWAS) have uncovered tens of thousands of unique associations between single nucleotide polymorphisms (SNPs) and complex diseases/traits [Buniello et al., 2019]. In the post-GWAS era, the next major challenge is to further understand the biological mechanisms underlying the observed associations and identify clinically actionable risk factors for various complex diseases/traits. Most of the disease/trait-associated SNPs have small effect sizes and reside in non-coding regions with unknown functions [Visscher et al., 2017; Maurano et al., 2012]. In order to elucidate their mechanisms and functions, many efforts have been made to integrate GWAS summary statistics with other information (e.g., eQTL statistics) and to identify genetically-regulated risk factors (e.g., gene expression levels) for complex diseases. Those methods include transcriptome-wide association studies (TWAS) [Gamazon et al., 2015; Zhu et al., 2016; Gusev et al., 2016; Barbeira et al., 2019], colocalization analyses [Giambartolomei et al., 2014, 2018; Wen et al., 2017; Hormozdiari et al., 2016], two-sample Mendelian Randomization (MR) analysis [Bowden et al., 2015; Qi and Chatterjee, 2019; Zhao et al., 2019, 2020] and others.

Compared to other integrative genomic analyses, MR analysis has its unique advantages. It steps beyond association towards causation, aiming to identify modifiable risk factors (exposures) for complex diseases while allowing unmeasured confounders affecting both exposures and disease outcomes of interest. Specifically, MR methods consider SNPs with known associations with an exposure of interest as instrumental variables (IVs) [Lawlor et al., 2008;

Smith and Ebrahim, 2003; Schadt et al., 2005; Chen et al., 2007]. Since SNP genotypes were 'Mendelian Randomized' from parents to offspring during meiosis, they are assumed to be generally unrelated to external confounders. Under certain assumptions, SNPs can be used as IVs to estimate and test for the causal effects of an exposure on a disease outcome from observational data. Two-sample MR methods refer to the MR methods requiring only two sets of summary statistics, IV-to-exposure and IV-to-outcome association statistics from two independent sets of samples, and thus are widely used to recapitalize on existing summary statistics.

Traditional MR methods imposed strong assumptions on the validity of IVs [Angrist et al., 1996]. A valid IV is a genetic variant that only affects the complex disease through the exposure of interest (no direct effect) and is independent of unmeasured confounders of the exposure and the disease outcome [Burgess et al., 2015]. That is, there is no 'horizontal pleiotropy' [Lawlor et al., 2008] (a phenomenon where a genetic variant also affects the complex trait via other pathways not through the exposure) nor 'correlated pleiotropy' [Morrison et al., 2020] (a phenomenon where a genetic variant affects both exposure and outcome through a heritable shared factor, i.e. IVs are associated with a confounder). See Figure 5.1A for an illustration. Note that valid IVs do not have to be the causal SNPs. Due to the pervasive pleiotropic effects of SNPs and linkage disequilibrium (LD) among SNPs in a region, it is commonly observed that SNPs may be associated with multiple molecular, intermediate and/or complex traits [Verbanck et al., 2018; Pierce et al., 2018; Gleason et al., 2019]. Both horizontal and correlated pleiotropy effects are prevalent in the genome. The inclusion of invalid IVs in traditional MR analyses may lead to biased causal effect estimation and inference. More recently, robust MR methods have been proposed to relax the assumptions by considering multiple IVs and allowing some to be invalid. Some methods allow up to half of the proportion of IVs to be invalid but require individual-level genotype and phenotype data, which may limit the applicability of the methods [Kang et al., 2016]. Some methods require IVs to be nearly independent [Qi and Chatterjee, 2019; Zhao et al.,

2019; Bowden et al., 2015; Zhao et al., 2020] and/or require the number of IVs to be large [Morrison et al., 2020; Cheng et al., 2020]. Those methods have been successfully applied to detect intermediate non-omics traits as exposures for complex diseases. For example, in detecting the protective effect of high-density lipoprotein cholesterol (HDL-C) on peripheral vascular disease, the suspected modifiable exposure HDL-C has many established GWAS SNPs as potential IVs [Cheng et al., 2020].

When applying MR methods to detect gene expression as an exposure for a disease outcome (termed as "transcriptome-wide MR" [Richardson et al., 2020; Barfield et al., 2018]), new challenges arise. First, few studies have genotype, gene expression and disease outcome data being measured on the same set of samples, and even when all data is available for the same set of subjects, sample sizes are generally limited. Thus, MR methods requiring individual-level data may have limited power and applicability. Second, invalid IVs can be quite prevalent when studying gene expression as the exposure. Many genetic variants may affect complex diseases not completely via gene expression levels of a cis-gene [Yang et al., 2017]. Recent studies have reported the existence of many GWAS SNPs being also multi-omics QTLs (i.e., SNPs affecting both cis-gene expression and methylation levels then affecting complex diseases) [Gleason et al., 2019; Pierce et al., 2018], and QTLs with effects on diseases mediated via splicing events [Li et al., 2018]. Methods allowing invalid IVs are necessary in studying gene expression as the exposure. Last but foremost, when treating cis-eQTLs as IVs, the numbers of independent cis-eQTLs for most genes in the genome are very limited. Existing robust two-sample MR methods allowing invalid IVs generally require either multiple independent IVs or a large number of (weakly correlated) IVs, and those existing methods would have limited applicability in analyzing most genes in the genome.

To address those challenges in analyzing gene expression as the exposure for a disease outcome, we propose a two-sample Mendelian Randomization method ROBust to correlated and some INvalid instruments, termed "MR-Robin". It requires only summary-level marginal GWAS and multi-tissue eQTL statistics as input, considers multi-tissue eQTL ef-

fects for multiple IVs of a gene, allows IVs to be correlated and some of them to be invalid, and can be applied to genes with only a small number of cis-eQTLs. Compared to existing two-sample MR methods allowing invalid IVs, MR-Robin lessens the required number of independent IVs by integrating GWAS statistics with multi-tissue eQTL statistics (i.e., multiple sets of IV-to-exposure summary statistics) in a mixed-model framework. Moreover, by carefully selecting cross-tissue eQTLs as IVs, MR-Robin also improves the robustness of IV effects across "two-samples" and may improve the reproducibility of estimation and inference based on two-sample MR analyses. Specifically, MR-Robin considers the estimated effect of a gene on a disease from each IV as an observed value of the true effect plus a SNP-specific bias. By jointly considering multiple IVs, MR-Robin decomposes the estimated effects of multiple IVs into two components – a concordant effect shared across IVs and a discordant component allowing some IVs to be invalid with SNP-specific deviations from the true effect. MR-Robin makes the estimation identifiable by taking advantage of the multi-tissue eQTL effects for multiple IVs of a gene and treating them as the response variable in a reverse regression, with GWAS effect estimates as the predictor. The rich multi-tissue eQTL effect information in the response variable allows the estimation of SNP-specific random-slopes (i.e. deviated effects) due to potentially invalid IVs. Thus, with only a limited number of potentially correlated IVs, MR-Robin can test the effect from a gene to a disease by testing the shared (fixed effects) correlation between eQTL and GWAS effects across IVs. We conducted extensive simulations to evaluate the performance of MR-Robin under various scenarios in analyzing gene expression as the exposure for a disease outcome in the presence of invalid IVs. And, we applied MR-Robin to identify gene expression levels affecting schizophrenia risk by leveraging multi-tissue eQTL summary statistics from 13 brain tissues in the Genotype-Tissue Expression (GTEx) project [Aguet et al., 2019] and GWAS summary statistics from the Psychiatric Genomics Consortium (PGC) [Ripke et al., 2014].

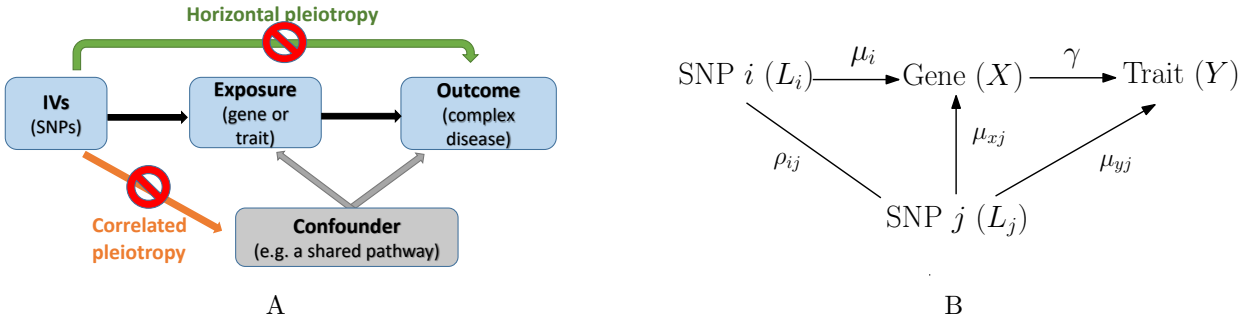


Figure 5.1: **Illustrations of Mendelian Randomization analysis and assumptions.** (A) When a SNP (or is in LD with a SNP that) is affecting the outcome not via the exposure of interest or is correlated with an unmeasured confounder for both the exposure and the outcome, the SNP is an invalid instrument. Note that the presence of unmeasured confounders is allowed in MR analysis, but instruments are assumed to be independent of the confounders. (B) An illustration of pleiotropy of SNP j in an LD block affecting the validity of SNP i of interest as an IV. A SNP j is in LD with an IV SNP i of interest. SNP j is an eQTL of the targeted gene and has a direct effect on the trait (horizontal pleiotropy). When conducting MR analysis with only marginal summary statistics, the effect of SNP j is not accounted for and will confound the relationships among the SNP i, the gene expression and the trait. That is, horizontal (and/or correlated) pleiotropy in a gene region will bias the effect estimate based on marginal statistics for SNP i, without conditioning on SNP j.

5.2 Methods

Let β_{xi} $(i = 1, \ldots, I)$ denote the marginal eQTL effect of a local eQTL/IV i for a gene, and β_{yi} denote the marginal GWAS association effect on a complex trait of the eQTL/IV i in the GWAS study. Note that both β_{xi} and β_{yi}'s are effects in the GWAS study, though β_{xi} is latent since expression data is not available for the GWAS samples, which is typical for most GWAS. Our goal is to test whether the effect of gene expression on the trait (γ in Figure 5.1B) is zero, $H_0 : \gamma = 0$ vs. $H_A : \gamma \neq 0$. Traditional two-sample MR methods often take the ratio β_{yi}/β_{xi} as an estimand for γ based on IV i. In the following subsections, we will first show that when there is a SNP j with a horizontal or correlated pleiotropic effect, and SNP j is in LD with the selected IV i, the ratio β_{yi}/β_{xi} is a biased estimate for γ with a SNP-specific bias depending on many factors. Then we will introduce MR-Robin with a mixed model framework based on reverse regressions taking multi-tissue eQTL summary statistics (multiple sets of IV-to-exposure statistics) as response and IV-to-outcome statistic

as predictor to test for a non-zero effect from gene to disease.

5.2.1 *Bias in β_{yi}/β_{xi} as an estimand for γ when SNP with pleiotropy is in LD with IV i*

Without loss of generality, we assume that there are two SNPs i and j in LD, and SNP i is a valid IV if conditioning on SNP j, and SNP j has a horizontal pleiotropic effect as depicted in Figure 5.1B. For multiple eQTLs in an LD block, one can consider them as being conditionally valid IVs and invalid IVs. Below are the data generating models in a GWAS:

$$X = \mu_{x0} + \mu_i L_i + \mu_{xj} L_j + \epsilon_x, \tag{5.1}$$

$$Y = \mu_{y0} + \gamma X + \mu_{yj} L_j + \epsilon_y, \tag{5.2}$$

where X is the gene expression levels and Y is the continuous complex trait of interest in a GWAS study; and L_i and L_j are the genotypes for SNPs i and j, respectively. As a valid IV given L_j, the genotype of SNP i (L_i) is independent of the error terms ϵ_x and ϵ_y. In the above models, the conditional association between X and L_i given L_j is captured by μ_i, and the conditional association between Y and L_i given L_j is $\gamma \cdot \mu_i$. And the ratio of the two, $\frac{\gamma\mu_i}{\mu_i}$, recovers the true effect of interest, γ.

Without adjusting for SNP j, the summary statistics for SNP i are calculated based on the following marginal models:

$$X = \beta_{x0} + \beta_{xi} L_i + \epsilon'_x, \tag{5.3}$$

$$Y = \beta_{y0} + \beta_{yi} L_i + \epsilon'_y, \tag{5.4}$$

where β_{xi} and β_{yi} are the marginal eQTL and GWAS association effects, respectively, in the GWAS study. Note that one could also adjust covariates in the above models (5.1)-(5.4) and that does not affect our conclusion. We ignore covariates for simplicity. Define

$\rho_{ij} = \frac{\text{Cov}(L_i,L_j)}{\text{Var}(L_i)}$, in terms of parameters in (5.1) and (5.2), it can be derived that the marginal effects $\beta_{xi} = \frac{\text{Cov}(X,L_i)}{\text{Var}(L_i)} = \mu_i + \mu_{xj}\rho_{ij}$, and $\beta_{yi} = \frac{\text{Cov}(Y,L_i)}{\text{Var}(L_i)} = [\gamma + (\gamma\mu_{xj} + \mu_{yj})\frac{\rho_{ij}}{\mu_i}]\mu_i$.

It can be seen that the bias of marginal eQTL effect estimate for SNP i on gene expression, β_{xi}, with respect to the true eQTL effect, μ_i, is $\mu_{xj}\rho_{ij}$. And the bias of marginal GWAS effect estimate for SNP i on complex trait, β_{yi}, with respect to the mediated effect from SNP to gene to trait, $\gamma\mu_i$, is $(\gamma\mu_{xj} + \mu_{yj})\rho_{ij}$. And it can be derived that the bias of the ratio of marginal GWAS to eQTL effect estimates, β_{yi}/β_{xi}, with respect to the true effect, γ, is given by $\frac{\mu_{yj}\rho_{ij}}{\mu_i+\mu_{xj}\rho_{ij}}$. All the biases are functions of SNP i's eQTL effect size, LD strength to the pleiotropic SNP j and effect size of the pleiotropy. Therefore, the bias will vary from SNP to SNP. Similarly, in the presence of correlated pleiotropic SNPs being in LD, the bias will also vary from SNP to SNP.

In the presence of horizontal or correlated pleiotropy in the LD region, an eQTL would be an invalid IV. And in such a case, the effect from gene to trait (γ) is not separable/identifiable from the direct effect of the eQTL nor confounding effects when only the total effect estimate (marginal summary statistic) is available. The presence of horizontal or correlated pleiotropy makes it challenging to infer the effect of a gene on a trait using single-IV-based MR approaches. When there are multiple eQTLs in the gene region, as shown in Figure 5.1B, the presence of one SNP with horizontal or correlated pleiotropic effect would also render all eQTLs invalid if they are in LD.

It should be noted that the above bias is derived for analyzing gene expression as exposure for disease outcome based on marginal eQTL statistics. Due to the fact that all IVs (cis-eQTLs) are from the same cis-region and are in LD, the bias caused by pleiotropy in the region is particularly pronounced. When analyzing intermediate non-omics trait as the exposure and there are many known susceptibility loci from different genomic regions being associated with the non-omics exposure of interest, the IVs are generally less dependent and the bias due to local pleiotropy is generally specific to each locus.

5.2.2 *MR-Robin – a reverse-regression-based mixed model framework with multi-tissue eQTL statistics as response*

Given the bias derived for β_{yi}/β_{xi} w.r.t γ, we model that $\beta_{yi}/\beta_{xi} = (\gamma + \gamma_i)$, where γ_i denotes the SNP-specific bias. The bias is zero if there is neither a horizontal nor correlated pleiotropic effect in the region. The bias is small to negligible for some eQTLs if those eQTLs themselves are valid IVs when adjusting for invalid IV L_j, those eQTLs are in moderate-to-weak LD with the invalid IV(s), and the pleiotropic effect of SNP j is not strong (i.e., small $\rho_{ij} \cdot \mu_{yj}$). It follows that

$$\beta_{yi} = (\gamma + \gamma_i)\beta_{xi}, \forall i = 1, \ldots, I. \tag{5.5}$$

And equivalently,

$$\beta_{xi} = (\theta + \theta_i)\beta_{yi}, \forall i = 1, \ldots, I. \tag{5.6}$$

where θ captures the dependence between β_{xi} and β_{yi}, and θ_i is the SNP-level deviation from the shared effect θ in the presence of pleiotropy.

In the above equation, β_{xi} is the marginal eQTL effect of SNP i to gene expression in the GWAS study and is often not available, since most GWAS studies do not have gene expression data measured. The availability of multi-tissue eQTL summary statistics from trait-relevant tissue types in a reference eQTL study such as GTEx provides a valuable resource to estimate β_{xi}, given that many cis-eQTL effects are shared across tissue types and are replicable across studies.

We model SNP i's eQTL effect in tissue k ($k = 1, \ldots, K$) in the reference multi-tissue eQTL data as a function of the eQTL effect in the GWAS data (β_{xi}) and an error term. Based on (5.6), we propose the following model of MR-Robin for testing trait-association of a gene using only summary statistics from GWAS and a multi-tissue eQTL reference study:

$$\hat{\beta}^R_{xik} = (\theta + \theta_i)\hat{\beta}_{yi} + \epsilon^R_{xik}, \tag{5.7}$$

where $\hat{\beta}^R_{xik}$ is the marginal eQTL effect estimate of the cross-tissue IV/eQTL i ($i = 1, \ldots, I$) in the k-th tissue with the cross-tissue effect in the reference eQTL data, and $\hat{\beta}_{yi}$ is the marginal GWAS effect estimate for SNP i; and θ captures the shared correlation of GWAS and eQTL statistics among all SNPs and is non-zero and bounded if and only if the true effect from the gene on the complex trait, γ, is non-zero and bounded; θ_i represents the SNP-specific bias due to horizontal or correlated pleiotropy in the region and is a SNP-specific random-slope; and ϵ^R_{xik} is a random error that follows a multivariate normal distribution $\mathrm{N}(0, \Sigma^R_x)$. Note that there are both SNP-SNP correlations due to LD and tissue-tissue correlations due to sample overlapping. In the P-value estimation procedure, we account for the correlated errors by resampling.

In the reverse regression (5.7), the eQTL effect estimates from multiple tissue types, $\hat{\beta}^R_{xik}$, are considered as the response variable while the GWAS association effects $\hat{\beta}_{yi}$ are considered as the predictor. This is mainly to take advantage of the rich information in multi-tissue eQTL datasets (i.e., variation in response). If there are multiple sets of correlated or independent GWAS summary statistics from the same population/ethnicity without study heterogeneity, often consortium-based meta-analysis may have been conducted with improved power and precision, and a single set of GWAS summary statistics would be made available. Each observation in the regression (5.7) is an estimated/observed marginal eQTL effect in a selected tissue type, with a maximum of $I \times K$ (SNP-by-tissue) observations. However, due to the fact that not all the IVs have effects shared across all K tissues, in our real data analysis, the model is mostly likely an unbalanced mixed model. By testing the shared correlation of tissue-specific eQTL effects and the corresponding GWAS association effects for multiple eQTLs in the same gene ($H_0 : \theta = 0$ vs. $H_A : \theta \neq 0$) while also allowing for SNP-level deviation, we can test the effect of gene expression on trait ($H_0 : \gamma = 0$ vs. $H_A : \gamma \neq 0$), allowing invalid and correlated IVs.

Many existing methods in the MR literature allowing invalid IVs [Morrison et al., 2020; Zhao et al., 2019; Kang et al., 2016] include an intercept or a random intercept in the model

to capture the direct effect from genotype to trait, i.e. horizontal pleiotropy. That is, SNP-to-disease association effects from GWAS are modeled as $\beta_{yi} = \gamma \cdot \beta_{xi} + \gamma_i, \forall i = 1, \ldots, I$. The model fits better when individual level data are available and statistics conditional on other SNPs in the region can be obtained or when summary statistics from joint models of multiple SNPs in the region are available. In contrast, in the MR-Robin model, there is no intercept nor random intercept. Instead, we include a random slope for each SNP to capture the effect due to potential pleiotropy in the region. This is because, by allowing correlated IVs and considering all eQTLs in a region, as shown above when there is a non-zero pleiotropic effect, most of the SNPs in the LD region would be affected with a non-zero (but possibly negligible) SNP-specific deviation θ_i. Allowing correlated IVs and some invalid IVs even when the number of IVs are limited is also a major innovation of our model. Due to limited numbers of eQTLs/IVs for most genes in the genome, a model with both an intercept and a random slope may not be identifiable and thus is not explored.

To account for uncertainty in the eQTL effect estimation, we perform a weighted mixed-effects regression analysis and weight each "observation" (i.e., a tissue-specific eQTL effect) by the reciprocal of the estimated standard error for $\hat{\beta}^R_{xik}$, i.e., $w_{ik} = 1/\left(\hat{\sigma}^R_{xik}\right)$. We obtain the t-statistic for testing the fixed effect of interest θ as our test statistic. To obtain the P-value while accounting for LD and tissue-tissue correlation as well as the uncertainty in the estimation of β_{yi}'s, we adopt a resampling-based approach to generate the null test statistics. In each resampling b ($b = 1, \ldots, B$), we sample a vector of GWAS effects from a multivariate distribution, $\beta_y^{0(b)} \sim \mathrm{N}(\mathbf{0}, \mathbf{\Sigma}_y^2)$, where the diagonal and off-diagonal elements are $\Sigma^2_{yii'} = \hat{\sigma}_{yi} r_{ii'} \hat{\sigma}_{yi'} \forall i, i'$ with $r_{ii'}$ being the genotype correlation and $\hat{\sigma}_{yi}$ being the estimated standard error for $\hat{\beta}_{yi}$. We apply the same weighted model (5.7) on data $\hat{\beta}^R_{xik}$'s and $\beta^{0(b)}_{yi}$'s to obtain a null statistic. We repeat the resampling process at least $B = 10,000$ times and calculate the P-value. The MR-Robin algorithm is summarized in the algorithm below.

Algorithm 2 MR-Robin for assessing the causal effect of gene expression of a gene on a complex trait with summary statistics from GWAS and a multi-tissue eQTL study

Step 1. Obtain the summary statistics from GWAS study and eQTL study. For each of I cis-eSNPs of the gene being selected, we obtain the association statistics between the SNP and the gene expression in the k-th tissues $\{\hat{\beta}^R_{xik}\}$ along with the standard errors $\{\hat{\sigma}^R_{xik}\}$ $(k = 1, \ldots, K)$ from the multi-tissue eQTL study. And we obtain the association statistics between the SNP and the complex trait $\{\hat{\beta}_{yi}\}$ and the standard error estimates $\{\hat{\sigma}_{yi}\}$ from the GWAS study.

Step 2. Obtain the test statistic. We perform a weighted analysis of the mixed-effects model (5.7) on data $\{\hat{\beta}^R_{xik}\}$ and $\{\hat{\beta}_{yi}\}$ with weight being $1/\hat{\sigma}^R_{xik}$ for each $\hat{\beta}^R_{xik}$ to obtain the test statistic t_{MR} for testing $H_0 : \theta = 0$ vs. $H_A : \theta \neq 0$.

Step 3. Calculate the MR-Robin *P*-value based on resampling. In each resampling b $(b = 1, \cdots, B)$, we generate a vector of GWAS effects $\beta_y^{0(b)}$ from $\text{N}(\mathbf{0}, \mathbf{\Sigma}_y^2)$ to account for GWAS effect estimation uncertainty and LD. We then apply the weighted analysis of the model (5.7) on data $\{\hat{\beta}^R_{xik}\}$ and $\{\beta^{0(b)}_{yi}\}$ with the weight of SNP i in the k-th tissue being $w_{ik} = 1/\hat{\sigma}^R_{xik}$ to obtain a null test statistic $t^{0(b)}_{\text{MR}}$. We then calculate the P-value of trait-association for the gene as, $P\text{-value} = \frac{1}{B}\sum_{b=1}^{B} \text{I}\left(|t^{0(b)}_{\text{MR}}| \geq |t_{\text{MR}}|\right)$, where $\text{I}(\cdot)$ is the indicator function.

5.3 Simulations

5.3.1 Simulations to evaluate the performance of MR-Robin when IVs are correlated, some being invalid, and/or limited in number

In this section, we conducted simulation studies to evaluate the performance of MR-Robin as a two-sample MR method in the settings where a limited number of potentially correlated and/or invalid genetic variants are available as candidate instrumental variables (IVs). We showed that with multi-tissue eQTL statistics as input, MR-Robin is robust to the inclusion

of correlated and some proportions of invalid IVs even when the number of IVs is small. We compared MR-Robin to several existing MR methods in the literature that are based on integrating GWAS summary statistics with summary statistics from a single exposure dataset and are robust to invalid IVs: MR-Egger [Bowden et al., 2015], MR-RAPS [Zhao et al., 2019], MRMix [Qi and Chatterjee, 2019], and BWMR (a Bayesian weighted Mendelian randomization method) [Zhao et al., 2020]. Note that those existing methods were developed for settings where a polygenic trait is analyzed as an exposure for other complex diseases and so many independent genetic variants associated with the exposure trait are available as candidate IVs. Those methods may not be suited for our target settings in which gene expression levels is considered as the exposure, and there are often only a limited number of correlated cis-eQTLs as IVs (trans-eQTLs are not considered as IVs in our two-sample MR analysis because trans-eQTL effects are less replicable across eQTL and GWAS samples). Some of those existing methods also do not allow the IVs to be correlated. Nonetheless, we included the methods for comparison. None of the existing methods were developed for taking multi-tissue eQTLs (multiple sets of IV-to-exposure association statistics) as input and that is an innovation of our method.

Data generation

In each simulation scenario, we simulated data for a total of $N = N_g + N_R = 10,300$ independent subjects: $N_g = 10,000$ subjects in a GWAS study, and $N_R = 300$ subjects in a reference multi-tissue eQTL study of $K = 10$ tissues.

First, we simulated an $N \times I$ genotype matrix $\mathbf{L}$ for each gene, comprised of Q independent LD blocks with 20 SNPs in each block (thus, a total of $I = 20 \times Q$ SNPs for each gene). The correlation between SNP index i and SNP index j in a given LD block is $r_{ij} = 0.95^{|i-j|}$, with the minor allele frequency (MAF) of SNP i, $\text{MAF}_i \sim \text{Unif}(0.05, 0.5)$. From each LD block, we randomly selected 1 SNP to be the true eQTL. The $N_g \times Q$ genotype matrix of the Q true eSNPs in the GWAS study is denoted $\mathbf{G}$. For M ($M \geq 0$) LD blocks, we randomly

selected 1 SNP to be an invalid IV having a direct effect on the complex trait (the value of M varies across simulation scenarios). The $N_g \times M$ genotype matrix of the M SNPs that are invalid IVs is denoted $\mathbf{H}$. We generated phenotypes in the GWAS study according to the following data generation models:

$$X = \mathbf{G}\boldsymbol{\mu_x} + \eta_x Z + \epsilon_x, \tag{5.8}$$

$$Y = \gamma X + \mathbf{H}\boldsymbol{\mu_y} + \eta_y Z + \epsilon_y, \tag{5.9}$$

In Model (5.8), X is a vector of gene expression levels; $\mathbf{G}$ are the genotypes of eSNPs; $\boldsymbol{\mu_x} \sim N_Q(\mathbf{0}, \Sigma_{\mu_x})$ are the eQTL effects of eSNPs from independent LD blocks, with Σ_{μ_x} a diagonal matrix with diagonal elements $\sigma^2_{\mu_x qq} = \frac{0.02}{\text{MAF}_q(1-\text{MAF}_q)}$; $Z \sim N(0,1)$ is a vector of a latent confounder; $\eta_x \sim \text{Unif}(0, 0.1)$ is the effect of the confounder on gene expression levels; and $\epsilon_x \sim N(0,1)$ are error terms. In Model (5.9), Y is a vector of a continuous complex trait; γ is the parameter of interest, the effect of gene X on trait Y, with $\gamma = 0$ under the null and $\gamma = 0.3$ under the alternative; $\mathbf{H}$ are the genotypes of SNPs having a direct effect on Y not through gene expression of X; $\boldsymbol{\mu_y} \sim N_M(\mathbf{0}, \Sigma_{\mu_y})$ are the direct effects on Y of M SNPs from independent LD blocks, with Σ_{μ_y} a diagonal matrix with diagonal elements $\sigma^2_{\mu_y mm} = \frac{0.002}{\text{MAF}_m(1-\text{MAF}_m)}$; $\eta_y \sim \text{Unif}(0, 0.1)$ is the effect of the confounder on the complex trait; and $\epsilon_y \sim N(0,1)$ are the error terms. Across scenarios we vary M, the number of LD blocks having an invalid IV.

Data from the eQTL study was generated based on the model:

$$\mathbf{X}^{\text{R}} = \mathbf{G}^{\text{R}}\boldsymbol{\mu}^{\text{R}}_{\boldsymbol{x}} + \epsilon^{\text{R}}_x, \tag{5.10}$$

where $\mathbf{X}^{\text{R}}$ is an $N_R \times K$ matrix of expression levels measured in K tissues; $\mathbf{G}^{\text{R}}$ is a $N_R \times Q$ genotype matrix of Q eSNPs in the eQTL study; $\boldsymbol{\mu}^{\text{R}}_{\boldsymbol{x}}$ is a $Q \times K$ matrix of the tissue-specific eQTL effects; and $\epsilon^{\text{R}}_x \sim N(0,1)$ are the error terms. Each column of $\boldsymbol{\mu}^{\text{R}}_{\boldsymbol{x}}$ is independently drawn from $N_Q(\boldsymbol{\mu_x}, 0.02 \cdot \mathbf{I})$, where $\boldsymbol{\mu_x}$ is from Model (5.8).

After individual-level data were generated in each simulation, we calculated the marginal eQTL and GWAS summary statistics. For two-sample MR analyses, we then obtained the marginal effect estimate of each SNP i on gene expression in tissue k in the reference eQTL study, $\hat{\beta}^R_{xik}$; and obtained the marginal effect estimate of each SNP i on its simulated trait in the GWAS study, $\hat{\beta}_{yi}$. We also obtained the standard error estimates for marginal eQTL and GWAS effects.

Results of simulation studies

In the first simulation setting, we evaluated the robustness of MR-Robin to the proportion of invalid IVs compared to existing two-sample MR methods. $P < 0.05$ was used as the significance criterion for each method. Table 5.1 shows the type I error rate and power comparison in the presence of 0, 10, ..., 50% invalid IVs, allowing IVs to be moderately correlated (pairwise LD $r^2 < 0.5$ or 0.3) over 10,000 simulations of $Q = 10$ LD blocks. Since our method allows for correlated IVs and it is hard to define invalid versus valid IVs when SNPs are correlated, the proportions of invalid IVs in the tables are the proportion of LD blocks with pleiotropy, and are only approximations of the invalid IVs among all selected ones. In each table, we also presented the average numbers of selected IVs that are from valid versus invalid LD blocks. For the competing methods, which were not developed for multi-tissue eQTL datasets, we used the eQTL summary statistics from one randomly selected tissue as input for the IV-exposure summary statistics. As shown in the table, whereas competing methods are unable to control the type I error rate when there are any invalid instruments and instruments are in LD, MR-Robin maintains reasonable control of the type I error rate if a majority of instruments are valid (e.g. up to 30% invalid IVs). The last three methods in the table were developed for independent instruments; since they do not account for correlation (LD) among the instruments, they do not control the type I error rate even when all instruments are valid. Power is reasonable for all methods when a majority of IVs are valid. In Appendix D, Table D.1, we compared the type I error rates and powers using

alternative LD selection criteria for the IVs (pairwise LD $r^2 < 0.1$ or 0.01).

Table 5.1: **Simulation results evaluating the performance of MR-Robin.** Averaged type I error rates and power over 10,000 simulations are shown by percentage of invalid instruments (using $P < 0.05$ as the significance criterion for each method). 10 LD blocks were simulated, with one true eQTL per LD block. Instruments were selected sequentially: the eSNP with the strongest association with gene expression was selected, and the next selected eSNP is the strongest-associated SNP remaining also with LD $r^2 < \rho$ with any already-selected eSNPs. Results shown for $\rho = 0.5$ (A) and $\rho = 0.3$ (B)

(A) pairwise LD $r^2 < 0.5$

Method	Proportion of invalid IV (%)					
	0	10	20	30	40	50
	Type I error rate					
MR-Robin	0.047	0.051	0.059	0.061	0.071	0.085
MR-Egger	0.032	0.057	0.087	0.110	0.122	0.138
MR-RAPS	0.255	0.303	0.348	0.380	0.399	0.426
MRMix	0.136	0.175	0.207	0.223	0.256	0.267
BWMR	0.279	0.349	0.396	0.440	0.452	0.479
	Power					
MR-Robin	0.800	0.763	0.725	0.685	0.659	0.615
MR-Egger	0.942	0.931	0.923	0.917	0.914	0.905
MR-RAPS	0.996	0.993	0.986	0.979	0.976	0.962
MRMix	0.511	0.499	0.498	0.502	0.491	0.493
BWMR	0.998	0.994	0.988	0.981	0.978	0.966
	Avg number of SNPs selected (valid/invalid)					
All Methods	30.4 /0.0	27.4 /3.0	24.4 /6.0	21.4 /9.2	18.3 /12.1	15.2 /15.2

(B) pairwise LD $r^2 < 0.3$

Method	Proportion of invalid IV (%)					
	0	10	20	30	40	50
	Type I error rate					
MR-Robin	0.047	0.051	0.055	0.051	0.061	0.062
MR-Egger	0.027	0.062	0.097	0.126	0.137	0.161
MR-RAPS	0.084	0.118	0.147	0.170	0.186	0.214
MRMix	0.130	0.199	0.237	0.273	0.288	0.307
BWMR	0.097	0.141	0.182	0.206	0.222	0.246
	Power					
MR-Robin	0.679	0.640	0.605	0.577	0.549	0.505
MR-Egger	0.894	0.886	0.878	0.870	0.860	0.853
MR-RAPS	0.986	0.979	0.969	0.955	0.944	0.925
MRMix	0.512	0.515	0.517	0.509	0.500	0.498
BWMR	0.992	0.985	0.975	0.962	0.952	0.936
	Avg number of SNPs selected (valid/invalid)					
All Methods	14.1 /0.0	12.7 /1.4	11.3 /2.8	9.9 /4.2	8.5 /5.6	7.0 /7.0

In the second simulation scenario, we evaluated the performance of MR-Robin when the number of selected IVs is small. We simulated the data using $Q = 3$ LD blocks, with two blocks without pleiotropy and one block with pleiotropy (thus the proportion of LD blocks with pleiotropic effects is fixed at 33.3%). Table 5.2 shows the type I error rates and power when the selection LD r^2 threshold is set to 0.5, 0.3, 0.2, 0.1 and 0.01. As shown in the table, MR-Robin performs reasonably well even when the number of IVs is very limited. Though in this setting, MR-Robin requires the IVs to be less dependent ($r^2 < 0.3$). MR-Robin outperforms competing methods in this setting.

Table 5.2: **Simulation results evaluating the performance of MR-Robin when there is a small number of IVs.** Averaged type I error rates and power over 10,000 simulations are shown by IV selection criteria. 3 LD blocks were simulated, with two blocks without pleiotropic effects (valid IVs) and one block with (invalid IV). Results shown for five IV selection criteria (LD $r^2 < 0.5$, 0.3, 0.2, 0.1, and 0.01).

Method	LD selection criteria (r^2)				
	0.5	0.3	0.2	0.1	0.01
	Type I error rate				
MR-Robin	0.084	0.056	0.048	0.039	0.032
MR-Egger	0.087	0.100	0.111	0.139	0.158
MR-RAPS	0.380	0.203	0.144	0.104	0.091
MRMix	0.254	0.240	0.239	0.227	0.223
BWMR	0.431	0.227	0.154	0.103	0.087
	Power				
MR-Robin	0.586	0.455	0.391	0.320	0.292
MR-Egger	0.676	0.625	0.602	0.553	0.504
MR-RAPS	0.860	0.799	0.773	0.740	0.726
MRMix	0.473	0.470	0.480	0.487	0.496
BWMR	0.884	0.811	0.769	0.722	0.695
	Avg # of SNPs selected				
All Methods	7.4 /3.7	3.5 /1.8	2.7 /1.4	2.2 /1.1	2.0 /1.0

The simulation results showed that MR-Robin is able to control the type I error using correlated instruments provided that a majority ($\geq 70\%$) of the instruments are valid IVs. Moreover, in Table 5.2, we showed that even when the number of available IVs is very small (3-10), the proposed MR-Robin can still yield reasonable results if the small number of IVs are relatively less dependent ($r^2 < 0.3$). Last but not least, we want to emphasize that when IVs are correlated, if one IV is an invalid IV, all the other correlated IVs are also affected to some degree, and as such the random-slope model of MR-Robin with its resampling-based inference procedure fits the need for allowing correlated IVs when considering the effect of gene expression on a complex trait.

5.4 Data Application

5.4.1 Application: Identifying schizophrenia (SCZ) risk-associated genes via MR-Robin

To detect genes with expression levels being associated with schizophrenia risk, we applied MR-Robin using summary statistics from two-samples: schizophrenia risk GWAS statistics from the second schizophrenia mega-analysis (SCZ2) conducted by the Psychiatric Genomics Consortium (PGC) [Ripke et al., 2014], and multi-tissue eQTL statistics from the 13 brain tissues in version 8 (V8) of the Genotype-Tissue Expression (GTEx) project [Aguet et al., 2019]. The GTEx project and V8 data are described in Section 4.4.1. The PGC SCZ2 GWAS was conducted using up to 36,989 cases and 113,075 controls. In the final analysis, 128 LD-independent SNPs in 108 loci were reported as surpassing the genome-wide significance threshold ($P < 5 \times 10^{-8}$). Additional details of the second PGC GWAS of schizophrenia-risk are reported elsewhere [Ripke et al., 2014].

Results

We first formed the set of instrumental variables (IVs) for each gene by selecting the cis-eSNPs/IVs (within 1 Mb of transcription start site) and the brain tissue types in which they have strong IV effects. All the cis-SNPs being selected are cross-tissue IVs (with median eQTL $P < 0.05$). However, it is well known in the IV literature that weak IVs, i.e., SNPs being only weakly associated with the genes, would result in high variance and misleading inferences even when they are valid IVs (Bound et al., 1995; Nelson and Startz, 1990). And therefore, we will choose cross-tissue eQTLs with significant eQTL effects of P-value ≤ 0.001 in at least three tissue types, i.e., being reasonably strong IVs to provide reliable inferences (Stock et al., 2002) in at least three tissue types. And we restrict the analysis to the cross-tissue IVs in the tissue types with strong cross-tissue (or shared) effects. Since this step involves only the selection of IV based on the strength of the eQTL effects, with no information regarding the outcome, the selection of IV and tissue types would not induce inflation in false positive findings.

While the analysis is restricted to strong IVs with $P < 0.001$ in at least 3 tissues, we iteratively selected the (next) best eSNP (i.e. lowest median eQTL P-value) satisfying the IV selection criteria and having pairwise LD $r^2 < 0.5$ with each of the eSNPs already selected. Note that here we conducted the primary analysis with a relatively liberal LD threshold to improve the power of the analysis. Following the primary analysis, we later conducted a sensitivity analysis on the implied genes to check the robustness of our results to the choice of IVs. If the gene has only 1 cis-eQTL, MR-Robin would be reduced to a single-IV analysis, which can be heavily affected by the validity of the IV with assumptions that cannot be adequately checked in general. Therefore, we restricted the MR-Robin analysis to 3,127 protein-coding genes with at least 5 IVs selected based on this criteria. For each SNP/IV used in the analysis, we used eQTL statistics only from those brain tissues where the SNP had eQTL $P < 0.001$ (with strong IV effects). Thus, each SNP/IV has 3-13 observed eQTL effect estimates from different tissue types in the unbalanced mixed-effects model.

At a false discovery rate (FDR) $< 5\%$, we identified 43 genes as showing evidence of a dependence between gene expression levels and SCZ risk. For the 43 genes whose expression showed an association with SCZ risk in the primary analysis, we performed a sensitivity analysis using different IV selection criteria. Specifically, for each gene, among the cross-tissue IVs with median eQTL $P < 0.05$ having strong IV effects in at least 3 tissues ($P < 0.001$), we iteratively dropped the eSNP with the highest correlation to others until all pairwise LD $r^2 < 0.3$ among remaining eSNPs or only 5 eSNPs remained. For each eSNP, we still only used eQTL statistics from tissues where that eSNP had eQTL $P < 0.001$. In the sensitivity analysis, there were 39 and 42 genes with MR-Robin $P < 0.05$ and $P < 0.1$, respectively, all of which had a fixed effect estimate matching the sign of the fixed effect estimate from the primary analysis.

Figure 5.2 plotted the multi-tissue eQTL effect sizes in the GTEx brain tissues against the GWAS effect sizes in the PGC dataset for two selected genes in the primary analysis (left column) versus the sensitivity analysis (right column). The gene *THOC7* (Figure 5.2A) showed consistent correlations between eQTL and GWAS effects based on two sets of correlated IVs in the primary and sensitivity analyses (both with $P < 5 \times 10^{-3}$). Despite some SNPs having a potentially larger deviation from the shared effect than the others – indicated by the random slopes (colored line segments) deviating from the fixed effect estimate (black line) – the plot shows a clear pattern of association between the magnitude of eQTL effects and magnitude of GWAS effects, implying that the expression levels of *THOC7* affect schizophrenia risk. The protein encoded by *THOC7* is a component of the THO complex of the TRanscription and EXport (TREX) complex which couples transcription to mRNA export, specifically associating with spliced mRNA [Masuda et al., 2005; Chi et al., 2013]. Mutations in subunits of TREX have been associated with neurodevelopmental disorders [Heath et al., 2016], and a recent TWAS study that imputed gene expression in brain tissues found an association between expression levels of *THOC7* in cerebellum and schizophrenia risk [Huckins et al., 2019]. In contrast, the gene *RNF149* (Figure 5.2B) was the only gene

no longer significant in the sensitivity analysis ($P = 0.15$), and prompts further exploration. The change in significance for *RNF149* may be at least partially due to an increase in the relative proportion among selected IVs that have potential pleiotropic effects (i.e. better fitted by a line with non-zero intercept in Figure 5.2B) when using more stringent LD r^2 selection criteria. In the Appendix E, we presented additional details and the scatterplots of multi-tissue eQTL effect estimates against SCZ GWAS effect estimates for selected IVs of all 42 genes identified by MR-Robin in the primary analysis having $P < 0.1$ the sensitivity analyses.

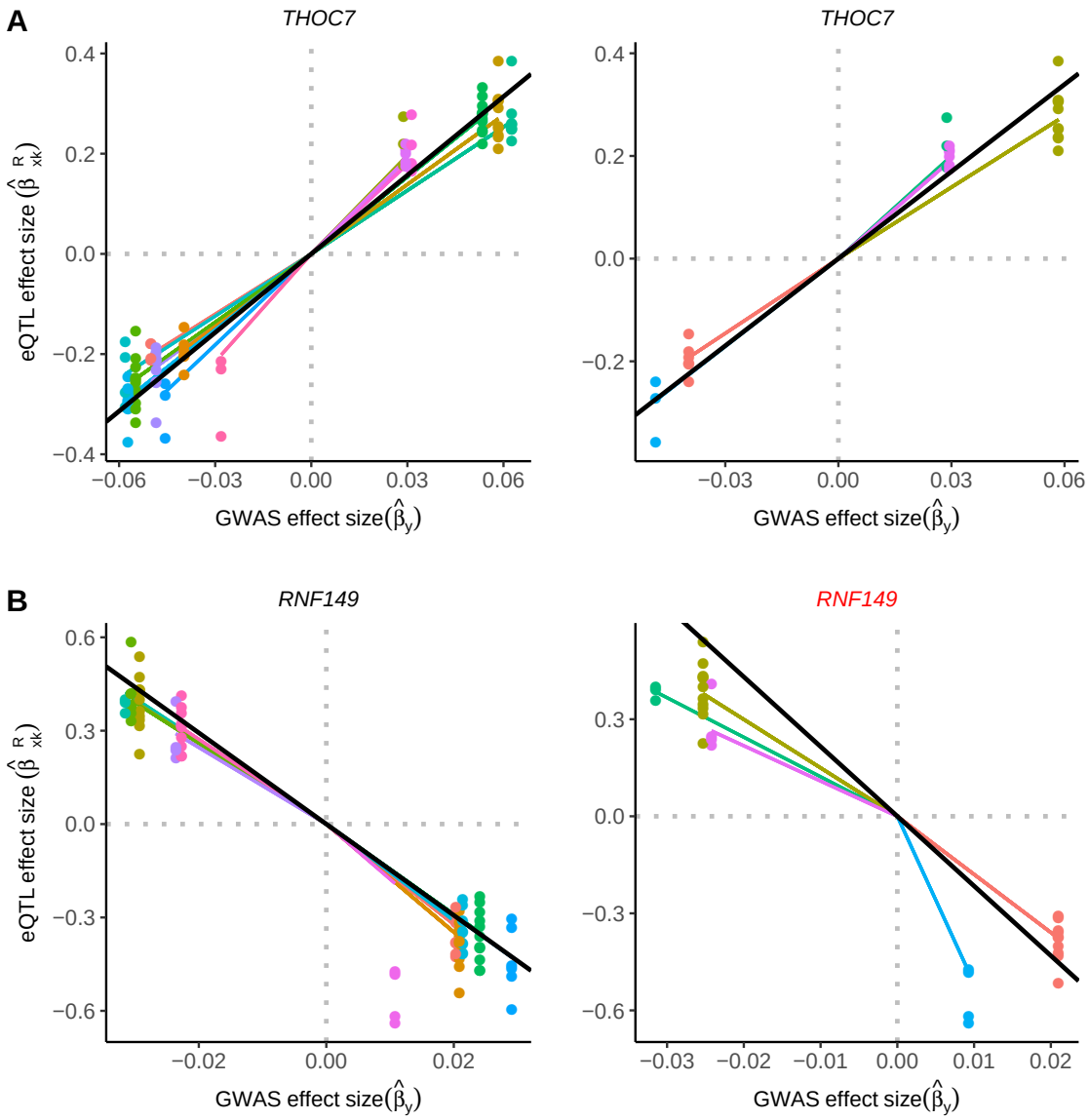


Figure 5.2: **Illustrations of two example genes in the MR-Robin primary (left column) and sensitivity (right column) analysis**. Multi-tissue eQTL effect sizes in the GTEx brain tissues were plotted against SCZ GWAS effect sizes in the PGC dataset for the genes, *THOC7* and *RNF149*. In the sensitivity analysis using alternative IV selection criteria, the SCZ risk association remained for *THOC7* (A) ($P < 0.1$) with consistency in the sign of the fixed effect estimate. The association was no longer significant between *RNF149* expression (B) and SCZ risk in the sensitivity analysis ($P = 0.15$). Points are colored by SNP. Colored line segments represent SNP-specific slope estimates. The slope of the black line is the fixed effect estimate from the MR-Robin reverse regression. The results imply a non-zero effect of the gene *THOC7* on schizophrenia risk.

In summary, we applied the newly proposed two-sample MR method, MR-Robin, to integrate multiple sets of brain tissue eQTL summary statistics from GTEx and SCZ GWAS summary statistics from PGC, and have identified 42 genes with potential causal associations to schizophrenia risk. These 42 genes demonstrated consistent dependencies between brain eQTL and SCZ GWAS association effects using two sets of SNPs as IVs based on different selection criteria. The results highlighted the value of MR-Robin as a robust two-sample MR method that allows moderately correlated and some invalid instrumental variables and identifies gene expression levels as causal exposures for complex diseases.

5.5 Discussion

In this work, we proposed a robust two-sample Mendelian randomization (MR) method – MR-Robin – allowing correlated and invalid instrumental variables (IVs). MR-Robin was motivated by analyses of gene expression levels as causal exposures for complex diseases/traits. In those settings, often only a limited number of potentially correlated cis-eQTLs are available as candidate IVs, posing new challenges to MR analyses. MR-Robin integrates GWAS statistics with multi-tissue eQTL statistics in a mixed model framework, considering the estimated effect of gene expression levels on disease from each IV as an observed value of the true effect plus a SNP-specific bias. Compared to existing robust two-sample MR methods, a major innovation of MR-Robin is the use of multi-tissue eQTL summary statistics (multiple sets of IV-to-exposure statistics). Based on a reverse regression framework with multi-tissue eQTL effects as response, the rich information in multi-tissue eQTL data allows the estimation of SNP-specific random slopes (due to being in LD with SNPs with horizontal and/or correlated pleiotropy) as well as the fixed-effects correlation of eQTL and GWAS effects across all IVs based on a limited number of IVs. In contrast, existing models and methods based on the deconvolution of mixture distributions or penalized regressions in general require a large number of IVs to achieve stability in estimation. To account for correlation among IVs due to LD and tissue-tissue correlations, MR-Robin utilizes a resam-

pling procedure when testing the effect from gene expression levels to the complex trait. We showed through simulations that MR-Robin was able to control the type I error rates using a limited number of moderately correlated IVs when the proportion of IVs that are invalid is moderate.

We applied MR-Robin to identify genes with expression levels affecting schizophrenia risk (SCZ) by integrating multiple sets of brain tissue eQTL statistics from GTEx and SCZ GWAS statistics from PGC. We identified 42 genes showing consistent dependencies between multi-tissue eQTL and GWAS association effects based on two different sets of IVs with different selection criteria from primary and sensitivity analyses. Our analysis illustrated that MR-Robin and two-sample MR methods, requiring only multi-tissue QTL and GWAS summary statistics as input, could be used as another integrative method in recapitalizing on existing summary statistics to further map gene expression levels or other omics traits affecting a complex trait of interest, to explain the potential mechanisms underlying trait susceptibility loci, and to identify clinically actionable targets with larger effects on complex diseases and traits.

There are several caveats and limitations to the current work. First, similar to other two-sample MR methods, MR-Robin cannot by itself prove a causal relationship from a gene to a complex trait but rather suggests instances consistent with a causal model. Nevertheless, analyses using MR-Robin may be useful in prioritizing candidate genes for additional follow-up and research. Second, MR-Robin requires summary statistics from a multi-tissue eQTL dataset as input. For some complex traits being considered as the outcome, it may not be obvious which tissues are most relevant to the trait being studied. Several recent works have proposed methods or provided resources to identify trait-relevant tissues [Cai et al., 2020; Jia et al., 2020; Hao et al., 2018], and these works may be useful in such cases. Third, to accurately estimate the SNP-specific bias, MR-Robin requires more than one SNP to be used as an IV. Depending on the dataset and IV selection criteria, there may be some genes whose association with the complex trait cannot be appropriately tested using MR-Robin.

MR-Robin was developed as a two-sample MR method to test for effects from the expression levels of a gene on a complex trait. MR-Robin can be applied to discover genes that may be causally associated with a complex trait of interest or to confirm that a putative gene demonstrates consistency with a model in which its gene expression causally affects the complex trait. The method may also be extended more generally to settings where a limited number of potentially correlated candidate IVs are present provided that multiple estimates of either the IV-exposure or IV-outcome statistics are available.

The R package MrRobin is freely available at https://github.com/kjgleason/MrRobin.

CHAPTER 6

SUMMARY AND FUTURE DIRECTIONS

6.1 Summary

In this work, we developed several statistical methods and computational tools to perform integrative multi-omics association analyses. Each method and tool requires only summary statistics as input, allowing for significant flexibility in applications of the methods while providing the opportunity for researchers to recapitalize on publicly available summary statistics and/or derive new insights from the results generated by their own research projects.

The methods were generally motivated by the need to elucidate the biological mechanisms that underlie associations between genetic variants and disease-related outcomes such as susceptibility. Better understanding of these mechanisms is necessary to identify clinically actionable therapeutic targets. But there are a number of challenges to conducting integrative multi-omics association analysis using summary statistics, including the possible presence of heterogeneity between studies or data types, potential correlations due to sample overlap, increasing complexity and computational burden when analyzing larger numbers of studies or traits, and attempting to distinguish associations which may be causal. Addressing these challenges was a major focus of the presented work.

We developed a flexible integrative association analysis framework – Primo – that jointly analyzes summary statistics from multiple studies/traits/conditions and, for each genetic variant analyzed, quantifies the probability of each possible association pattern (i.e. with which studies/traits/conditions is the genetic variant associated). Requiring only summary statistics as input, we developed versions of Primo for integration of t-statistics, z-scores, and P-values and other second order statistics such as χ^2 statistics. Primo could also be adapted to other types of statistics as well, provided that a suitable distribution can be identified for each of the null and alternative hypotheses.

A major innovation of Primo is that it allows for summary statistics to be estimated

in overlapping samples since it accounts for sample correlations in its estimation algorithm. This innovation helps maximize the utility of large-scale multi-omics projects, such as The Cancer Genome Atlas TCGA)/Clinical Proteomics Tumor Analysis Consortium (CPTAC) or the Genotype-Tissue Expression (GTEX) project, since such project often collect omics data for different molecular traits and/or in multiple tissue-/cell-types for overlapping sets of subjects. Indeed, in this work, we integrated summary statistics studying the effects of copy number alterations (CNAs) and germline genetic variants on multiple omics traits collected in overlapping sets of subjects.

To move beyond identifying biological correlations towards detecting possible causal associations, we made several tailored developments to the Primo framework. In one such adaptation, we tailored the Primo framework to study mediation test statistics to identify possible cases of cis-mediated trans-associations. Such innovation may be useful in identifying which gene in a correlated cis-region may be the putative gene affecting a trans-omics trait, as may be the case when studying trans-associations of CNAs. In another development, we tailored Primo to study the effects of germline variants on omics and complex traits by accounting for correlation due to linkage disequilbirum (LD) through conditional analysis, conditioning on lead omics-SNPs in a cis region. We applied the method to identify cis-omics associations of genetic variants with reported associations to complex traits (GWAS SNPs) as well as to identify possible pleiotropic associations of GWAS SNPs affecting multiple complex traits while elucidating the biological mechanisms through which they may be affecting those complex traits.

We also developed another integrative association analysis method – MR-Robin – with the goal of identifying potential causal associations of a gene's expression on a complex trait such as disease susceptibility. A two-sample Mendelian randomization (MR) analysis method, MR-Robin employs a mixed-effects reverse regression framework to identify genes with potentially causal associations to a complex trait. A challenge when using two-sample MR to analyze the effects of gene expression levels as the exposure on an outcome such as

disease risk is that a limited number of often correlated SNPs are available as candidate instrumental variables (IVs). Existing robust two-sample MR methods that attempt to account for the presence of pleiotropy, a violation of traditional MR analysis assumptions, require a large number and/or relatively independent SNPs to be used as IVs. By leveraging the rich variation in multi-tissue eQTL datasets in estimating random effects and employing a resampling procedure to account for correlation between SNPs, the mixed-effects implementation of the MR-Robin algorithm allows a fewer number of moderately correlated SNPs to be used as IVs in two-sample MR analysis. This innovation allows MR-Robin to evaluate whether a gene is consistent with a causal model in which its expression is causally associated with an outcome of interest, such as susceptibility to a particular disease.

Using the proposed methods and accompanying software packages, we performed a series of analyses that produced several interesting findings. We demonstrated that cis-effects of CNAs are frequently shared across omics traits and tumor types, and identified dozens of cis-mediated trans-protein associations of CNAs shared between breast and ovary tumors. Several of these associations occurred by mediation through the protein abundance of "cis-hubs" with known associations with cancer-related phenotypes. In analyzing the effects of germline genetic variants, we detected cis-omics associations for many reported breast cancer risk SNPs and identified dozens of genetic variants with pleiotropic effects on multiple complex traits, many of which associations replicated in independent studies. And by integrating multi-tissue eQTL summary statistics from GTEx brain tissues with summary statistics from a GWAS of schizophrenia risk, we identified dozens of genes consistent with a causal model in which their expression levels are causally associated with schizophrenia risk.

Each of the proposed methods was incorporated into R packages (i.e. Primo and Mr-Robin) developed as part of this work. The packages are freely available for download through the following web address: https://github.com/kjgleason.

6.2 Future Directions

There are many interesting future directions for the presented research. In this section, we discuss a few promising areas for further development.

As the amount of data to be integrated by Primo increases, so too does the computational burden – linearly with an increasing number of genetic variants (and/or omics outcomes, if a variant can be mapped to multiple genes, CpG targets, etc.) and exponentially with an increasing number of studies/traits/conditions (since the number of association patterns for which to estimate the pattern proportions, π, and posterior probabilities is 2^J for J studies). When considering the computational burden in the presented work, a key development was converting the EM algorithm estimating the π vector to a series of matrix operations and using Rcpp to perform the estimation, which substantially reduced computation time. For example, in estimating π for 10 million observations of summary statistics measured in $J = 6$ studies, the median time for a single E-step iteration was 304 seconds using R but 64 seconds using Rcpp (representing an $\sim$80% decrease in computation time). Further improvements in performance or reductions in computational burden could further improve efficiency and add to the potential applications of the work. One idea is to use resampling when estimating the π vector, taking a random sample of the sets of summary statistics during each iteration of the EM algorithm to speed up computation. Even if not used for the full estimation, resampling could be used in the early iterations to move the initialized values of the π vector closer to the "neighborhood" of the true π, requiring fewer of the slower iterations using the full amount of data. Resampling in estimation of π and other considerations of ways to improve computation will continue to be explored.

As noted in Chapter 4, incorporating functional annotations into Primo could be another way to improve the method. Such annotations could include predicted effects on protein functions (e.g. SIFT [Kumar et al., 2009], PolyPhen [Adzhubei et al., 2010], or CADD [Kircher et al., 2014] scores) or location in relationship to coding regions and regulatory elements (exons, introns, enhancers, promoters, transcription factor binding sites, etc.). Annotations

could be used by Primo as prior weights of the variants. Or, Primo could estimate different π vectors for genetic variants depending on their annotation. For example, a separate π vector could be estimated for exonic SNPs in coding regions than for intronic or intergenic SNPs, etc. Ways to incorporate functional annotations will be explored in future research.

Another future direction for additional research involves adapting how Primo handles missing data. Despite the use of genotype imputation, some important SNPs may be missing in one or more data sets, or an important omics outcome could be missing from the data. For example, gene expression could be measured for a SNP's cis-genes, but not the abundance of some or all proteins translated for those genes; or DNA methylation CpG sites could be missing for sections of the genome (e.g. if the 27k array is used). The missingness may be ignorable or non-ignorable [Rubin, 1976]. An example of the latter may occur, for example, in proteomics/pQTL datasets since the probability of missingness for a protein may be related to the abundance or number of peptides that the protein consists of [Chen et al., 2014, 2017]. Though not discussed in the previous chapters, the current version of the Primo R package incorporates the following conservative approach to handling cases of non-ignorable missingness. Say SNP i is missing in data type j'. We first estimate key parameters (e.g. scaling factors v_j) using the Primo method with the set of all genetic variants that are not missing any data types. We identify each association pattern k^* such that the missing data type j' comes from the null distribution ($q_{k^*j'} = 0$). We then quantify the probability of SNP i coming from pattern k^* as: $P(a_i = k^*|T_i, \pi_{k^*}) = \frac{\pi_{k^*} D_{k^*}(T_i)}{\sum_{k^*} \pi_{k^*} D_{k^*}(T_i)}$, where, T_i is a $J-1$ vector of t-statistics from the non-missing data types for SNP i. We estimate the joint density D_{k^*} following the Primo algorithm by removing row j' and column j' from the variance and correlation matrices (Σ_{k^*} and Γ). This approach effectively treats the effect of SNP i on data type j' as if it were known to be null, allowing for inference regarding the remaining data types.

On the other hand, if the missing-data mechanism is known or can be modeled, then one may also impute the summary statistics based on the missing-data mechanism. To impute

the missing summary statistics for the effect of SNP i in data type j', we can utilize summary statistics from SNPs present in data type j' and their correlations with SNP i (due to LD) to impute the missing summary statistics for the effect of SNP i in data type j'. Such an approach has been explored by other works in the literature. For example, ImpG-Summary [Pasaniuc et al., 2014] derives the posterior mean of z-scores for missing SNPs conditional on z-scores of non-missing SNPs and a correlation matrix estimated from a reference panel (e.g. 1000 Genomes [Auton et al., 2015]). DISSCO follows a similar framework to impute missing summary statistics, allowing for adjustment of the effects of potential confounding covariates [Xu et al., 2015]. Incorporating imputation of missing summary statistics into the Primo framework is a promising direction for future research that will improve the applicability of the method.

Finally, several of the methods could benefit from exploration of jointly modeling multiple outcomes and/or predictors. In its present implementation, Primo(med) models the effect of one independent variable (predictor) on a single dependent variable (outcome) mediated through a single mediator. Similarly, MR-Robin models the effect of a single exposure on a single outcome. Future directions to explore for these two methods could include incorporating multiple mediators, exposures and/or outcomes into the models. Such innovation could boost power of the methods or increase confidence in significant findings given the consistency of effects observed across multiple variables and outcomes.

6.3 Conclusion

In conclusion, the work presented in this dissertation included the development of multiple integrative multi-omics association analysis methods and computational tools. We demonstrated good performance of the methods through the use of simulations and showed through several data analyses a wide range of applications yielding interesting results. We hope the methods will be used to elucidate biological mechanisms underlying disease processes, potentially highlighting therapeutic targets and other noteworthy associations in the genome.

APPENDIX A

TUTORIAL FOR PRIMO R PACKAGE

In this appendix, I provide a tutorial to the Primo R package, which is incorporated as a vignette within the package. The package (with vignette included) can be downloaded at https://github.com/kjgleason/Primo.

Primo: Package in R for Integrative Multi-Omics association analysis

...

Tutorial

The Primo package can be used to integrate summary statistics to detect joint associations across multiple studies, allowing for the possibility of sample overlap. Here, a "study" refers to associations to a particular trait in a particular condition/cell-type/tissue-type or associations measured in a particular source/sample.

General framework

Following the method described by Gleason *et al.*[1], Primo takes as input m sets of summary statistics from each of J studies and then:

1. Estimates null and alternative density functions for each study.
2. For each of m sets of summary statistics, estimates the posterior probability of coming from each of 2^J association patterns representing the binary combinations of association status (null or alternative) between set i and the J studies.
3. Estimates false discovery rates (FDR) to allow for selection of a posterior probability threshold to make inferences about association patterns.

The inference about association patterns may involve either a particular pattern of interest (e.g. study1+study2, but not study3 or study4) or group of combined patterns (e.g. study1+"at least 1 of studies 2/3/4").

1.1 Integrate summary statistics to estimate posterior probabilities

To illustrate, let's say we have obtained summary statistics from $J = 4$ studies. From each study, these summary statistics include:

- `betas`:coefficient estimates
- `sds`: standard error estimates (of coefficients)
- `dfs`: degrees of freedom for each analysis
- `pvalues`: nominal P-values from each study

Here, each of the above sets of summary statistics are formed into matrices ($m \times J$), though `dfs` may also be a vector (length J) if the degrees of freedom never vary within any study.

We demonstrate Primo using the t-statistics version, which takes as input `betas`, `sds` and `dfs`. For a demonstration of the P-value version of Primo, see Primo for integrating P-values.

For each observation, we are interested in identifying its underlying association pattern. That is, we wish to identify the set of studies (e.g. traits) with which it is associated. To quantify the probability of each association pattern for each observation, we run an integrative analysis using Primo:

```
Primo_results <- Primo(betas=betas,sds=sds,dfs=dfs,alt_props=c(1e-5,rep(1e-3,3)))
```

In addition to the data previously described, Primo also requires the specification of `alt_props`, the estimated proportion of statistics that comes from the alternative distribution for each study. Here we specified 10^{-5} for the first study and 10^{-3} for the other 3 studies.

[1] Gleason et al.: https://www.biorxiv.org/content/10.1101/579581v3

If the observations are SNPs, we may also have obtained minor allele frequencies (MAF) in the form of either a matrix ($m \times J$) or a vector (length m). In such cases, MAF and the number of subjects (N) may also be passed to the function to further adjust the error variance of the moderated t-statistics:

```
Primo_results <- Primo(betas=betas,sds=sds,dfs=dfs,alt_props=c(1e-5,rep(1e-3,3)),mafs=MAF,N=N)
```

Results

`Primo_results` now holds a list of 10 items. The primary elements of interest are:

- `Primo_results$post_prob`: the posterior probabilites of each association pattern for each observation ($m \times 2^J$ matrix)
- `Primo_results$pis`: the estimated proportions of all observations belonging to each association pattern (vector of length 2^J)

The remaining elements are returned largely for use by other functions.

1.2 Combining association patterns into interpretable results

From the results of Primo, we can combine posterior probabilities into interpretable results by summing over association patterns. E.g., the following will calculate the posterior probabilities of being associated with:

- at least one study
- at least two studies
- at least three studies
- all four studies

```
postprob_atLeastN <- Primo::collapse_pp_num(post_prob=Primo_results$post_prob)
```

And the following will provide the posterior probability of being associated with the first study and:

- also the second
- also the third
- also the fourth

```
postprob_traitX <- Primo::collapse_pp_trait(post_prob=Primo_results$post_prob,req_idx=1)
```

1.3 Estimating the false discovery rate (FDR)

Primo estimates the false discovery rate (FDR) at specified posterior probability threshold(s) to guide selection of a threshold for inference. For probability threshold λ and a vector $\hat{P}$ of the estimated probabilities of each variant belonging to the (possibly collapsed) pattern of interest, the estimated FDR is given by

$$estFDR(\lambda) = \frac{\sum_i (1 - \hat{P}_i) 1(\hat{P}_i \geq \lambda)}{\#\{\hat{P}_i \geq \lambda\}}$$

where the index i represents an observation. The following would estimate the FDR if we used a threshold of 0.8 to identify observations associated with all four studies:

```
Primo::calc_fdr(Primo_results$post_prob[,16],thresh=0.8)
```

We can also estimate the FDR for collapsed probabilities and/or use a grid of possible thresholds to guide selection of an appropriate threshold. For example:

```
sapply(seq(0.95,0.75,-0.05),
       function(th) Primo::calc_fdr(postprob_atLeastN[,"PP_ge2"],thresh=th))
```

Primo tailored to provide mechanistic interpretations of trait-associated SNPs

Beyond its use as a general integrative analysis tool, Primo incorporates tailored developments to provide molecular mechanistic interpretations of known complex trait-associated SNPs by integrating summary statistics from GWAS and QTL studies. Primo takes as input m sets of summary statistics from J complex trait and omics studies, and then:

1. Estimates key parameters and results as described in the general Primo framework (e.g. `post_prob` and `pis`) using all SNPs in the genome across all J complex and omics traits.
2. Focuses on the S regions harboring GWAS SNPs to obtain the probability of association for SNPs in those regions, and identifies distinct lead omics SNPs.
3. Performs conditional analysis of GWAS SNPs adjusting for distinct lead omics SNPs in each omics trait.
4. Reports which GWAS SNPs are still associated with omics traits after conditional analysis adjusting for lead omics SNPs.
5. Calculates estimated FDR for collapsed patterns of interests (GWAS+at least 1 omics, GWAS+at least 2 omics, etc).

Note that m may be larger than the total number of SNPs if a SNP can be mapped to multiple outcomes (e.g. genes) within the same omics study.

2.1 Estimating key parameters

As an illustrative example, let's assume we have obtained m sets of summary statistics from the associations of genetic variants with 1 complex trait and 3 omics traits (for a total of $J = 4$ traits). As in the general version of Primo, Primo takes as input matrices ($m \times J$) of summary statistics:

- `betas`:coefficient estimates
- `sds`: standard error estimates (of coefficients)
- `dfs`: degrees of freedom for each analysis
- `mafs`: minor allele frequencies
- `N`: number of subjects

Note that `dfs`, `mafs`, and `N` may also be vectors (of length J, m and J, respectively).

We estimate key parameters using all SNPs in the genome by running an integrative analysis using Primo:

```
Primo_results <- Primo(betas=betas,sds=sds,dfs=dfs,alt_props=c(1e-5,rep(1e-3,3)),mafs=mafs,N=N)
```

Here, for `alt_props` (the estimated proportion of statistics that come from the alternative distribution), we specified 10^{-5} for the complex trait ($j = 1$) and 10^{-3} for the 3 omics traits ($j \in \{2, 3, 4\}$).

While not needed by the main Primo function, it is also important to store the associated identifiers for variants and traits forming the m rows of our data. Here, we store them in a data.frame called `myID`:

```
head(myID,5)
#>    SNP  study1 study2         study3   study4
#> 1 SNP1 complex  geneA CpG1-for-geneA proteinA
#> 2 SNP2 complex  geneA CpG1-for-geneA proteinA
#> 3 SNP3 complex  geneA CpG1-for-geneA proteinA
#> 4 SNP4 complex  geneB CpG1-for-geneB proteinB
#> 5 SNP5 complex  geneB CpG1-for-geneB proteinB
```

2.2 Focus on regions harboring GWAS SNPs

Now we can subset the Primo results to the S regions harboring GWAS SNPs. If `myGenes` holds the names of genes in the GWAS regions, then we subset by:

```
gwas_region_idx <- which(myID$study2 %in% myGenes)
Primo_gwas <- Primo::subset_Primo_obj(Primo_results,gwas_region_idx)
myID_gwas <- myID[gwas_region_idx,]
```

(See Identifying gene regions harboring GWAS loci for tip to identify genes in GWAS regions.)

2.3 Conditional analysis

Primo performs conditional analysis to assess whether the trait-association of a particular variant may be due to being in LD with a nearby variant that is a lead SNP for one (or more) of the traits. To conduct conditional analysis, Primo needs:

- `Primo_obj`: list of Primo results, possibly subset by `Primo::subset_Primo_obj`
- `IDs`: data.frame of identifiers for each observation in the Primo results
- `gwas_snps`: character vector of known trait-associated SNPs
- `pvals`: matrix of nominal P-values for the marginal associations in each study
- `LD_mat`: matrix of genotype correlation coefficients (r) for SNPs in GWAS regions
- `snp_info`: data.frame of chromosome/position information for each SNP
- `pp_thresh`: a posterior probability threshold at which to calculate FDR

The `LD_mat` can be estimated using genotypes from one of the studies or a reference dataset (e.g. 1000 Genomes)[2]. Row and column names of `LD_mat` should match the corresponding SNP names in `IDs`.

`snp_info` should be a data.frame with at least three columns:

```
head(snp_info,3)
#>    SNP CHR  POS
#> 1 SNP1   1 1000
#> 2 SNP2   1 1003
#> 3 SNP3   1 1006
```

Note that if `pvals` is from the full results (m rows), then it should also be subset to the GWAS regions if the Primo results were subset to the GWAS regions:

```
pvals <- pvals[gwas_region_idx,]
```

Now we run conditional analysis for the known complex trait-associated SNPs:

```
conditional_results <-
  Primo::run_conditional_gwas(Primo_obj=Primo_gwas,IDs=myID_gwas,
                              gwas_snps=gwas_snps,pvals=pvals,
                              LD_mat=LD_mat,snp_info=snp_info,
                              pp_thresh=0.8, LD_thresh=0.9,
                              dist_thresh=5e3, pval_thresh=1e-2)
```

For each known trait-associated SNP, the function will condition on any lead SNPs that are > 5 Mb away from the trait-associated SNP, provided that those lead SNPs are not in high LD with the trait-associated SNP ($r^2 < 0.9$) and demonstrate some marginal association with the phenotype for which they are the lead SNP ($p < 10^{-2}$). The function will determine omics associations and FDR using posterior probability > 0.8.

`Primo::run_conditional_gwas` returns a list containing two elements: `pp_grouped` and `fdr`, described in the next two sections.

[2] 1000 Genomes Project: http://www.internationalgenome.org/

2.4 GWAS SNPs still associated with omics traits after conditional analysis

We use the list element `pp_grouped` returned by `Primo::run_conditional_gwas` to determine which trait-associated SNPs are still associated with omics traits after conditional analysis. Note that only results for the SNPs specified in `gwas_snps` are returned. The first $J + 1$ columns of `conditional_results$pp_grouped` hold the SNP and trait identifiers:

```
head(conditional_results$pp_grouped[,1:5],2)
#>    SNP  study1 study2         study3   study4
#> 1 SNP1 complex  geneA CpG1-for-geneA proteinA
#> 2 SNP5 complex  geneB CpG1-for-geneB proteinB
```

The remaining columns of `pp_grouped` hold:

- posterior probabilities of the collapsed association patterns ("GWAS + at least x omics trait(s)")
- # omics traits with which the SNP was associated before conditional analysis (post. prob. $>$ `pp_thresh`)
- # omics traits with which the SNP is associated after conditional analysis
- top association pattern before conditional analysis
- top association pattern after conditional analysis
- omics trait associations for SNP based on top patterns before and after conditional analysis

```
head(conditional_results$pp_grouped[,6:ncol(conditional_results$pp_grouped)],2)
#>   pp_nQTL_ge1 pp_nQTL_ge2 pp_nQTL_ge3 nQTL_orig nQTL_final
#> 1        0.94        0.86        0.48         2          2
#> 2        0.88        0.81        0.34         2          1
#>   top_pattern_precond top_pattern_postcond poss_QTL_assoc
#> 1                  13                   13  study2;study4
#> 2                  12                    6         study2
```

Note that because a known trait-associated SNP may be mappable to multiple outcomes (e.g. genes) for the same omics trait, there may be more than one row in `conditional_results$pp_grouped` for a given trait-associated (GWAS) SNP. This allows the user to identify all outcomes (e.g. genes) with which the SNP may be associated. Often, a SNP-level summary will be desirable. The following will provide a SNP-level summary of the number of omics associations across all outcomes:

```
pp_grouped_maxN <- conditional_results$pp_grouped %>%
                     dplyr::group_by(SNP) %>%
                     dplyr::slice(which.max(nQTL_final))
pp_grouped_maxN <- data.frame(pp_grouped_maxN)[,c("SNP","nQTL_final")]
```

2.5 Estimating the false discovery rate (FDR)

In the list element `fdr`, the function `Primo::run_conditional_gwas` returns a named vector of the estimated false discovery rates (FDR) for each of the collapsed association patterns ("GWAS + at least x omics trait(s)", for $x \in \{1, ..., J\}$).

The false discovery rate (FDR) is estimated in similar fashion to the general version of Primo. However, after conditional analysis, we adjust the numerator to account for SNPs which change pattern after conditional analysis and no longer match the collapsed pattern description since we consider them to be estimated false discoveries. For SNPs which change pattern after conditional analysis such that they no longer match the collapsed association pattern (e.g. "GWAS + at least x omics trait(s)"), their contribution to the numerator of the following equation is corrected to be 1:

$$estFDR(\lambda) = \frac{\sum_i (1 - \hat{P}_i) 1(\hat{P}_i \geq \lambda)}{\#\{\hat{P}_i \geq \lambda\}}$$

Primo for integrating *P*-values

3.1 Integrating *P*-values

In addition to integrating effect sizes and standard errors (i.e. t-statistics), Primo can also integrate P-values or other second-order association statistics. For example, if effect sizes and standard errors are not available, Primo can perform integrative analysis of m sets of P-values from J studies as in the following:

```
Primo_results <- Primo(pvals=pvals,alt_props=c(1e-5,rep(1e-3,3)),use_method="pval")
```

Here, `pvals` is a matrix ($m \times J$) of the (marginal) association P-values. For `alt_props` (the estimated proportion of statistics that come from the alternative distribution), we specified 10^{-5} for the first study and 10^{-3} for the other 3 studies (thus $J = 4$ in the example). We specified `use_method="pval"` so that Primo did not try to run the default t-statistics version.

The P-value version of Primo returns a list of 7 elements, the first four of which have the same interpretation as when running `Primo` with effect sizes and standard errors (i.e. the t-statistic version). The primary elements of interest are again:

- `Primo_results$post_prob`: the posterior probabilites of each association pattern for each observation ($m \times 2^J$ matrix)
- `Primo_results$pis`: the estimated proportions of all observations belonging to each association pattern (vector of length 2^J)

The remaining elements are returned largely for use by other functions.

Tips and tricks

4.1 Creating input matrices

In many cases, the summary statistics from different studies or traits will be stored in multiple places. To create the necessary inputs for Primo, we recommend utilizing functions from the data.table[3] package to read and align the data.

For example, lets's say we are interested in integrating complex trait-GWAS summary statistics (stored in a file titled "GWAS_results.txt") with eQTL summary statistics (stored in "expression_results.txt"). We start by reading in the data:

```
library(data.table)
gwas_stats <- data.table::fread("GWAS_results.txt")
eqtl_stats <- data.table::fread("expression_results.txt")
```

which may contain the following fields:

```
colnames(gwas_stats)
#> [1] "SNP"   "trait" "beta"  "sd"    "pval"  "df"    "maf"   "N"
colnames(eqtl_stats)
#> [1] "SNP"  "gene" "beta" "sd"   "pval" "df"   "maf"  "N"
```

Now we align the data by merging:

```
colnames(gwas_stats)[3:ncol(gwas_stats)] <-
  paste(colnames(gwas_stats)[3:ncol(gwas_stats)],"g",sep="_")
colnames(eqtl_stats)[3:ncol(eqtl_stats)] <-
  paste(colnames(eqtl_stats)[3:ncol(eqtl_stats)],"e",sep="_")

data.table::setkey(gwas_stats,SNP)
data.table::setkey(eqtl_stats,SNP)
merged_stats <- merge(gwas_stats,eqtl_stats)
```

While the first two commands aren't necessary, appending identifiers (e.g. "g" for GWAS; "e" for expression) to common variable names can make later processing clearer and easier (rather than letting `merge` append the defaults ".x" and ".y"). From our merged dataset `merged_stats`, it is easy to create the set of input matrices since the data is now properly aligned:

```
myID <- subset(merged_stats, select=c(SNP,trait,gene))

betas <- as.matrix(subset(merged_stats, select=paste("beta",c("g","e"),sep="_")))
sds <- as.matrix(subset(merged_stats, select=paste("sd",c("g","e"),sep="_")))
pvals <- as.matrix(subset(merged_stats, select=paste("pval",c("g","e"),sep="_")))
dfs <- as.matrix(subset(merged_stats, select=paste("df",c("g","e"),sep="_")))
mafs <- as.matrix(subset(merged_stats, select=paste("maf",c("g","e"),sep="_")))
```

[3]data.table package: https://cran.r-project.org/web/packages/data.table/index.html

There may be situations where we wish to merge/align the data by more than just SNP. For example, we may wish to match pairs of gene expression and protein abundance sets of summary statistics (so that a protein is aligned with the gene from which it is translated). After ensuring that the name(s) of the additional matching variable(s) match across the datasets, the merge step can be modified thusly:

```
data.table::setkeyv(eqtl_stats,c("SNP","gene"))
data.table::setkeyv(pqtl_stats,c("SNP","gene"))
merged_stats <- merge(eqtl_stats,pqtl_stats)
```

4.2 Providing mechanistic interpretations of trait-associated SNPs

4.2.1 Identifying gene regions harboring GWAS loci

If the genes in cis-regions harboring GWAS loci are not provided or known in advance, we can utilize the data to identify such regions. Since the trait-associated SNPs *are* known in advance, we store their identifiers in a vector: `gwas_snps`. Next we identify the indices of our identifier data.frame (`myID`), where the SNP is one of the known trait-associated SNPs, and use that information to identify genes in GWAS regions:

```
head(myID,4)
#>    SNP  study1 study2         study3   study4
#> 1 SNP1 complex  geneA CpG1-for-geneA proteinA
#> 2 SNP2 complex  geneA CpG1-for-geneA proteinA
#> 3 SNP3 complex  geneA CpG1-for-geneA proteinA
#> 4 SNP4 complex  geneB CpG1-for-geneB proteinB
```

```
gwas_snps_idx <- which(myID$SNP %in% gwas_snps)
myGenes <- unique(myID$study2[gwas_snps_idx])
```

Now `myGenes` holds the names of genes in the GWAS regions.

APPENDIX B

ADDITIONAL SIMULATIONS FOR PRIMO(MED)

In this appendix, we present additional simulation results studying the behavior of product of coefficient test statistics, which are used as input in the Primo(med) algorithm.

B.1 The "genome-wide" distribution of product of coefficient test statistics

To motivate the Primo algorithm for integrating product of coefficient test statistics, we conducted simulations to examine the distributions of the Sobel test statistic $z_{\text{Sobel}} = \frac{\hat{\alpha}\hat{\beta}}{\hat{\sigma}_{\hat{\alpha}\hat{\beta}}}$ across the genome. Here, $\hat{\sigma}_{\hat{\alpha}\hat{\beta}} = \sqrt{\hat{\alpha}^2\hat{\sigma}^2_{\hat{\beta}} + \hat{\beta}^2\hat{\sigma}^2_{\hat{\alpha}}}$.

In each simulation, we simulated α and β by random draws of $\alpha \sim N(0, \sigma^2_\alpha)$ and $\beta \sim N(0, \sigma^2_\beta)$, and then generated data for N subjects according to the models:

$$\begin{aligned} X_i &\sim N(0,1) \\ M_i &= \alpha X_i + \varepsilon_i \\ Y_i &= \beta M_i + e_i \end{aligned} \tag{B.1}$$

where $i = 1, ..., N$ is the subject index, $\varepsilon_i \sim N(0,1)$ and $e_i \sim N(0,1)$. Then we analyzed the data using regression equations 3.2 and 3.3 (see Section 3.2.1) to obtain estimates to calculate the Sobel test statistic $z_{\text{Sobel}} = \frac{\hat{\alpha}\hat{\beta}}{\hat{\sigma}_{\hat{\alpha}\hat{\beta}}}$. For each setting, we repeated the simulation one million times to obtain a distribution of Sobel test statistics for the given setting.

In the first set of simulations, we fix $\beta = 0$ and vary both the sample size N and the value of σ_α such that each setting simulates statistics for an N, σ_α pair. Since $\beta = 0$, all settings represent distributions under the null. The density curves over one million simulations for each setting are shown by the red curves in Figure B.1 (the dotted black curves show the standard normal distribution, for reference). Note that $\sigma_\alpha = 0 \Rightarrow \alpha = 0$ with probability

1. All distributions are symmetric, centered on 0. The distribution has much thinner tails than the standard normal distribution when both $\alpha = 0$ and $\beta = 0$, regardless of sample size (first column in Figure B.1). Note that this special case was recognized by Sobel in his landmark paper since if both α and β are zero then "the variance term vanishes and the initial conditions for application of the delta method are not met" [Sobel, 1982]. When one of the effects is non-zero (in this case, $\alpha \neq 0$ while $\beta = 0$), then the distribution converges towards the standard normal distribution as either the sample size N increases (shown by row, top to bottom) or as the average magnitude of the non-zero effect increases (in this case, α; shown by column, left to right). Simulations where we fixed $\alpha = 0$ and varied the value of σ_β yielded similar conclusions (not shown). Thus, the standard normal distribution may be a conservative but appropriate distribution to model the null distribution of the Sobel test-statistic.

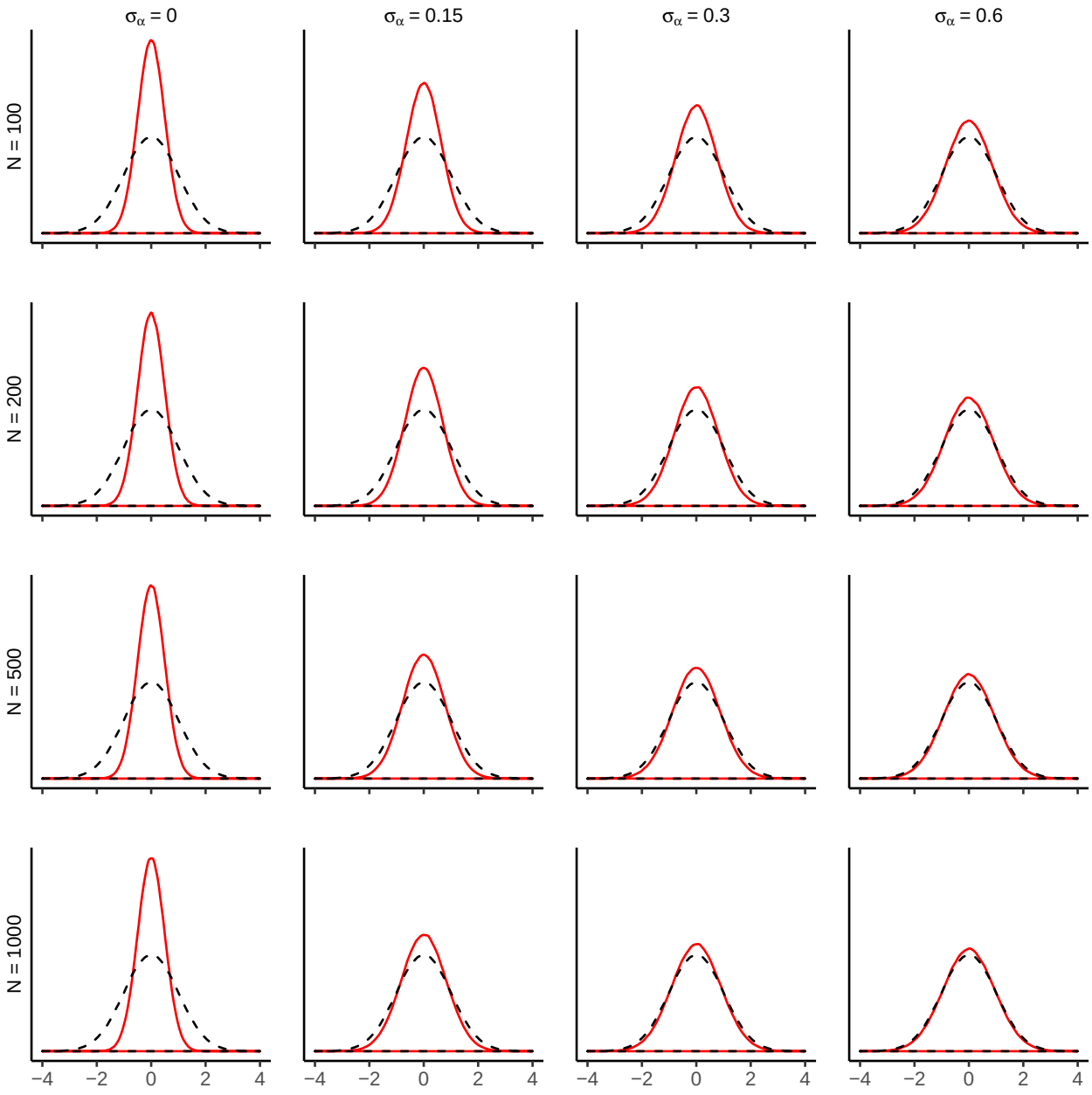


Figure B.1: **Distribution of Sobel test statistics under the null for normally distributed α and varying sample sizes.** The dotted black curve plots the density of $N(0,1)$ for reference. Each red curve shows the density of $z_{\text{Sobel}} = \frac{\hat{\alpha}\hat{\beta}}{\hat{\sigma}_{\hat{\alpha}\hat{\beta}}}$ over one million simulations for pairs of σ_α and sample size N, where $\alpha \sim N(0, \sigma_\alpha^2)$ and $\beta = 0$. In each simulation, we generated data for N subjects using Equations B.1 and obtained estimates of z_{Sobel} by the regression Equations 3.2 and 3.3 in Section 3.2.1. The plots reveal that under the null, the distribution of Sobel test statistics converges to a $N(0,1)$ distribution as either σ_α increases (left to right by row) or the sample size N increases (top to bottom by column).

In the second set of simulations, we fixed the sample size ($N = 200$) and varied the values of σ_α and σ_β such that each setting simulates statistics for a σ_α, σ_β pair. The density curves over one million simulations for each setting are shown by the red curves in Figure B.2 (the dotted black curves show the standard normal distribution, for reference). Note that $\sigma_\cdot = 0$ implies the Dirac delta function (e.g $\sigma_\alpha = 0 \Rightarrow \alpha = 0$ with probability 1). The first row and first column of plots in Figure B.2 represent distributions under the null (where either $\alpha = 0$ or $\beta = 0$ or both) while the nine plots in the blue box in the bottom right corner represent distributions under the alternative (where both $\alpha \neq 0$ and $\beta \neq 0$). All distributions are symmetric, centered on 0. Under the null, the distribution has much thinner tails than the standard normal when both $\alpha = 0$ and $\beta = 0$, and converges towards the standard normal distribution as the average magnitude of the non-zero effect increases (i.e. as either σ_α or σ_β increases in magnitude), consistent with simulations in the previous setting. Under the alternative, the distribution has fatter tails than the standard normal distribution, with a scaling factor that increases in magnitude as σ_α and/or σ_β increase in magnitude. Since the distributions are symmetric and centered on 0 with fatter tails than the standard normal, it may be appropriate to model the alternative distribution using a normal distribution with variance greater than 1: $z|z \neq 0 \sim N(0, \sigma^2_{\text{alt}})$, $\sigma_{\text{alt}} > 1$.

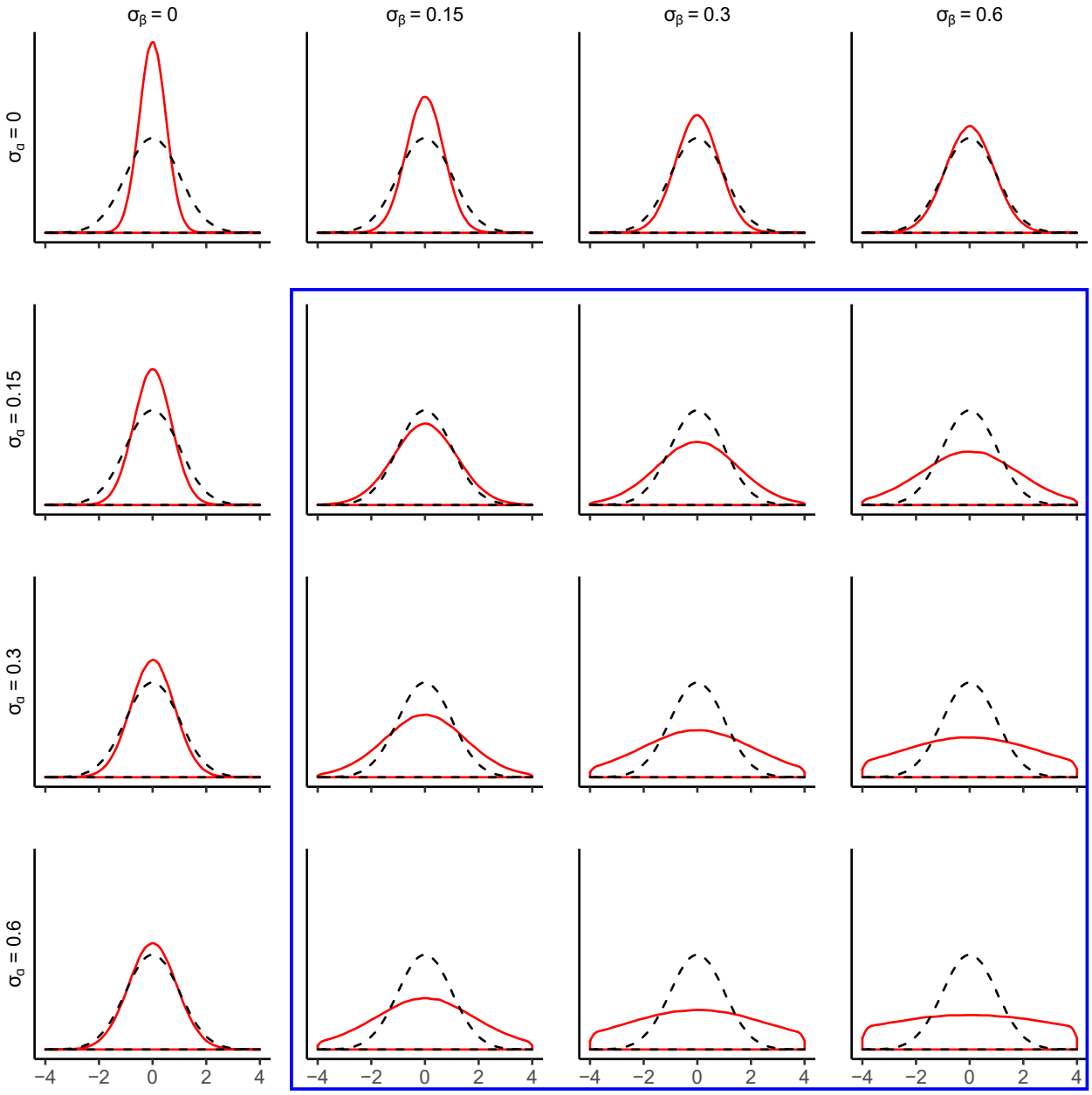


Figure B.2: **Distribution of Sobel test statistics for normally distributed α, β.** The dotted black curve plots the density of $N(0,1)$ for reference. Each red curve shows the density of $z_{\text{Sobel}} = \frac{\hat{\alpha}\hat{\beta}}{\hat{\sigma}_{\hat{\alpha}\hat{\beta}}}$ over one million simulations for pairs of σ_α, σ_β, where $\alpha \sim N(0, \sigma_\alpha^2)$ and $\beta \sim N(0, \sigma_\beta^2)$. In each simulation, we generated data for $N = 200$ subjects using Equations B.1 and obtained estimates of z_{Sobel} by the regression Equations 3.2 and 3.3 in Section 3.2.1. The plots reveal that z_{Sobel} follows a symmetric distribution centered on zero for all pairs. Under the null (top row and first column), the distribution converges to $N(0,1)$ as the average magnitude of the non-zero coefficient increases. Under the alternative (nine graphs in the blue box in the bottom right), the distribution is symmetric around zero with heavier tails than a $N(0,1)$ distribution.

B.2 Using principal components, surrogate variables or PEER factors to adjust for systematic variation and reduce spurious mediation associations due to confounding

In this simulation setting, we illustrate the benefits of using principal components, surrogate variables (SVs) [Leek and Storey, 2007], PEER factors [Stegle et al., 2012], or other dimension reduction techniques to account for common variations in the data. As discussed in Section 3.2.1, adjusting for PCs, SVs or PEER factors created from matrices of mediators and/or dependent variables has the potential to reduce spurious associations due to confounding.

We simulated data for $J = 4$ studies of $m = 100k$ mediation trios based on the following process. In study j, trio i for subject n_j is simulated according to the models:

$$\begin{aligned} X_{n_j ij} &\sim N(0,1) \\ M_{n_j ij} &= \alpha_{ij} X_{n_j ij} + \gamma_{2ij} H_{n_j} + \varepsilon_{M,n_j ij} \\ Y_{n_j ij} &= \beta_{ij} M_{n_j ij} + \gamma_{3ij} H_{n_j} + \varepsilon_{Y,n_j ij} \end{aligned} \tag{B.2}$$

where $H_{n_j} \sim N(0,1)$ is a potential confounding variable, γ's are confounding effects, and other terms have the same interpretation as in Equations 3.4 in Section 3.3. In this scenario, non-zero effects are drawn from normal distributions with mean 0 and variance of 0.5 (e.g. $\gamma_{\cdot ij} | \gamma_{\cdot ij} \neq 0 \sim N(0, 0.5)$). We simulated $\pi_k = (4 \times 10^{-3}, 2 \times 10^{-3}, 1 \times 10^{-3}, 5 \times 10^{-4})$ for patterns with trans-association mediated by cis being present in only one, exactly two, exactly three, and all four studies, respectively. Thus the true alternative proportion $\theta_j^1 = 0.0135$, $\forall j$. Under the null hypothesis, 50% of trios had both $\alpha_{ij} = 0$ and $\beta_{ij} = 0$, while 25% had $\alpha_{ij} \neq 0$ and 25% had $\beta_{ij} \neq 0$. In each study j, 10% of M_{ij}/Y_{ij} pairs were selected to have non-zero confounding effects ($\gamma_{2ij} \neq 0$ and $\gamma_{3ij} \neq 0$).

We are interested in comparing the false discovery rate (FDR) with and without adjustment for variables capturing potential sources of unmeasured confounding. For the un-

adjusted analysis, we calculated Sobel test statistics using the estimators from regression Equations 3.2 and 3.3 in Section 3.2.1 with no adjustment for confounders. In the adjusted analysis, in each regression we also adjust for the top 3 principal components (PCs) generated from a matrix of the mediators and dependent variables (mimicking the process of generating dimension reduction-based variables from a gene expression matrix of all genes or protein abundance matrix of all proteins, for example). More specifically, in each study j, we generate PCs by combining the mediator vectors (M_{ij}, $\forall i$) and dependent variable vectors (Y_{ij}, $\forall i$) into a single matrix and perform principal components analysis on the matrix to extract the loadings.

Figure B.3 compares the false discovery rates (FDR) over 1000 simulations with (solid lines) and without (dotted lines) adjustment for the top 3 PCs created from the matrix combining M_j and Y_j. As shown in the figure, failing to adjust for the top PCs results in inflated FDRs due to confounding by covariate H while adjusting for the PCs allows for better control of the FDR. The simulation demonstrates the benefit of using dimension reduction techniques, such as PCs, SVs or PEER factors, to reduce spurious mediation associations due to confounding.

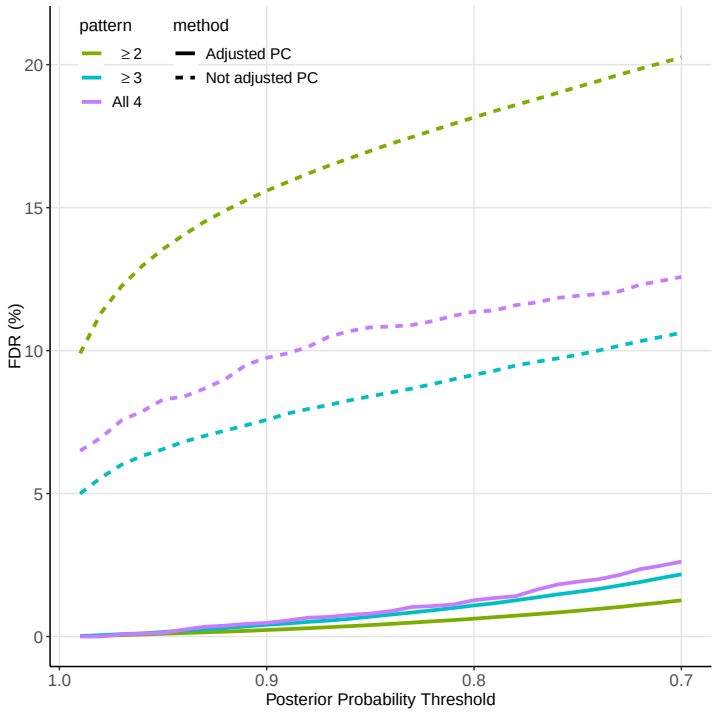
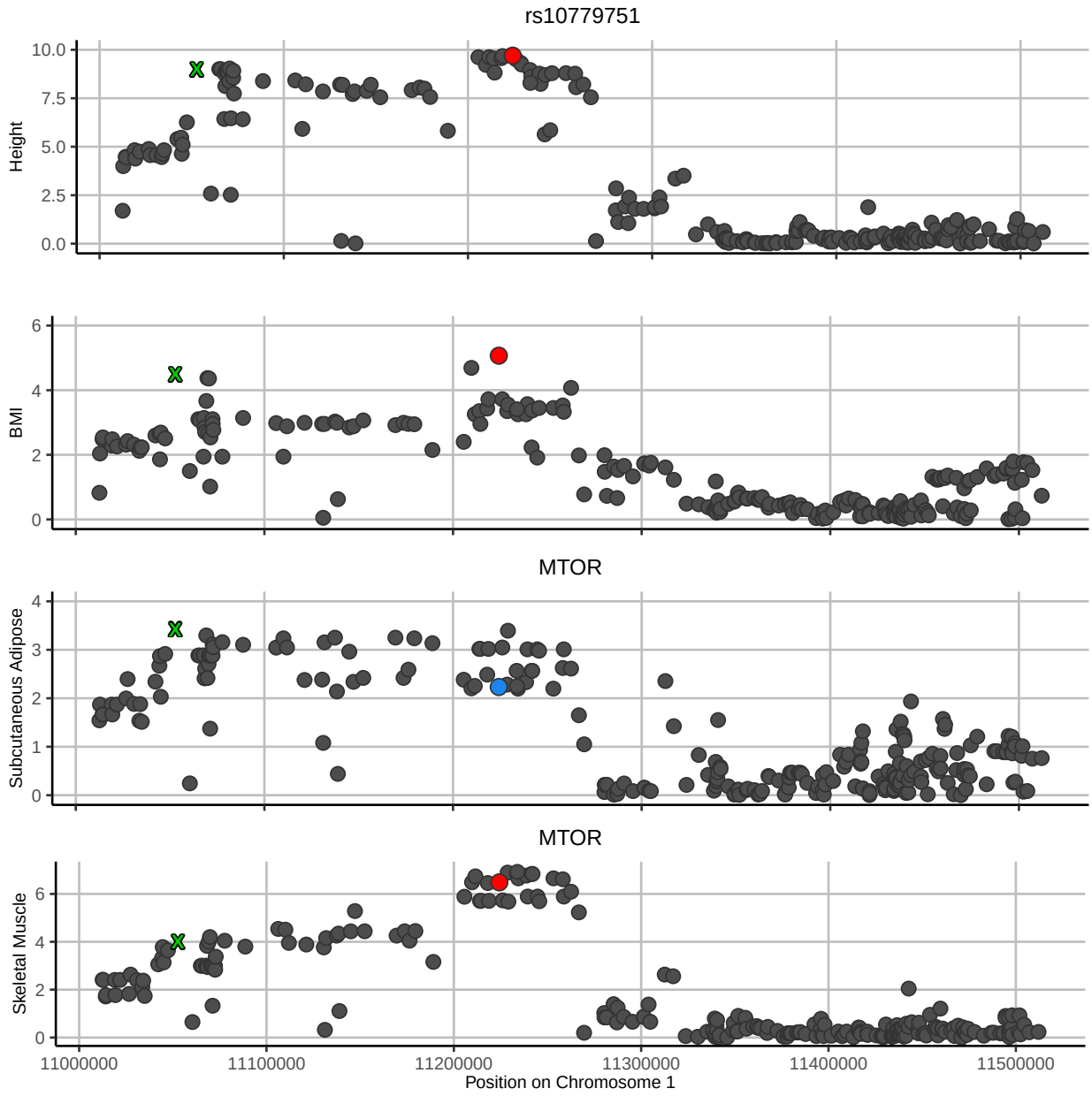


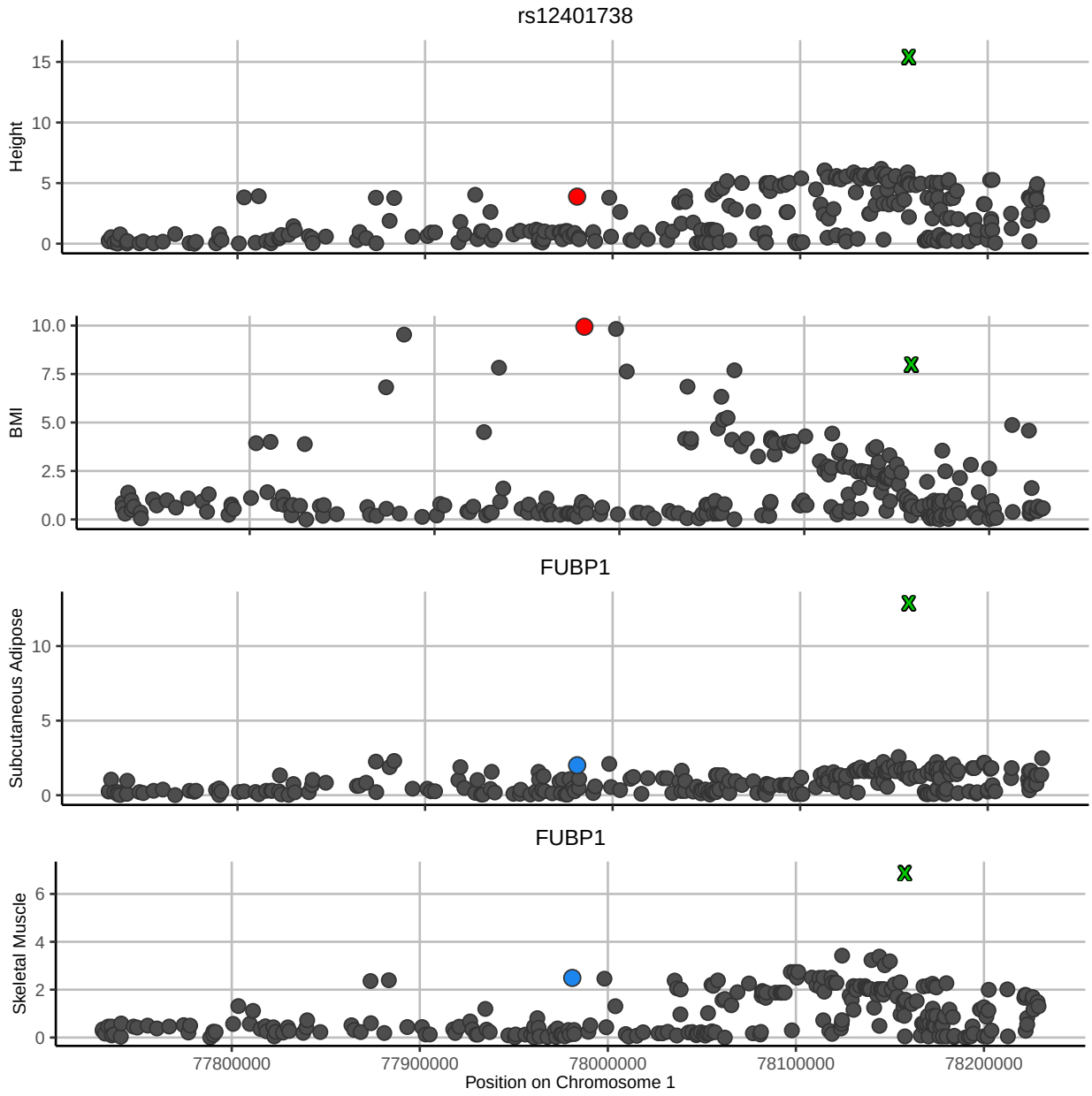
Figure B.3: **Comparing performance of Primo(z)-Sobel with and without adjustment for PCs in the presence of confounding.** Observed false discovery rate (FDR; y-axis) by posterior probability threshold (x-axis) in the presence of a confounder H of the relationship between some mediator (M) and dependent variable (Y) pairs shown over 1000 simulations for Primo(z)-Sobel with (solid lines) and without (dotted lines) adjustment for PCs. Line color represents grouped association pattern ("associated with at least # studies"). Failing to adjust for PCs results in inflated FDRs due to confounding by H while adjusting for PCs allows for better control of the FDR.

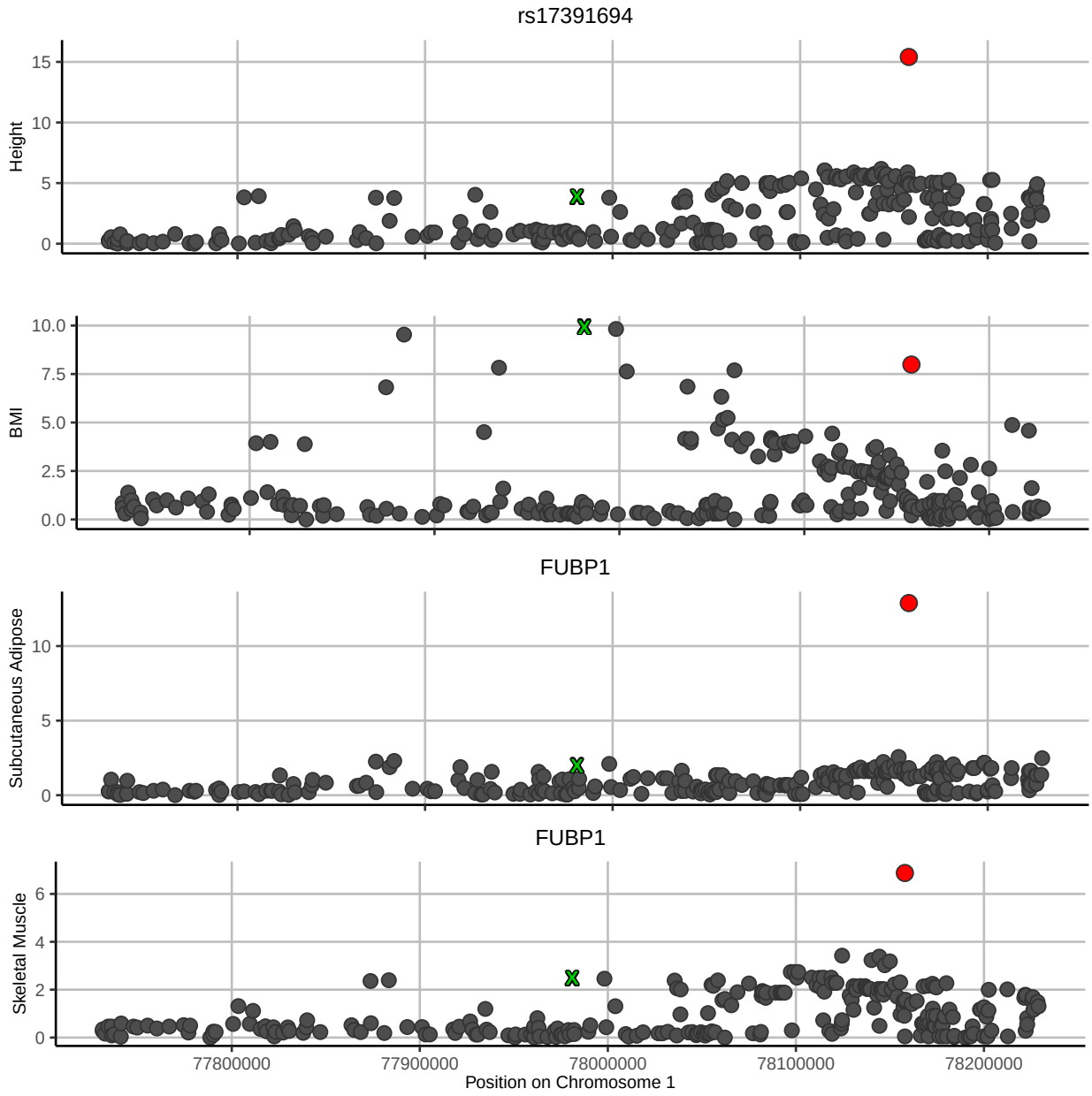
APPENDIX C

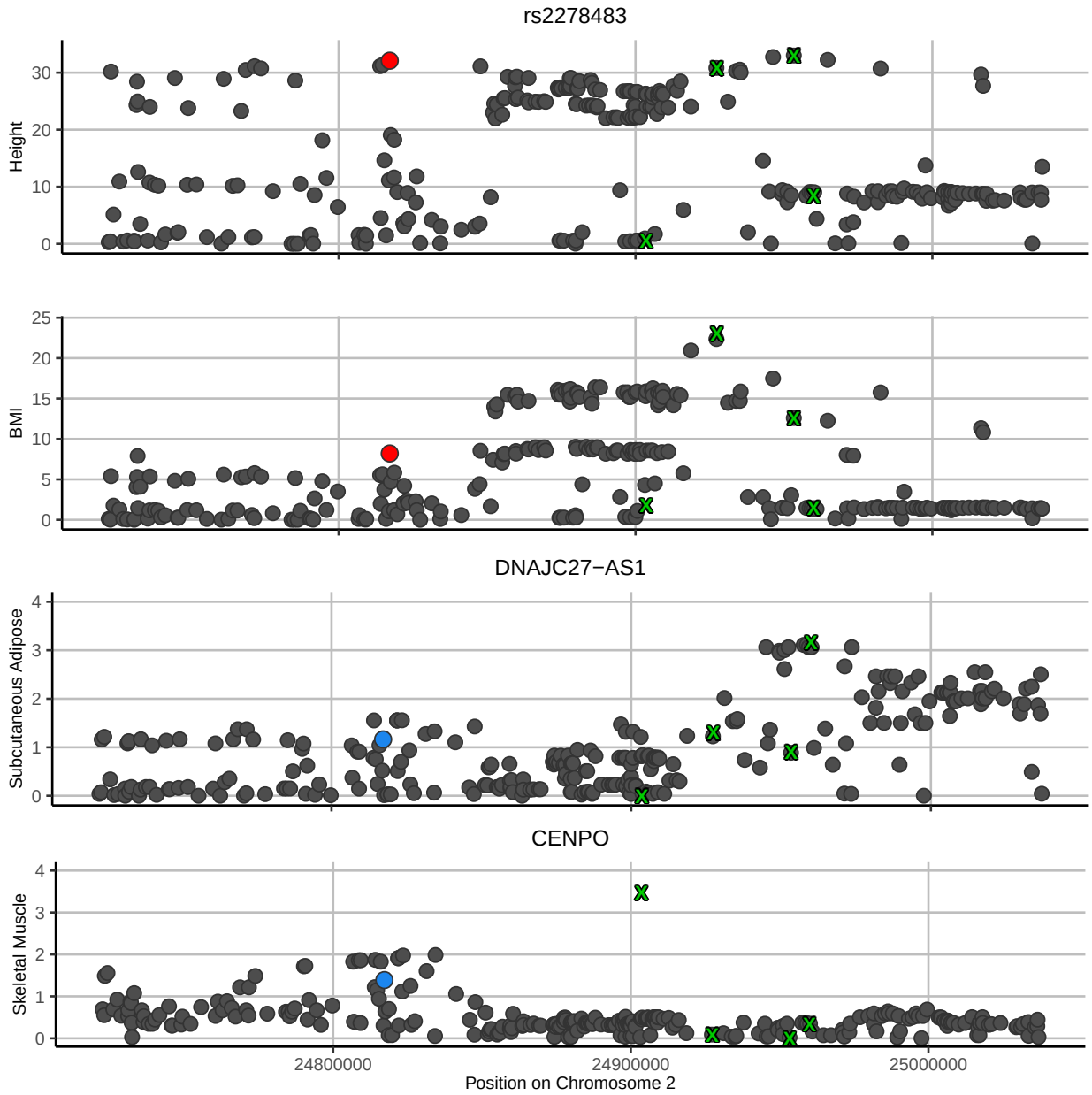
SIGNIFICANT LOCI FROM ANALYSES OF PLEIOTROPY OF HEIGHT AND BMI

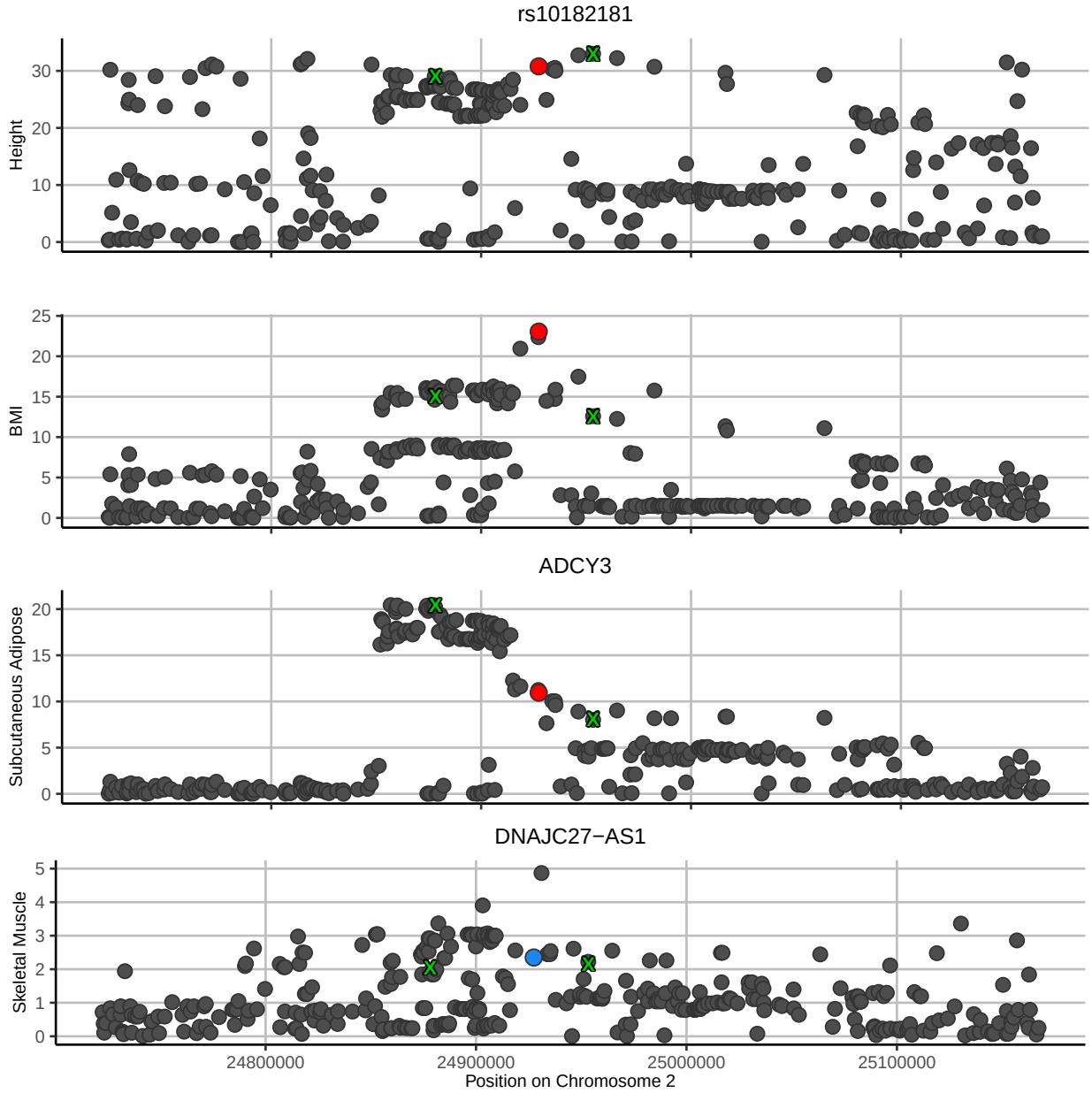
Figure C.1: **Locus-zoom plots of $-\log_{10}(P)$-values for associations with Height, BMI (from GWAS) and gene expression levels (from GTEx eQTL analysis) for gene regions with pleiotropic SNPs being replicated.** Each page shows a set of four association plots for a locus: one for Height, one for BMI, and one for gene expression in each of the two tissue types in the analysis – subcutaneous adipose and skeletal muscle. In each plot, the GWAS-reported SNP is colored red if associated with the given trait and colored blue if not associated with the given trait (at an 80% posterior probability threshold and after conditional association analysis accounting for LD). Lead omics SNPs which were adjusted for in conditional association analysis are shown by a green "X". For any SNP which was associated with expression of multiple genes, the gene with which it has the strongest association is presented in each tissue.

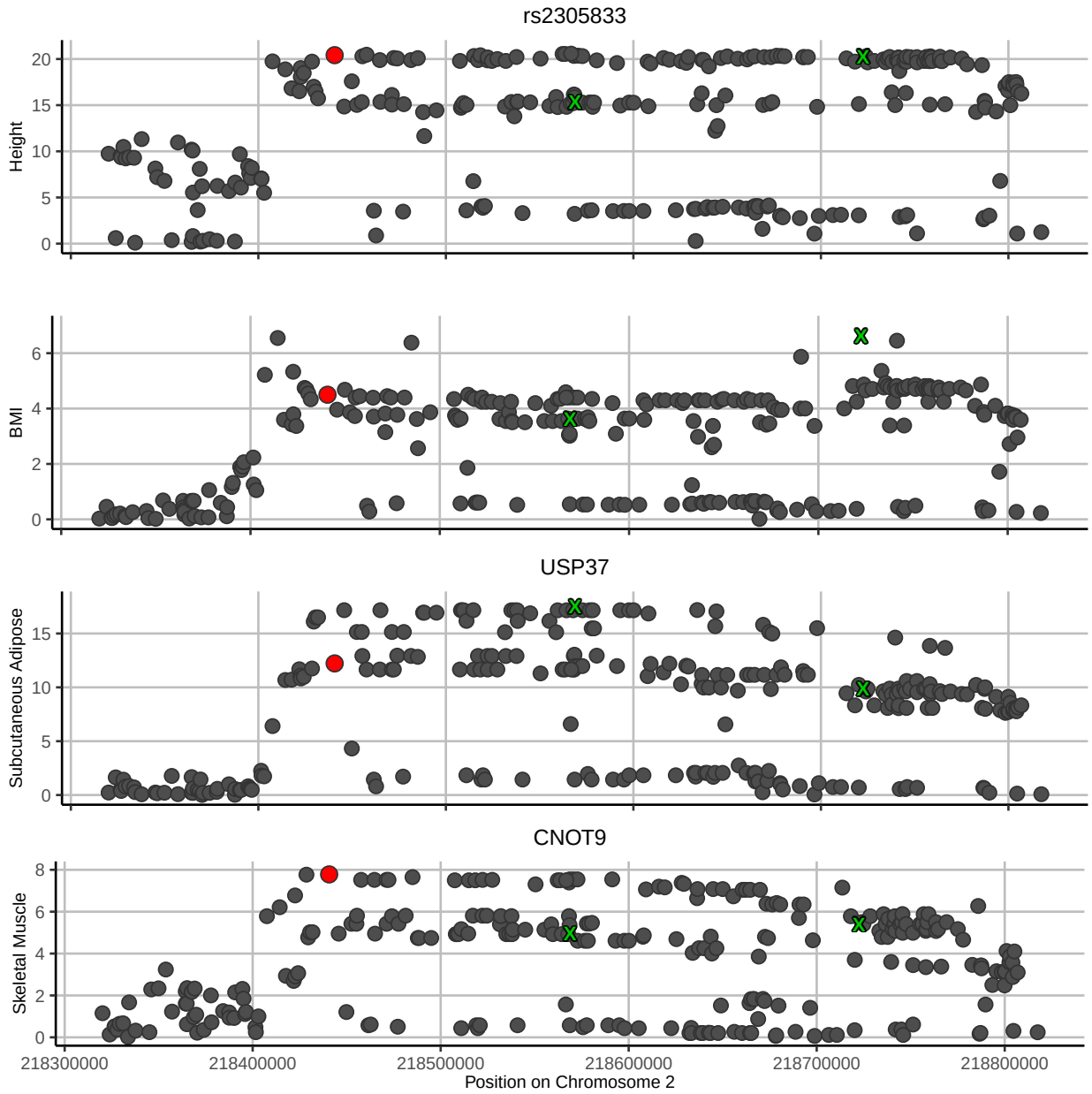

rs10779751
Height
10.0
7.5
5.0
2.5
0.0
BMI
6
4
2
0
MTOR
Subcutaneous Adipose
4
3
2
1
0
MTOR
Skeletal Muscle
6
4
2
0
11000000
11100000
11200000
11300000
11400000
11500000
Position on Chromosome 1

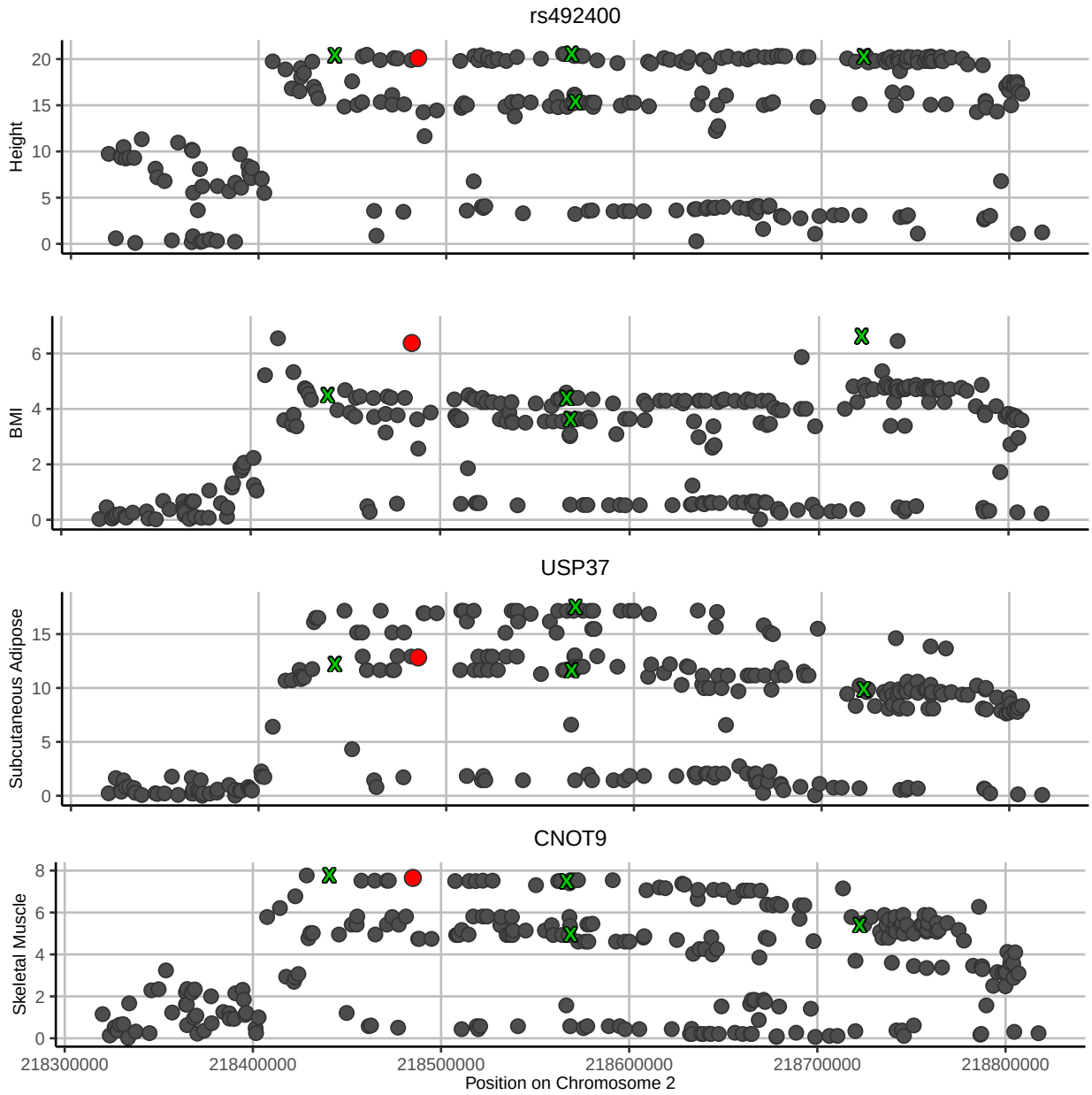

rs12401738
Height
15
10
5
0
BMI
10.0
7.5
5.0
2.5
0.0
FUBP1
Subcutaneous Adipose
10
5
0
FUBP1
Skeletal Muscle
6
4
2
0
77800000
77900000
78000000
78100000
78200000
Position on Chromosome 1

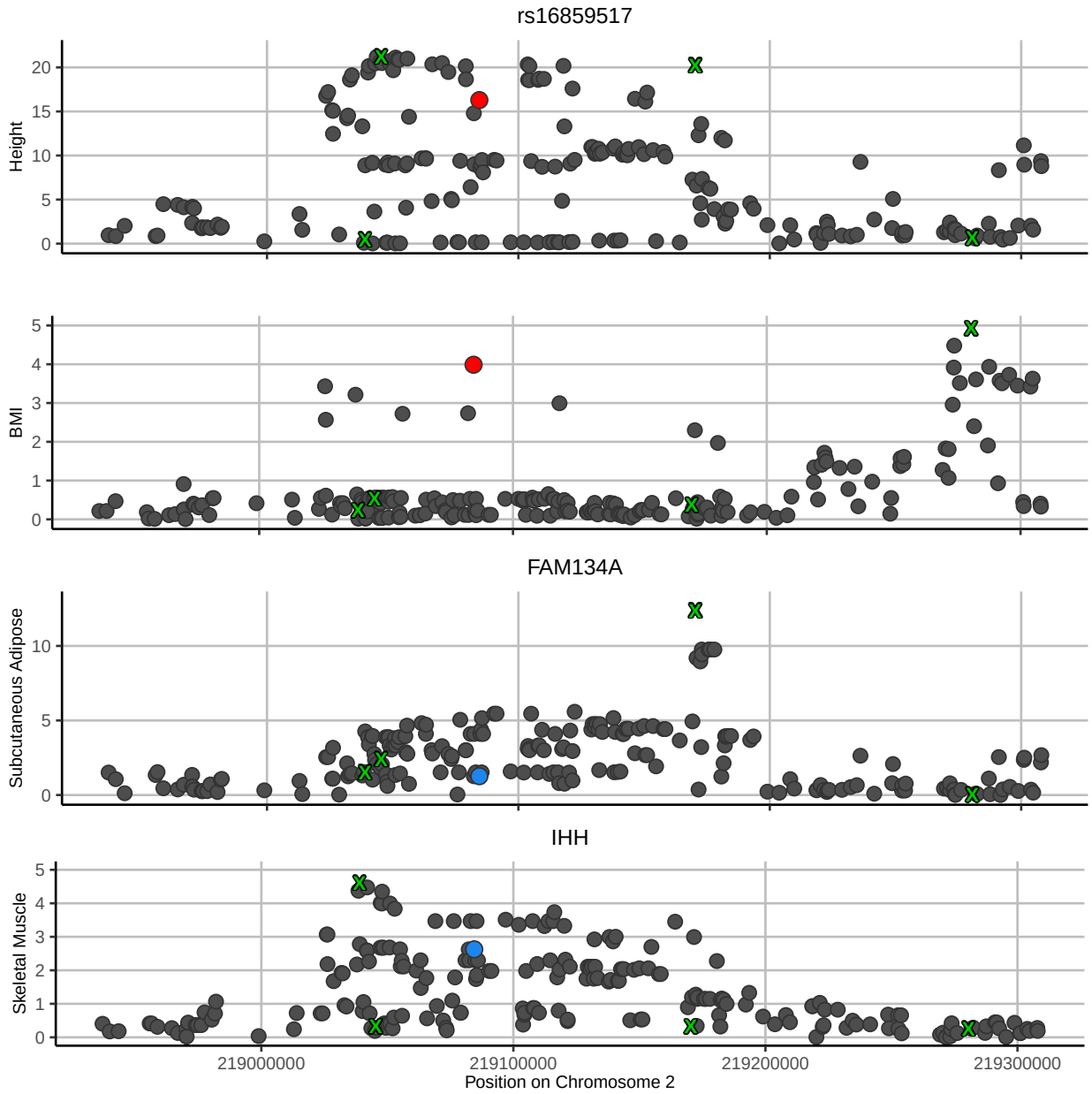

rs17391694
Height
15
10
5
0
BMI
10.0
7.5
5.0
2.5
0.0
FUBP1
Subcutaneous Adipose
10
5
0
FUBP1
Skeletal Muscle
6
4
2
0
77800000
77900000
78000000
78100000
78200000
Position on Chromosome 1

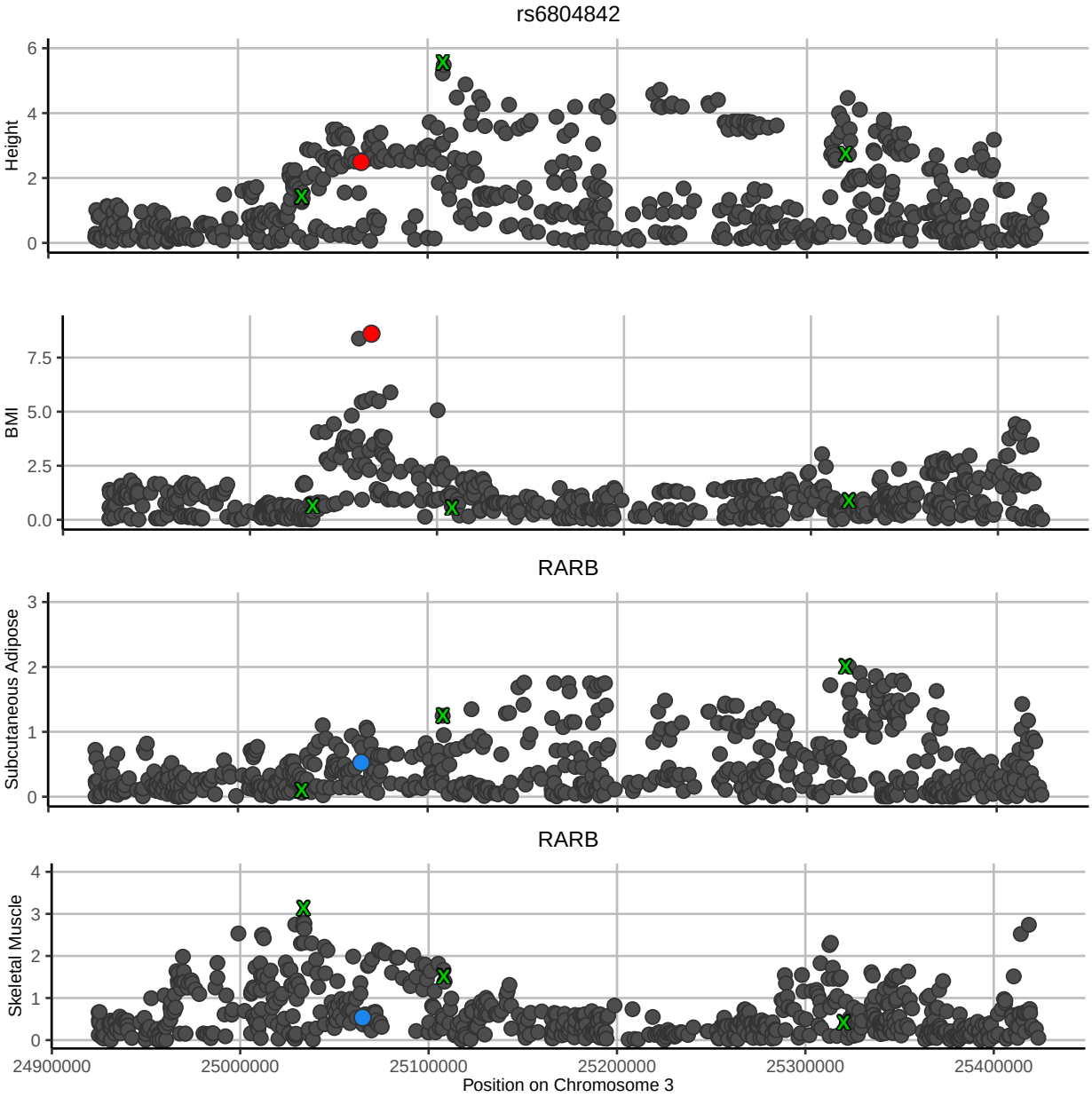

rs2278483
Height
30
20
10
0
BMI
25
20
15
10
5
0
DNAJC27−AS1
Subcutaneous Adipose
4
3
2
1
0
CENPO
Skeletal Muscle
4
3
2
1
0
24800000
24900000
25000000
Position on Chromosome 2

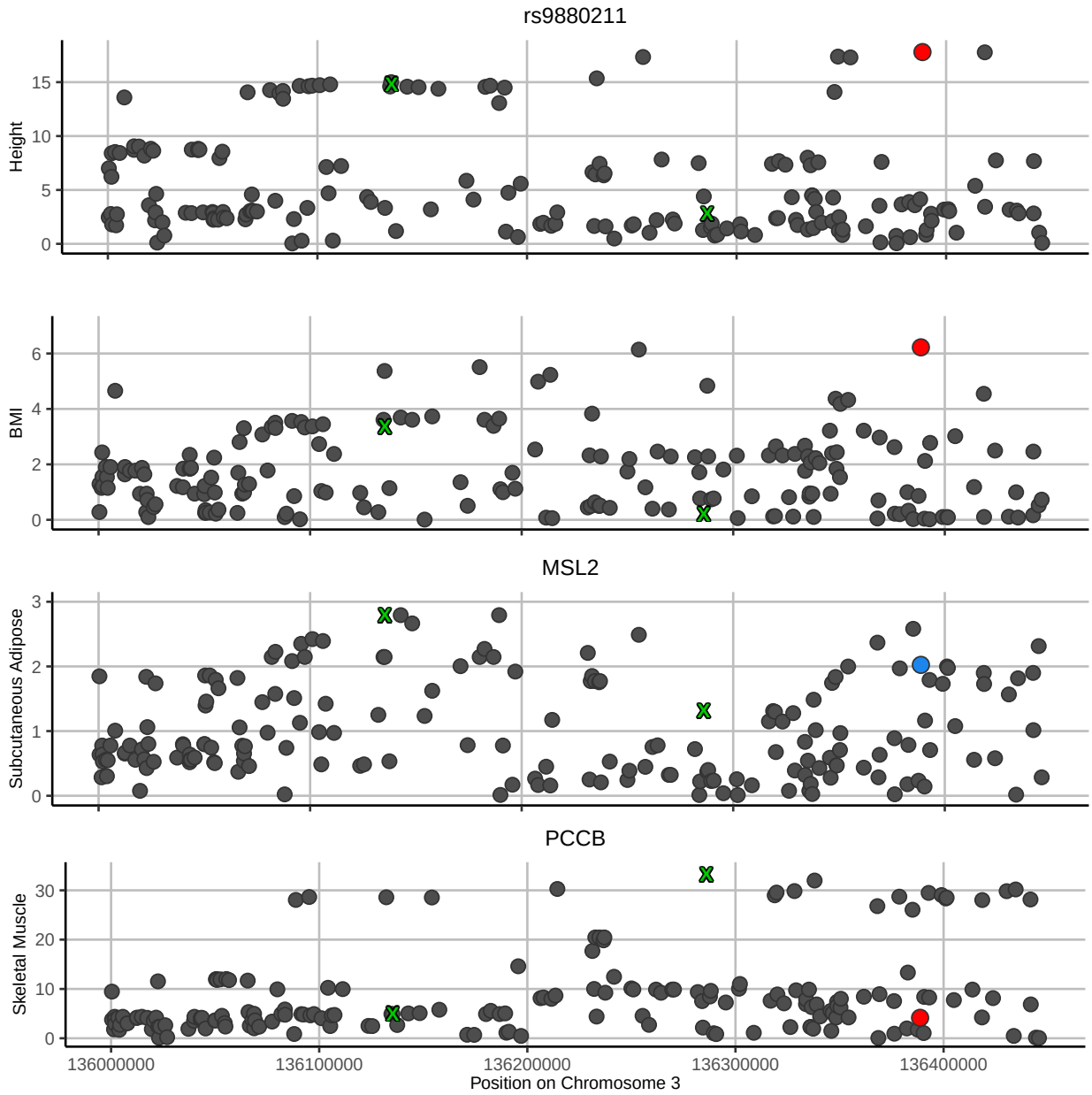

rs10182181
Height
30
20
10
0
BMI
25
20
15
10
5
0
ADCY3
Subcutaneous Adipose
20
15
10
5
0
DNAJC27−AS1
Skeletal Muscle
5
4
3
2
1
0
24800000
24900000
25000000
25100000
Position on Chromosome 2

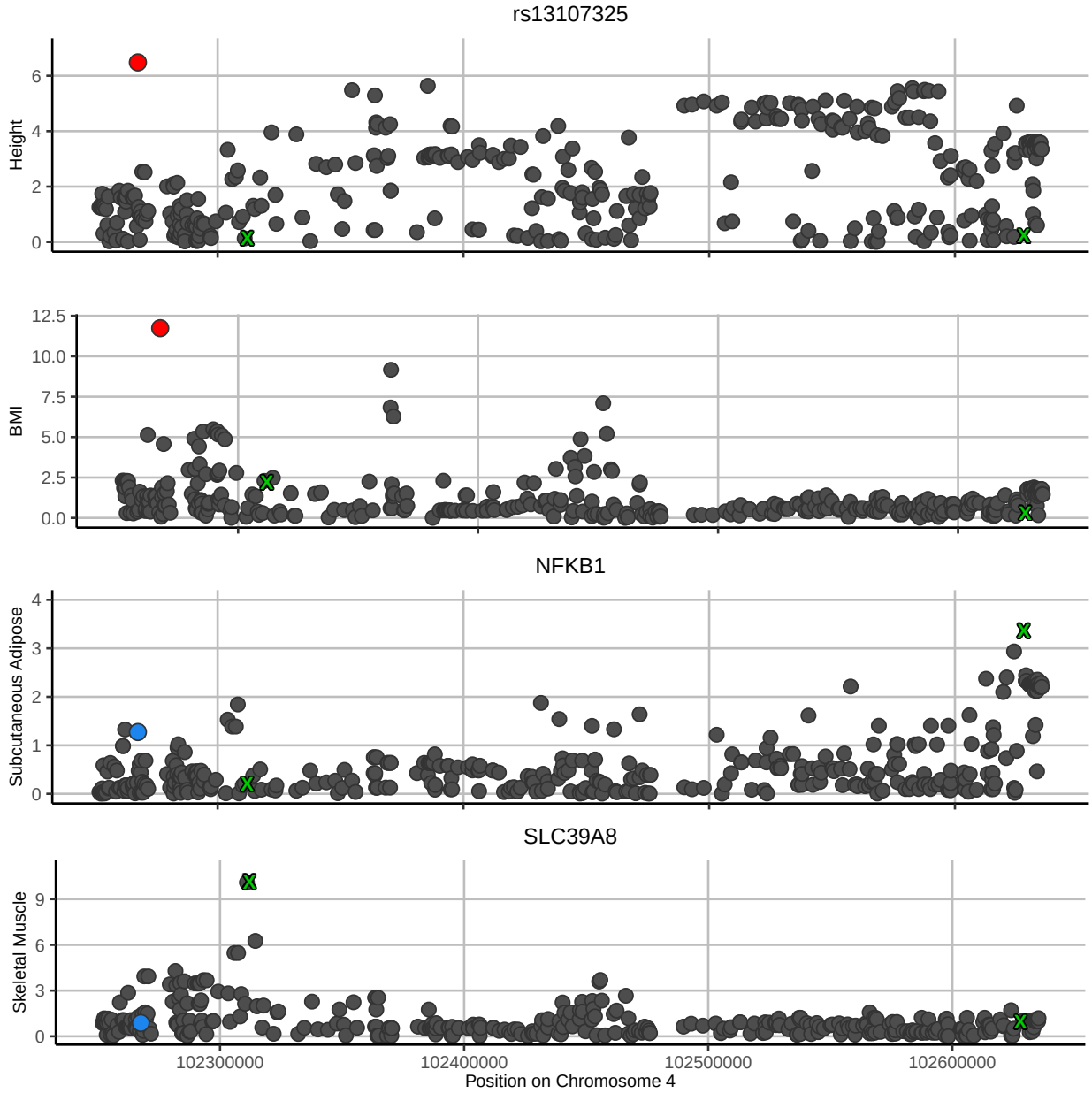

rs2305833
Height
BMI
USP37
Subcutaneous Adipose
CNOT9
Skeletal Muscle
218300000
218400000
218500000
218600000
218700000
218800000
Position on Chromosome 2

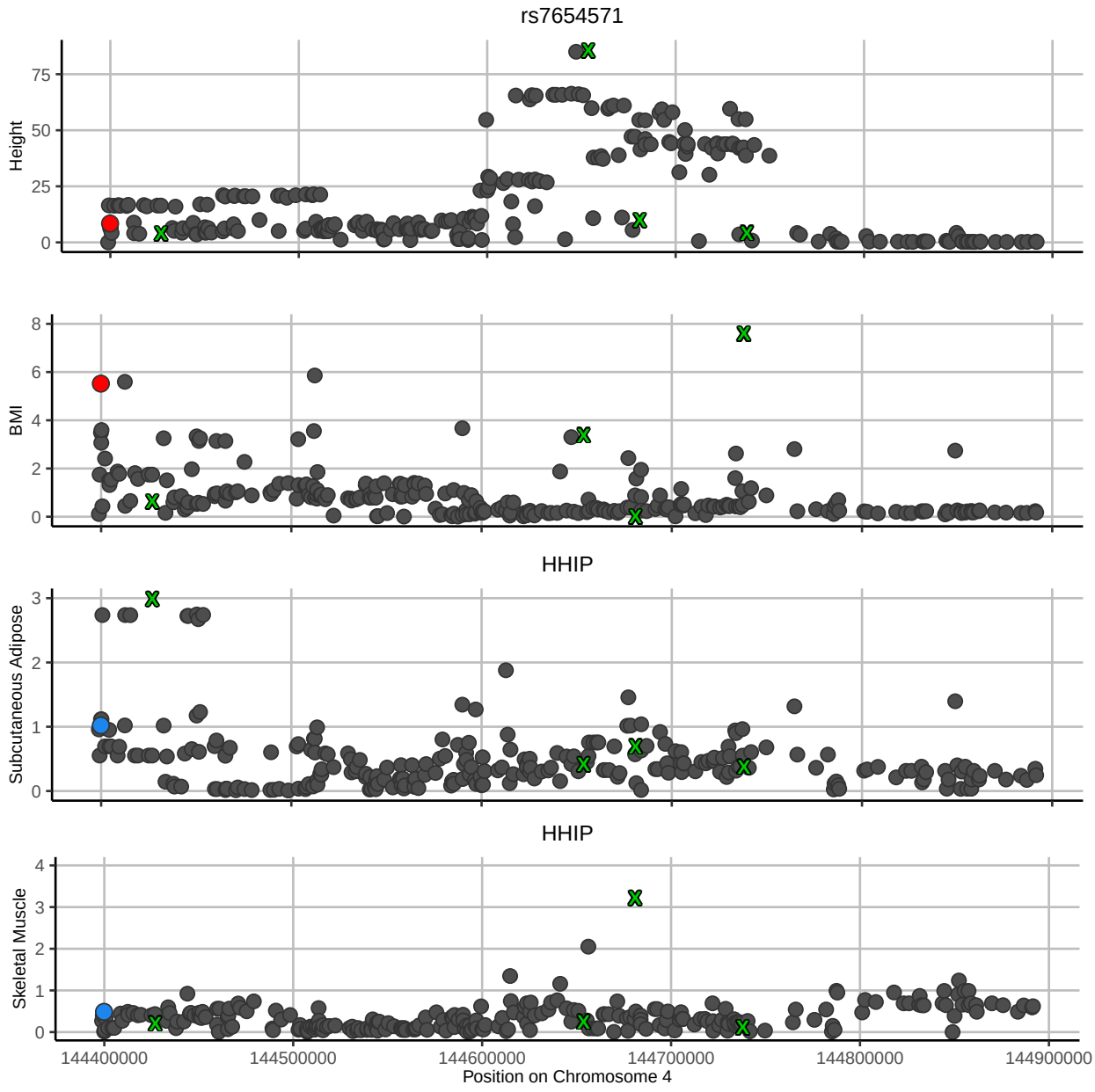

rs492400
Height
BMI
USP37
Subcutaneous Adipose
CNOT9
Skeletal Muscle
218300000
218400000
218500000
218600000
218700000
218800000
Position on Chromosome 2

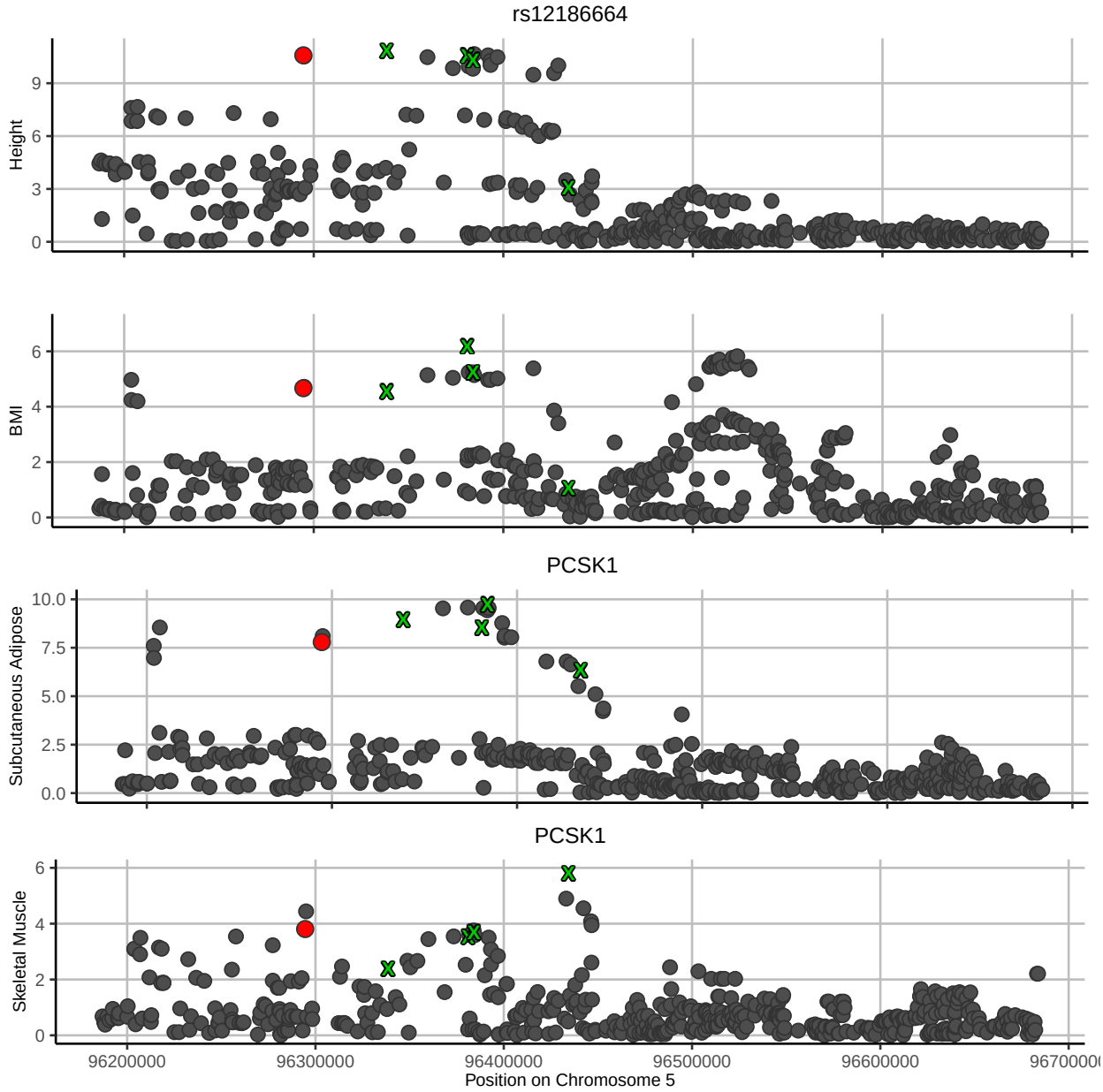

rs16859517
Height
BMI
FAM134A
Subcutaneous Adipose
IHH
Skeletal Muscle
219000000
219100000
219200000
219300000
Position on Chromosome 2

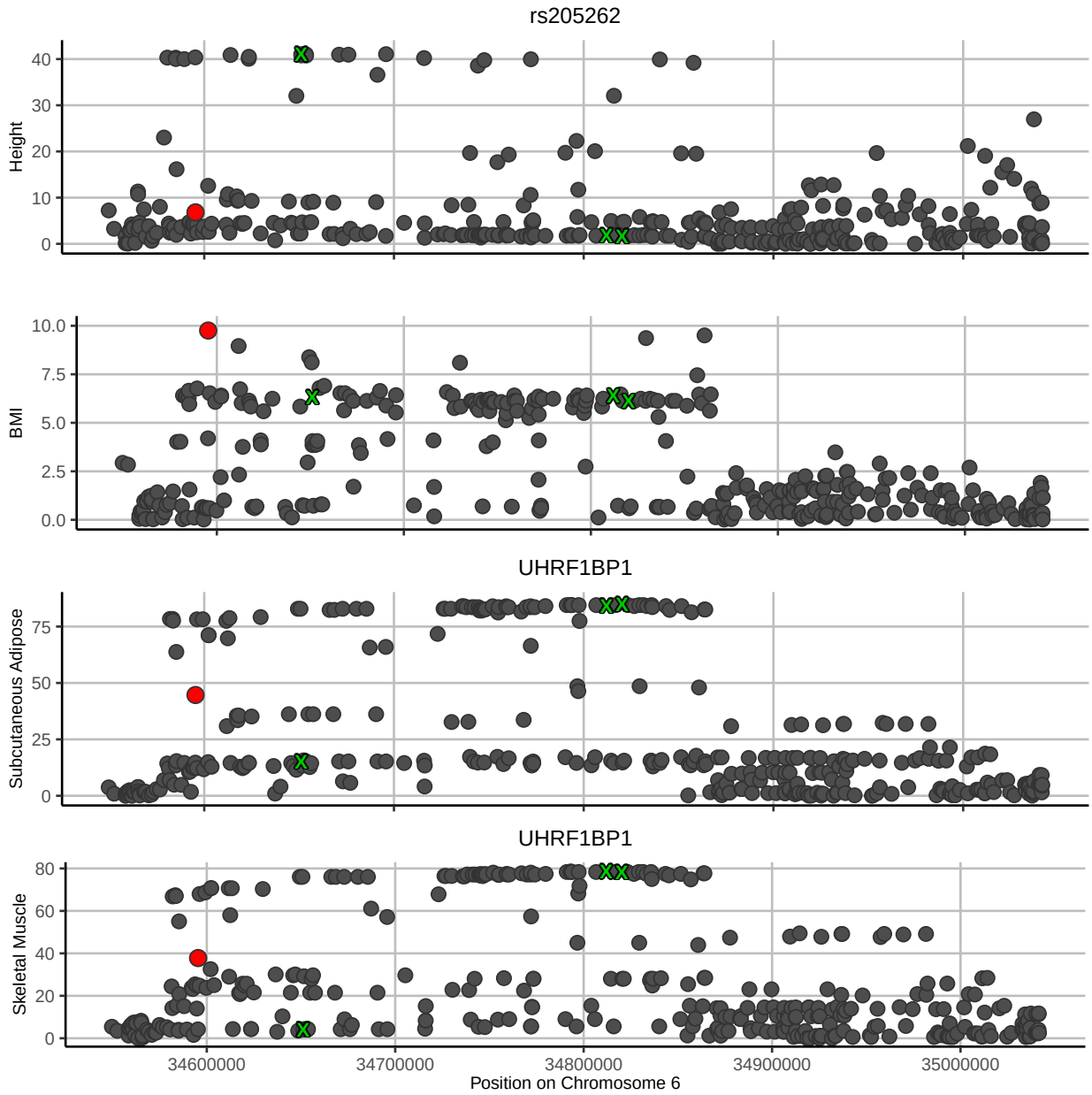

rs6804842
Height
BMI
RARB
Subcutaneous Adipose
RARB
Skeletal Muscle
24900000
25000000
25100000
25200000
25300000
25400000
Position on Chromosome 3

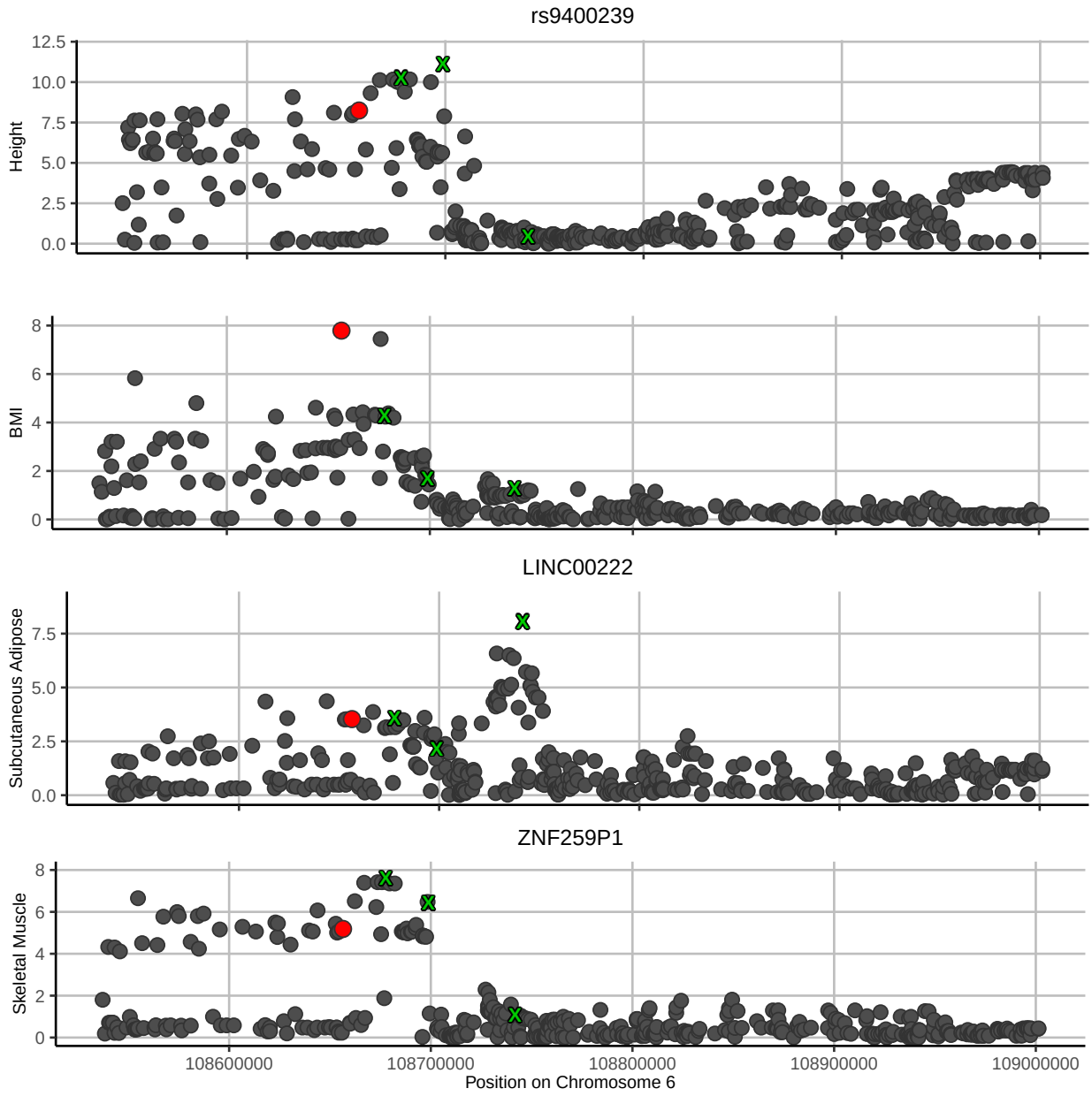

rs9880211
Height
BMI
MSL2
Subcutaneous Adipose
PCCB
Skeletal Muscle
136000000
136100000
136200000
136300000
136400000
Position on Chromosome 3

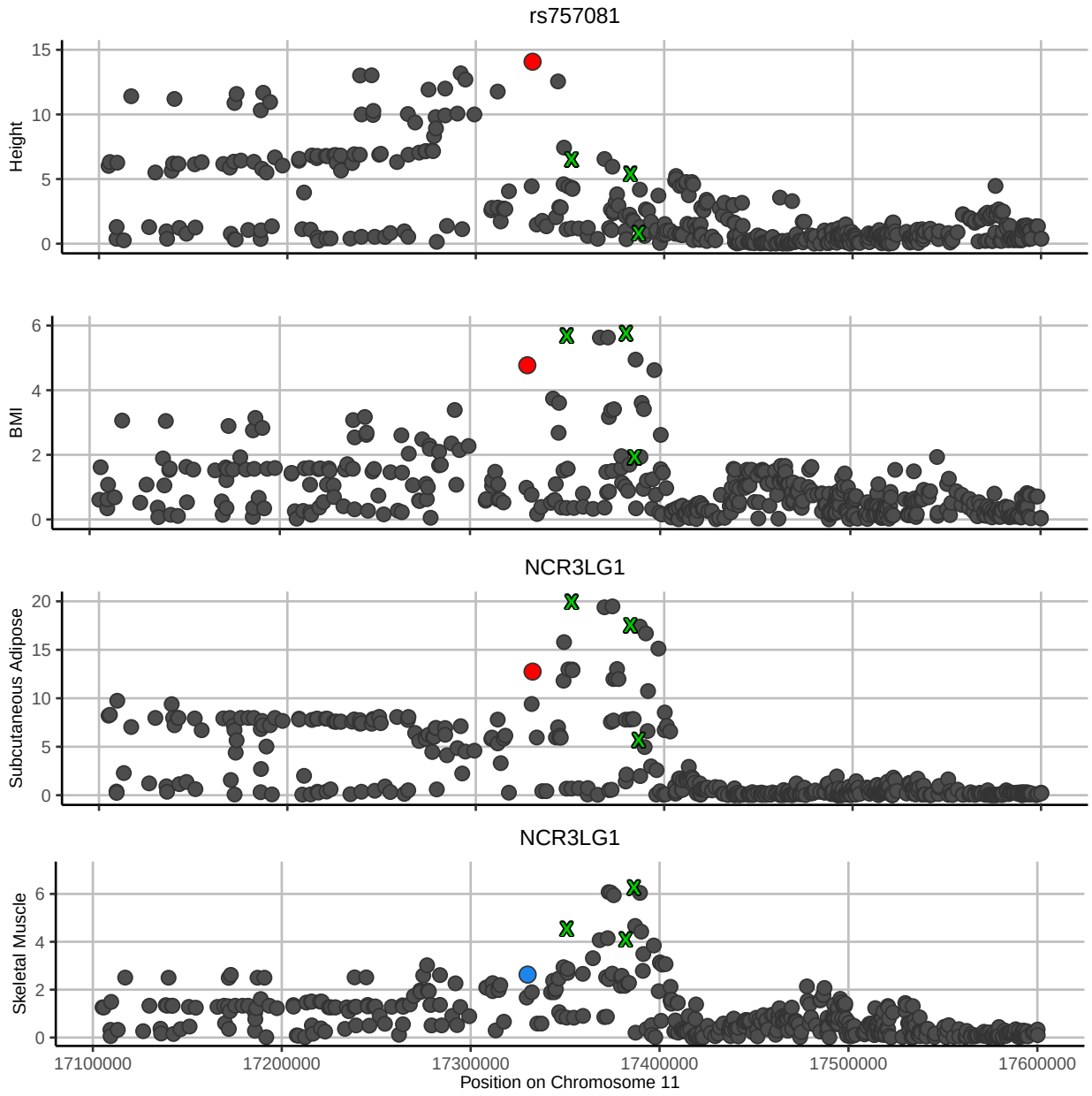

rs13107325
Height
BMI
NFKB1
Subcutaneous Adipose
SLC39A8
Skeletal Muscle
102300000
102400000
102500000
102600000
Position on Chromosome 4

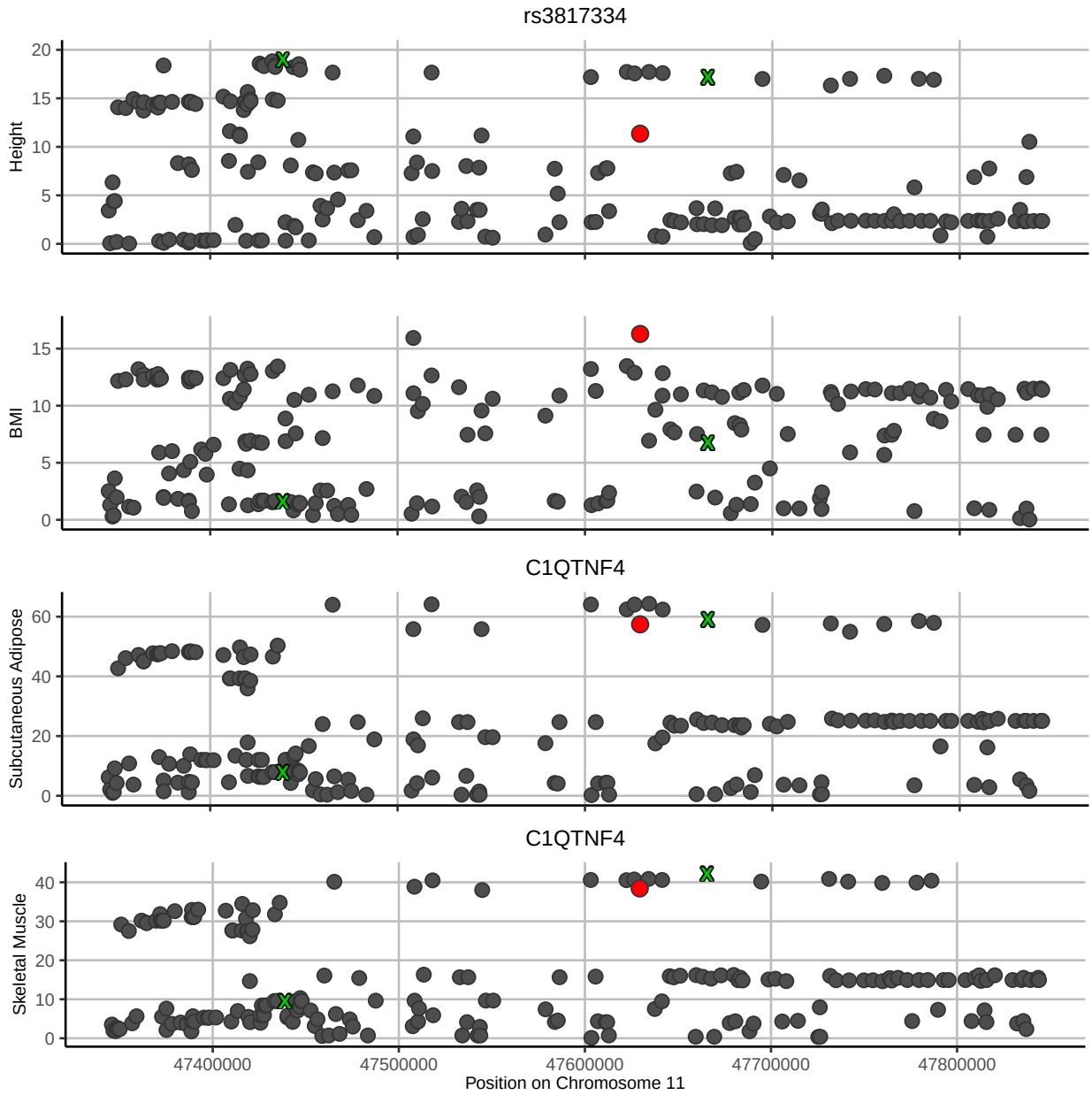

rs7654571
Height
75
50
25
0
BMI
8
6
4
2
0
HHIP
Subcutaneous Adipose
3
2
1
0
HHIP
Skeletal Muscle
4
3
2
1
0
144400000
144500000
144600000
144700000
144800000
144900000
Position on Chromosome 4

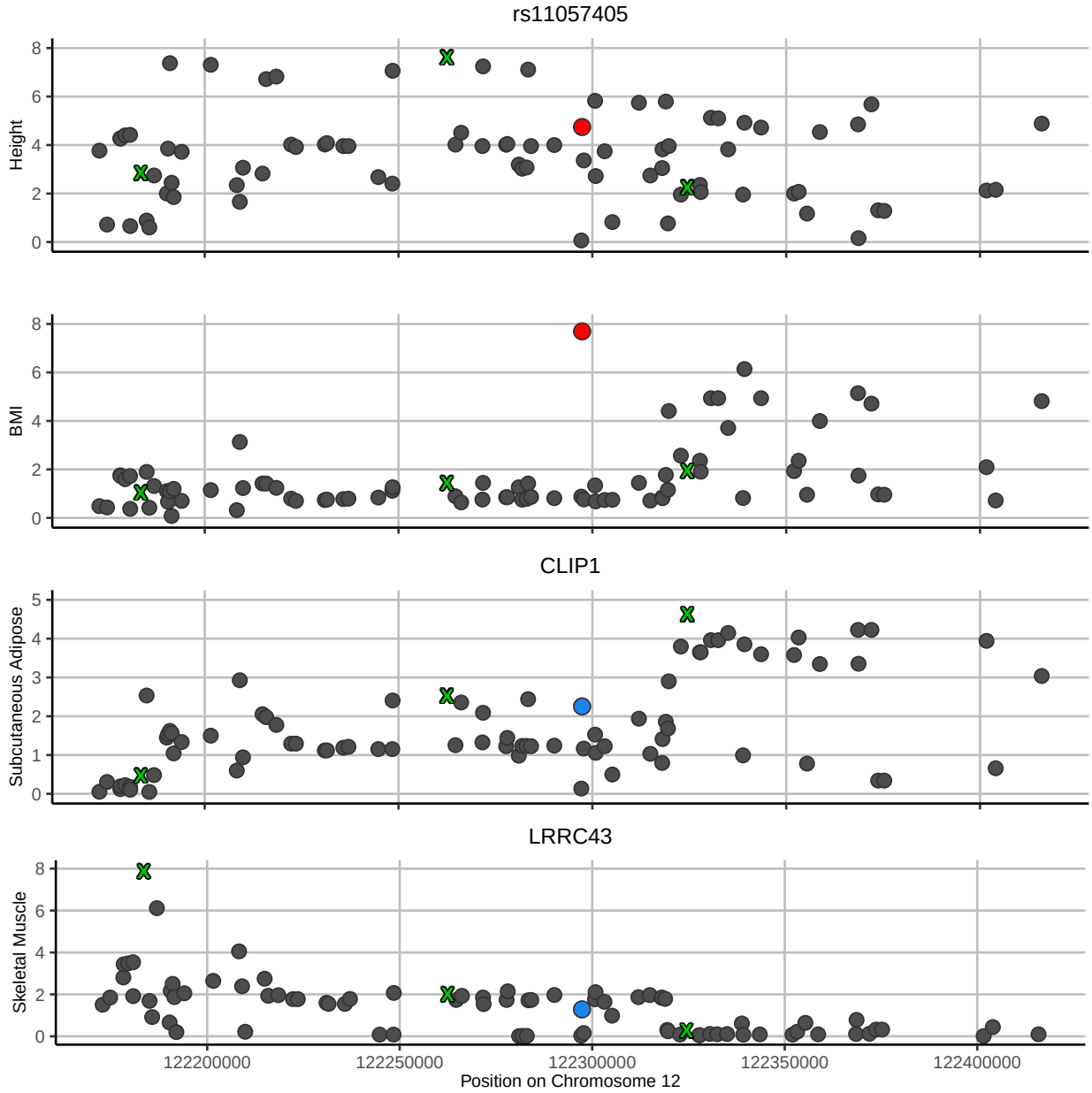

rs12186664
Height
9
6
3
0
BMI
6
4
2
0
PCSK1
Subcutaneous Adipose
10.0
7.5
5.0
2.5
0.0
PCSK1
Skeletal Muscle
6
4
2
0
96200000
96300000
96400000
96500000
96600000
Position on Chromosome 5

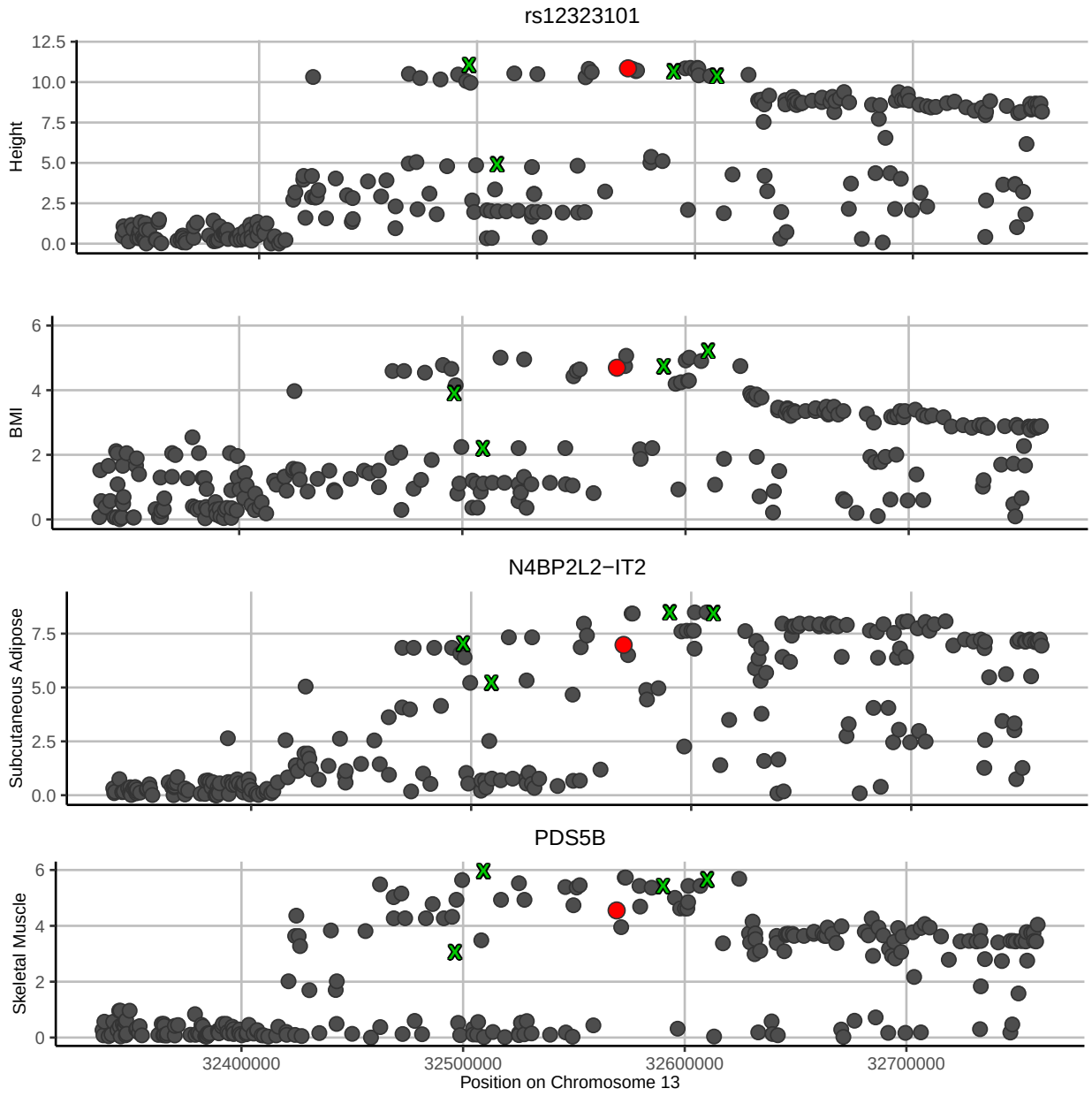

rs205262
Height
BMI
UHRF1BP1
Subcutaneous Adipose
UHRF1BP1
Skeletal Muscle
34600000
34700000
34800000
34900000
35000000
Position on Chromosome 6

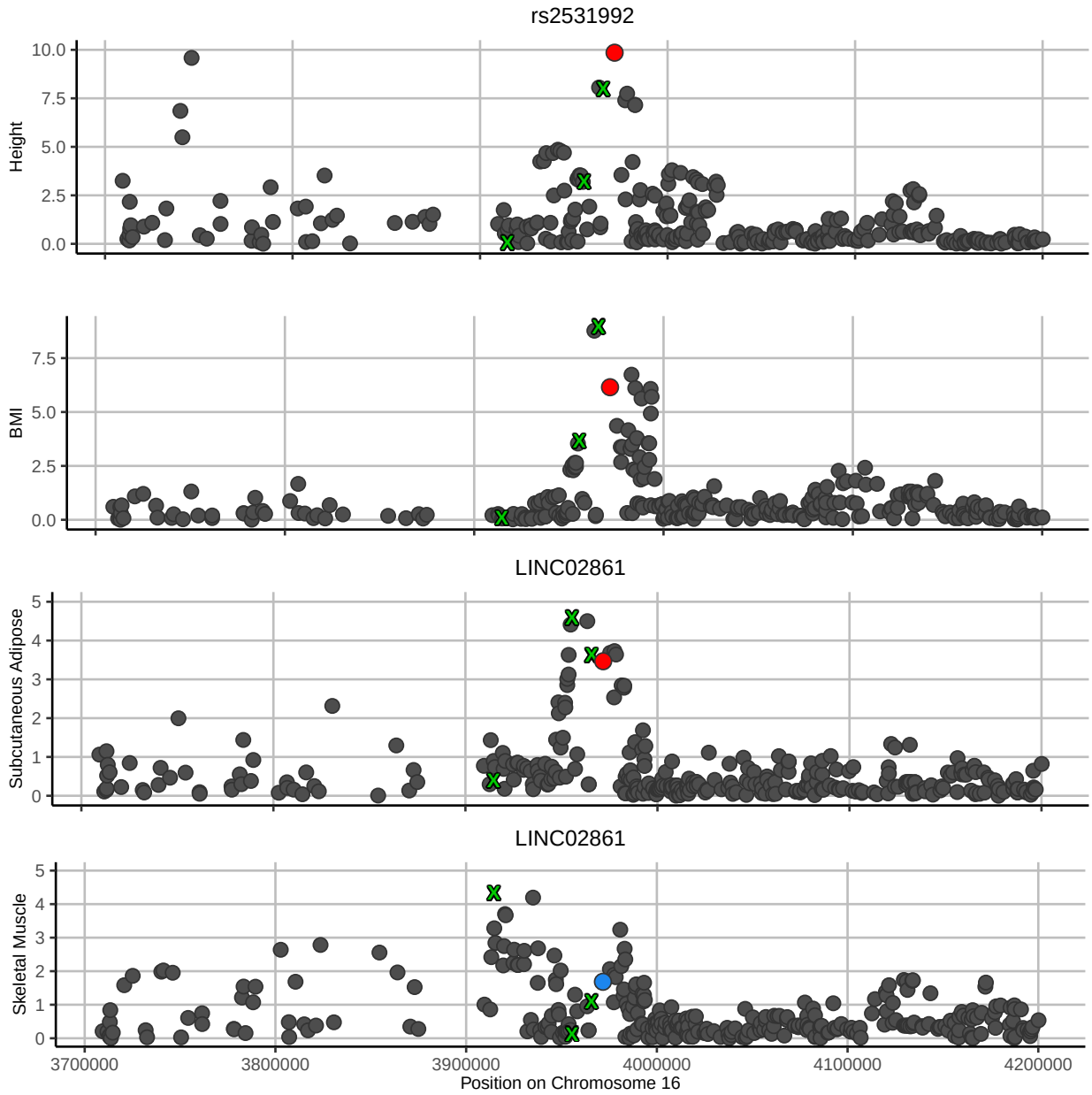

rs9400239
Height
12.5
10.0
7.5
5.0
2.5
0.0
BMI
8
6
4
2
0
LINC00222
Subcutaneous Adipose
7.5
5.0
2.5
0.0
ZNF259P1
Skeletal Muscle
8
6
4
2
0
108600000
108700000
108800000
108900000
109000000
Position on Chromosome 6

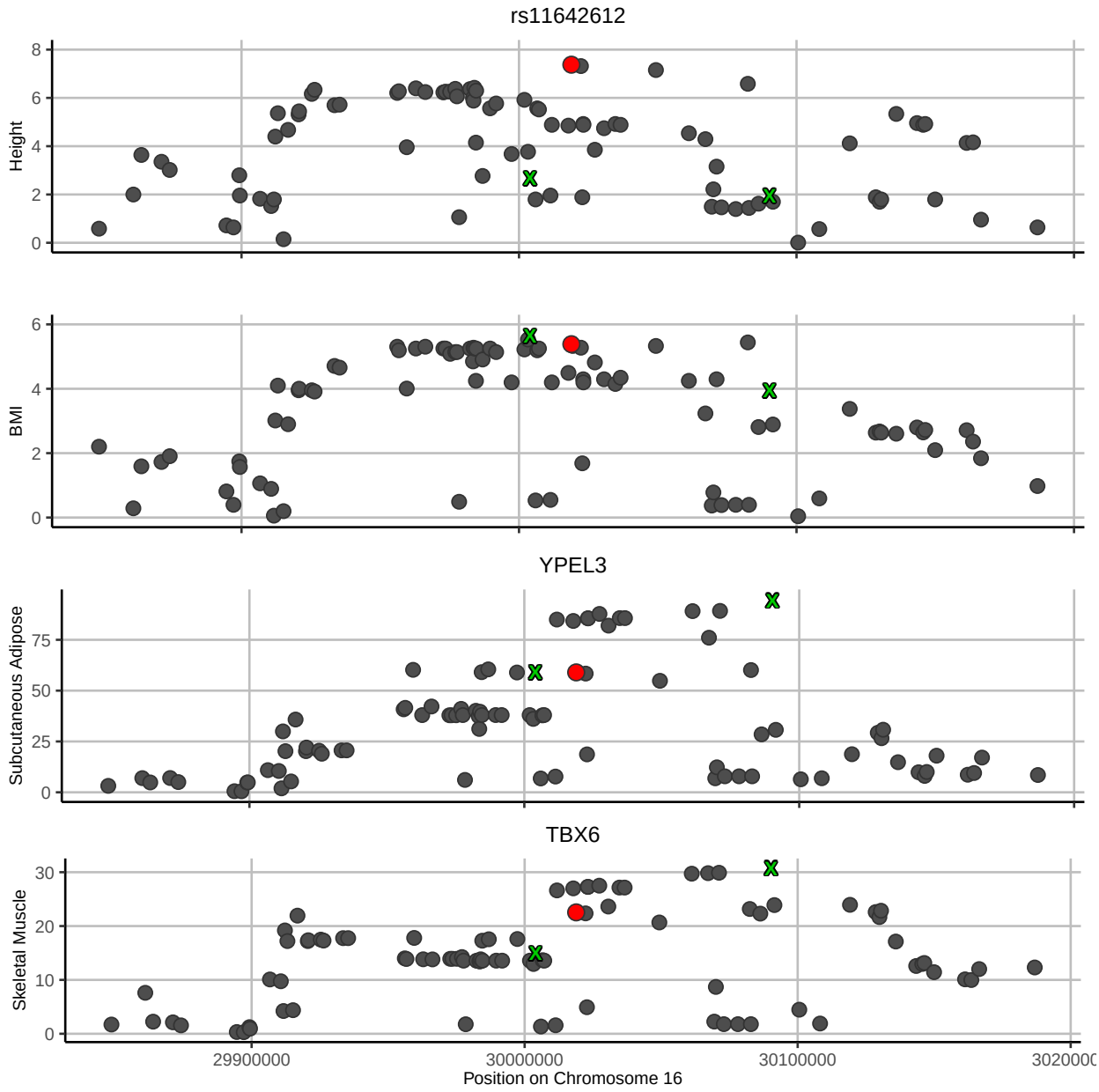

rs757081
Height
BMI
NCR3LG1
Subcutaneous Adipose
NCR3LG1
Skeletal Muscle
17100000
17200000
17300000
17400000
17500000
17600000
Position on Chromosome 11

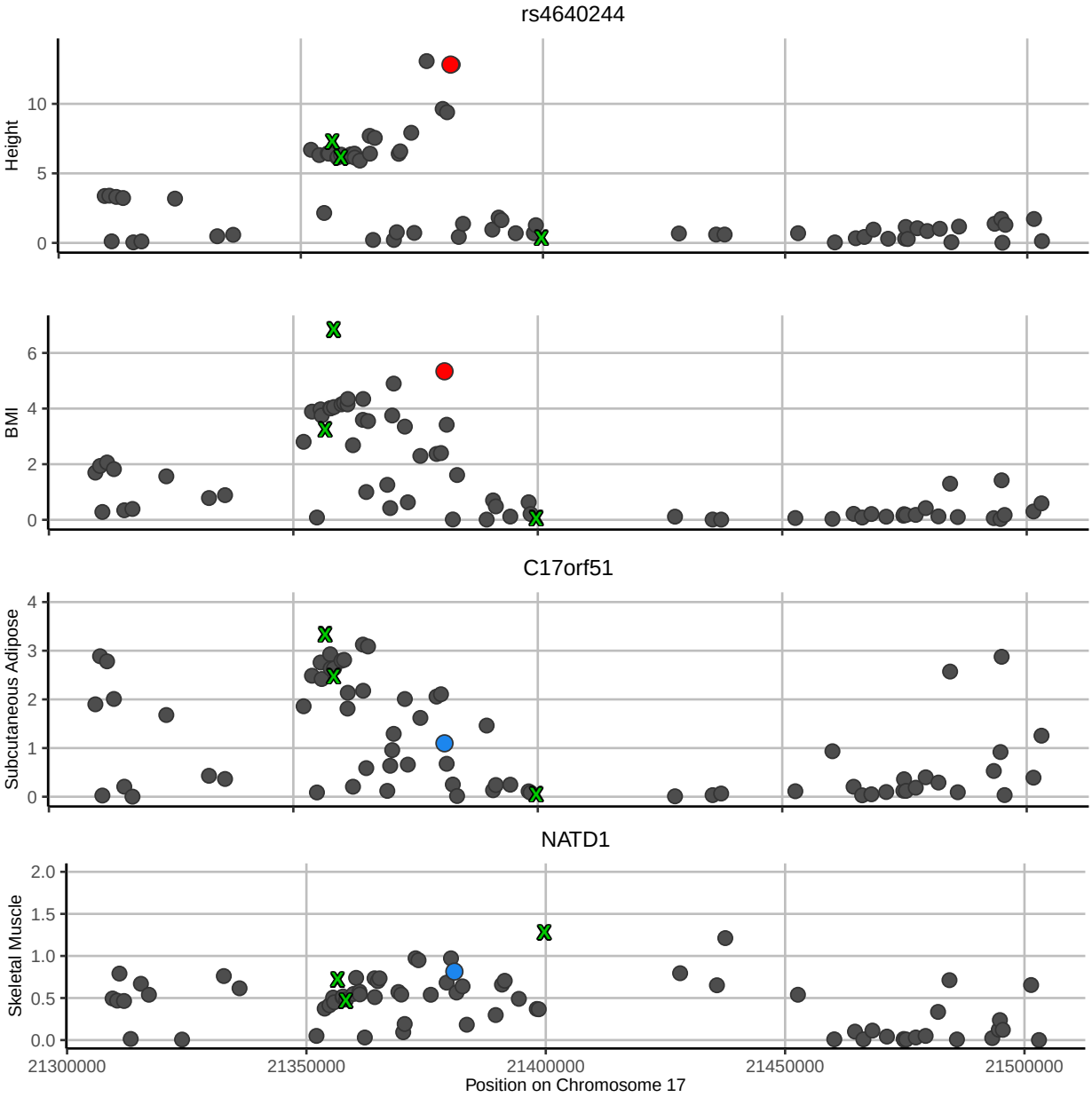

rs3817334
Height
BMI
C1QTNF4
Subcutaneous Adipose
C1QTNF4
Skeletal Muscle
47400000
47500000
47600000
47700000
47800000
Position on Chromosome 11

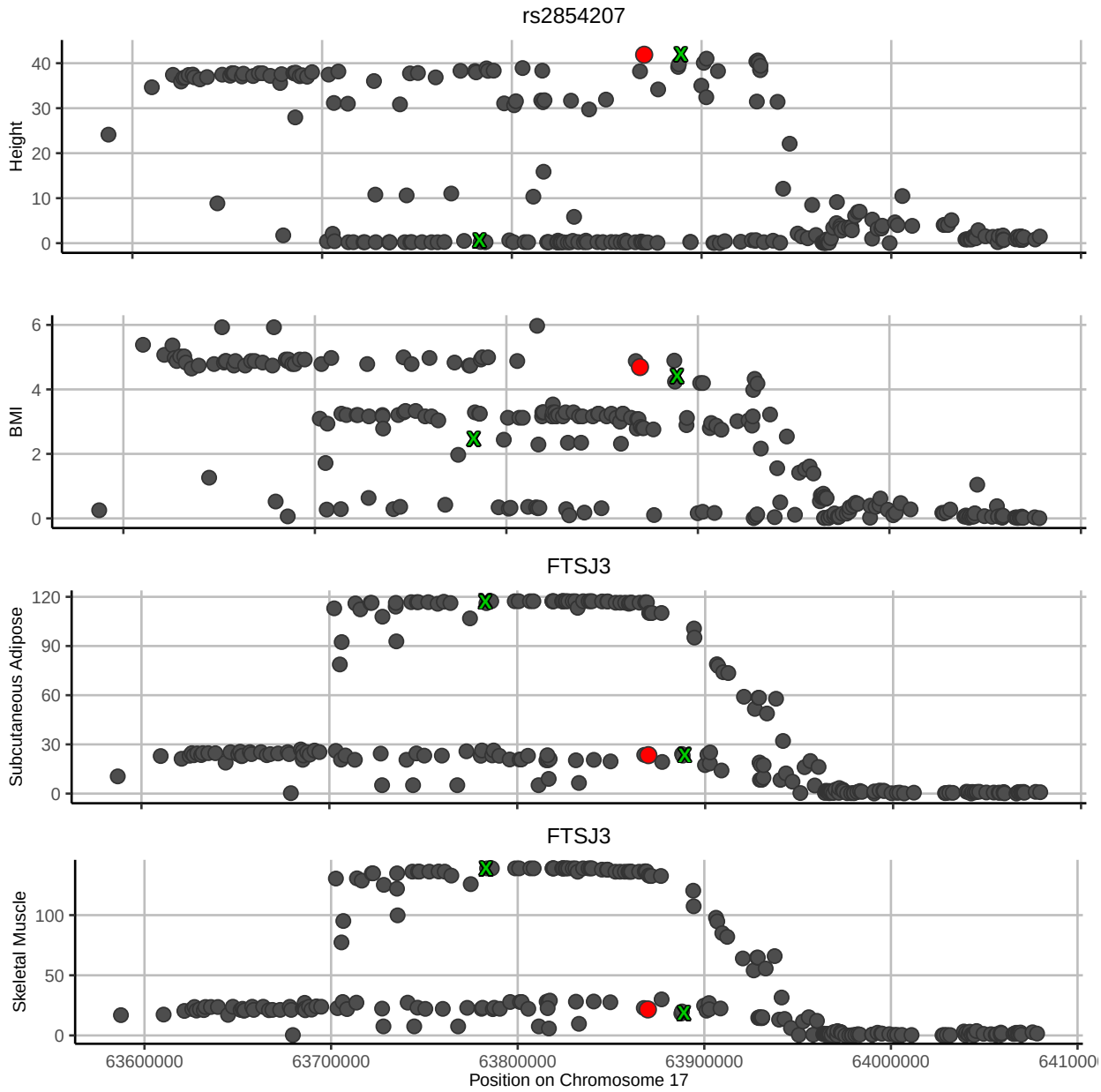

rs11057405
Height
BMI
CLIP1
Subcutaneous Adipose
LRRC43
Skeletal Muscle
122200000
122250000
122300000
122350000
122400000
Position on Chromosome 12

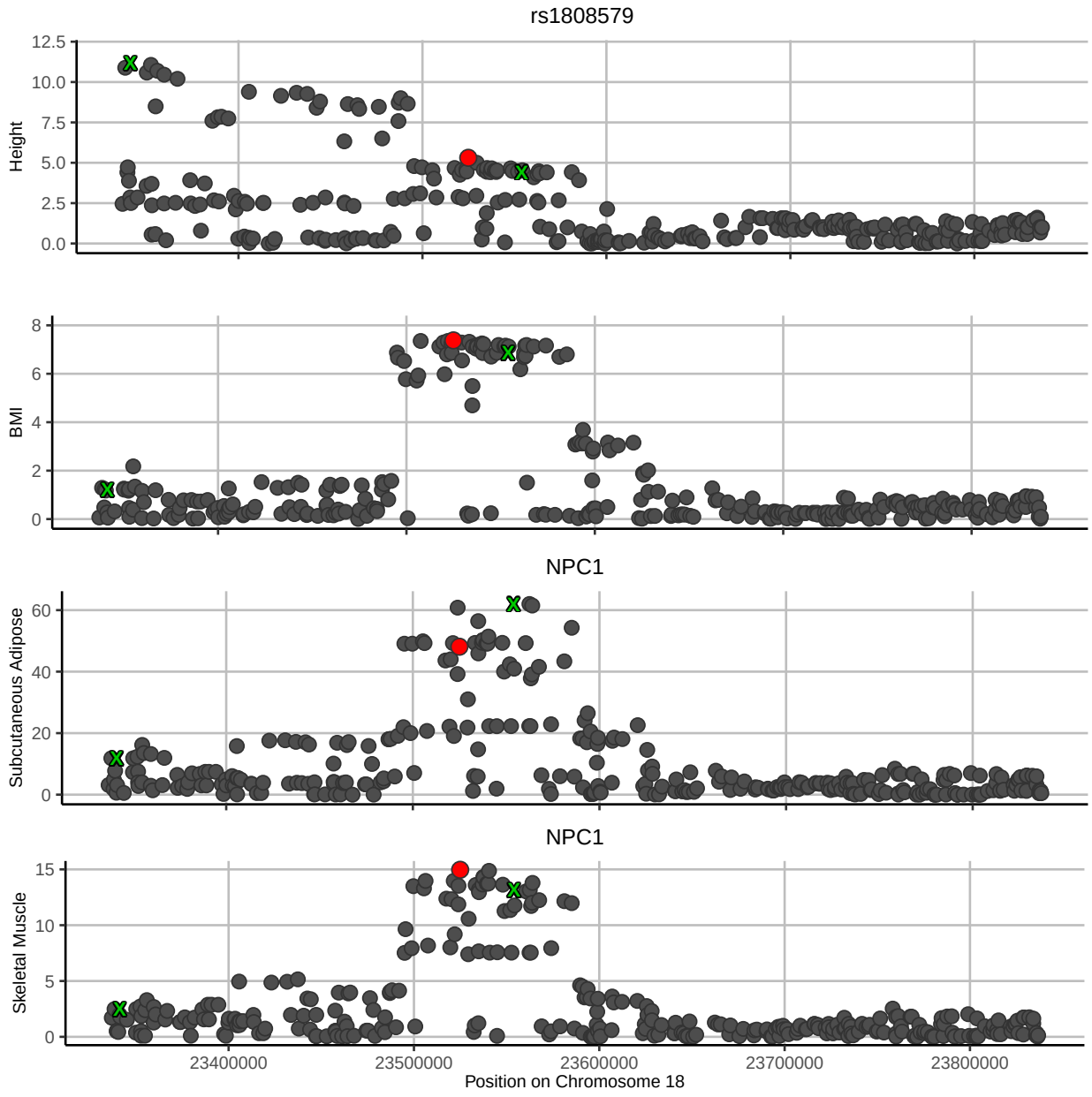

rs12323101
Height
12.5
10.0
7.5
5.0
2.5
0.0
BMI
6
4
2
0
N4BP2L2−IT2
Subcutaneous Adipose
7.5
5.0
2.5
0.0
PDS5B
Skeletal Muscle
6
4
2
0
32400000
32500000
32600000
32700000
Position on Chromosome 13

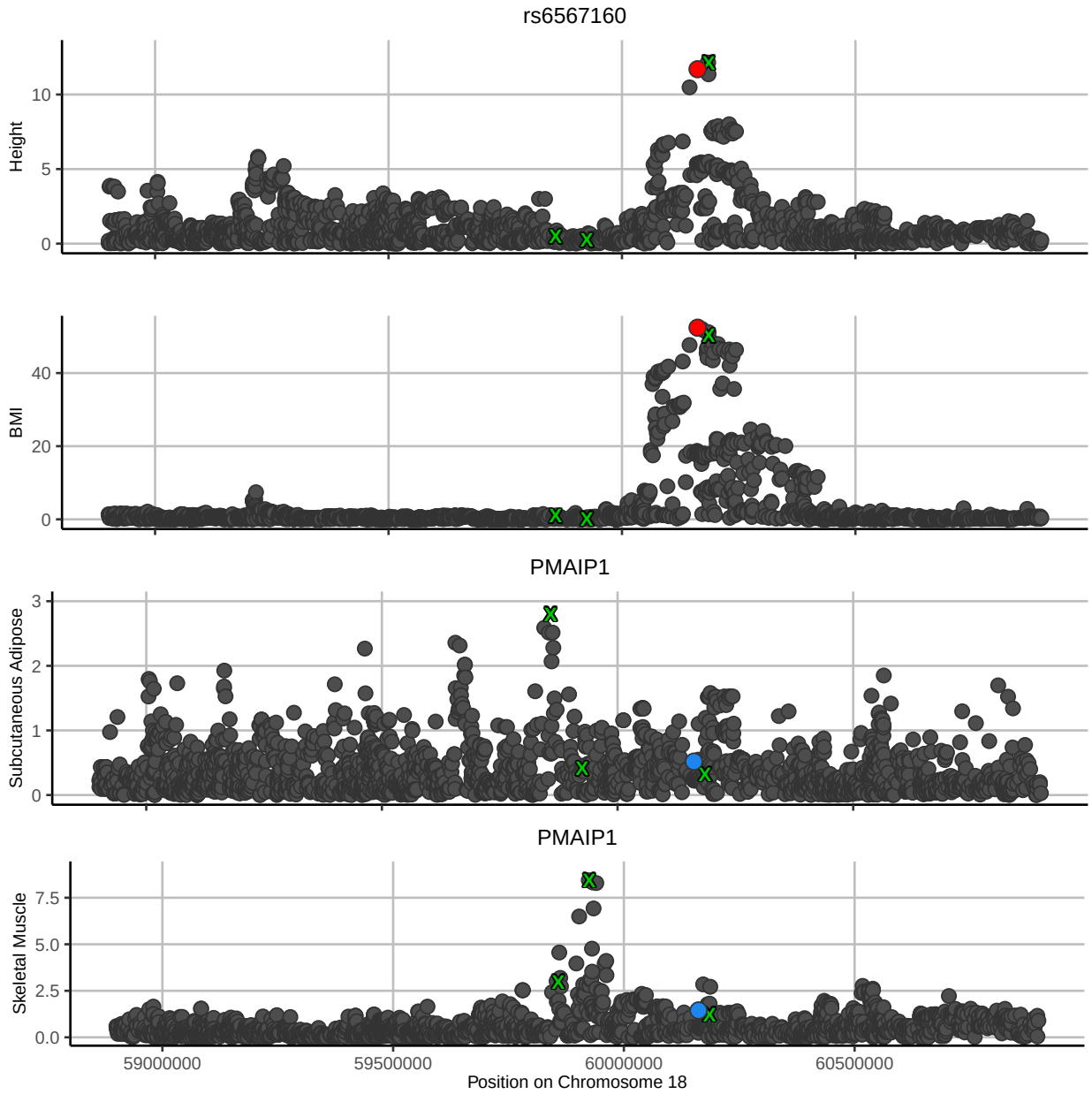

rs2531992
Height
10.0
7.5
5.0
2.5
0.0
BMI
7.5
5.0
2.5
0.0
LINC02861
Subcutaneous Adipose
5
4
3
2
1
0
LINC02861
Skeletal Muscle
5
4
3
2
1
0
3700000
3800000
3900000
4000000
4100000
4200000
Position on Chromosome 16

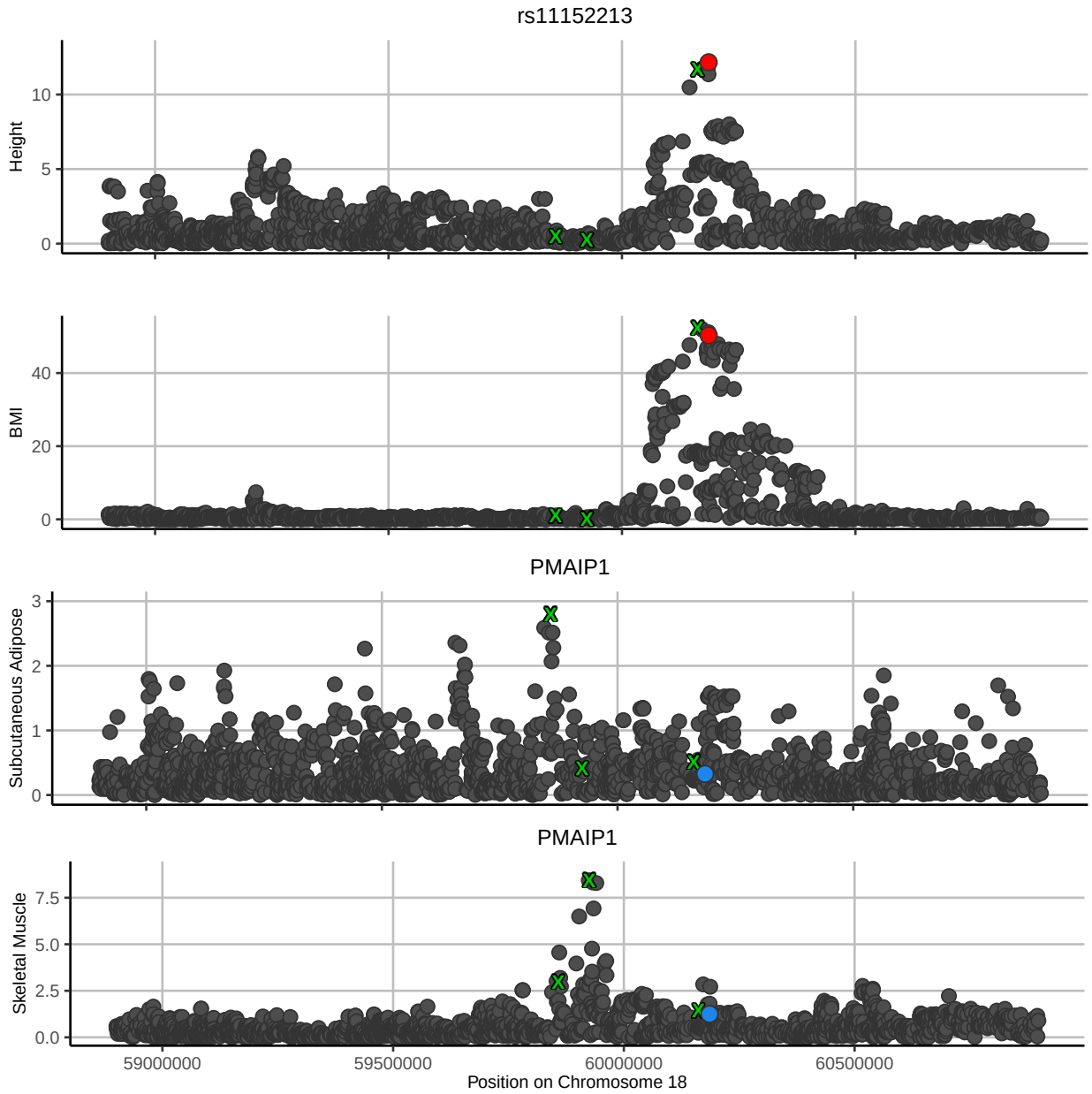

rs11642612
Height
BMI
YPEL3
Subcutaneous Adipose
TBX6
Skeletal Muscle
29900000
30000000
30100000
Position on Chromosome 16

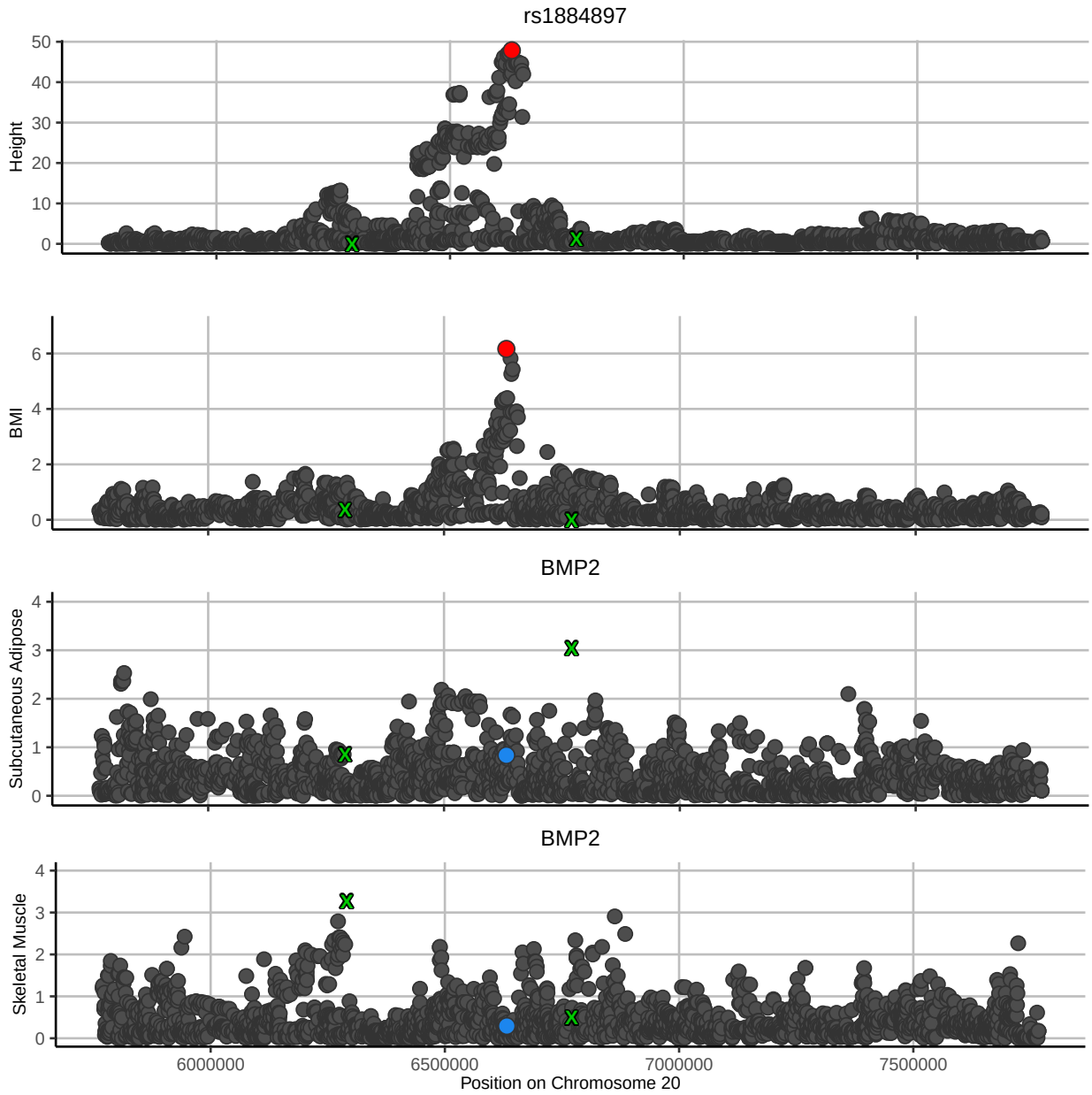

rs4640244
Height
10
5
0
BMI
6
4
2
0
C17orf51
Subcutaneous Adipose
4
3
2
1
0
NATD1
Skeletal Muscle
2.0
1.5
1.0
0.5
0.0
21300000
21350000
21400000
21450000
21500000
Position on Chromosome 17


rs2854207
Height
BMI
FTSJ3
Subcutaneous Adipose
FTSJ3
Skeletal Muscle
63600000
63700000
63800000
63900000
64000000
Position on Chromosome 17


rs1808579
Height
BMI
NPC1
Subcutaneous Adipose
NPC1
Skeletal Muscle
23400000
23500000
23600000
23700000
23800000
Position on Chromosome 18


rs6567160
Height
10
5
0
BMI
40
20
0
PMAIP1
Subcutaneous Adipose
3
2
1
0
PMAIP1
Skeletal Muscle
7.5
5.0
2.5
0.0
59000000
59500000
60000000
60500000
Position on Chromosome 18


rs11152213
Height
10
5
0
BMI
40
20
0
PMAIP1
Subcutaneous Adipose
3
2
1
0
PMAIP1
Skeletal Muscle
7.5
5.0
2.5
0.0
59000000
59500000
60000000
60500000
Position on Chromosome 18


rs1884897
Height
50
40
30
20
10
0
BMI
6
4
2
0
BMP2
Subcutaneous Adipose
4
3
2
1
0
BMP2
Skeletal Muscle
4
3
2
1
0
6000000
6500000
7000000
7500000
Position on Chromosome 20

APPENDIX D

ADDITIONAL SIMULATIONS EVALUATING THE PERFORMANCE OF MR-ROBIN COMPARED TO EXISTING METHODS

In Section 5.3.1, we have presented the performance of MR-Robin when the selected IVs were in moderate LD ($r^2 < 0.5$ and $r^2 < 0.3$). In this section, using simulation studies we evaluated the performance of MR-Robin when the selected IVs are in weak LD or are nearly independent, and compared to competing methods. In each simulation scenario, we simulated data for a total of $N = N_g + N_R = 10,300$ independent subjects: $N_g = 10,000$ subjects in a GWAS study, and $N_R = 300$ subjects in a reference multi-tissue eQTL study of $K = 10$ tissues. Details of the simulations are described in Section 5.3.1.

D.1 MR-Robin controls type I error rate with moderate proportion of invalid IVs

We evaluated the robustness of MR-Robin to the proportion of invalid IVs. We simulated the data using $Q = 10$ LD blocks with 20 SNPs in each, varying the proportion of invalid IVs across settings. That is, we varied the proportion of eSNPs having direct effects on the complex trait Y (i.e. effects not mediated through gene expression X). Over 10,000 simulations, we compare the type I error rate and power of MR-Robin to existing two-sample MR methods. $P < 0.05$ was used as the significance criterion for each method. Tables D.1A and D.1B compare the methods when the selection LD r^2 threshold is set to 0.1 and 0.01, respectively (results using selection LD r^2 thresholds of 0.5 and 0.3 are reported in Table 5.2 in Chapter 5).

Based on the results, we observe that competing methods are generally unable to control the type I error rate when there are a moderate proportion of invalid IVs (e.g > 20%)

and IVs are in weak LD. This is mostly because their estimating algorithms may require a relatively large number of IVs to perform well, and in our target setting there is often only a limited number of independent cis-eQTLs in a region. On the other hand, MR-Robin is able to control the type I error rate when a majority of IVs are valid (with up to 50% invalid IVs) if the selected IVs are in weak LD. See Chapter 5, Table 5.1 for simulation results with alternative LD selection criteria (i.e. LD r^2 thresholds of 0.5 and 0.3) and Table 5.2 for simulation results when the number of candidate IVs is very small (i.e. $Q = 3$ LD blocks each with 20 SNPs).

Since our method allows for correlated IVs and it is hard to define invalid versus valid IVs when SNPs are correlated, the proportions of invalid IVs in the tables are the proportion of LD blocks with pleiotropy, and are only approximations of the invalid IVs among all selected ones. In each table, we also presented the average numbers of selected IVs that are from valid versus invalid LD blocks.

Table D.1: **Simulation results evaluating the performance of MR-Robin for IVs that are in weak LD or nearly independent**. Averaged type I error rates and power over 10,000 simulations are shown by percentage of invalid instruments. 10 LD blocks were simulated, with one true eQTL per LD block. Instruments were selected sequentially: the eSNP with the strongest association with gene expression was selected, and the next selected eSNP is the strongest-associated SNP remaining also with LD $r^2 < 0.1$ (A) or $r^2 < 0.01$ (B) with any already-selected eSNPs.

(A) pairwise LD $r^2 < 0.1$

Method	Proportion of invalid IV (%)					
	0	10	20	30	40	50
	Type I error rate					
MR-Robin	0.052	0.049	0.047	0.047	0.054	0.049
MR-Egger	0.023	0.078	0.129	0.162	0.190	0.221
MR-RAPS	0.030	0.051	0.061	0.074	0.090	0.100
MRMix	0.113	0.190	0.240	0.275	0.301	0.313
BWMR	0.037	0.059	0.073	0.086	0.101	0.106
	Power					
MR-Robin	0.588	0.553	0.522	0.493	0.484	0.435
MR-Egger	0.809	0.798	0.786	0.778	0.767	0.762
MR-RAPS	0.962	0.950	0.937	0.919	0.903	0.878
MRMix	0.564	0.563	0.550	0.545	0.544	0.542
BWMR	0.975	0.961	0.947	0.931	0.914	0.889
	Avg number of SNPs selected (valid/invalid)					
All Methods	7.8 /0.0	7.1 /0.8	6.3 /1.6	5.5 /2.4	4.7 /3.1	3.9 /3.9

(B) pairwise LD $r^2 < 0.01$

Method	Proportion of invalid IV (%)					
	0	10	20	30	40	50
	Type I error rate					
MR-Robin	0.052	0.049	0.050	0.047	0.048	0.045
MR-Egger	0.020	0.077	0.127	0.170	0.192	0.236
MR-RAPS	0.024	0.045	0.058	0.072	0.086	0.096
MRMix	0.106	0.182	0.230	0.263	0.305	0.318
BWMR	0.031	0.052	0.066	0.078	0.088	0.098
	Power					
MR-Robin	0.650	0.608	0.573	0.541	0.523	0.484
MR-Egger	0.770	0.754	0.740	0.733	0.725	0.717
MR-RAPS	0.957	0.947	0.931	0.915	0.898	0.876
MRMix	0.590	0.586	0.580	0.572	0.578	0.566
BWMR	0.969	0.954	0.936	0.925	0.902	0.880
	Avg number of SNPs selected (valid/invalid)					
All Methods	6.3 /0.0	5.7 /0.6	5.0 /1.3	4.4 /1.9	3.8 /2.5	3.1 /3.1

APPENDIX E

SCHIZOPHRENIA RISK ASSOCIATED GENES FROM MR-ROBIN ANALYSIS

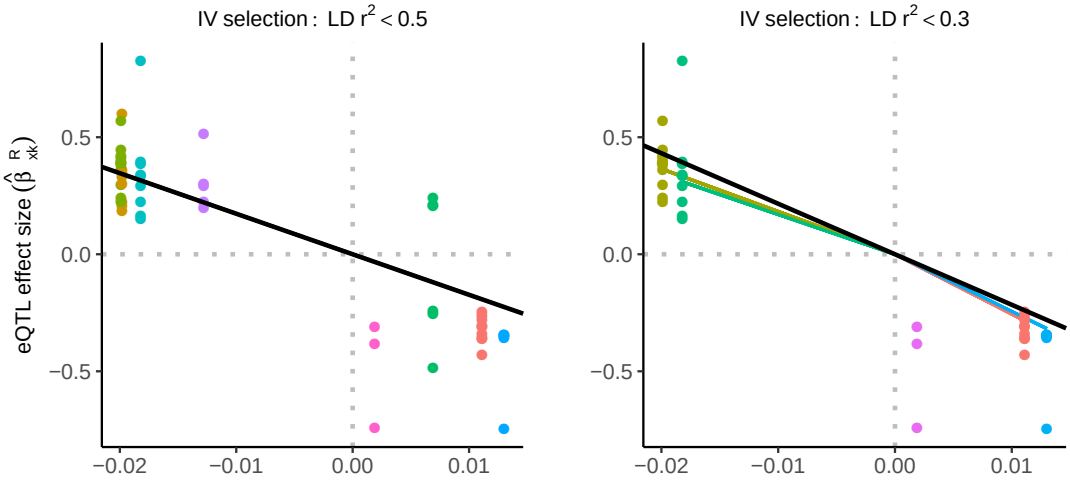
Detailed information about the 42 schizophrenia risk-associated genes identified by MR-Robin is presented in Table E.1. Scatterplots plotting the multi-tissue eQTL effect size estimates against the schizophrenia risk GWAS effect size estimates for the 42 genes are presented in Figure E.1.

Table E.1: **Detailed information on the 42 schizophrenia risk-associated genes identified by MR-Robin.** Presented genes had FDR $< 5\%$ in the primary analysis and MR-Robin $P < 0.1$ in the sensitivity analysis. IVs were selected from cross-tissue eSNPs (median eQTL $P < 0.05$) that were strong IVs ($P < 0.001$) in at least 3 tissues. In the primary analysis, the strongest eSNP having pairwise LD $r^2 < 0.5$ was iteratively selected. In the sensitivity analysis, the eSNP with the highest pairwise correlation with other SNPs was iteratively removed until pairwise LD $r^2 < 0.3$ among all eSNPs or only 5 eSNPs remained. For a given IV, only summary statistics from tissues with eQTL $P < 0.001$ were used in the MR-Robin analysis.

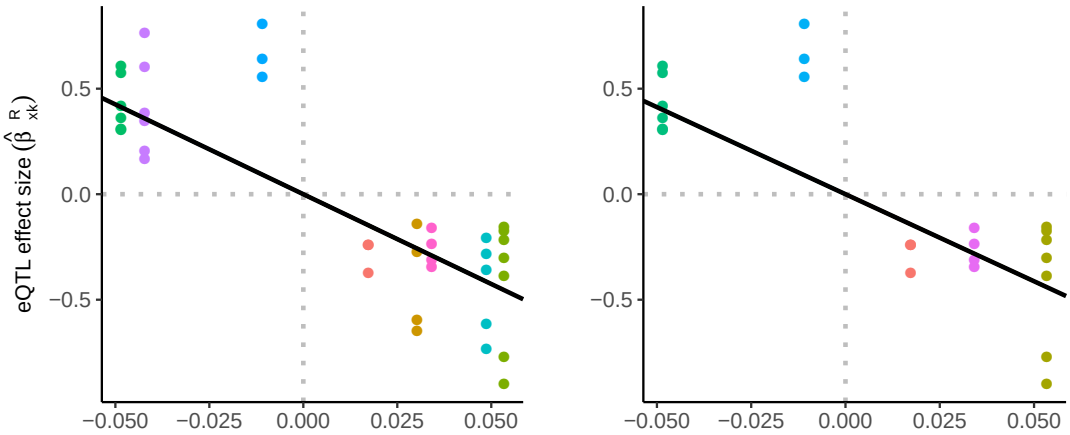
Gene info			MR-Robin P-values	
Ensembl ID	Gene Symbol	Chromosome	Primary	Sensitivity
ENSG00000048544	*MRPS10*	6	4.2×10^{-4}	3.5×10^{-2}
ENSG00000089486	*CDIP1*	16	4.5×10^{-6}	4.2×10^{-4}
ENSG00000090263	*MRPS33*	7	2.7×10^{-4}	1.6×10^{-3}
ENSG00000100731	*PCNX1*	14	1.6×10^{-5}	1.5×10^{-4}
ENSG00000115649	*CNPPD1*	2	4.6×10^{-5}	4.3×10^{-5}
ENSG00000117601	*SERPINC1*	1	5.7×10^{-4}	1.2×10^{-2}
ENSG00000120451	*SNX19*	11	4.7×10^{-5}	5.0×10^{-4}
ENSG00000123643	*SLC36A1*	5	9.7×10^{-4}	7.7×10^{-3}
ENSG00000125611	*CHCHD5*	2	4.8×10^{-4}	6.0×10^{-2}
ENSG00000126464	*PRR12*	19	2.9×10^{-4}	7.0×10^{-3}
ENSG00000129925	*TMEM8A*	16	$< 1.0 \times 10^{-6}$	$< 1.0 \times 10^{-6}$
ENSG00000130304	*SLC27A1*	19	1.5×10^{-4}	1.4×10^{-3}
ENSG00000141013	*GAS8*	16	4.0×10^{-4}	3.7×10^{-3}
ENSG00000141127	*PRPSAP2*	17	7.0×10^{-4}	6.5×10^{-2}
ENSG00000142233	*NTN5*	19	4.5×10^{-5}	4.3×10^{-5}
ENSG00000142534	*RPS11*	19	3.0×10^{-4}	2.1×10^{-4}
ENSG00000142599	*RERE*	1	2.8×10^{-4}	6.7×10^{-4}
ENSG00000146966	*DENND2A*	7	$< 1.0 \times 10^{-6}$	$< 1.0 \times 10^{-6}$
ENSG00000147403	*RPL10*	X	2.7×10^{-4}	3.3×10^{-3}
ENSG00000159199	*ATP5G1*	17	6.4×10^{-4}	3.5×10^{-3}
ENSG00000162753	*SLC9C2*	1	5.2×10^{-6}	1.9×10^{-3}
ENSG00000163040	*CCDC74A*	2	1.1×10^{-4}	7.1×10^{-3}
ENSG00000163634	*THOC7*	3	4.7×10^{-4}	5.0×10^{-3}
ENSG00000163938	*GNL3*	3	3.1×10^{-6}	7.2×10^{-4}
ENSG00000167468	*GPX4*	19	3.2×10^{-4}	9.6×10^{-2}
ENSG00000167535	*CACNB3*	12	$< 1.0 \times 10^{-6}$	4.1×10^{-5}
ENSG00000169220	*RGS14*	5	6.3×10^{-5}	5.3×10^{-4}
ENSG00000170802	*FOXN2*	2	1.0×10^{-4}	2.9×10^{-4}
ENSG00000171928	*TVP23B*	17	7.2×10^{-4}	1.7×10^{-4}
ENSG00000173273	*TNKS*	8	6.7×10^{-4}	4.2×10^{-3}
ENSG00000177595	*PIDD1*	11	2.1×10^{-4}	1.9×10^{-3}
ENSG00000177707	*NECTIN3*	3	1.6×10^{-4}	7.0×10^{-4}
ENSG00000180921	*FAM83H*	8	4.4×10^{-4}	9.2×10^{-3}
ENSG00000182093	*WRB*	21	$< 1.0 \times 10^{-6}$	$< 1.0 \times 10^{-6}$
ENSG00000196268	*ZNF493*	19	$< 1.0 \times 10^{-6}$	7.1×10^{-6}
ENSG00000203499	*FAM83H-AS1*	8	8.8×10^{-4}	3.6×10^{-3}
ENSG00000204257	*HLA-DMA*	6	$< 1.0 \times 10^{-6}$	3.1×10^{-6}
ENSG00000204962	*PCDHA8*	5	9.9×10^{-4}	8.0×10^{-3}
ENSG00000214013	*GANC*	15	3.3×10^{-5}	4.5×10^{-3}
ENSG00000224389	*C4B*	6	7.8×10^{-4}	2.3×10^{-2}
ENSG00000225190	*PLEKHM1*	17	5.2×10^{-5}	4.0×10^{-5}
ENSG00000244731	*C4A*	6	4.5×10^{-5}	1.7×10^{-4}

Figure E.1: **Scatterplots of schizophrenia risk associated genes identified by MR-Robin.** Multi-tissue eQTL effect size estimates in the GTEx brain tissues (y-axis) are plotted against GWAS effect size estimates in the PGC dataset (x-axis) for SNPs used in the primary analysis (left column) and sensitivity analysis (right column). Points are colored by SNP. Colored line segments represent SNP-specific slope estimates. The slope of the black line is the fixed effect estimate from the MR-Robin reverse regression.

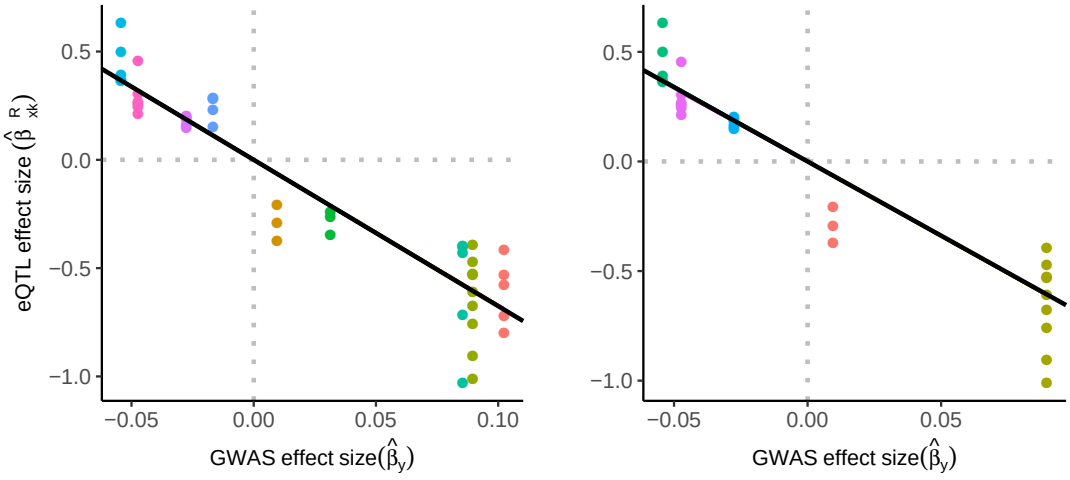
MRPS10



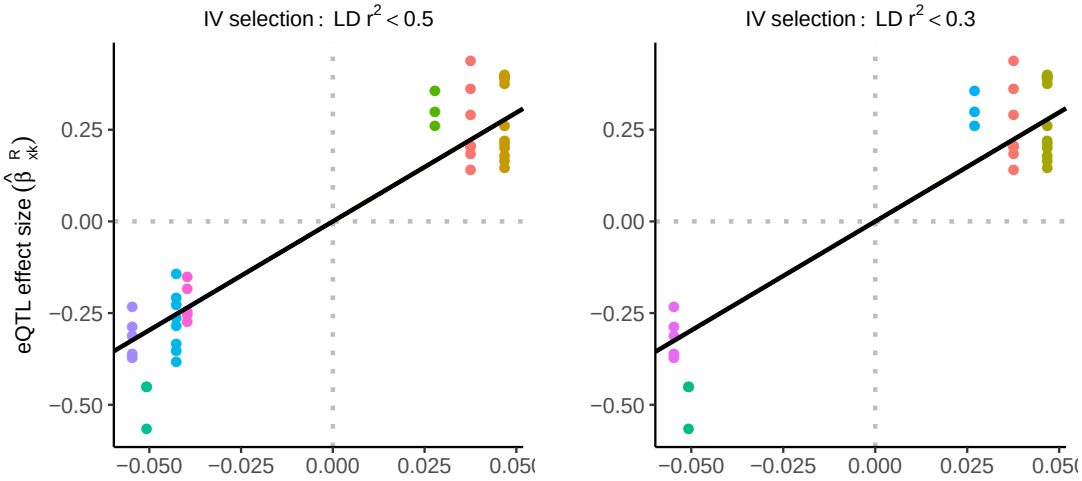
CDIP1



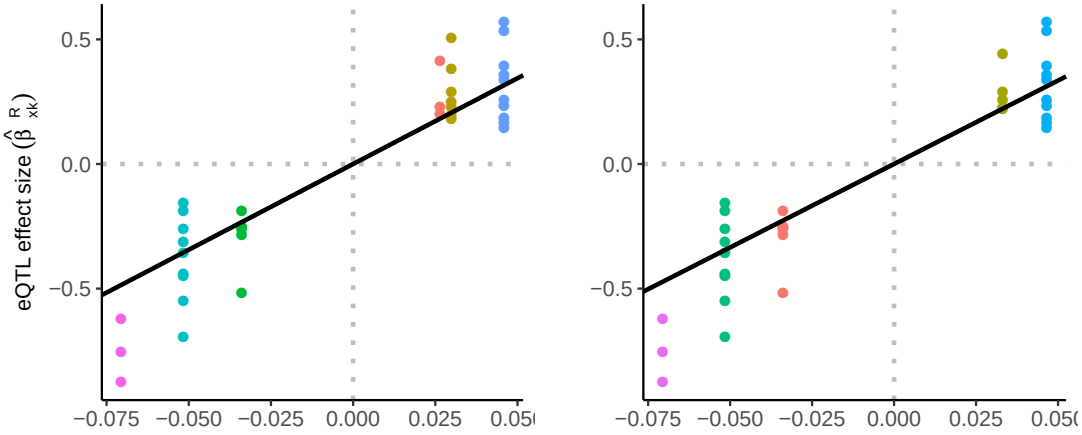
MRPS33



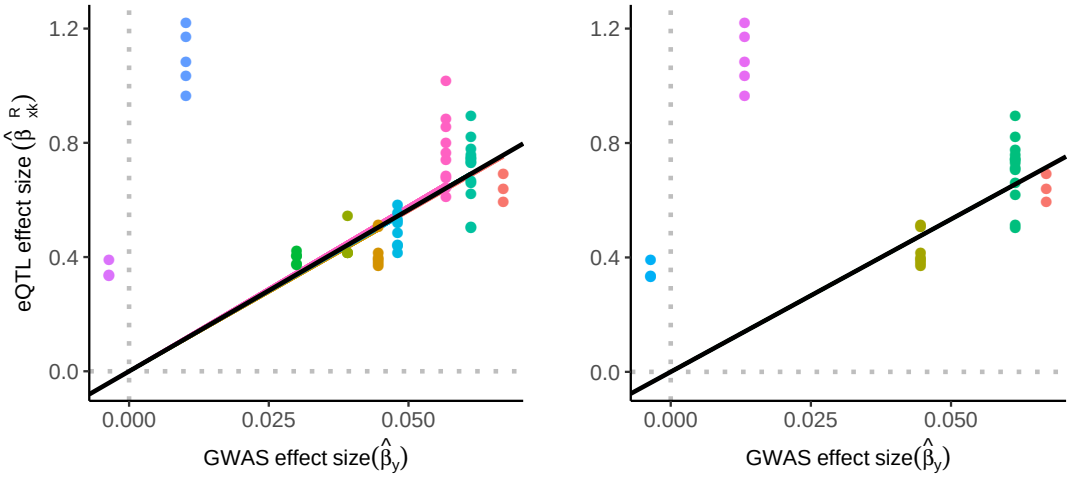
PCNX1

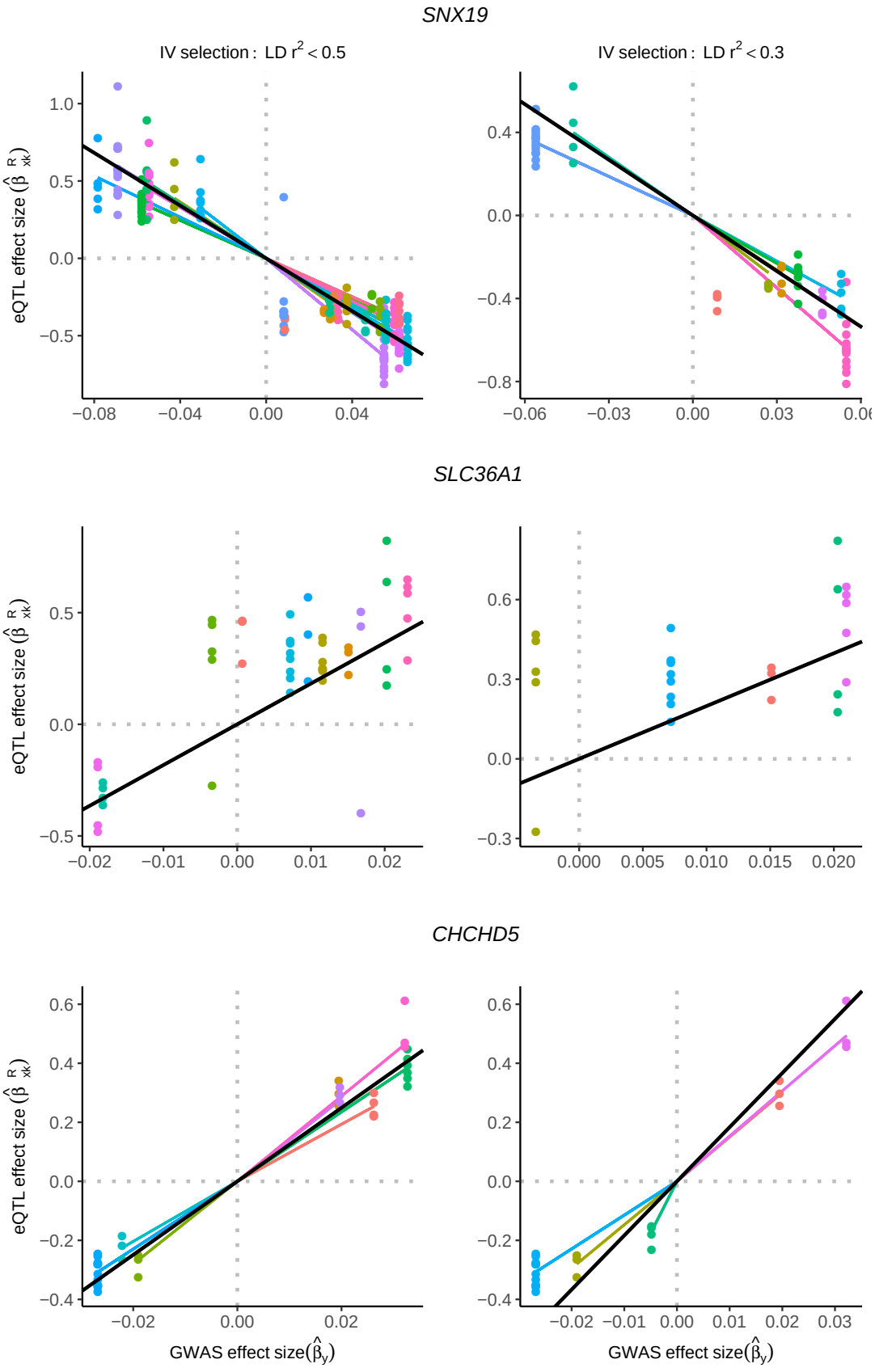


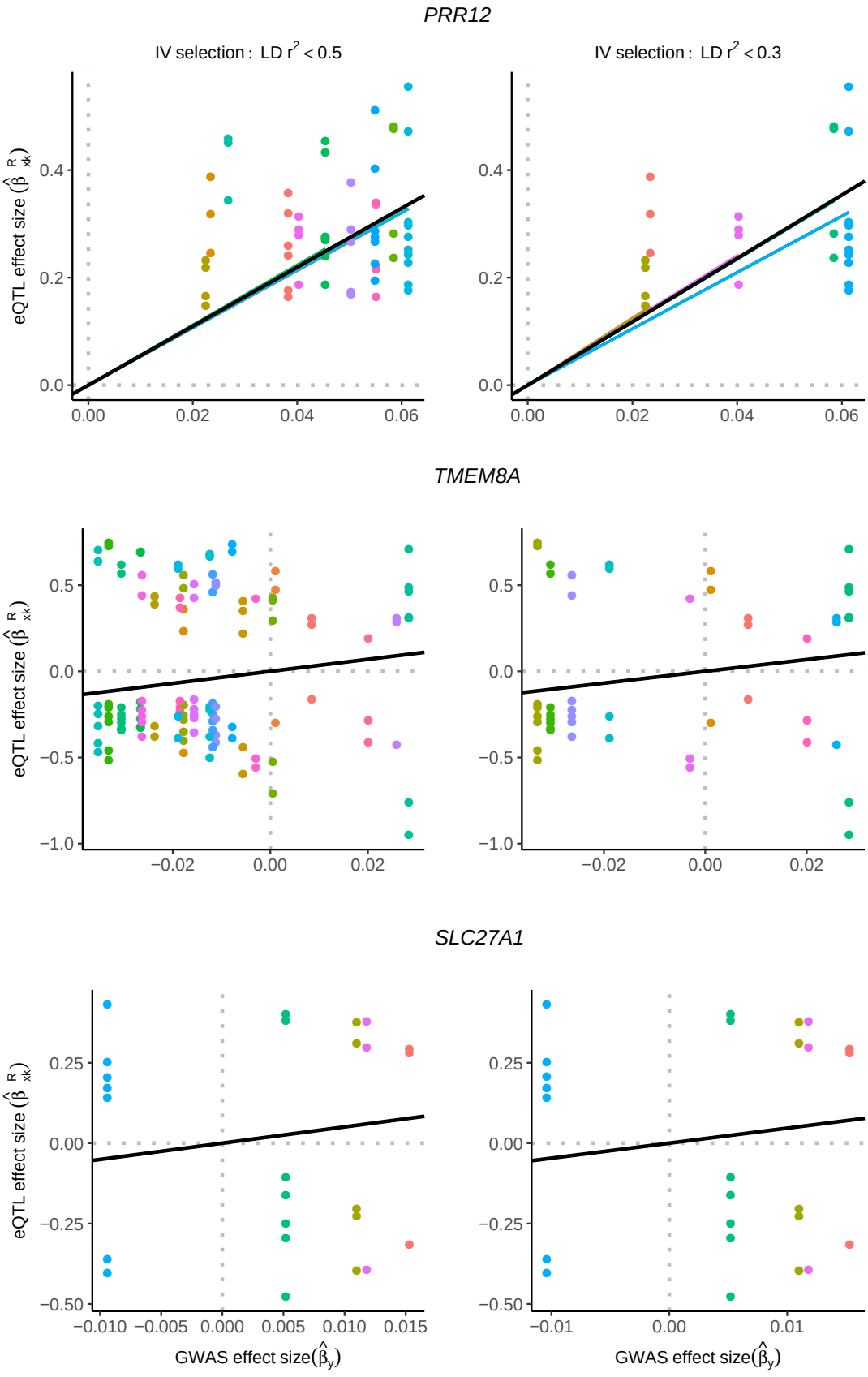
CNPPD1

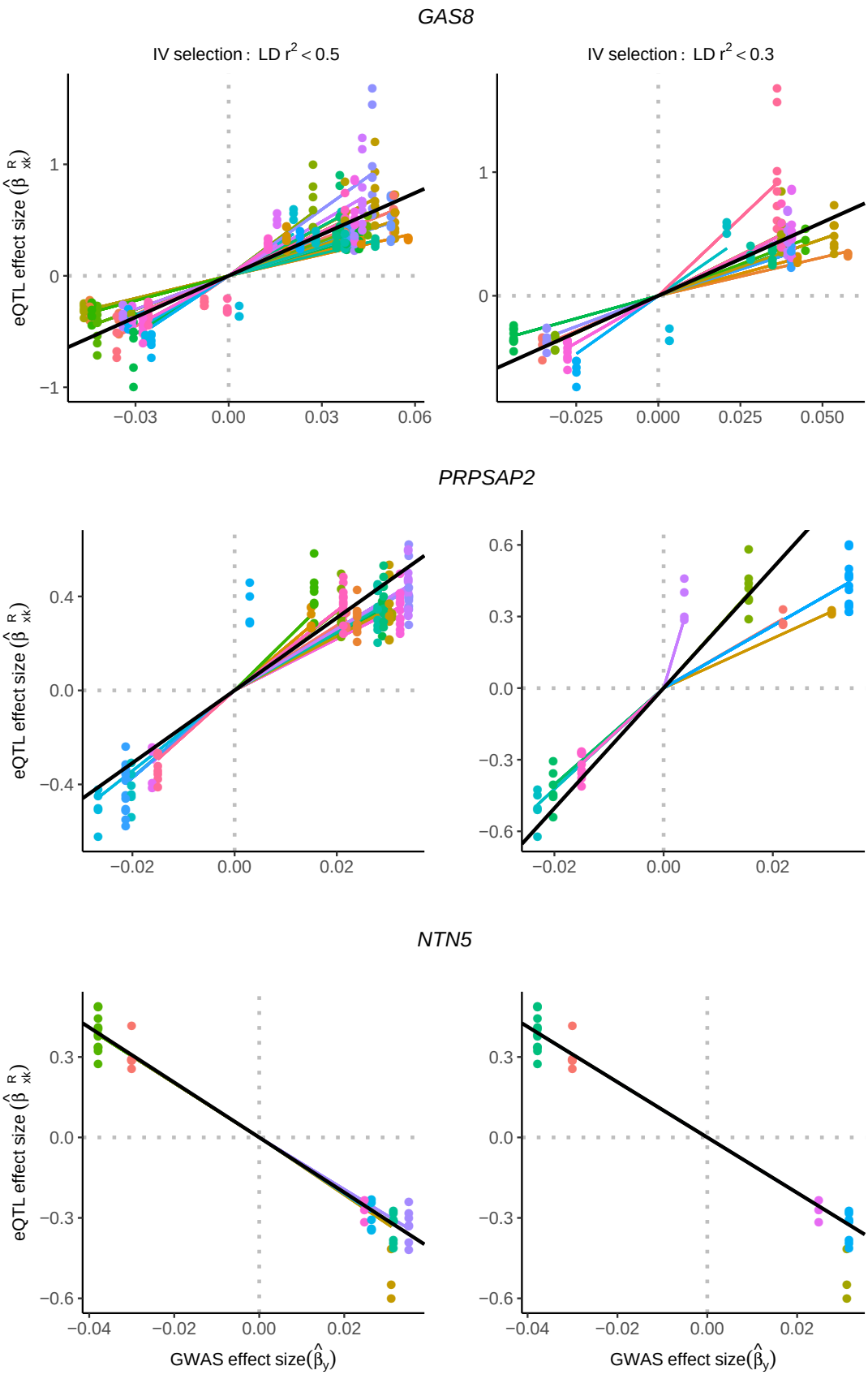


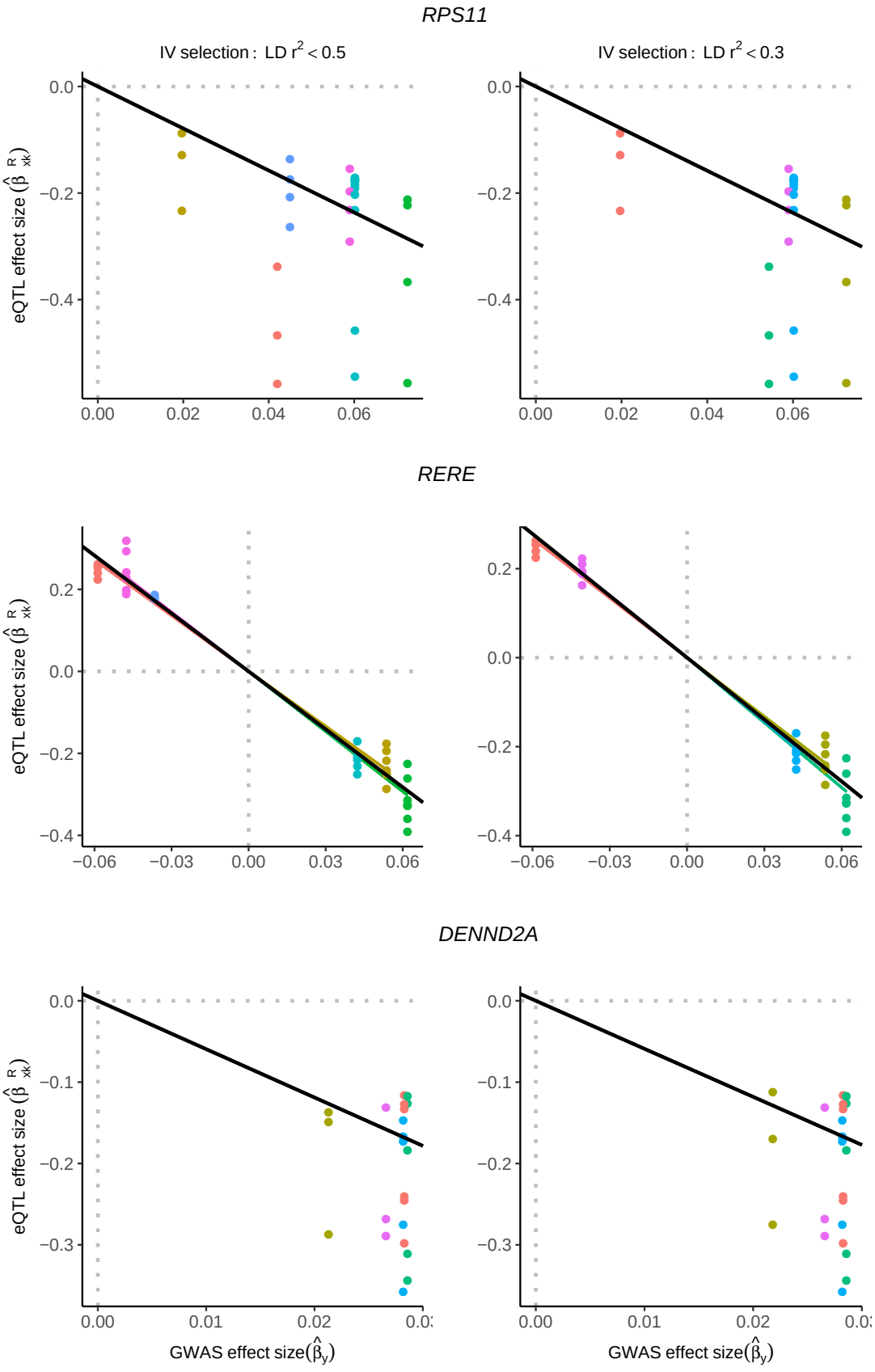
SERPINC1

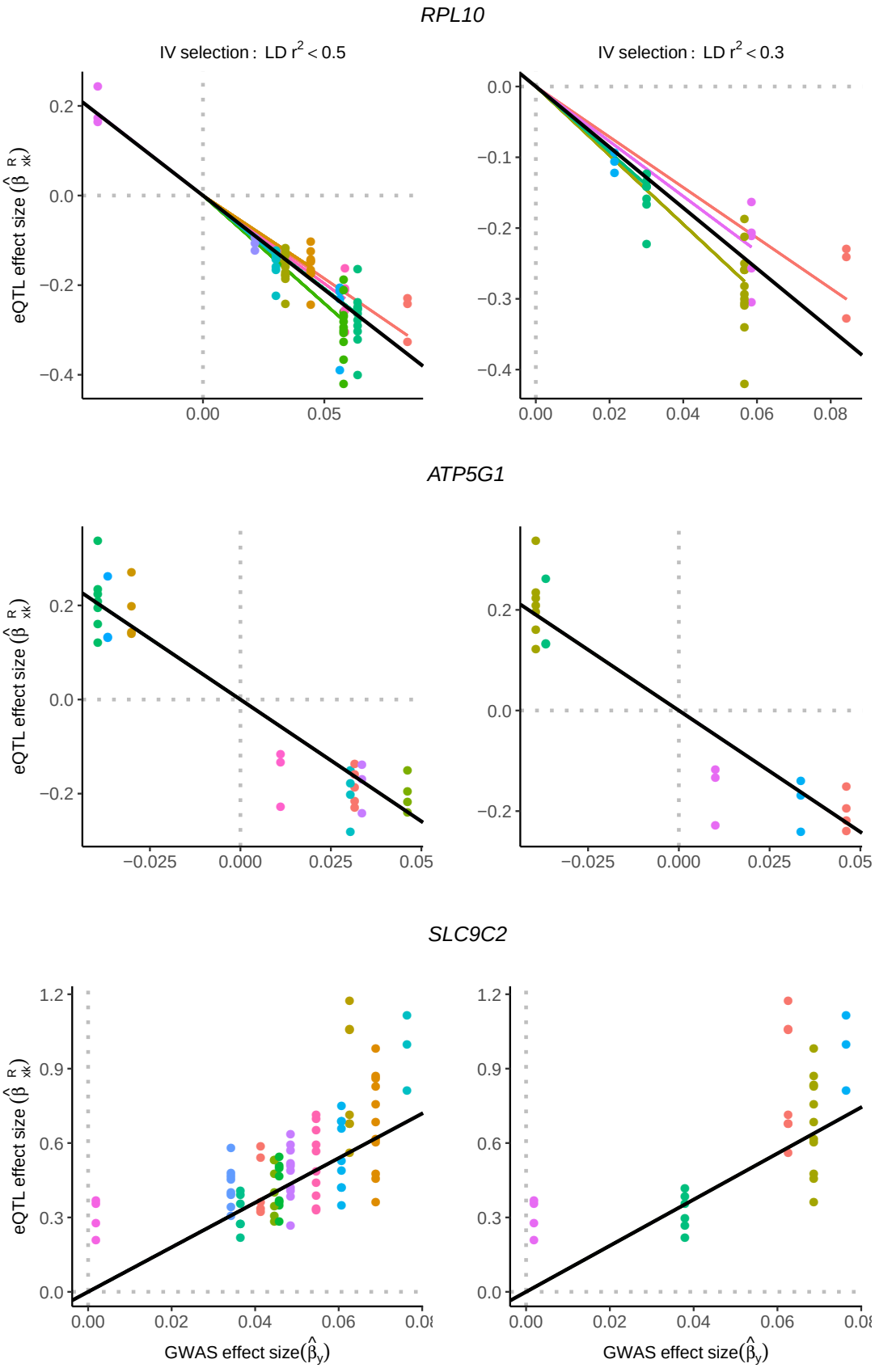


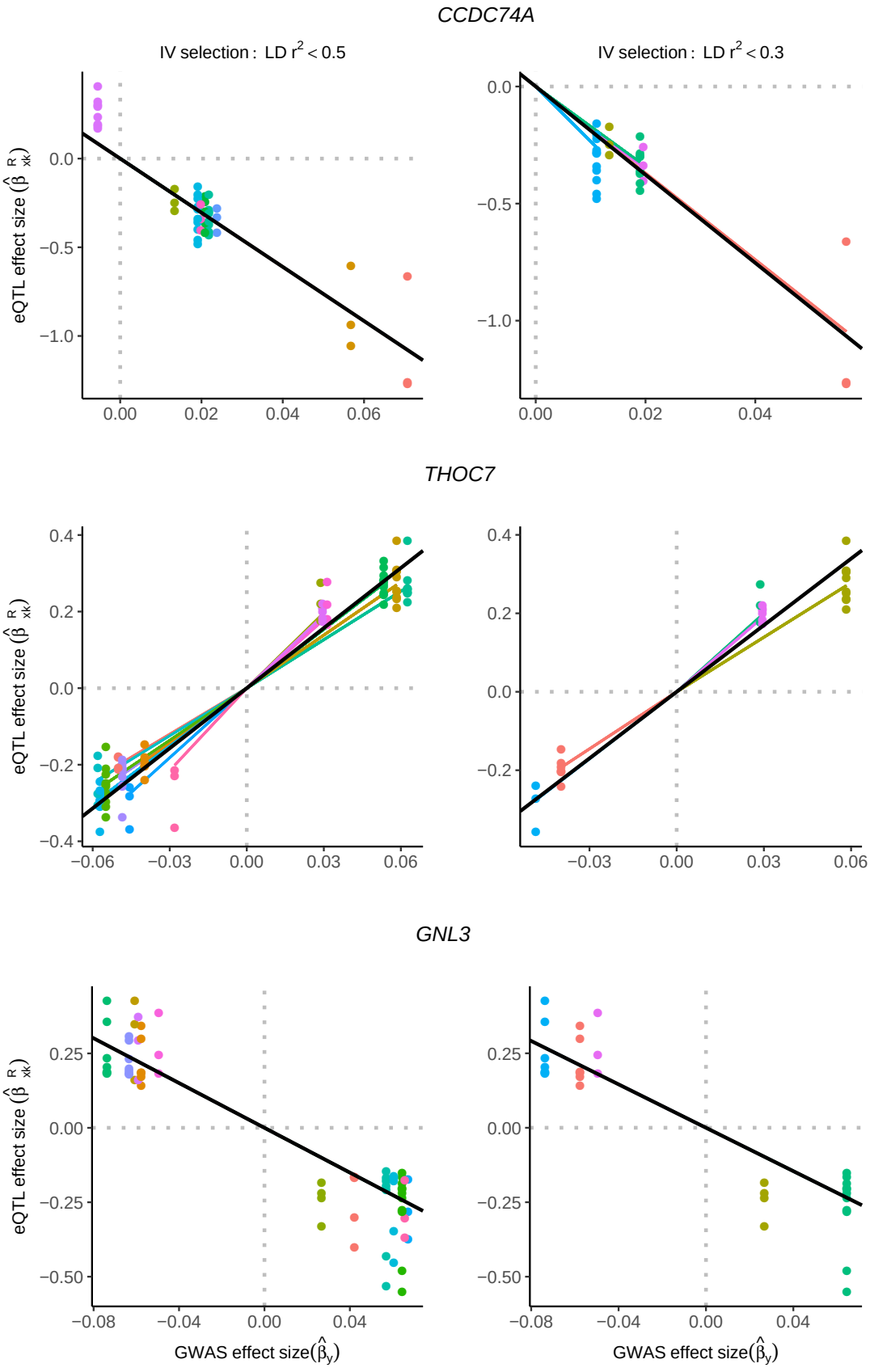

SNX19
IV selection : LD r2 < 0.5
IV selection : LD r2 < 0.3
eQTL effect size (β̂ Rxk)
SLC36A1
CHCHD5
GWAS effect size(β̂y)

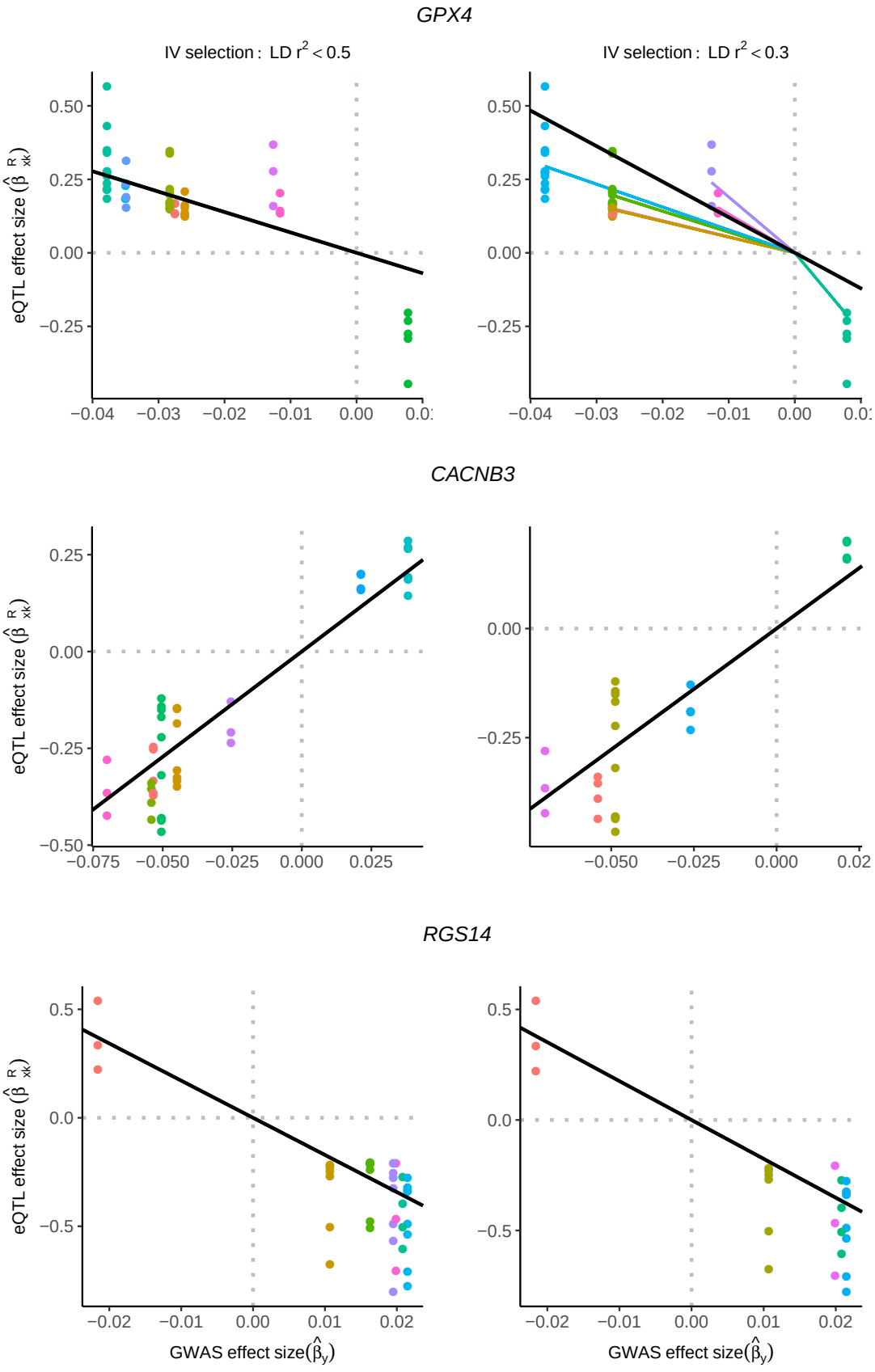

PRR12
IV selection : LD r2 < 0.5
IV selection : LD r2 < 0.3
eQTL effect size (β̂ R xk)
0.4
0.2
0.0
0.00
0.02
0.04
0.06
TMEM8A
0.5
0.0
−0.5
−1.0
−0.02
0.00
0.02
SLC27A1
0.25
0.00
−0.25
−0.50
−0.010 −0.005 0.000 0.005 0.010 0.015
−0.01
0.00
0.01
GWAS effect size(β̂y)

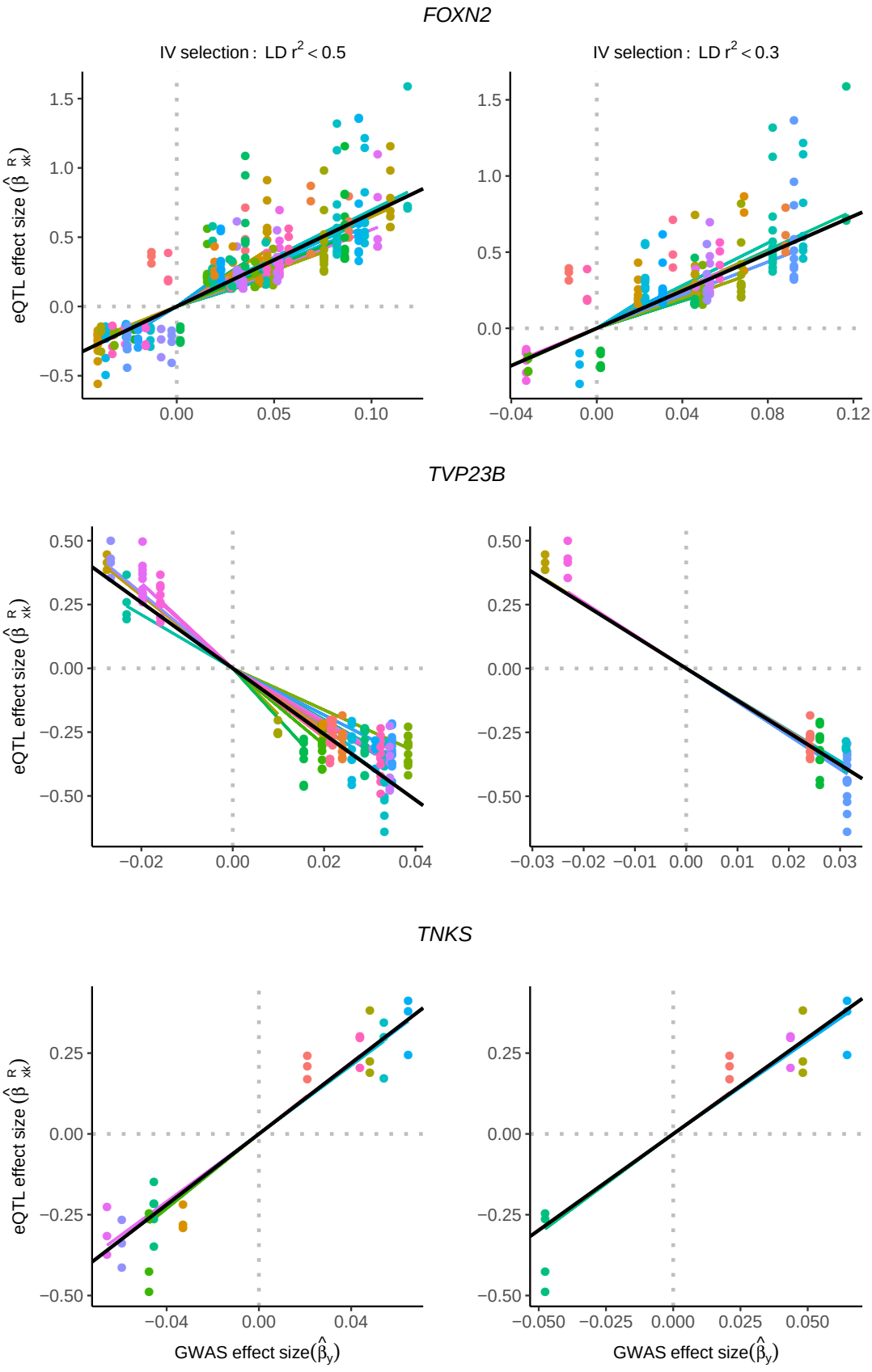

GAS8
IV selection : LD r2 < 0.5
IV selection : LD r2 < 0.3
eQTL effect size (β̂ Rxk)
PRPSAP2
NTN5
GWAS effect size(β̂y)


RPS11
IV selection : LD r2 < 0.5
IV selection : LD r2 < 0.3
eQTL effect size (β̂ R xk)
0.0
−0.2
−0.4
0.00
0.02
0.04
0.06
RERE
0.2
0.0
−0.2
−0.4
−0.06
−0.03
0.00
0.03
0.06
DENND2A
−0.1
−0.3
0.01
GWAS effect size(β̂y)

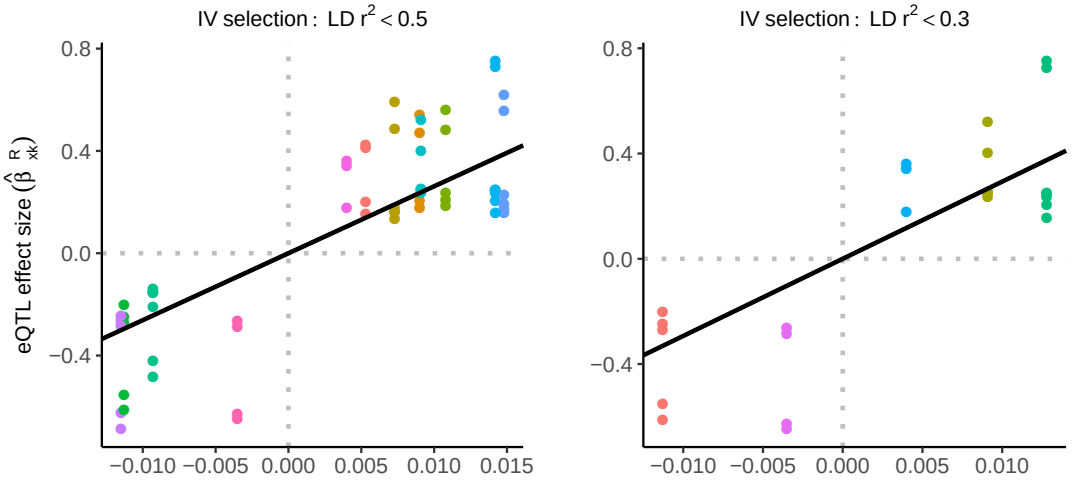

RPL10
IV selection : LD r2 < 0.5
IV selection : LD r2 < 0.3
eQTL effect size (β̂ R xk)
ATP5G1
SLC9C2
GWAS effect size(β̂y)


CCDC74A
IV selection : LD r2 < 0.5
IV selection : LD r2 < 0.3
eQTL effect size (β̂ Rxk)
0.0
−0.5
−1.0
0.00
0.02
0.04
0.06
THOC7
0.4
0.2
−0.2
−0.4
−0.06
−0.03
0.03
GNL3
0.25
0.00
−0.25
−0.50
−0.08
−0.04
GWAS effect size(β̂y)

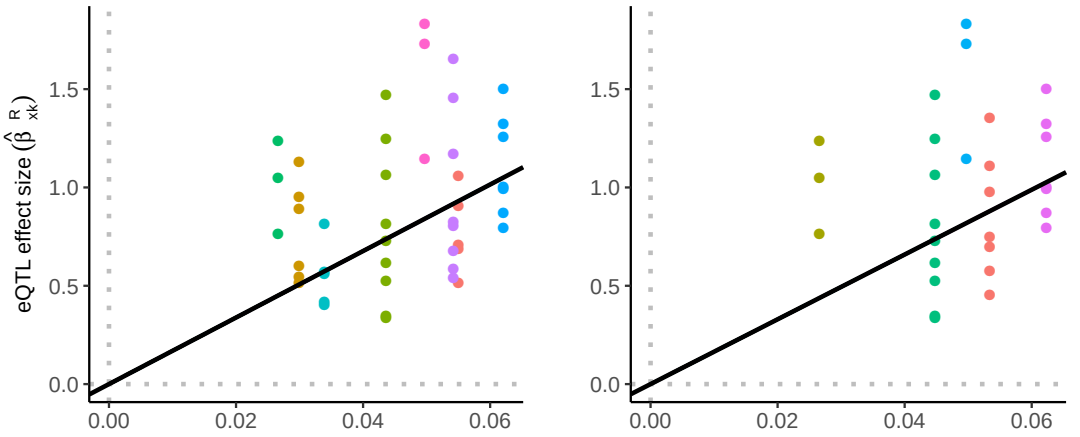

GPX4
IV selection : LD r2 < 0.5
IV selection : LD r2 < 0.3
eQTL effect size (β̂ R xk)
0.50
0.25
0.00
−0.25
−0.04
−0.03
−0.02
−0.01
0.00
CACNB3
0.25
0.00
−0.25
−0.50
−0.075
−0.050
−0.025
0.000
0.025
0.02
RGS14
0.5
0.0
−0.5
−0.02
−0.01
0.00
0.01
0.02
GWAS effect size(β̂y)


FOXN2
IV selection : LD r2 < 0.5
IV selection : LD r2 < 0.3
eQTL effect size (β̂ R xk)
TVP23B
TNKS
GWAS effect size(β̂y)

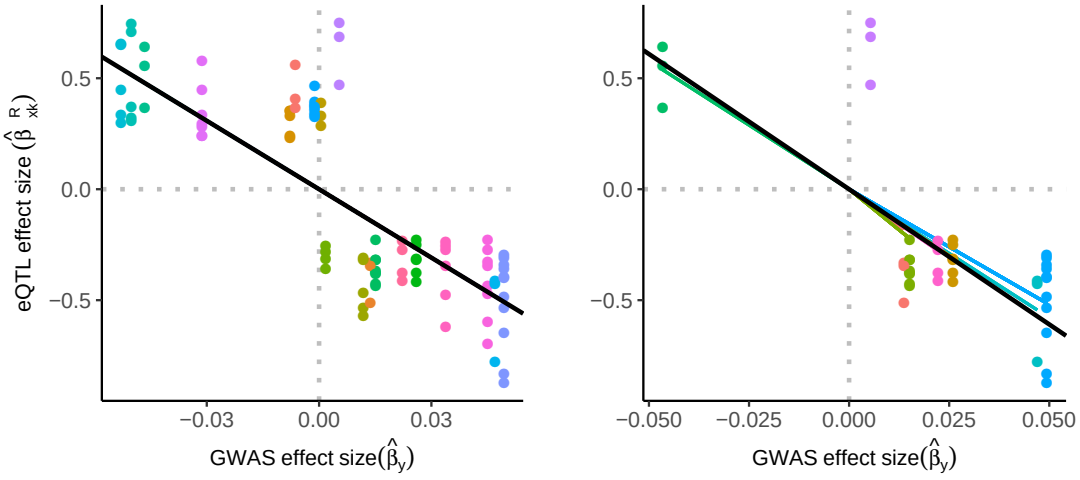
PIDD1

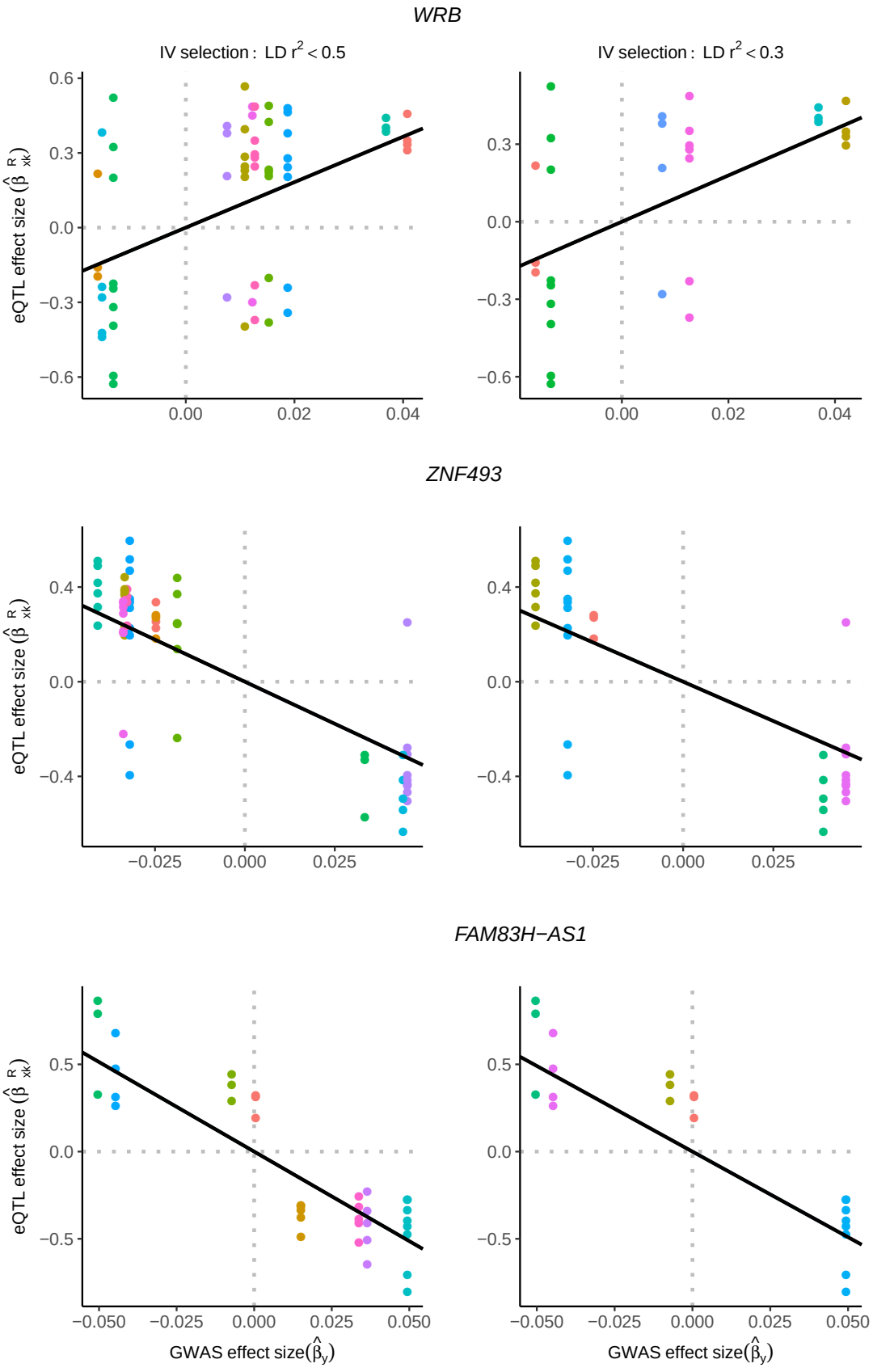


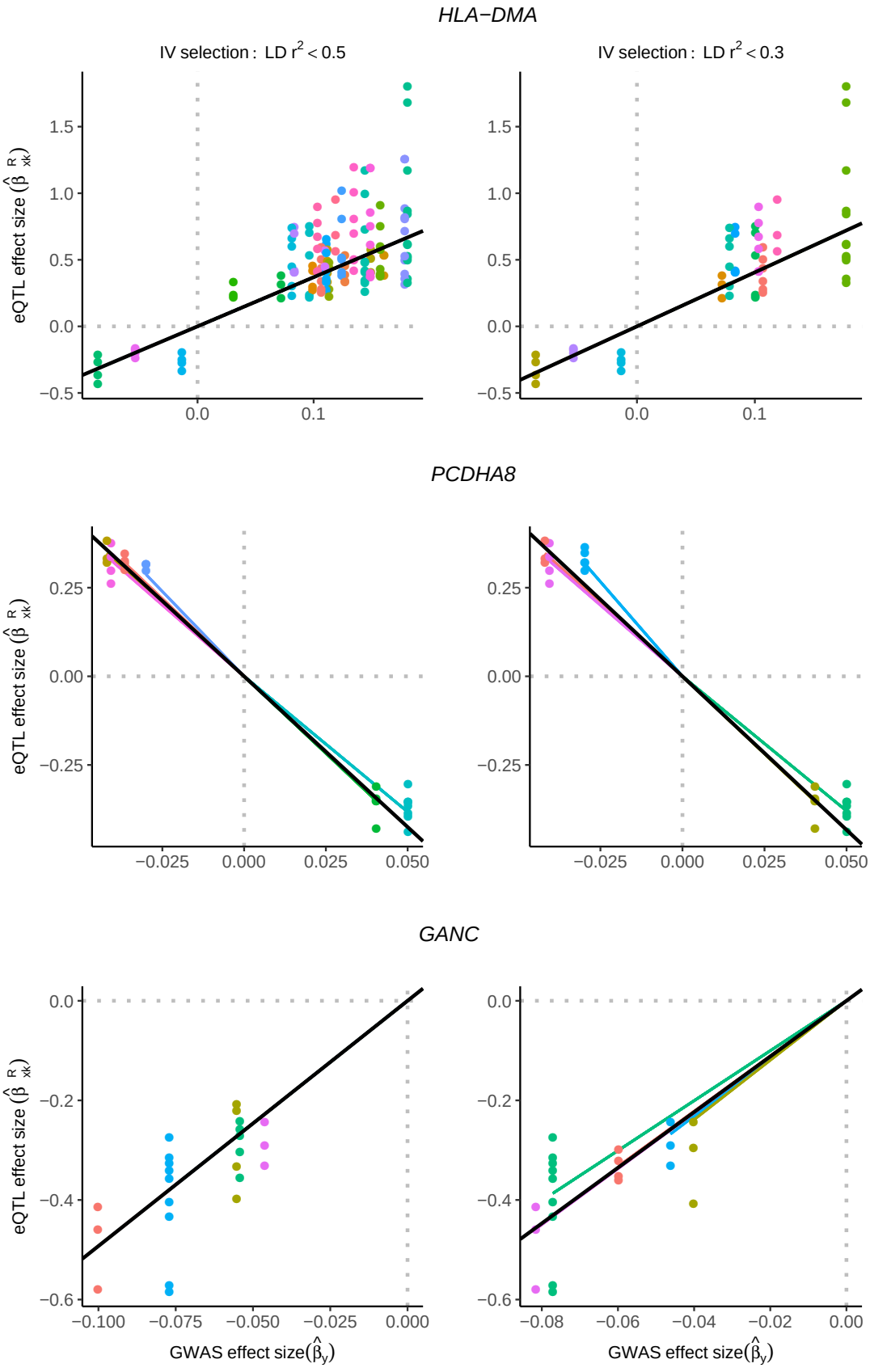
NECTIN3

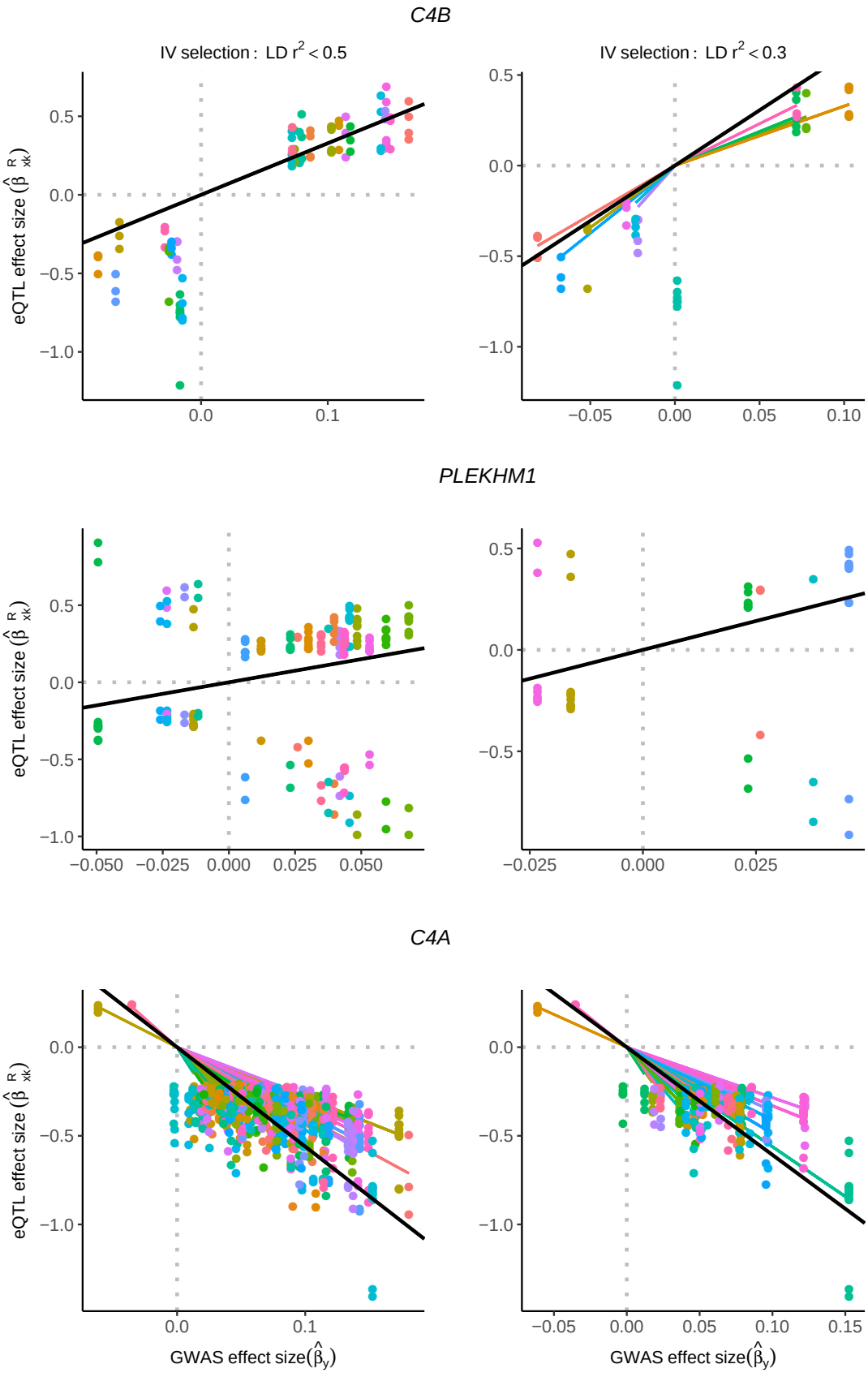


FAM83H




WRB
IV selection : LD r2 < 0.5
IV selection : LD r2 < 0.3
eQTL effect size (β̂ R xk)
0.6
0.3
0.0
−0.3
−0.6
0.00
0.02
0.04
ZNF493
0.4
0.0
−0.4
−0.025
0.000
0.025
FAM83H−AS1
0.5
0.0
−0.5
−0.050
−0.025
0.000
0.025
0.050
GWAS effect size(β̂y)


HLA−DMA
IV selection : LD r2 < 0.5
IV selection : LD r2 < 0.3
eQTL effect size (β̂ Rxk)
1.5
1.0
0.5
0.0
−0.5
0.0
0.1
PCDHA8
0.25
0.00
−0.25
−0.025
0.000
0.025
0.050
GANC
−0.2
−0.4
−0.6
−0.100
−0.075
−0.050
−0.025
0.000
−0.08
−0.06
−0.04
−0.02
0.00
GWAS effect size(β̂y)


C4B
IV selection : LD r2 < 0.5
IV selection : LD r2 < 0.3
eQTL effect size (β̂ R xk)
PLEKHM1
C4A
GWAS effect size(β̂y)

REFERENCES

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

François Aguet, Alvaro N Barbeira, Rodrigo Bonazzola, Andrew Brown, Stephane E Castel, Brian Jo, Silva Kasela, Sarah Kim-Hellmuth, Yanyu Liang, Meritxell Oliva, Princy E Parsana, Elise Flynn, Laure Fresard, Eric R Gaamzon, Andrew R Hamel, Yuan He, Farhad Hormozdiari, Pejman Mohammadi, Manuel Muñoz-Aguirre, YoSon Park, Ashis Saha, Ayellet V Segre, Benjamin J Strober, Xiaoquan Wen, Valentin Wucher, Sayantan Das, Diego Garrido-Martín, Nicole R Gay, Robert E Handsaker, Paul J Hoffman, Seva Kashin, Alan Kwong, Xiao Li, Daniel MacArthur, John M Rouhana, Matthew Stephens, Ellen Todres, Ana Viñuela, Gao Wang, Yuxin Zou, The GTEx Consortium, Christopher D Brown, Nancy Cox, Emmanouil Dermitzakis, Barbara E Engelhardt, Gad Getz, Roderic Guigo, Stephen B Montgomery, Barbara E Stranger, Hae Kyung Im, Alexis Battle, Kristin G Ardlie, and Tuuli Lappalainen. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *bioRxiv*, doi: 10.1101/787903, 2019.

U. D. Akavia, O. Litvin, J. Kim, F. Sanchez-Garcia, D. Kotliar, H. C. Causton, P. Pochanard, E. Mozes, L. A. Garraway, and D. Pe'er. An integrated approach to uncover drivers of cancer. *Cell*, 143(6):1005–1017, Dec 2010.

H. A. Al-Ahmadie, G. Iyer, B. H. Lee, S. N. Scott, R. Mehra, A. Bagrodia, E. J. Jordan, S. P. Gao, R. Ramirez, E. K. Cha, N. B. Desai, E. C. Zabor, I. Ostrovnaya, A. Gopalan, Y. B. Chen, S. W. Fine, S. K. Tickoo, A. Gandhi, J. Hreiki, A. Viale, M. E. Arcila, G. Dalbagni, J. E. Rosenberg, B. H. Bochner, D. F. Bajorin, M. F. Berger, V. E. Reuter, B. S. Taylor, and D. B. Solit. Frequent somatic CDH1 loss-of-function mutations in plasmacytoid variant bladder cancer. *Nat. Genet.*, 48(4):356–358, Apr 2016.

F. W. Albert and L. Kruglyak. The role of regulatory variation in complex traits and disease. *Nat. Rev. Genet.*, 16(4):197–212, Apr 2015.

S. Anders, P. T. Pyl, and W. Huber. HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–169, Jan 2015.

Joshua D. Angrist, Guido W. Imbens, and Donald B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434): 444–455, 1996.

D. Aran, M. Sirota, and A. J. Butte. Systematic pan-cancer analysis of tumour purity. *Nat Commun*, 6:8971, Dec 2015.

L.A. Aroian. The probability function of the product of two normally distributed variables. *Annals Math Stat*, 18:265–271, 1944.

T. M. Ashton, W. G. McKenna, L. A. Kunz-Schughart, and G. S. Higgins. Oxidative Phosphorylation as an Emerging Target in Cancer Therapy. *Clin. Cancer Res.*, 24(11): 2482–2490, 06 2018.

A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, G. R. Abecasis, A. Auton, G. R. Abecasis, D. M. Altshuler, R. M. Durbin, G. R. Abecasis, D. R. Bentley, A. Chakravarti, A. G. Clark, P. Donnelly, E. E. Eichler, P. Flicek, S. B. Gabriel, R. A. Gibbs, E. D. Green, M. E. Hurles, B. M. Knoppers, J. O. Korbel, E. S. Lander, C. Lee, H. Lehrach, E. R. Mardis, G. T. Marth, G. A. McVean, D. A. Nickerson, J. P. Schmidt, S. T. Sherry, J. Wang, R. K. Wilson, R. A. Gibbs, E. Boerwinkle, H. Doddapaneni, Y. Han, V. Korchina, C. Kovar, S. Lee, D. Muzny, J. G. Reid, Y. Zhu, J. Wang, Y. Chang, Q. Feng, X. Fang, X. Guo, M. Jian, H. Jiang, X. Jin, T. Lan, G. Li, J. Li, Y. Li, S. Liu, X. Liu, Y. Lu, X. Ma, M. Tang, B. Wang, G. Wang, H. Wu, R. Wu, X. Xu, Y. Yin, D. Zhang, W. Zhang, J. Zhao, M. Zhao, X. Zheng, E. S. Lander, D. M. Altshuler, S. B. Gabriel, N. Gupta, N. Gharani, L. H. Toji, N. P. Gerry, A. M. Resch, P. Flicek, J. Barker, L. Clarke, L. Gil, S. E. Hunt, G. Kelman, E. Kulesha, R. Leinonen, W. M. McLaren, R. Radhakrishnan, A. Roa, D. Smirnov, R. E. Smith, I. Streeter, A. Thormann, I. Toneva, B. Vaughan, X. Zheng-Bradley, D. R. Bentley, R. Grocock, S. Humphray, T. James, Z. Kingsbury, H. Lehrach, R. Sudbrak, M. W. Albrecht, V. S. Amstislavskiy, T. A. Borodina, M. Lienhard, F. Mertes, M. Sultan, B. Timmermann, M. L. Yaspo, E. R. Mardis, R. K. Wilson, L. Fulton, R. Fulton, S. T. Sherry, V. Ananiev, Z. Belaia, D. Beloslyudtsev, N. Bouk, C. Chen, D. Church, R. Cohen, C. Cook, J. Garner, T. Hefferon, M. Kimelman, C. Liu, J. Lopez, P. Meric, C. O'Sullivan, Y. Ostapchuk, L. Phan, S. Ponomarov, V. Schneider, E. Shekhtman, K. Sirotkin, D. Slotta, H. Zhang, G. A. McVean, R. M. Durbin, S. Balasubramaniam, J. Burton, P. Danecek, T. M. Keane, A. Kolb-Kokocinski, S. McCarthy, J. Stalker, M. Quail, R. M. Durbin, S. Balasubramaniam, J. Burton, P. Danecek, T. M. Keane, A. Kolb-Kokocinski, S. McCarthy, J. Stalker, M. Quail, J. P. Schmidt, C. J. Davies, J. Gollub, T. Webster, B. Wong, Y. Zhan, A. Auton, C. L. Campbell, Y. Kong, A. Marcketta, R. A. Gibbs, F. Yu, L. Antunes, M. Bainbridge, D. Muzny, A. Sabo, Z. Huang, J. Wang, L. J. Coin, L. Fang, X. Guo, X. Jin, G. Li, Q. Li, Y. Li, Z. Li, H. Lin, B. Liu, R. Luo, H. Shao, Y. Xie, C. Ye, C. Yu, F. Zhang, H. Zheng, H. Zhu, C. Alkan, E. Dal, F. Kahveci, G. T. Marth, E. P. Garrison, D. Kural, W. P. Lee, W. F. Leong, M. Stromberg, A. N. Ward, J. Wu, M. Zhang, M. J. Daly, M. A. DePristo, R. E. Handsaker, D. M. Altshuler, E. Banks, G. Bhatia, G. Del Angel, S. B. Gabriel, G. Genovese, N. Gupta, H. Li, S. Kashin, E. S. Lander, S. A. McCarroll, J. C. Nemesh, R. E. Poplin, S. C. Yoon, J. Lihm, V. Makarov, A. G. Clark, S. Gottipati, A. Keinan, J. L. Rodriguez-Flores, J. O. Korbel, T. Rausch, M. H. Fritz, A. M. Stutz, P. Flicek, K. Beal, L. Clarke, A. Datta, J. Herrero, W. M. McLaren, G. R. Ritchie, R. E. Smith, D. Zerbino, X. Zheng-Bradley, P. C. Sabeti, I. Shlyakhter, S. F. Schaffner, J. Vitti, D. N. Cooper, E. V. Ball, P. D. Stenson, D. R. Bentley, B. Barnes, M. Bauer, R. K. Cheetham, A. Cox, M. Eberle, S. Humphray, S. Kahn, L. Murray, J. Peden, R. Shaw, E. E. Kenny, M. A. Batzer, M. K. Konkel, J. A. Walker, D. G. MacArthur, M. Lek, R. Sudbrak, V. S. Amstislavskiy, R. Herwig, E. R. Mardis, L. Ding, D. C. Koboldt, D. Larson, K. Ye, S. Gravel, A. Swaroop, E. Chew, T. Lappalainen, Y. Erlich, M. Gymrek, T. F. Willems, J. T. Simpson, M. D. Shriver, J. A. Rosenfeld, C. D. Bustamante, S. B. Montgomery, F. M. De La Vega, J. K. Byrnes, A. W. Carroll, M. K. DeGorter, P. Lacroute, B. K. Maples, A. R. Martin, A. Moreno-Estrada, S. S. Shringarpure, F. Zakharia, E. Halperin, Y. Baran, C. Lee, E. Cerveira, J. Hwang, A. Malhotra, D. Plewczynski, K. Radew, M. Romanovitch, C. Zhang, F. C. Hyland, D. W.

Craig, A. Christoforides, N. Homer, T. Izatt, A. A. Kurdoglu, S. A. Sinari, K. Squire, S. T. Sherry, C. Xiao, J. Sebat, D. Antaki, M. Gujral, A. Noor, K. Ye, E. G. Burchard, R. D. Hernandez, C. R. Gignoux, D. Haussler, S. J. Katzman, W. J. Kent, B. Howie, A. Ruiz-Linares, E. T. Dermitzakis, S. E. Devine, G. R. Abecasis, H. M. Kang, J. M. Kidd, T. Blackwell, S. Caron, W. Chen, S. Emery, L. Fritsche, C. Fuchsberger, G. Jun, B. Li, R. Lyons, C. Scheller, C. Sidore, S. Song, E. Sliwerska, D. Taliun, A. Tan, R. Welch, M. K. Wing, X. Zhan, P. Awadalla, A. Hodgkinson, Y. Li, X. Shi, A. Quitadamo, G. Lunter, G. A. McVean, J. L. Marchini, S. Myers, C. Churchhouse, O. Delaneau, A. Gupta-Hinch, W. Kretzschmar, Z. Iqbal, I. Mathieson, A. Menelaou, A. Rimmer, D. K. Xifara, T. K. Oleksyk, Y. Fu, X. Liu, M. Xiong, L. Jorde, D. Witherspoon, J. Xing, E. E. Eichler, B. L. Browning, S. R. Browning, F. Hormozdiari, P. H. Sudmant, E. Khurana, R. M. Durbin, M. E. Hurles, C. Tyler-Smith, C. A. Albers, Q. Ayub, S. Balasubramaniam, Y. Chen, V. Colonna, P. Danecek, L. Jostins, T. M. Keane, S. McCarthy, K. Walter, Y. Xue, M. B. Gerstein, A. Abyzov, S. Balasubramanian, J. Chen, D. Clarke, Y. Fu, A. O. Harmanci, M. Jin, D. Lee, J. Liu, X. J. Mu, J. Zhang, Y. Zhang, M. B. Gerstein, A. Abyzov, S. Balasubramanian, J. Chen, D. Clarke, Y. Fu, A. O. Harmanci, M. Jin, D. Lee, J. Liu, X. J. Mu, J. Zhang, Y. Zhang, Y. Li, R. Luo, H. Zhu, C. Alkan, E. Dal, F. Kahveci, G. T. Marth, E. P. Garrison, D. Kural, W. P. Lee, A. N. Ward, J. Wu, M. Zhang, S. A. McCarroll, R. E. Handsaker, D. M. Altshuler, E. Banks, G. Del Angel, G. Genovese, C. Hartl, H. Li, S. Kashin, J. C. Nemesh, K. Shakir, S. C. Yoon, J. Lihm, V. Makarov, J. Degenhardt, J. O. Korbel, M. H. Fritz, S. Meiers, B. Raeder, T. Rausch, A. M. Stutz, P. Flicek, F. P. Casale, L. Clarke, R. E. Smith, O. Stegle, X. Zheng-Bradley, D. R. Bentley, B. Barnes, R. K. Cheetham, M. Eberle, S. Humphray, S. Kahn, L. Murray, R. Shaw, E. W. Lameijer, M. A. Batzer, M. K. Konkel, J. A. Walker, L. Ding, I. Hall, K. Ye, P. Lacroute, C. Lee, E. Cerveira, A. Malhotra, J. Hwang, D. Plewczynski, K. Radew, M. Romanovitch, C. Zhang, D. W. Craig, N. Homer, D. Church, C. Xiao, J. Sebat, D. Antaki, V. Bafna, J. Michaelson, K. Ye, S. E. Devine, E. J. Gardner, G. R. Abecasis, J. M. Kidd, R. E. Mills, G. Dayama, S. Emery, G. Jun, X. Shi, A. Quitadamo, G. Lunter, G. A. McVean, K. Chen, X. Fan, Z. Chong, T. Chen, D. Witherspoon, J. Xing, E. E. Eichler, M. J. Chaisson, F. Hormozdiari, J. Huddleston, M. Malig, B. J. Nelson, P. H. Sudmant, N. F. Parrish, E. Khurana, M. E. Hurles, B. Blackburne, S. J. Lindsay, Z. Ning, K. Walter, Y. Zhang, M. B. Gerstein, A. Abyzov, J. Chen, D. Clarke, H. Lam, X. J. Mu, C. Sisu, J. Zhang, Y. Zhang, M. B. Gerstein, A. Abyzov, J. Chen, D. Clarke, H. Lam, X. J. Mu, C. Sisu, J. Zhang, Y. Zhang, R. A. Gibbs, F. Yu, M. Bainbridge, D. Challis, U. S. Evani, C. Kovar, J. Lu, D. Muzny, U. Nagaswamy, J. G. Reid, A. Sabo, J. Yu, X. Guo, W. Li, Y. Li, R. Wu, G. T. Marth, E. P. Garrison, W. F. Leong, A. N. Ward, G. Del Angel, M. A. DePristo, S. B. Gabriel, N. Gupta, C. Hartl, R. E. Poplin, A. G. Clark, J. L. Rodriguez-Flores, P. Flicek, L. Clarke, R. E. Smith, X. Zheng-Bradley, D. G. MacArthur, E. R. Mardis, R. Fulton, D. C. Koboldt, S. Gravel, C. D. Bustamante, D. W. Craig, A. Christoforides, N. Homer, T. Izatt, S. T. Sherry, C. Xiao, E. T. Dermitzakis, G. R. Abecasis, H. Min Kang, G. A. McVean, M. B. Gerstein, S. Balasubramanian, L. Habegger, M. B. Gerstein, S. Balasubramanian, L. Habegger, H. Yu, P. Flicek, L. Clarke, F. Cunningham, I. Dunham, D. Zerbino, X. Zheng-Bradley, K. Lage, J. B. Jespersen, H. Horn, S. B. Montgomery, M. K. DeGorter, E. Khurana, C. Tyler-Smith, Y. Chen, V. Colonna, Y. Xue, M. B. Gerstein, S. Balasubramanian, Y. Fu, D. Kim, M. B. Gerstein, S. Bala-

subramanian, Y. Fu, D. Kim, A. Auton, A. Marcketta, R. Desalle, A. Narechania, M. A. Sayres, E. P. Garrison, R. E. Handsaker, S. Kashin, S. A. McCarroll, J. L. Rodriguez-Flores, P. Flicek, L. Clarke, X. Zheng-Bradley, Y. Erlich, M. Gymrek, T. F. Willems, C. D. Bustamante, F. L. Mendez, G. D. Poznik, P. A. Underhill, C. Lee, E. Cerveira, A. Malhotra, M. Romanovitch, C. Zhang, G. R. Abecasis, L. Coin, H. Shao, D. Mittelman, C. Tyler-Smith, Q. Ayub, R. Banerjee, M. Cerezo, Y. Chen, T. W. Fitzgerald, S. Louzada, A. Massaia, S. McCarthy, G. R. Ritchie, Y. Xue, F. Yang, C. Tyler-Smith, Q. Ayub, R. Banerjee, M. Cerezo, Y. Chen, T. W. Fitzgerald, S. Louzada, A. Massaia, S. McCarthy, G. R. Ritchie, Y. Xue, F. Yang, R. A. Gibbs, C. Kovar, D. Kalra, W. Hale, D. Muzny, J. G. Reid, J. Wang, X. Dan, X. Guo, G. Li, Y. Li, C. Ye, X. Zheng, D. M. Altshuler, P. Flicek, L. Clarke, X. Zheng-Bradley, D. R. Bentley, A. Cox, S. Humphray, S. Kahn, R. Sudbrak, M. W. Albrecht, M. Lienhard, D. Larson, D. W. Craig, T. Izatt, A. A. Kurdoglu, S. T. Sherry, C. Xiao, D. Haussler, G. R. Abecasis, G. A. McVean, R. M. Durbin, S. Balasubramaniam, T. M. Keane, S. McCarthy, J. Stalker, R. M. Durbin, S. Balasubramaniam, T. M. Keane, S. McCarthy, J. Stalker, A. Chakravarti, B. M. Knoppers, G. R. Abecasis, K. C. Barnes, C. Beiswanger, E. G. Burchard, C. D. Bustamante, H. Cai, H. Cao, R. M. Durbin, N. P. Gerry, N. Gharani, R. A. Gibbs, C. R. Gignoux, S. Gravel, B. Henn, D. Jones, L. Jorde, J. S. Kaye, A. Keinan, A. Kent, A. Kerasidou, Y. Li, R. Mathias, G. A. McVean, A. Moreno-Estrada, P. N. Ossorio, M. Parker, A. M. Resch, C. N. Rotimi, C. D. Royal, K. Sandoval, Y. Su, R. Sudbrak, Z. Tian, S. Tishkoff, L. H. Toji, C. Tyler-Smith, M. Via, Y. Wang, H. Yang, L. Yang, J. Zhu, W. Bodmer, G. Bedoya, A. Ruiz-Linares, Z. Cai, Y. Gao, J. Chu, L. Peltonen, A. Garcia-Montero, A. Orfao, J. Dutil, J. C. Martinez-Cruzado, T. K. Oleksyk, K. C. Barnes, R. A. Mathias, A. Hennis, H. Watson, C. McKenzie, F. Qadri, R. LaRocque, P. C. Sabeti, J. Zhu, X. Deng, P. C. Sabeti, D. Asogun, O. Folarin, C. Happi, O. Omoniwa, M. Stremlau, R. Tariyal, M. Jallow, F. S. Joof, T. Corrah, K. Rockett, D. Kwiatkowski, J. Kooner, T. T. Hien, S. J. Dunstan, N. T. Hang, R. Fonnie, R. Garry, L. Kanneh, L. Moses, P. C. Sabeti, J. Schieffelin, D. S. Grant, C. Gallo, G. Poletti, D. Saleheen, A. Rasheed, D. Saleheen, A. Rasheed, L. D. Brooks, A. L. Felsenfeld, J. E. McEwen, Y. Vaydylevich, E. D. Green, A. Duncanson, M. Dunn, J. A. Schloss, J. Wang, H. Yang, A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. Min Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, G. R. Abecasis, A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. Min Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, and G. R. Abecasis. A global reference for human genetic variation. *Nature*, 526(7571):68–74, Oct 2015.

S.C. Baker. Next-Generation Desktop Sequencers. *Gen. Eng. and Biotech. News*, 32(15), Sept 2012.

A. N. Barbeira, S. P. Dickinson, R. Bonazzola, J. Zheng, H. E. Wheeler, J. M. Torres, E. S. Torstenson, K. P. Shah, T. Garcia, T. L. Edwards, E. A. Stahl, L. M. Huckins, D. L. Nicolae, N. J. Cox, and H. K. Im. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat Commun*, 9(1): 1825, May 2018.

A. N. Barbeira, M. Pividori, J. Zheng, H. E. Wheeler, D. L. Nicolae, and H. K. Im. Integrat-

ing predicted transcriptome from multiple tissues improves association detection. *PLoS Genet.*, 15(1):e1007889, 01 2019.

R. Barfield, H. Feng, A. Gusev, L. Wu, W. Zheng, B. Pasaniuc, and P. Kraft. Transcriptome-wide association studies accounting for colocalization using Egger regression. *Genet. Epidemiol.*, 42(5):418–433, 07 2018.

A. Battle, Z. Khan, S. H. Wang, A. Mitrano, M. J. Ford, J. K. Pritchard, and Y. Gilad. Genomic variation. Impact of regulatory variation from RNA to protein. *Science*, 347 (6222):664–667, Feb 2015.

D. Bell, A. Berchuck, M. Birrer, J. Chien, D. Cramer, F. Dao, R. Dhir, P. DiSaia, H. Gabra, P. Glenn, A. Godwin, J. Gross, L. Hartmann, M. Huang, D. Huntsman, M. Iacocca, M. Imielinski, S. Kalloger, B. Karlan, D. Levine, G. Mills, C. Morrison, D. Mutch, N. Olvera, S. Orsulic, K. Park, N. Petrelli, B. Rabeno, J. Rader, B. Sikic, K. Smith-McCune, A. Sood, D. Bowtell, R. Penny, J. Testa, K. Chang, H. Dinh, J. Drummond, G. Fowler, P. Gunaratne, A. Hawes, C. Kovar, L. Lewis, M. Morgan, I. Newsham, J. Santibanez, J. Reid, L. Trevino, Y. Wu, M. Wang, D. Muzny, D. Wheeler, R. Gibbs, G. Getz, M. Lawrence, K. Cibulskis, A. Sivachenko, C. Sougnez, D. Voet, J. Wilkinson, T. Bloom, K. Ardlie, T. Fennell, J. Baldwin, S. Gabriel, E. Lander, L. L. Ding, R. Fulton, D. Koboldt, M. McLellan, T. Wylie, J. Walker, M. O'Laughlin, D. Dooling, L. Fulton, R. Abbott, N. Dees, Q. Zhang, C. Kandoth, M. Wendl, W. Schierding, D. Shen, C. Harris, H. Schmidt, J. Kalicki, K. Delehaunty, C. Fronick, R. Demeter, L. Cook, J. Wallis, L. Lin, V. Magrini, J. Hodges, J. Eldred, S. Smith, C. Pohl, F. Vandin, B. Raphael, G. Weinstock, E. Mardis, R. Wilson, M. Meyerson, W. Winckler, G. Getz, R. Verhaak, S. Carter, C. Mermel, G. Saksena, H. Nguyen, R. Onofrio, M. Lawrence, D. Hubbard, S. Gupta, A. Crenshaw, A. Ramos, K. Ardlie, L. Chin, A. Protopopov, J. Zhang, T. Kim, I. Perna, Y. Xiao, H. Zhang, G. Ren, N. Sathiamoorthy, R. Park, E. Lee, P. Park, R. Kucherlapati, M. Absher, L. Waite, G. Sherlock, J. Brooks, J. Li, J. Xu, R. Myers, P. W. Laird, L. Cope, J. Herman, H. Shen, D. Weisenberger, H. Noushmehr, F. Pan, T. Triche, B. Berman, D. Van Den Berg, J. Buckley, S. Baylin, P. Spellman, E. Purdom, P. Neuvial, H. Bengtsson, L. Jakkula, S. Durinck, J. Han, S. Dorton, H. Marr, Y. Choi, V. Wang, N. Wang, J. Ngai, J. Conboy, B. Parvin, H. Feiler, T. Speed, J. Gray, A. Levine, N. Socci, Y. Liang, B. Taylor, N. Schultz, L. Borsu, A. Lash, C. Brennan, A. Viale, C. Sander, M. Ladanyi, K. Hoadley, S. Meng, Y. Du, Y. Shi, L. Li, Y. Turman, D. Zang, E. Helms, S. Balu, X. Zhou, J. Wu, M. Topal, D. Hayes, C. Perou, G. Getz, D. Voet, G. Saksena, J. Zhang, H. Zhang, C. Wu, S. Shukla, K. Cibulskis, M. Lawrence, A. Sivachenko, R. Jing, R. Park, Y. Liu, P. Park, M. Noble, L. Chin, H. Carter, D. Kim, R. Karchin, P. Spellman, E. Purdom, P. Neuvial, H. Bengtsson, S. Durinck, J. Han, J. Korkola, L. Heiser, R. Cho, Z. Hu, B. Parvin, T. Speed, J. Gray, N. Schultz, E. Cerami, B. Taylor, A. Olshen, B. Reva, Y. Antipin, R. Shen, P. Mankoo, R. Sheridan, G. Ciriello, W. Chang, J. Bernanke, L. Borsu, D. Levine, M. Ladanyi, C. Sander, D. Haussler, C. Benz, J. Stuart, S. Benz, J. Sanborn, C. Vaske, J. Zhu, C. Szeto, G. Scott, C. Yau, K. Hoadley, Y. Du, S. Balu, D. Hayes, C. Perou, M. Wilkerson, N. Zhang, R. Akbani, K. Baggerly, W. Yung, G. Mills, J. Weinstein, R. Penny, T. Shelton, D. Grimm, M. Hatfield, S. Morris, P. Yena, P. Rhodes, M. Sherman, J. Paulauskis, S. Millis, A. Kahn, J. Greene, R. Sfeir, M. Jensen, J. Chen,

J. Whitmore, S. Alonso, J. Jordan, A. Chu, J. Zhang, A. Barker, C. Compton, G. Eley, M. Ferguson, P. Fielding, D. Gerhard, R. Myles, C. Schaefer, K. Mills Shaw, J. Vaught, J. Vockley, P. Good, M. Guyer, B. Ozenberger, J. Peterson, and E. Thomson. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609–615, Jun 2011.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *JRSS, Series B*, 57(1):289–300, 1995.

R. Beroukhim, G. Getz, L. Nghiemphu, J. Barretina, T. Hsueh, D. Linhart, I. Vivanco, J. C. Lee, J. H. Huang, S. Alexander, J. Du, T. Kau, R. K. Thomas, K. Shah, H. Soto, S. Perner, J. Prensner, R. M. Debiasi, F. Demichelis, C. Hatton, M. A. Rubin, L. A. Garraway, S. F. Nelson, L. Liau, P. S. Mischel, T. F. Cloughesy, M. Meyerson, T. A. Golub, E. S. Lander, I. K. Mellinghoff, and W. R. Sellers. Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc. Natl. Acad. Sci. U.S.A.*, 104(50): 20007–20012, Dec 2007.

R. Beroukhim, C. H. Mermel, D. Porter, G. Wei, S. Raychaudhuri, J. Donovan, J. Barretina, J. S. Boehm, J. Dobson, M. Urashima, K. T. Mc Henry, R. M. Pinchback, A. H. Ligon, Y. J. Cho, L. Haery, H. Greulich, M. Reich, W. Winckler, M. S. Lawrence, B. A. Weir, K. E. Tanaka, D. Y. Chiang, A. J. Bass, A. Loo, C. Hoffman, J. Prensner, T. Liefeld, Q. Gao, D. Yecies, S. Signoretti, E. Maher, F. J. Kaye, H. Sasaki, J. E. Tepper, J. A. Fletcher, J. Tabernero, J. Baselga, M. S. Tsao, F. Demichelis, M. A. Rubin, P. A. Janne, M. J. Daly, C. Nucera, R. L. Levine, B. L. Ebert, S. Gabriel, A. K. Rustgi, C. R. Antonescu, M. Ladanyi, A. Letai, L. A. Garraway, M. Loda, D. G. Beer, L. D. True, A. Okamoto, S. L. Pomeroy, S. Singer, T. R. Golub, E. S. Lander, G. Getz, W. R. Sellers, and M. Meyerson. The landscape of somatic copy-number alteration across human cancers. *Nature*, 463 (7283):899–905, Feb 2010.

P. Bobko and A. Rieck. Large sample estimators for standard errors of functions of correlation coefficients. *Appl Psychol Meas*, 4:385–398, 1980.

M. J. Bonder, R. Luijk, D. V. Zhernakova, M. Moed, P. Deelen, M. Vermaat, M. van Iterson, F. van Dijk, M. van Galen, J. Bot, R. C. Slieker, P. M. Jhamai, M. Verbiest, H. E. Suchiman, M. Verkerk, R. van der Breggen, J. van Rooij, N. Lakenberg, W. Arindrarto, S. M. Kielbasa, I. Jonkers, P. van 't Hof, I. Nooren, M. Beekman, J. Deelen, D. van Heemst, A. Zhernakova, E. F. Tigchelaar, M. A. Swertz, A. Hofman, A. G. Uitterlinden, R. Pool, J. van Dongen, J. J. Hottenga, C. D. Stehouwer, C. J. van der Kallen, C. G. Schalkwijk, L. H. van den Berg, E. W. van Zwet, H. Mei, Y. Li, M. Lemire, T. J. Hudson, P. E. Slagboom, C. Wijmenga, J. H. Veldink, M. M. van Greevenbroek, C. M. van Duijn, D. I. Boomsma, A. Isaacs, R. Jansen, J. B. van Meurs, P. A. 't Hoen, L. Franke, and B. T. Heijmans. Disease variants alter transcription factor levels and methylation of their binding sites. *Nat. Genet.*, 49(1):131–138, 01 2017.

D. Boomsma, A. Busjahn, and L. Peltonen. Classical twin studies and beyond. *Nat. Rev. Genet.*, 3(11):872–882, Nov 2002.

D. Botstein, R. L. White, M. Skolnick, and R. W. Davis. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.*, 32 (3):314–331, May 1980.

John Bound, David A Jaeger, and Regina M Baker. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American statistical association*, 90(430):443–450, 1995.

Jack Bowden, George Davey Smith, and Stephen Burgess. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International journal of epidemiology*, 44(2):512–525, 2015.

Broad Institute. Picard tools, 2019. Available from: `http://broadinstitute.github.io/picard/`.

M. Brown. A method for combining non-independent, one-sided tests of significance. *Biometrics*, 31(4):987–992, 1975.

J. Bryois, A. Buil, D. M. Evans, J. P. Kemp, S. B. Montgomery, D. F. Conrad, K. M. Ho, S. Ring, M. Hurles, P. Deloukas, G. Davey Smith, and E. T. Dermitzakis. Cis and trans effects of human genomic variants on gene expression. *PLoS Genet.*, 10(7):e1004461, Jul 2014.

A. Buniello, J. A. L. MacArthur, M. Cerezo, L. W. Harris, J. Hayhurst, C. Malangone, A. McMahon, J. Morales, E. Mountjoy, E. Sollis, D. Suveges, O. Vrousgou, P. L. Whetzel, R. Amode, J. A. Guillen, H. S. Riat, S. J. Trevanion, P. Hall, H. Junkins, P. Flicek, T. Burdett, L. A. Hindorff, F. Cunningham, and H. Parkinson. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.*, 47(D1):D1005–D1012, 2019.

S. Burgess, A. Butterworth, and S. G. Thompson. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet. Epidemiol.*, 37(7):658–665, Nov 2013.

S. Burgess, F. Dudbridge, and S. G. Thompson. Combining information on multiple instrumental variables in Mendelian randomization: comparison of allele score and summarized data methods. *Stat Med*, 35(11):1880–1906, May 2016.

S. Burgess, D. S. Small, and S. G. Thompson. A review of instrumental variable estimators for Mendelian randomization. *Stat Methods Med Res*, 26(5):2333–2355, Oct 2017. Epub 2015 Aug.

Stephen Burgess, Robert A Scott, Nicholas J Timpson, George Davey Smith, Simon G Thompson, and EPIC-InterAct Consortium. Using published data in mendelian randomization: a blueprint for efficient identification of causal risk factors. *European journal of epidemiology*, 30(7):543–552, 2015.

M. Cai, L. S. Chen, J. Liu, and C. Yang. IGREX for quantifying the impact of genetically regulated expression on phenotypes. *NAR Genom Bioinform*, 2(1):lqaa010, Mar 2020.

I. G. Campbell, S. E. Russell, D. Y. Choong, K. G. Montgomery, M. L. Ciavarella, C. S. Hooi, B. E. Cristiano, R. B. Pearson, and W. A. Phillips. Mutation of the PIK3CA gene in ovarian and breast cancer. *Cancer Res.*, 64(21):7678–7681, Nov 2004.

V. L. Cannataro, S. G. Gaffney, and J. P. Townsend. Effect Sizes of Somatic Mutations in Cancer. *J. Natl. Cancer Inst.*, 110(11):1171–1177, 11 2018.

C. S. Carlson, M. A. Eberle, L. Kruglyak, and D. A. Nickerson. Mapping complex disease loci in whole-genome association studies. *Nature*, 429(6990):446–452, May 2004.

C. C. Chang, C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*, 4:7, 2015.

L. Chen, B. Ge, F. P. Casale, L. Vasquez, T. Kwan, D. Garrido-Martin, S. Watt, Y. Yan, K. Kundu, S. Ecker, A. Datta, D. Richardson, F. Burden, D. Mead, A. L. Mann, J. M. Fernandez, S. Rowlston, S. P. Wilder, S. Farrow, X. Shao, J. J. Lambourne, A. Redensek, C. A. Albers, V. Amstislavskiy, S. Ashford, K. Berentsen, L. Bomba, G. Bourque, D. Bujold, S. Busche, M. Caron, S. H. Chen, W. Cheung, O. Delaneau, E. T. Dermitzakis, H. Elding, I. Colgiu, F. O. Bagger, P. Flicek, E. Habibi, V. Iotchkova, E. Janssen-Megens, B. Kim, H. Lehrach, E. Lowy, A. Mandoli, F. Matarese, M. T. Maurano, J. A. Morris, V. Pancaldi, F. Pourfarzad, K. Rehnstrom, A. Rendon, T. Risch, N. Sharifi, M. M. Simon, M. Sultan, A. Valencia, K. Walter, S. Y. Wang, M. Frontini, S. E. Antonarakis, L. Clarke, M. L. Yaspo, S. Beck, R. Guigo, D. Rico, J. H. A. Martens, W. H. Ouwehand, T. W. Kuijpers, D. S. Paul, H. G. Stunnenberg, O. Stegle, K. Downes, T. Pastinen, and N. Soranzo. Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell*, 167(5):1398–1414, 11 2016.

L. S. Chen, F. Emmert-Streib, and J. D. Storey. Harnessing naturally randomized transcription to infer regulatory relationships among genes. *Genome Biol.*, 8(10):R219, 2007.

L. S. Chen, R. L. Prentice, and P. Wang. A penalized EM algorithm incorporating missing data mechanism for Gaussian parameter estimation. *Biometrics*, 70(2):312–322, Jun 2014.

L. S. Chen, J. Wang, X. Wang, and P. Wang. A Mixed-Effects Model for Incomplete Data from Labeling-based Quantitative Proteomics Experiments. *Ann Appl Stat*, 11(1):114–138, Mar 2017.

Qing Cheng, Yi Yang, Xingjie Shi, Can Yang, Heng Peng, and Jin Liu. MR-LDP: a two-sample mendelian randomization for GWAS summary statistics accounting linkage disequilibrium and horizontal pleiotropy. *NAR Genomics and Bioinformatics*, 2(2):lqaa028, 2020.

B. Chi, Q. Wang, G. Wu, M. Tan, L. Wang, M. Shi, X. Chang, and H. Cheng. Aly and THO are required for assembly of the human TREX complex and association of TREX components with the spliced mRNA. *Nucleic Acids Res.*, 41(2):1294–1306, Jan 2013.

J. M. Chick, S. C. Munger, P. Simecek, E. L. Huttlin, K. Choi, D. M. Gatti, N. Raghupathy, K. L. Svenson, G. A. Churchill, and S. P. Gygi. Defining the consequences of genetic variation on a proteome-wide scale. *Nature*, 534(7608):500–505, 06 2016.

C. Churchhouse and B. Neale. Rapid GWAS of thousands of phenotypes for 337,000 samples in the UK Biobank, 2017. Available from: `http://www.nealelab.is/blog/2017/7/19/rapid-gwas-of-thousands-of-phenotypes-for-337000-samples-in-the-uk-biobank`. Accessed: 4 Feb 2019.

M. Civelek and A. J. Lusis. Systems genetics approaches to understand complex traits. *Nat. Rev. Genet.*, 15(1):34–48, Jan 2014.

A. Colaprico, T. C. Silva, C. Olsen, L. Garofano, C. Cava, D. Garolini, T. S. Sabedot, T. M. Malta, S. M. Pagnotta, I. Castiglioni, M. Ceccarelli, G. Bontempi, and H. Noushmehr. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.*, 44(8):e71, 05 2016.

Cross-Disorder Group of the Psychiatric Genomics Consortium. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet*, 381(9875):1371–1379, Apr 2013.

C. V. Dang. MYC on the path to cancer. *Cell*, 149(1):22–35, Mar 2012.

C. A. Davis, B. C. Hitz, C. A. Sloan, E. T. Chan, J. M. Davidson, I. Gabdank, J. A. Hilton, K. Jain, U. K. Baymuradov, A. K. Narayanan, K. C. Onate, K. Graham, S. R. Miyasato, T. R. Dreszer, J. S. Strattan, O. Jolanki, F. Y. Tanaka, and J. M. Cherry. The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.*, 46 (D1):D794–D801, 01 2018.

D. S. DeLuca, J. Z. Levin, A. Sivachenko, T. Fennell, M. D. Nazaire, C. Williams, M. Reich, W. Winckler, and G. Getz. RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics*, 28(11):1530–1532, Jun 2012.

A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *JRSS, Series B*, 39(1):1–38, 1977.

A. Dimitromanolakis, J. Xu, A. Krol, and L. Briollais. sim1000G: a user-friendly genetic variant simulator in R for unrelated individuals and family-based designs. *BMC Bioinformatics*, 20(1):26, Jan 2019.

A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1): 15–21, Jan 2013.

I. Dunham, A. Kundaje, S. F. Aldred, P. J. Collins, C. A. Davis, F. Doyle, C. B. Epstein, S. Frietze, J. Harrow, R. Kaul, J. Khatun, B. R. Lajoie, S. G. Landt, B. K. Lee, F. Pauli, K. R. Rosenbloom, P. Sabo, A. Safi, A. Sanyal, N. Shoresh, J. M. Simon, L. Song, N. D. Trinklein, R. C. Altshuler, E. Birney, J. B. Brown, C. Cheng, S. Djebali, X. Dong,

I. Dunham, J. Ernst, T. S. Furey, M. Gerstein, B. Giardine, M. Greven, R. C. Hardison, R. S. Harris, J. Herrero, M. M. Hoffman, S. Iyer, M. Kellis, J. Khatun, P. Kheradpour, A. Kundaje, T. Lassmann, Q. Li, X. Lin, G. K. Marinov, A. Merkel, A. Mortazavi, S. C. Parker, T. E. Reddy, J. Rozowsky, F. Schlesinger, R. E. Thurman, J. Wang, L. D. Ward, T. W. Whitfield, S. P. Wilder, W. Wu, H. S. Xi, K. Y. Yip, J. Zhuang, M. J. Pazin, R. F. Lowdon, L. A. Dillon, L. B. Adams, C. J. Kelly, J. Zhang, J. R. Wexler, E. D. Green, P. J. Good, E. A. Feingold, B. E. Bernstein, E. Birney, G. E. Crawford, J. Dekker, L. Elnitski, P. J. Farnham, M. Gerstein, M. C. Giddings, T. R. Gingeras, E. D. Green, R. Guigo, R. C. Hardison, T. J. Hubbard, M. Kellis, W. Kent, J. D. Lieb, E. H. Margulies, R. M. Myers, M. Snyder, J. A. Stamatoyannopoulos, S. A. Tenenbaum, Z. Weng, K. P. White, B. Wold, J. Khatun, Y. Yu, J. Wrobel, B. A. Risk, H. P. Gunawardena, H. C. Kuiper, C. W. Maier, L. Xie, X. Chen, M. C. Giddings, B. E. Bernstein, C. B. Epstein, N. Shoresh, J. Ernst, P. Kheradpour, T. S. Mikkelsen, S. Gillespie, A. Goren, O. Ram, X. Zhang, L. Wang, R. Issner, M. J. Coyne, T. Durham, M. Ku, T. Truong, L. D. Ward, R. C. Altshuler, M. L. Eaton, M. Kellis, S. Djebali, C. A. Davis, A. Merkel, A. Dobin, T. Lassmann, A. Mortazavi, A. Tanzer, J. Lagarde, W. Lin, F. Schlesinger, C. Xue, G. K. Marinov, J. Khatun, B. A. Williams, C. Zaleski, J. Rozowsky, M. Roder, F. Kokocinski, R. F. Abdelhamid, T. Alioto, I. Antoshechkin, M. T. Baer, P. Batut, I. Bell, K. Bell, S. Chakrabortty, X. Chen, J. Chrast, J. Curado, T. Derrien, J. Drenkow, E. Dumais, J. Dumais, R. Duttagupta, M. Fastuca, K. Fejes-Toth, P. Ferreira, S. Foissac, M. J. Fullwood, H. Gao, D. Gonzalez, A. Gordon, H. P. Gunawardena, C. Howald, S. Jha, R. Johnson, P. Kapranov, B. King, C. Kingswood, G. Li, O. J. Luo, E. Park, J. B. Preall, K. Presaud, P. Ribeca, B. A. Risk, D. Robyr, X. Ruan, M. Sammeth, K. S. Sandhu, L. Schaeffer, L. H. See, A. Shahab, J. Skancke, A. M. Suzuki, H. Takahashi, H. Tilgner, D. Trout, N. Walters, H. Wang, J. Wrobel, Y. Yu, Y. Hayashizaki, J. Harrow, M. Gerstein, T. J. Hubbard, A. Reymond, S. E. Antonarakis, G. J. Hannon, M. C. Giddings, Y. Ruan, B. Wold, P. Carninci, R. Guigo, T. R. Gingeras, K. R. Rosenbloom, C. A. Sloan, K. Learned, V. S. Malladi, M. C. Wong, G. P. Barber, M. S. Cline, T. R. Dreszer, S. G. Heitner, D. Karolchik, W. Kent, V. M. Kirkup, L. R. Meyer, J. C. Long, M. Maddren, B. J. Raney, T. S. Furey, L. Song, L. L. Grasfeder, P. G. Giresi, B. K. Lee, A. Battenhouse, N. C. Sheffield, J. M. Simon, K. A. Showers, A. Safi, D. London, A. A. Bhinge, C. Shestak, M. R. Schaner, S. K. Kim, Z. Z. Zhang, P. A. Mieczkowski, J. O. Mieczkowska, Z. Liu, R. M. McDaniell, Y. Ni, N. U. Rashid, M. J. Kim, S. Adar, Z. Zhang, T. Wang, D. Winter, D. Keefe, E. Birney, V. R. Iyer, J. D. Lieb, G. E. Crawford, G. Li, K. S. Sandhu, M. Zheng, P. Wang, O. J. Luo, A. Shahab, M. J. Fullwood, X. Ruan, Y. Ruan, R. M. Myers, F. Pauli, B. A. Williams, J. Gertz, G. K. Marinov, T. E. Reddy, J. Vielmetter, E. Partridge, D. Trout, K. E. Varley, C. Gasper, A. Bansal, S. Pepke, P. Jain, H. Amrhein, K. M. Bowling, M. Anaya, M. K. Cross, B. King, M. A. Muratet, I. Antoshechkin, K. M. Newberry, K. McCue, A. S. Nesmith, K. I. Fisher-Aylor, B. Pusey, G. DeSalvo, S. L. Parker, S. Balasubramanian, N. S. Davis, S. K. Meadows, T. Eggleston, C. Gunter, J. Newberry, S. E. Levy, D. M. Absher, A. Mortazavi, W. H. Wong, B. Wold, M. J. Blow, A. Visel, L. A. Pennachio, L. Elnitski, E. H. Margulies, S. C. Parker, H. M. Petrykowska, A. Abyzov, B. Aken, D. Barrell, G. Barson, A. Berry, A. Bignell, V. Boychenko, G. Bussotti, J. Chrast, C. Davidson, T. Derrien, G. Despacio-Reyes, M. Diekhans, I. Ezkurdia, A. Frankish, J. Gilbert, J. M. Gonzalez, E. Griffiths, R. Harte, D. A. Hendrix, C. Howald, T. Hunt, I. Jungreis, M. Kay, E. Khu-

rana, F. Kokocinski, J. Leng, M. F. Lin, J. Loveland, Z. Lu, D. Manthravadi, M. Mariotti, J. Mudge, G. Mukherjee, C. Notredame, B. Pei, J. M. Rodriguez, G. Saunders, A. Sboner, S. Searle, C. Sisu, C. Snow, C. Steward, A. Tanzer, E. Tapanari, M. L. Tress, M. J. van Baren, N. Walters, S. Washietl, L. Wilming, A. Zadissa, Z. Zhang, M. Brent, D. Haussler, M. Kellis, A. Valencia, M. Gerstein, A. Reymond, R. Guigo, J. Harrow, T. J. Hubbard, S. G. Landt, S. Frietze, A. Abyzov, N. Addleman, R. P. Alexander, R. K. Auerbach, S. Balasubramanian, K. Bettinger, N. Bhardwaj, A. P. Boyle, A. R. Cao, P. Cayting, A. Charos, Y. Cheng, C. Cheng, C. Eastman, G. Euskirchen, J. D. Fleming, F. Grubert, L. Habegger, M. Hariharan, A. Harmanci, S. Iyengar, V. X. Jin, K. J. Karczewski, M. Kasowski, P. Lacroute, H. Lam, N. Lamarre-Vincent, J. Leng, J. Lian, M. Lindahl-Allen, R. Min, B. Miotto, H. Monahan, Z. Moqtaderi, X. J. Mu, H. O'Geen, Z. Ouyang, D. Patacsil, B. Pei, D. Raha, L. Ramirez, B. Reed, J. Rozowsky, A. Sboner, M. Shi, C. Sisu, T. Slifer, H. Witt, L. Wu, X. Xu, K. K. Yan, X. Yang, K. Y. Yip, Z. Zhang, K. Struhl, S. M. Weissman, M. Gerstein, P. J. Farnham, M. Snyder, S. A. Tenenbaum, L. O. Penalva, F. Doyle, S. Karmakar, S. G. Landt, R. R. Bhanvadia, A. Choudhury, M. Domanus, L. Ma, J. Moran, D. Patacsil, T. Slifer, A. Victorsen, X. Yang, M. Snyder, T. Auer, L. Centanin, M. Eichenlaub, F. Gruhl, S. Heermann, B. Hoeckendorf, D. Inoue, T. Kellner, S. Kirchmaier, C. Mueller, R. Reinhardt, L. Schertel, S. Schneider, R. Sinn, B. Wittbrodt, J. Wittbrodt, Z. Weng, T. W. Whitfield, J. Wang, P. J. Collins, S. F. Aldred, N. D. Trinklein, E. C. Partridge, R. M. Myers, J. Dekker, G. Jain, B. R. Lajoie, A. Sanyal, G. Balasundaram, D. L. Bates, R. Byron, T. K. Canfield, M. J. Diegel, D. Dunn, A. K. Ebersol, T. Frum, K. Garg, E. Gist, R. Hansen, L. Boatman, E. Haugen, R. Humbert, G. Jain, A. K. Johnson, E. M. Johnson, T. V. Kutyavin, B. R. Lajoie, K. Lee, D. Lotakis, M. T. Maurano, S. J. Neph, F. V. Neri, E. D. Nguyen, H. Qu, A. P. Reynolds, V. Roach, E. Rynes, P. Sabo, M. E. Sanchez, R. S. Sandstrom, A. Sanyal, A. O. Shafer, A. B. Stergachis, S. Thomas, R. E. Thurman, B. Vernot, J. Vierstra, S. Vong, H. Wang, M. A. Weaver, Y. Yan, M. Zhang, J. M. Akey, M. Bender, M. O. Dorschner, M. Groudine, M. J. MacCoss, P. Navas, G. Stamatoyannopoulos, R. Kaul, J. Dekker, J. A. Stamatoyannopoulos, I. Dunham, K. Beal, A. Brazma, P. Flicek, J. Herrero, N. Johnson, D. Keefe, M. Lukk, N. M. Luscombe, D. Sobral, J. M. Vaquerizas, S. P. Wilder, S. Batzoglou, A. Sidow, N. Hussami, S. Kyriazopoulou-Panagiotopoulou, M. W. Libbrecht, M. A. Schaub, A. Kundaje, R. C. Hardison, W. Miller, B. Giardine, R. S. Harris, W. Wu, P. J. Bickel, B. Banfai, N. P. Boley, J. B. Brown, H. Huang, Q. Li, J. J. Li, W. S. Noble, J. A. Bilmes, O. J. Buske, M. M. Hoffman, A. D. Sahu, P. V. Kharchenko, P. J. Park, D. Baker, J. Taylor, Z. Weng, S. Iyer, X. Dong, M. Greven, X. Lin, J. Wang, H. S. Xi, J. Zhuang, M. Gerstein, R. P. Alexander, S. Balasubramanian, C. Cheng, A. Harmanci, L. Lochovsky, R. Min, X. J. Mu, J. Rozowsky, K. K. Yan, K. Y. Yip, and E. Birney. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, Sep 2012.

J. H. Edwards. Familial predisposition in man. *Br. Med. Bull.*, 25(1):58–64, Jan 1969.

S. L. Edwards, J. Beesley, J. D. French, and A. M. Dunning. Beyond GWASs: illuminating the dark road from association to function. *Am. J. Hum. Genet.*, 93(5):779–797, Nov 2013.

M. J. Ellis, M. Gillette, S. A. Carr, A. G. Paulovich, R. D. Smith, K. K. Rodland, R. R.

Townsend, C. Kinsinger, M. Mesri, H. Rodriguez, and D. C. Liebler. Connecting genomic alterations to cancer biology with proteomics: the NCI Clinical Proteomic Tumor Analysis Consortium. *Cancer Discov*, 3(10):1108–1112, Oct 2013.

N. J. Epsi, S. Panja, S. R. Pine, and A. Mitrofanova. pathCHEMO, a generalizable computational framework uncovers molecular pathways of chemoresistance in lung adenocarcinoma. *Commun Biol*, 2:334, 2019.

R. P. Erickson. Somatic gene mutation and human disease other than cancer. *Mutat. Res.*, 543(2):125–136, Mar 2003.

D.S. Falconer. The inheritance of liability to diseases with variable age of onser, with particular reference to diabetes mellitus. *Am Hum Genet*, 31(1):1–20, Aug 1967.

R.A. Fisher. The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Philosophical Transactions of the Royal Society of Edinburgh*, 52:399–433, Apr 1918.

R.A. Fisher. *Statistical Methods for Research Workers, 4th Edition.* Oliver and Boyd, Edinburgh, 1932.

S. A. Forbes, D. Beare, N. Bindal, S. Bamford, S. Ward, C. G. Cole, M. Jia, C. Kok, H. Boutselakis, T. De, Z. Sondka, L. Ponting, R. Stefancsik, B. Harsha, J. Tate, E. Dawson, S. Thompson, H. Jubb, and P. J. Campbell. COSMIC: High-Resolution Cancer Genetics Using the Catalogue of Somatic Mutations in Cancer. *Curr Protoc Hum Genet*, 91:1–10, 10 2016.

P. A. Futreal, L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, N. Rahman, and M. R. Stratton. A census of human cancer genes. *Nat. Rev. Cancer*, 4(3):177–183, Mar 2004.

F. Galton. Hereditary stature. *Nature*, 33:295–298, 1886.

Eric R Gamazon, Heather E Wheeler, Kaanan P Shah, Sahar V Mozaffari, Keston Aquino-Michaels, Robert J Carroll, Anne E Eyler, Joshua C Denny, GTEx Consortium, Dan L Nicolae, Nancy J Cox, and Hae Kyung Im. A gene-based association method for mapping traits using reference transcriptome data. *Nature genetics*, 47(9):1091, 2015.

H. Gao, M. Zheng, S. Sun, H. Wang, Z. Yue, Y. Zhu, X. Han, J. Yang, Y. Zhou, Y. Cai, and W. Hu. Chaperonin containing TCP1 subunit 5 is a tumor associated antigen of non-small cell lung cancer. *Oncotarget*, 8(38):64170–64179, Sep 2017.

C. Giambartolomei, D. Vukcevic, E. E. Schadt, L. Franke, A. D. Hingorani, C. Wallace, and V. Plagnol. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.*, 10(5):e1004383, May 2014.

C. Giambartolomei, J. Zhenli Liu, W. Zhang, M. Hauberg, H. Shi, J. Boocock, J. Pickrell, A. E. Jaffe, B. Pasaniuc, and P. Roussos. A Bayesian Framework for Multiple Trait Colocalization from Summary Association Statistics. *Bioinformatics*, Mar 2018.

Y. Gilad, S. A. Rifkin, and J. K. Pritchard. Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends Genet.*, 24(8):408–415, Aug 2008.

Kevin J. Gleason, Fan Yang, Brandon L. Pierce, Xin He, and Lin S. Chen. Primo: integration of multiple gwas and omics qtl summary statistics for elucidation of molecular mechanisms of trait-associated snps and detection of pleiotropy in complex traits. *bioRxiv*, 579581, 2019.

W. M. Gombert and A. Krumm. Targeted deletion of multiple CTCF-binding elements in the human C-MYC gene reveals a requirement for CTCF in C-MYC expression. *PLoS ONE*, 4(7):e6109, Jul 2009.

J. Gong, S. Mei, C. Liu, Y. Xiang, Y. Ye, Z. Zhang, J. Feng, R. Liu, L. Diao, A. Y. Guo, X. Miao, and L. Han. PancanQTL: systematic identification of cis-eQTLs and trans-eQTLs in 33 cancer types. *Nucleic Acids Res.*, 46(D1):D971–D976, 01 2018.

L.A. Goodman. On the exact variance of products. *J Am Stat Assoc*, 55:708–713, 1960.

R. L. Grossman, A. P. Heath, V. Ferretti, H. E. Varmus, D. R. Lowy, W. A. Kibbe, and L. M. Staudt. Toward a Shared Vision for Cancer Genomic Data. *N. Engl. J. Med.*, 375 (12):1109–1112, Sep 2016.

F. Grubert, J. B. Zaugg, M. Kasowski, O. Ursu, D. V. Spacek, A. R. Martin, P. Greenside, R. Srivas, D. H. Phanstiel, A. Pekowska, N. Heidari, G. Euskirchen, W. Huber, J. K. Pritchard, C. D. Bustamante, L. M. Steinmetz, A. Kundaje, and M. Snyder. Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. *Cell*, 162(5):1051–1065, Aug 2015.

S. T. Guest, Z. R. Kratche, A. Bollig-Fischer, R. Haddad, and S. P. Ethier. Two members of the TRiC chaperonin complex, CCT2 and TCP1 are essential for survival of breast cancer cells and are linked to driving oncogenes. *Exp. Cell Res.*, 332(2):223–235, Mar 2015.

X. Guo, W. Lin, J. Bao, Q. Cai, X. Pan, M. Bai, Y. Yuan, J. Shi, Y. Sun, M. R. Han, J. Wang, Q. Liu, W. Wen, B. Li, J. Long, J. Chen, and W. Zheng. A Comprehensive cis-eQTL Analysis Revealed Target Genes in Breast Cancer Susceptibility Loci Identified in Genome-wide Association Studies. *Am. J. Hum. Genet.*, 102(5):890–903, 05 2018.

A. Gusev, A. Ko, H. Shi, G. Bhatia, W. Chung, B. W. Penninx, R. Jansen, E. J. de Geus, D. I. Boomsma, F. A. Wright, P. F. Sullivan, E. Nikkola, M. Alvarez, M. Civelek, A. J. Lusis, T. Lehtimaki, E. Raitoharju, M. Kahonen, I. Seppala, O. T. Raitakari, J. Kuusisto, M. Laakso, A. L. Price, P. Pajukanta, and B. Pasaniuc. Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.*, 48(3):245–252, Mar 2016.

S. Hallal, B. P. Russell, H. Wei, M. Y. T. Lee, C. W. Toon, J. Sy, B. Shivalingam, M. E. Buckland, and K. L. Kaufman. Extracellular Vesicles from Neurosurgical Aspirates Identifies Chaperonin Containing TCP1 Subunit 6A as a Potential Glioblastoma Biomarker with Prognostic Significance. *Proteomics*, 19(1-2):e1800157, 01 2019.

L. Han, L. Diao, S. Yu, X. Xu, J. Li, R. Zhang, Y. Yang, H. M. J. Werner, A. K. Eterovic, Y. Yuan, J. Li, N. Nair, R. Minelli, Y. H. Tsang, L. W. T. Cheung, K. J. Jeong, J. Roszik, Z. Ju, S. E. Woodman, Y. Lu, K. L. Scott, J. B. Li, G. B. Mills, and H. Liang. The Genomic Landscape and Clinical Relevance of A-to-I RNA Editing in Human Cancers. *Cancer Cell*, 28(4):515–528, Oct 2015.

S. Hansford, P. Kaurah, H. Li-Chang, M. Woo, J. Senz, H. Pinheiro, K. A. Schrader, D. F. Schaeffer, K. Shumansky, G. Zogopoulos, T. A. Santos, I. Claro, J. Carvalho, C. Nielsen, S. Padilla, A. Lum, A. Talhouk, K. Baker-Lange, S. Richardson, I. Lewis, N. M. Lindor, E. Pennell, A. MacMillan, B. Fernandez, G. Keller, H. Lynch, S. P. Shah, P. Guilford, S. Gallinger, G. Corso, F. Roviello, C. Caldas, C. Oliveira, P. D. Pharoah, and D. G. Huntsman. Hereditary Diffuse Gastric Cancer Syndrome: CDH1 Mutations and Beyond. *JAMA Oncol*, 1(1):23–32, Apr 2015.

X. Hao, P. Zeng, S. Zhang, and X. Zhou. Identifying and exploiting trait-relevant tissues with multiple functional annotations in genome-wide association studies. *PLoS Genet.*, 14 (1):e1007186, 01 2018.

Y. Hasin, M. Seldin, and A. Lusis. Multi-omics approaches to disease. *Genome Biol.*, 18(1): 83, 05 2017.

C. M. Hearne, S. Ghosh, and J. A. Todd. Microsatellites for linkage analysis of genetic traits. *Trends Genet.*, 8(8):288–294, Aug 1992.

C. G. Heath, N. Viphakone, and S. A. Wilson. The role of TREX in gene expression and disease. *Biochem. J.*, 473(19):2911–2935, 10 2016.

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

J. N. Hirschhorn and M. J. Daly. Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.*, 6(2):95–108, Feb 2005.

F. Hormozdiari, M. van de Bunt, A. V. Segre, X. Li, J. W. J. Joo, M. Bilow, J. H. Sul, S. Sankararaman, B. Pasaniuc, and E. Eskin. Colocalization of GWAS and eQTL Signals Detects Target Genes. *Am. J. Hum. Genet.*, 99(6):1245–1260, Dec 2016.

B. N. Howie, P. Donnelly, and J. Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.*, 5(6): e1000529, Jun 2009.

Q. P. Hu, J. Y. Kuang, Q. K. Yang, X. W. Bian, and S. C. Yu. Beyond a tumor suppressor: Soluble E-cadherin promotes the progression of cancer. *Int. J. Cancer*, 138(12):2804–2812, Jun 2016.

L. M. Huckins, A. Dobbyn, D. M. Ruderfer, G. Hoffman, W. Wang, A. F. Pardiñas, V. M. Rajagopal, T. D. Als, H. T Nguyen, K. Girdhar, J. Boocock, P. Roussos, M. Fromer, R. Kramer, E. Domenici, E. R. Gamazon, S. Purcell, D. Demontis, A. D. Børglum, J. T. R. Walters, M. C. O'Donovan, P. Sullivan, M. J. Owen, B. Devlin, S. K. Sieberts, N. J. Cox, H. K. Im, P. Sklar, E. A. Stahl, J. S. Johnson, H. R. Shah, L. L. Klein, K. K. Dang, B. A. Logsdon, M. C. Mahajan, L. M. Mangravite, H. Toyoshiba, R. E. Gur, C. G. Hahn, E. Schadt, D. A. Lewis, V. Haroutunian, M. A. Peters, B. K. Lipska, J. D. Buxbaum, K. Hirai, T. M. Perumal, L. Essioux, S. Ripke, B. M. Neale, A. Corvin, J. T. R. Walters, K. H. Farh, P. A. Holmans, P. Lee, B. Bulik-Sullivan, D. A. Collier, H. Huang, T. H. Pers, I. Agartz, E. Agerbo, M. Albus, M. Alexander, F. Amin, S. A. Bacanu, M. Begemann, R. A. Belliveau, J. Bene, S. E. Bergen, E. Bevilacqua, T. B. Bigdeli, D. W. Black, R. Bruggeman, N. G. Buccola, R. L. Buckner, W. Byerley, W. Cahn, G. Cai, D. Campion, R. M. Cantor, V. J. Carr, N. Carrera, S. V. Catts, K. D. Chambert, R. C. K. Chan, R. Y. L. Chen, E. Y. H. Chen, W. Cheng, E. F. C. Cheung, S. A. Chong, C. R. Cloninger, D. Cohen, N. Cohen, P. Cormican, N. Craddock, J. J. Crowley, D. Curtis, M. Davidson, K. L. Davis, F. Degenhardt, J. Del Favero, D. Demontis, D. Dikeos, T. Dinan, S. Djurovic, G. Donohoe, E. Drapeau, J. Duan, F. Dudbridge, N. Durmishi, P. Eichhammer, J. Eriksson, V. Escott-Price, L. Essioux, A. H. Fanous, M. S. Farrell, J. Frank, L. Franke, R. Freedman, N. B. Freimer, M. Friedl, J. I. Friedman, M. Fromer, G. Genovese, L. Georgieva, I. Giegling, P. Giusti-Rodríguez, S. Godard, J. I. Goldstein, V. Golimbet, S. Gopal, J. Gratten, L. de Haan, C. Hammer, M. L. Hamshere, M. Hansen, T. Hansen, V. Haroutunian, A. M. Hartmann, F. A. Henskens, S. Herms, J. N. Hirschhorn, P. Hoffmann, A. Hofman, M. V. Hollegaard, D. M. Hougaard, M. Ikeda, I. Joa, A. Julia, R. S. Kahn, L. Kalaydjieva, S. Karachanak-Yankova, J. Karjalainen, D. Kavanagh, M. C. Keller, J. L. Kennedy, A. Khrunin, Y. Kim, J. Klovins, J. A. Knowles, B. Konte, V. Kucinskas, Z. A. Kucinskiene, H. Kuzelova-Ptackova, A. K. Kahler, C. Laurent, J. L. C. Keong, S. H. Lee, S. E. Legge, B. Lerer, M. Li, T. Li, K. Y. Liang, J. Lieberman, S. Limborska, C. M. Loughland, J. Lubinski, J. Lonnqvist, M. Macek, P. K. E. Magnusson, B. S. Maher, W. Maier, J. Mallet, S. Marsal, M. Mattheisen, M. Mattingsdal, R. W. McCarley, C. McDonald, A. M. McIntosh, S. Meier, C. J. Meijer, B. Melegh, I. Melle, R. I. Mesholam-Gately, A. Metspalu, P. T. Michie, L. Milani, V. Milanova, Y. Mokrab, D. W. Morris, O. Mors, K. C. Murphy, R. M. Murray, I. Myin-Germeys, B. Muller-Myhsok, M. Nelis, I. Nenadic, D. A. Nertney, G. Nestadt, K. K. Nicodemus, L. Nikitina-Zake, L. Nisenbaum, A. Nordin, E. O'Callaghan, C. O'Dushlaine, F. A. O'Neill, S. Y. Oh, A. Olincy, L. Olsen, J. Van Os, C. Pantelis, G. N. Papadimitriou, S. Papiol, E. Parkhomenko, M. T. Pato, T. Paunio, M. Pejovic-Milovancevic, D. O. Perkins, O. Pietiläinen, J. Pimm, A. J. Pocklington, J. Powell, A. Price, A. E. Pulver, S. M. Purcell, D. Quested, H. B. Rasmussen, A. Reichenberg, M. A. Reimers, A. L. Richards, J. L. Roffman, P. Roussos, D. M. Ruderfer, V. Salomaa, A. R. Sanders, U. Schall, C. R. Schubert, T. G. Schulze, S. G. Schwab, E. M. Scolnick, R. J. Scott, L. J. Seidman, J. Shi, E. Sigurdsson, T. Silagadze, J. M. Silverman, K. Sim, P. Slominsky, J. W. Smoller, H. C. So, C. C. A. Spencer, E. A. Stahl, H. Stefansson, S. Steinberg, E. Stogmann, R. E. Straub, E. Strengman, J. Strohmaier, T. S. Stroup, M. Subramaniam, J. Suvisaari, D. M. Svrakic, J. P. Szatkiewicz, E. Soderman, S. Thirumalai, D. Toncheva, S. Tosato, J. Veijola, J. Waddington, D. Walsh, D. Wang, Q. Wang, B. T. Webb, M. Weiser, D. B. Wildenauer, N. M. Williams, S. Williams, S. H. Witt, A. R.

Wolen, E. H. M. Wong, B. K. Wormley, H. S. Xi, C. C. Zai, X. Zheng, F. Zimprich, N. R. Wray, K. Stefansson, P. M. Visscher, R. Adolfsson, O. A. Andreassen, D. H. R. Blackwood, E. Bramon, J. D. Buxbaum, A. D. Børglum, S. Cichon, A. Darvasi, E. Domenici, H. Ehrenreich, T. Esko, P. V. Gejman, M. Gill, H. Gurling, C. M. Hultman, N. Iwata, A. V. Jablensky, E. G. Jonsson, K. S. Kendler, G. Kirov, J. Knight, T. Lencz, D. F. Levinson, Q. S. Li, J. Liu, A. K. Malhotra, S. A. McCarroll, A. McQuillin, J. L. Moran, P. B. Mortensen, B. J. Mowry, M. M. Nothen, R. A. Ophoff, M. J. Owen, A. Palotie, C. N. Pato, T. L. Petryshen, D. Posthuma, M. Rietschel, B. P. Riley, D. Rujescu, P. C. Sham, P. Sklar, D. S. Clair, D. R. Weinberger, J. R. Wendland, T. Werge, M. J. Daly, P. F. Sullivan, M. C. O'Donovan, A. D. Børglum, D. Demontis, V. M. Rajagopal, T. D. Als, M. Mattheisen, J. Grove, T. Werge, P. B. Mortensen, C. B. Pedersen, E. Agerbo, M. G. Pedersen, O. Mors, M. Nordentoft, D. M. Hougaard, J. Bybjerg-Grauholm, M. Bækvad-Hansen, and C. S. Hansen. Gene expression imputation across multiple brain regions provides insights into schizophrenia risk. *Nat. Genet.*, 51(4):659–674, 04 2019.

C. Hutter and J. C. Zenklusen. The Cancer Genome Atlas: Creating Lasting Value beyond Its Data. *Cell*, 173(2):283–285, 04 2018.

N. Inohara and G. Nuñez. NODs: intracellular proteins involved in inflammation and apoptosis. *Nat. Rev. Immunol.*, 3(5):371–382, May 2003.

P. Jia and Z. Zhao. Impacts of somatic mutations on gene expression: an association perspective. *Brief. Bioinformatics*, 18(3):413–425, 05 2017.

P. Jia, Y. Dai, R. Hu, G. Pei, A. M. Manuel, and Z. Zhao. TSEA-DB: a trait-tissue association map for human complex traits and diseases. *Nucleic Acids Res.*, 48(D1):D1022–D1030, Jan 2020.

A. Johansson, S. Enroth, M. Palmblad, A. M. Deelder, J. Bergquist, and U. Gyllensten. Identification of genetic variants influencing the human plasma proteome. *Proc. Natl. Acad. Sci. U.S.A.*, 110(12):4673–4678, Mar 2013.

Steven G. Johnson. The nlopt nonlinear-optimization package, 2018. Available from: `http://ab-initio.mit.edu/nlopt`. Accessed: 7 Mar 2019.

Hyunseung Kang, Anru Zhang, T Tony Cai, and Dylan S Small. Instrumental variables estimation with some invalid instruments and its application to mendelian randomization. *Journal of the American Statistical Association*, 111(513):132–144, 2016.

D. Karolchik, A. S. Hinrichs, T. S. Furey, K. M. Roskin, C. W. Sugnet, D. Haussler, and W. J. Kent. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.*, 32(Database issue):D493–496, Jan 2004.

M. B. Katan. Apolipoprotein E isoforms, serum cholesterol, and cancer. *Lancet*, 1(8479): 507–508, Mar 1986.

L. N. Kent and G. Leone. The broken cycle: E2F dysfunction in cancer. *Nat. Rev. Cancer*, 19(6):326–338, 06 2019.

J. Kettunen, A. Demirkan, P. Wurtz, H. H. Draisma, T. Haller, R. Rawal, A. Vaarhorst, A. J. Kangas, L. P. Lyytikainen, M. Pirinen, R. Pool, A. P. Sarin, P. Soininen, T. Tukiainen, Q. Wang, M. Tiainen, T. Tynkkynen, N. Amin, T. Zeller, M. Beekman, J. Deelen, K. W. van Dijk, T. Esko, J. J. Hottenga, E. M. van Leeuwen, T. Lehtimaki, E. Mihailov, R. J. Rose, A. J. de Craen, C. Gieger, M. Kahonen, M. Perola, S. Blankenberg, M. J. Savolainen, A. Verhoeven, J. Viikari, G. Willemsen, D. I. Boomsma, C. M. van Duijn, J. Eriksson, A. Jula, M. R. Jarvelin, J. Kaprio, A. Metspalu, O. Raitakari, V. Salomaa, P. E. Slagboom, M. Waldenberger, S. Ripatti, and M. Ala-Korpela. Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. *Nat Commun*, 7: 11122, Mar 2016.

T. M. Kim, R. Xi, L. J. Luquette, R. W. Park, M. D. Johnson, and P. J. Park. Functional genomic analysis of chromosomal aberrations in a compendium of 8000 cancer genomes. *Genome Res.*, 23(2):217–227, Feb 2013.

M. Kircher, D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper, and J. Shendure. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.*, 46(3):310–315, Mar 2014.

D. C. Koboldt, R. S. Fulton, M. D. McLellan, H. Schmidt, J. Kalicki-Veizer, J. F. McMichael, L. L. Fulton, D. J. Dooling, L. Ding, E. R. Mardis, R. K. Wilson, A. Ally, M. Balasundaram, Y. S. Butterfield, R. Carlsen, C. Carter, A. Chu, E. Chuah, H. J. Chun, R. J. Coope, N. Dhalla, R. Guin, C. Hirst, M. Hirst, R. A. Holt, D. Lee, H. I. Li, M. Mayo, R. A. Moore, A. J. Mungall, E. Pleasance, A. Robertson, J. E. Schein, A. Shafiei, P. Sipahimalani, J. R. Slobodan, D. Stoll, A. Tam, N. Thiessen, R. J. Varhol, N. Wye, T. Zeng, Y. Zhao, I. Birol, S. J. Jones, M. A. Marra, A. D. Cherniack, G. Saksena, R. C. Onofrio, N. H. Pho, S. L. Carter, S. E. Schumacher, B. Tabak, B. Hernandez, J. Gentry, H. Nguyen, A. Crenshaw, K. Ardlie, R. Beroukhim, W. Winckler, G. Getz, S. B. Gabriel, M. Meyerson, L. Chin, P. J. Park, R. Kucherlapati, K. A. Hoadley, J. Auman, C. Fan, Y. J. Turman, Y. Shi, L. Li, M. D. Topal, X. He, H. H. Chao, A. Prat, G. O. Silva, M. D. Iglesia, W. Zhao, J. Usary, J. S. Berg, M. Adams, J. Booker, J. Wu, A. Gulabani, T. Bodenheimer, A. P. Hoyle, J. V. Simons, M. G. Soloway, L. E. Mose, S. R. Jefferys, S. Balu, J. S. Parker, D. Hayes, C. M. Perou, S. Malik, S. Mahurkar, H. Shen, D. J. Weisenberger, T. Triche, P. H. Lai, M. S. Bootwalla, D. T. Maglinte, B. P. Berman, D. J. Van Den Berg, S. B. Baylin, P. W. Laird, C. J. Creighton, L. A. Donehower, G. Getz, M. Noble, D. Voet, G. Saksena, N. Gehlenborg, D. DiCara, J. Zhang, H. Zhang, C. J. Wu, S. Y. Liu, M. S. Lawrence, L. Zou, A. Sivachenko, P. Lin, P. Stojanov, R. Jing, J. Cho, R. Sinha, R. W. Park, M. D. Nazaire, J. Robinson, H. Thorvaldsdottir, J. Mesirov, P. J. Park, L. Chin, S. Reynolds, R. B. Kreisberg, B. Bernard, R. Bressler, T. Erkkila, J. Lin, V. Thorsson, W. Zhang, I. Shmulevich, G. Ciriello, N. Weinhold, N. Schultz, J. Gao, E. Cerami, B. Gross, A. Jacobsen, R. Sinha, B. Aksoy, Y. Antipin, B. Reva, R. Shen, B. S. Taylor, M. Ladanyi, C. Sander, P. Anur, P. T. Spellman, Y. Lu, W. Liu, R. R. Verhaak, G. B. Mills, R. Akbani, N. Zhang, B. M. Broom, T. D. Casasent, C. Wakefield, A. K. Unruh, K. Baggerly, K. Coombes, J. N. Weinstein, D. Haussler, C. C. Benz, J. M. Stuart, S. C. Benz, J. Zhu, C. C. Szeto, G. K. Scott, C. Yau, E. O. Paull, D. Carlin,

C. Wong, A. Sokolov, J. Thusberg, S. Mooney, S. Ng, T. C. Goldstein, K. Ellrott, M. Grifford, C. Wilks, S. Ma, B. Craft, C. Yan, Y. Hu, D. Meerzaman, J. M. Gastier-Foster, J. Bowen, N. C. Ramirez, A. D. Black, R. E. Pyatt, P. White, E. J. Zmuda, J. Frick, T. M. Lichtenberg, R. Brookens, M. M. George, M. A. Gerken, H. A. Harper, K. M. Leraas, L. J. Wise, T. R. Tabler, C. McAllister, T. Barr, M. Hart-Kothari, K. Tarvin, C. Saller, G. Sandusky, C. Mitchell, M. V. Iacocca, J. Brown, B. Rabeno, C. Czerwinski, N. Petrelli, O. Dolzhansky, M. Abramov, O. Voronina, O. Potapova, J. R. Marks, W. M. Suchorska, D. Murawa, W. Kycler, M. Ibbs, K. Korski, A. Spychała, P. Murawa, J. J. Brzeziński, H. Perz, R. Łaźniak, M. Teresiak, H. Tatka, E. Leporowska, M. Bogusz-Czerniewicz, J. Malicki, A. Mackiewicz, M. Wiznerowicz, X. V. Le, B. Kohl, V. T. Nguyen, R. Thorp, V. B. Nguyen, H. Sussman, D. P. Bui, R. Hajek, P. H. Nguyen, V. T. Tran, Q. T. Huynh, K. Z. Khan, R. Penny, D. Mallery, E. Curley, C. Shelton, P. Yena, J. N. Ingle, F. J. Couch, W. L. Lingle, T. A. King, A. M. Gonzalez-Angulo, G. B. Mills, M. D. Dyer, S. Liu, X. Meng, M. Patangan, F. Waldman, H. Stoppler, W. Rathmell, L. Thorne, M. Huang, L. Boice, A. Hill, C. Morrison, C. Gaudioso, W. Bshara, K. Daily, S. C. Egea, M. Pegram, C. Gomez-Fernandez, R. Dhir, R. Bhargava, A. Brufsky, C. D. Shriver, J. A. Hooke, J. L. Campbell, R. J. Mural, H. Hu, S. Somiari, C. Larson, B. Deyarmin, L. Kvecher, A. J. Kovatich, M. J. Ellis, T. A. King, H. Hu, F. J. Couch, R. J. Mural, T. Stricker, K. White, O. Olopade, J. N. Ingle, C. Luo, Y. Chen, J. R. Marks, F. Waldman, M. Wiznerowicz, R. Bose, L. W. Chang, A. H. Beck, A. M. Gonzalez-Angulo, T. Pihl, M. Jensen, R. Sfeir, A. Kahn, A. Chu, P. Kothiyal, Z. Wang, E. Snyder, J. Pontius, B. Ayala, M. Backus, J. Walton, J. Baboud, D. Berton, M. Nicholls, D. Srinivasan, R. Raman, S. Girshik, P. Kigonya, S. Alonso, R. Sanbhadti, S. Barletta, D. Pot, M. Sheth, J. A. Demchok, K. R. Shaw, L. Yang, G. Eley, M. L. Ferguson, R. W. Tarnuzzer, J. Zhang, L. A. Dillon, K. Buetow, P. Fielding, B. A. Ozenberger, M. S. Guyer, H. J. Sofia, and J. D. Palchik. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, Oct 2012.

J. Koopman, M. Howe, J. R. Hollenbeck, and H. P. Sin. Small sample mediation testing: misplaced confidence in bootstrapped confidence intervals. *J Appl Psychol*, 100(1):194–202, Jan 2015.

H. Koso, H. Yi, P. Sheridan, S. Miyano, Y. Ino, T. Todo, and S. Watanabe. Identification of RNA-Binding Protein LARP4B as a Tumor Suppressor in Glioma. *Cancer Res.*, 76(8): 2254–2264, 04 2016.

J Kost and M. McDermott. Combining dependent P-values. *Stat. and Prob. Letters*, 60(2): 183–190, 2002.

N. Koundouros and G. Poulogiannis. Reprogramming of fatty acid metabolism in cancer. *Br. J. Cancer*, Dec 2019.

V. N. Kristensen, O. C. Lingjærde, H. G. Russnes, H. K. Vollan, A. Frigessi, and A. L. Børresen-Dale. Principles and methods of integrative genomic analyses in cancer. *Nat. Rev. Cancer*, 14(5):299–313, May 2014.

M. E. Kroehl, S. Lutz, and B. D. Wagner. Permutation-based methods for mediation analysis in studies with small sample sizes. *PeerJ*, 8:e8246, 2020.

P. Kumar, S. Henikoff, and P. C. Ng. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc*, 4(7):1073–1081, 2009.

A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M. J. Ziller, V. Amin, J. W. Whitaker, M. D. Schultz, L. D. Ward, A. Sarkar, G. Quon, R. S. Sandstrom, M. L. Eaton, Y. C. Wu, A. R. Pfenning, X. Wang, M. Claussnitzer, Y. Liu, C. Coarfa, R. A. Harris, N. Shoresh, C. B. Epstein, E. Gjoneska, D. Leung, W. Xie, R. D. Hawkins, R. Lister, C. Hong, P. Gascard, A. J. Mungall, R. Moore, E. Chuah, A. Tam, T. K. Canfield, R. S. Hansen, R. Kaul, P. J. Sabo, M. S. Bansal, A. Carles, J. R. Dixon, K. H. Farh, S. Feizi, R. Karlic, A. R. Kim, A. Kulkarni, D. Li, R. Lowdon, G. Elliott, T. R. Mercer, S. J. Neph, V. Onuchic, P. Polak, N. Rajagopal, P. Ray, R. C. Sallari, K. T. Siebenthall, N. A. Sinnott-Armstrong, M. Stevens, R. E. Thurman, J. Wu, B. Zhang, X. Zhou, A. E. Beaudet, L. A. Boyer, P. L. De Jager, P. J. Farnham, S. J. Fisher, D. Haussler, S. J. Jones, W. Li, M. A. Marra, M. T. McManus, S. Sunyaev, J. A. Thomson, T. D. Tlsty, L. H. Tsai, W. Wang, R. A. Waterland, M. Q. Zhang, L. H. Chadwick, B. E. Bernstein, J. F. Costello, J. R. Ecker, M. Hirst, A. Meissner, A. Milosavljevic, B. Ren, J. A. Stamatoyannopoulos, T. Wang, M. Kellis, A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M. J. Ziller, V. Amin, J. W. Whitaker, M. D. Schultz, L. D. Ward, A. Sarkar, G. Quon, R. S. Sandstrom, M. L. Eaton, Y. C. Wu, A. Pfenning, X. Wang, M. Claussnitzer, Y. Liu, C. Coarfa, R. A. Harris, N. Shoresh, C. B. Epstein, E. Gjoneska, D. Leung, W. Xie, R. D. Hawkins, R. Lister, C. Hong, P. Gascard, A. J. Mungall, R. Moore, E. Chuah, A. Tam, T. K. Canfield, R. S. Hansen, R. Kaul, P. J. Sabo, M. S. Bansal, A. Carles, J. R. Dixon, K. H. Farh, S. Feizi, R. Karlic, A. R. Kim, A. Kulkarni, D. Li, R. Lowdon, G. Elliott, T. R. Mercer, S. J. Neph, V. Onuchic, P. Polak, N. Rajagopal, P. Ray, R. C. Sallari, K. T. Siebenthall, N. A. Sinnott-Armstrong, M. Stevens, R. E. Thurman, J. Wu, B. Zhang, X. Zhou, N. Abdennur, M. Adli, M. Akerman, L. Barrera, J. Antosiewicz-Bourget, T. Ballinger, M. J. Barnes, D. Bates, R. J. Bell, D. A. Bennett, K. Bianco, C. Bock, P. Boyle, J. Brinchmann, P. Caballero-Campo, R. Camahort, M. J. Carrasco-Alfonso, T. Charnecki, H. Chen, Z. Chen, J. B. Cheng, S. Cho, A. Chu, W. Y. Chung, C. Cowan, Q. Athena Deng, V. Deshpande, M. Diegel, B. Ding, T. Durham, L. Echipare, L. Edsall, D. Flowers, O. Genbacev-Krtolica, C. Gifford, S. Gillespie, E. Giste, I. A. Glass, A. Gnirke, M. Gormley, H. Gu, J. Gu, D. A. Hafler, M. J. Hangauer, M. Hariharan, M. Hatan, E. Haugen, Y. He, S. Heimfeld, S. Herlofsen, Z. Hou, R. Humbert, R. Issner, A. R. Jackson, H. Jia, P. Jiang, A. K. Johnson, T. Kadlecek, B. Kamoh, M. Kapidzic, J. Kent, A. Kim, M. Kleinewietfeld, S. Klugman, J. Krishnan, S. Kuan, T. Kutyavin, A. Y. Lee, K. Lee, J. Li, N. Li, Y. Li, K. L. Ligon, S. Lin, Y. Lin, J. Liu, Y. Liu, C. J. Luckey, Y. P. Ma, C. Maire, A. Marson, J. S. Mattick, M. Mayo, M. McMaster, H. Metsky, T. Mikkelsen, D. Miller, M. Miri, E. Mukamel, R. P. Nagarajan, F. Neri, J. Nery, T. Nguyen, H. O'Geen, S. Paithankar, T. Papayannopoulou, M. Pelizzola, P. Plettner, N. E. Propson, S. Raghuraman, B. J. Raney, A. Raubitschek, A. P. Reynolds, H. Richards, K. Riehle, P. Rinaudo, J. F. Robinson, N. B. Rockweiler,

E. Rosen, E. Rynes, J. Schein, R. Sears, T. Sejnowski, A. Shafer, L. Shen, R. Shoemaker, M. Sigaroudinia, I. Slukvin, S. Stehling-Sun, R. Stewart, S. L. Subramanian, K. Suknuntha, S. Swanson, S. Tian, H. Tilden, L. Tsai, M. Urich, I. Vaughn, J. Vierstra, S. Vong, U. Wagner, H. Wang, T. Wang, Y. Wang, A. Weiss, H. Whitton, A. Wildberg, H. Witt, K. J. Won, M. Xie, X. Xing, I. Xu, Z. Xuan, Z. Ye, C. A. Yen, P. Yu, X. Zhang, X. Zhang, J. Zhao, Y. Zhou, J. Zhu, Y. Zhu, S. Ziegler, A. E. Beaudet, L. A. Boyer, P. L. De Jager, P. J. Farnham, S. J. Fisher, D. Haussler, S. J. Jones, W. Li, M. A. Marra, M. T. McManus, S. Sunyaev, J. A. Thomson, T. D. Tlsty, L. H. Tsai, W. Wang, R. A. Waterland, M. Q. Zhang, L. H. Chadwick, B. E. Bernstein, J. F. Costello, J. R. Ecker, M. Hirst, A. Meissner, A. Milosavljevic, B. Ren, J. A. Stamatoyannopoulos, T. Wang, M. Kellis, B. E. Bernstein, J. F. Costello, J. R. Ecker, M. Hirst, A. Meissner, A. Milosavljevic, B. Ren, J. A. Stamatoyannopoulos, T. Wang, and M. Kellis. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, Feb 2015.

T. LaFramboise. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acids Res.*, 37(13):4181–4193, Jul 2009.

Debbie A Lawlor, Roger M Harbord, Jonathan AC Sterne, Nic Timpson, and George Davey Smith. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in medicine*, 27(8):1133–1163, 2008.

M. S. Lawrence, P. Stojanov, C. H. Mermel, J. T. Robinson, L. A. Garraway, T. R. Golub, M. Meyerson, S. B. Gabriel, E. S. Lander, and G. Getz. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, 505(7484):495–501, Jan 2014.

K. Lawrenson, S. Kar, K. McCue, K. Kuchenbaeker, K. Michailidou, J. Tyrer, J. Beesley, S. J. Ramus, Q. Li, M. K. Delgado, J. M. Lee, K. Aittomaki, I. L. Andrulis, H. Anton-Culver, V. Arndt, B. K. Arun, B. Arver, E. V. Bandera, M. Barile, R. B. Barkardottir, D. Barrowdale, M. W. Beckmann, J. Benitez, A. Berchuck, M. Bisogna, L. Bjorge, C. Blomqvist, W. Blot, N. Bogdanova, A. Bojesen, S. E. Bojesen, M. K. Bolla, B. Bonanni, A. L. Børresen-Dale, H. Brauch, P. Brennan, H. Brenner, F. Bruinsma, J. Brunet, S. A. Buhari, B. Burwinkel, R. Butzow, S. S. Buys, Q. Cai, T. Caldes, I. Campbell, R. Canniotto, J. Chang-Claude, J. Chiquette, J. Y. Choi, K. B. Claes, L. S. Cook, A. Cox, D. W. Cramer, S. S. Cross, C. Cybulski, K. Czene, M. B. Daly, F. Damiola, A. Dansonka-Mieszkowska, H. Darabi, J. Dennis, P. Devilee, O. Diez, J. A. Doherty, S. M. Domchek, C. M. Dorfling, T. Dork, M. Dumont, H. Ehrencrona, B. Ejlertsen, S. Ellis, C. Engel, E. Lee, D. G. Evans, P. A. Fasching, L. Feliubadalo, J. Figueroa, D. Flesch-Janys, O. Fletcher, H. Flyger, L. Foretova, F. Fostira, W. D. Foulkes, B. L. Fridley, E. Friedman, D. Frost, G. Gambino, P. A. Ganz, J. Garber, M. Garcia-Closas, A. Gentry-Maharaj, M. Ghoussaini, G. G. Giles, R. Glasspool, A. K. Godwin, M. S. Goldberg, D. E. Goldgar, A. Gonzalez-Neira, E. L. Goode, M. T. Goodman, M. H. Greene, J. Gronwald, P. Guenel, C. A. Haiman, P. Hall, E. Hallberg, U. Hamann, T. V. Hansen, P. A. Harrington, M. Hartman, N. Hassan, S. Healey, F. Heitz, J. Herzog, E. Høgdall, C. K. Høgdall, F. B. Hogervorst, A. Hollestelle, J. L. Hopper, P. J. Hulick, T. Huzarski, E. N. Imyanitov, C. Isaacs, H. Ito, A. Jakubowska, R. Janavicius, A. Jensen, E. M. John, N. Johnson, M. Kabisch, D. Kang, M. Kapuscinski, B. Y. Karlan, S. Khan, L. A. Kiemeney, S. K.

Kjaer, J. A. Knight, I. Konstantopoulou, V. M. Kosma, V. Kristensen, J. Kupryjanczyk, A. Kwong, M. de la Hoya, Y. Laitman, D. Lambrechts, N. Le, K. De Leeneer, J. Lester, D. A. Levine, J. Li, A. Lindblom, J. Long, A. Lophatananon, J. T. Loud, K. Lu, J. Lubinski, A. Mannermaa, S. Manoukian, L. Le Marchand, S. Margolin, F. Marme, L. F. Massuger, K. Matsuo, S. Mazoyer, L. McGuffog, C. McLean, I. McNeish, A. Meindl, U. Menon, A. R. Mensenkamp, R. L. Milne, M. Montagna, K. B. Moysich, K. Muir, A. M. Mulligan, K. L. Nathanson, R. B. Ness, S. L. Neuhausen, H. Nevanlinna, S. Nord, R. L. Nussbaum, K. Odunsi, K. Offit, E. Olah, O. I. Olopade, J. E. Olson, C. Olswold, D. O'Malley, I. Orlow, N. Orr, A. Osorio, S. K. Park, C. L. Pearce, T. Pejovic, P. Peterlongo, G. Pfeiler, C. M. Phelan, E. M. Poole, K. Pylkas, P. Radice, J. Rantala, M. U. Rashid, G. Rennert, V. Rhenius, K. Rhiem, H. A. Risch, G. Rodriguez, M. A. Rossing, A. Rudolph, H. B. Salvesen, S. Sangrajrang, E. J. Sawyer, J. M. Schildkraut, M. K. Schmidt, R. K. Schmutzler, T. A. Sellers, C. Seynaeve, M. Shah, C. Y. Shen, X. O. Shu, W. Sieh, C. F. Singer, O. M. Sinilnikova, S. Slager, H. Song, P. Soucy, M. C. Southey, M. Stenmark-Askmalm, D. Stoppa-Lyonnet, C. Sutter, A. Swerdlow, S. Tchatchou, M. R. Teixeira, S. H. Teo, K. L. Terry, M. B. Terry, M. Thomassen, M. G. Tibiletti, L. Tihomirova, S. Tognazzo, A. E. Toland, I. Tomlinson, D. Torres, T. Truong, C. C. Tseng, N. Tung, S. S. Tworoger, C. Vachon, A. M. van den Ouweland, H. C. van Doorn, E. J. van Rensburg, L. J. Van't Veer, A. Vanderstichele, I. Vergote, J. Vijai, Q. Wang, S. Wang-Gohrke, J. N. Weitzel, N. Wentzensen, A. S. Whittemore, H. Wildiers, R. Winqvist, A. H. Wu, D. Yannoukakos, S. Y. Yoon, J. C. Yu, W. Zheng, Y. Zheng, K. K. Khanna, J. Simard, A. N. Monteiro, J. D. French, F. J. Couch, M. L. Freedman, D. F. Easton, A. M. Dunning, P. D. Pharoah, S. L. Edwards, G. Chenevix-Trench, A. C. Antoniou, S. A. Gayther, M. A. Collonge-Rame, A. Damette, E. Barouk-Simonet, F. Bonnet, V. Bubien, N. Sevenet, M. Longy, P. Berthet, D. Vaur, L. Castera, S. F. Ferrer, Y. J. Bignon, N. Uhrhammer, F. Coron, L. Faivre, A. Baurand, C. Jacquot, G. Bertolone, S. Lizard, D. Leroux, H. Dreyfus, C. Rebischung, M. Peysselon, J. P. Peyrat, J. Fournier, F. Revillion, C. Adenis, L. Venat-Bouvet, M. Leone, N. Boutry-Kryza, A. Calender, S. Giraud, C. Verny-Pierre, C. Lasset, V. Bonadona, L. Barjhoux, H. Sobol, V. Bourdon, T. Noguchi, A. Remenieras, I. Coupier, P. Pujol, J. Sokolowska, M. Bronner, C. Delnatte, S. Bezieau, V. Mari, M. Gauthier-Villars, B. Buecher, E. Rouleau, L. Golmard, V. Moncoutier, M. Belotti, A. de Pauw, C. Elan, E. Fourme, A. M. Birot, C. Saule, M. Laurent, C. Houdayer, F. Lesueur, N. Mebirouk, F. Coulet, C. Colas, F. Soubrier, M. Warcoin, F. Prieur, M. Lebrun, C. Kientz, D. Muller, J. P. Fricker, C. Toulas, R. Guimbaud, L. Gladieff, V. Feillel, I. Mortemousque, B. Bressac-de Paillerets, O. Caron, M. Guillaud-Bataille, H. Gregory, Z. Miedzybrodzka, P. J. Morrison, A. Donaldson, M. T. Rogers, M. J. Kennedy, M. E. Porteous, A. Brady, J. Barwell, C. Foo, F. Lalloo, L. E. Side, J. Eason, A. Henderson, L. Walker, J. Cook, K. Snape, A. Murray, E. McCann, M. A. Rookus, F. E. van Leeuwen, L. E. van der Kolk, M. K. Schmidt, N. S. Russell, J. L. de Lange, R. Wijnands, J. M. Collee, M. J. Hooning, C. Seynaeve, C. H. van Deurzen, I. M. Obdeijn, C. J. van Asperen, R. A. Tollenaar, T. C. van Cronenburg, C. M. Kets, M. G. Ausems, C. C. van der Pol, T. A. van Os, Q. Waisfisz, H. E. Meijers-Heijboer, E. B. Gomez-Garcia, J. C. Oosterwijk, M. J. Mourits, G. H. de Bock, H. F. Vasen, S. Siesling, J. Verloop, L. I. Overbeek, S. Fox, J. Kirk, G. Lindeman, M. Price, D. Bowtell, A. deFazio, and P. Webb. Functional mechanisms underlying pleiotropic risk alleles at the 19p13.1 breast-ovarian cancer susceptibility

locus. *Nat Commun*, 7:12675, 09 2016.

J. T. Leek and J. D. Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.*, 3(9):1724–1735, Sep 2007.

Jeffrey T. Leek, W. Evan Johnson, Hilary S. Parker, Elana J. Fertig, Andrew E. Jaffe, John D. Storey, Yuqing Zhang, and Leonardo Collado Torres. sva: Surrogate variable analysis, 2019. R package version 3.32.1.

L. J. Li, L. S. Zhang, Z. J. Han, Z. Y. He, H. Chen, and Y. M. Li. Chaperonin containing TCP-1 subunit 3 is critical for gastric cancer growth. *Oncotarget*, 8(67):111470–111481, Dec 2017a.

Y. Li, B. van de Geijn, A. Raj, D. A. Knowles, A. A. Petti, D. Golan, Y. Gilad, and J. K. Pritchard. RNA splicing is a primary link between genetic variation and disease. *Science*, 352(6285), 2016.

Y. Li, N. Sahni, R. Pancsa, D. J. McGrail, J. Xu, X. Hua, J. Coulombe-Huntington, M. Ryan, B. Tychhon, D. Sudhakar, L. Hu, M. Tyers, X. Jiang, S. Y. Lin, M. M. Babu, and S. Yi. Revealing the Determinants of Widespread Alternative Splicing Perturbation in Cancer. *Cell Rep*, 21(3):798–812, Oct 2017b.

Y. Li, Y. Jiao, Y. Li, and Y. Liu. Expression of La Ribonucleoprotein Domain Family Member 4B (LARP4B) in Liver Cancer and Their Clinical and Prognostic Significance. *Dis. Markers*, 2019:1569049, 2019.

Yang I Li, David A Knowles, Jack Humphrey, Alvaro N Barbeira, Scott P Dickinson, Hae Kyung Im, and Jonathan K Pritchard. Annotation-free quantification of rna splicing using leafcutter. *Nature genetics*, 50(1):151, 2018.

A. Liberzon, C. Birger, H. Thorvaldsdottir, M. Ghandi, J. P. Mesirov, and P. Tamayo. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst*, 1(6): 417–425, Dec 2015.

J. Z. Liu, S. van Sommeren, H. Huang, S. C. Ng, R. Alberts, A. Takahashi, S. Ripke, J. C. Lee, L. Jostins, T. Shah, S. Abedian, J. H. Cheon, J. Cho, N. E. Dayani, L. Franke, Y. Fuyuno, A. Hart, R. C. Juyal, G. Juyal, W. H. Kim, A. P. Morris, H. Poustchi, W. G. Newman, V. Midha, T. R. Orchard, H. Vahedi, A. Sood, J. Y. Sung, R. Malekzadeh, H. J. Westra, K. Yamazaki, S. K. Yang, J. C. Barrett, B. Z. Alizadeh, M. Parkes, T. Bk, M. J. Daly, M. Kubo, C. A. Anderson, R. K. Weersma, S. Abedian, C. Abraham, J. P. Achkar, T. Ahmad, R. Alberts, B. Alizadeh, L. Amininejad, A. N. Anathakrishnan, V. Andersen, C. A. Anderson, J. M. Andrews, V. Annese, G. Aumais, L. Baidoo, R. N. Baldassano, T. Balschun, P. A. Bampton, M. Barclay, J. C. Barrett, T. M. Bayless, J. Bethge, C. Bewshea, J. C. Bis, A. Bitton, T. B K, G. Boucher, O. Brain, S. Brand, S. R. Brant, C. Buning, J. H. Cheon, A. Chew, J. H. Cho, I. Cleynen, A. Cohain, R. Cooney, A. Croft, M. J. Daly, M. D'Amato, S. Danese, N. E. Daryani, D. De Jong, K. M. de Lange, M. De Vos, G. Denapiene, L. A. Denson, K. L. Devaney, O. Dewit, R. D'Inca, H. E.

Drummond, M. Dubinsky, R. H. Duerr, C. Edwards, D. Ellinghaus, M. Esaki, J. Essers, L. R. Ferguson, E. A. Festen, P. Fleshner, T. Florin, D. Franchimont, A. Franke, K. Fransen, Y. Fuyano, R. Gearry, M. Georges, C. Gieger, J. Glas, P. Goyette, T. Green, A. M. Griffiths, S. L. Guthery, H. Hakonarson, J. Halfvarson, K. Hanigan, T. Haritunians, A. Hart, C. Hawkey, N. K. Hayward, M. Hedl, P. Henderson, G. L. Hold, X. Hu, H. Huang, K. Y. Hui, M. Imielinski, A. Ippoliti, O. Jazayeri, L. Jonaitis, L. Jostins, G. Juyal, R. C. Juyal, R. Kalla, T. H. Karlsen, T. Kawaguchi, N. A. Kennedy, M. A. Khan, W. H. Kim, T. Kitazono, G. Kiudelis, M. Kubo, S. Kugathasan, L. Kupcinskas, C. A. Lamb, A. Latiano, D. Laukens, I. C. Lawrance, J. C. Lee, C. W. Lees, M. Leja, N. Lewis, J. Van Limbergen, P. Lionetti, J. Z. Liu, E. Louis, Y. Luo, G. Mahy, M. M. Malekzadeh, R. Malekzadeh, J. Mansfield, S. Marriott, D. Massey, C. G. Mathew, T. Matsui, D. P. McGovern, V. Midha, R. Milgrom, S. Mirzaei, M. Mitrovic, G. W. Montgomery, S. Motoya, C. Mowat, W. G. Newman, A. Ng, S. C. Ng, S. M. Ng, S. Nikolaus, E. R. Nimmo, K. Ning, M. Nothen, I. Oikonomou, T. R. Orchard, O. Palmieri, M. Parkes, A. Phillips, C. Y. Ponsioen, U. Potocnik, H. Poustchi, N. J. Prescott, D. D. Proctor, G. Radford-Smith, J. F. Rahier, S. Raychaudhuri, M. Regueiro, F. Rieder, J. D. Rioux, S. Ripke, R. Roberts, R. K. Russell, J. D. Sanderson, M. Sans, J. Satsangi, E. E. Schadt, S. Schreiber, L. P. Schumm, R. Scott, M. Seielstad, T. Shah, Y. Sharma, M. S. Silverberg, A. Simmons, L. A. Simms, A. Singh, J. Skieceviciene, A. Sood, S. L. Spain, A. H. Steinhart, J. M. Stempak, L. Stronati, J. J. Sung, Y. Suzuki, J. Sventoraityte, A. Takahashi, M. Takazoe, H. Tanaka, K. M. Taylor, A. ter Velde, E. Theatre, L. Torkvist, M. Tremelling, H. H. Uhlig, H. Vahedi, A. van der Meulen, S. van Sommeren, E. Vasiliauskas, N. T. Ventham, S. Vermeire, H. W. Verspaget, T. Walters, K. Wang, M. H. Wang, R. K. Weersma, Z. Wei, D. Whiteman, C. Wijmenga, D. C. Wilson, J. Winkelmann, R. J. Xavier, T. Yamada, K. Yamazaki, S. Zeissig, B. Zhang, C. K. Zhang, H. Zhang, W. Zhang, H. Zhao, and Z. Z. Zhao. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.*, 47(9):979–986, Sep 2015.

P. Liu, L. Kong, H. Jin, Y. Wu, X. Tan, and B. Song. Differential secretome of pancreatic cancer cells in serum-containing conditioned medium reveals CCT8 as a new biomarker of pancreatic cancer invasion and metastasis. *Cancer Cell Int.*, 19:262, 2019a.

X. Liu, Y. I. Li, and J. K. Pritchard. Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. *Cell*, 177(4):1022–1034, 05 2019b.

Y. Liu, X. Zhang, J. Lin, Y. Chen, Y. Qiao, S. Guo, Y. Yang, G. Zhu, Q. Pan, J. Wang, and F. Sun. CCT3 acts upstream of YAP and TFCP2 as a potential target and tumour biomarker in liver cancer. *Cell Death Dis*, 10(9):644, 09 2019c.

A. E. Locke, B. Kahali, S. I. Berndt, A. E. Justice, T. H. Pers, F. R. Day, C. Powell, S. Vedantam, M. L. Buchkovich, J. Yang, D. C. Croteau-Chonka, T. Esko, T. Fall, T. Ferreira, S. Gustafsson, Z. Kutalik, J. Luan, R. Magi, J. C. Randall, T. W. Winkler, A. R. Wood, T. Workalemahu, J. D. Faul, J. A. Smith, J. H. Zhao, and W. et al. Zhao. Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518(7538):197–206, Feb 2015.

D. P. MacKinnon, C. M. Lockwood, J. M. Hoffman, S. G. West, and V. Sheets. A comparison of methods to test mediation and other intervening variable effects. *Psychol Methods*, 7 (1):83–104, Mar 2002.

D.P. MacKinnon and C. Lockwood. Distribution of products tests for the mediated effect. Unpublished manuscript, 2001.

D.P. MacKinnon, C. Lockwood, and J. Hoffman. A new method to test for mediation. Paper presented at the annual meeting of the Society for Prevention Research; Park City, UT, Jun 1998.

T. A. Manolio. Genomewide association studies and assessment of the risk of disease. *N. Engl. J. Med.*, 363(2):166–176, Jul 2010.

S. Masuda, R. Das, H. Cheng, E. Hurt, N. Dorman, and R. Reed. Recruitment of the human TREX complex to mRNA during splicing. *Genes Dev.*, 19(13):1512–1517, Jul 2005.

Matthew T. Maurano, Richard Humbert, Eric Rynes, Robert E. Thurman, Eric Haugen, Hao Wang, Alex P. Reynolds, Richard Sandstrom, Hongzhu Qu, Jennifer Brody, Anthony Shafer, Fidencio Neri, Kristen Lee, Tanya Kutyavin, Sandra Stehling-Sun, Audra K. Johnson, Theresa K. Canfield, Erika Giste, Morgan Diegel, Daniel Bates, R. Scott Hansen, Shane Neph, Peter J. Sabo, Shelly Heimfeld, Antony Raubitschek, Steven Ziegler, Chris Cotsapas, Nona Sotoodehnia, Ian Glass, Shamil R. Sunyaev, Rajinder Kaul, and John A. Stamatoyannopoulos. Systematic localization of common disease-associated variation in regulatory DNA. *Science*, 337(6099):1190–1195, 2012.

J. L. McClay, A. A. Shabalin, M. G. Dozmorov, D. E. Adkins, G. Kumar, S. Nerella, S. L. Clark, S. E. Bergen, C. M. Hultman, P. K. Magnusson, P. F. Sullivan, K. A. Aberg, and E. J. van den Oord. High density methylation QTL analysis in human blood via next-generation sequencing of the methylated genomic DNA fraction. *Genome Biol.*, 16:291, Dec 2015.

G. McVicker, B. van de Geijn, J. F. Degner, C. E. Cain, N. E. Banovich, A. Raj, N. Lewellen, M. Myrthil, Y. Gilad, and J. K. Pritchard. Identification of genetic variants that affect histone modifications in human cells. *Science*, 342(6159):747–749, Nov 2013.

P. Mertins, D. R. Mani, K. V. Ruggles, M. A. Gillette, K. R. Clauser, P. Wang, X. Wang, J. W. Qiao, S. Cao, F. Petralia, E. Kawaler, F. Mundt, K. Krug, Z. Tu, J. T. Lei, M. L. Gatza, M. Wilkerson, C. M. Perou, V. Yellapantula, K. L. Huang, C. Lin, M. D. McLellan, P. Yan, S. R. Davies, R. R. Townsend, S. J. Skates, J. Wang, B. Zhang, C. R. Kinsinger, M. Mesri, H. Rodriguez, L. Ding, A. G. Paulovich, D. Fenyo, M. J. Ellis, S. A. Carr, S. A. Carr, M. A. Gillette, K. R. Clauser, E. Kuhn, D. R. Mani, P. Mertins, K. A. Ketchum, R. R. Thangudu, S. Cai, M. Oberti, A. G. Paulovich, J. R. Whiteaker, X. Wang, C. Lin, Y. Ping, N. J. Edwards, S. Madhavan, P. B. McGarvey, P. Wang, F. Petralia, Z. Tu, D. Chan, A. Pandey, L. M. Shih, H. Zhang, Z. Zhang, S. Thomas, H. Zhu, G. A. Whiteley, S. J. Skates, F. M. White, D. A. Levine, E. S. Boja, C. R. Kinsinger, T. Hiltke, M. Mesri, R. C. Rivers, H. Rodriguez, K. M. Shaw, S. E. Stein, D. Fenyo, T. Liu, J. E.

McDermott, S. H. Payne, K. D. Rodland, R. D. Smith, P. Rudnick, M. Snyder, Y. Zhao, X. Chen, D. F. Ransohoff, A. N. Hoofnagle, D. C. Liebler, M. E. Sanders, Z. Shi, R. J. Slebos, D. L. Tabb, B. Zhang, L. J. Zimmerman, Y. Wang, S. Li, S. R. Davies, L. Ding, C. Maher, R. Townsend, M. J. Ellis, J. T. Lei, and J. Luo. Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature*, 534(7605):55–62, 06 2016.

K. Michailidou, S. Lindstrom, J. Dennis, J. Beesley, S. Hui, S. Kar, A. Lemacon, P. Soucy, D. Glubb, A. Rostamianfar, M. K. Bolla, Q. Wang, J. Tyrer, E. Dicks, A. Lee, Z. Wang, J. Allen, R. Keeman, U. Eilber, J. D. French, X. Qing Chen, L. Fachal, K. McCue, A. E. McCart Reed, M. Ghoussaini, J. S. Carroll, X. Jiang, H. Finucane, M. Adams, M. A. Adank, H. Ahsan, K. Aittomaki, H. Anton-Culver, N. N. Antonenkova, V. Arndt, K. J. Aronson, B. Arun, P. L. Auer, F. Bacot, M. Barrdahl, C. Baynes, M. W. Beckmann, S. Behrens, J. Benitez, M. Bermisheva, L. Bernstein, C. Blomqvist, N. V. Bogdanova, S. E. Bojesen, B. Bonanni, A. L. Børresen-Dale, J. S. Brand, H. Brauch, P. Brennan, H. Brenner, L. Brinton, P. Broberg, I. W. Brock, A. Broeks, A. Brooks-Wilson, S. Y. Brucker, T. Bruning, B. Burwinkel, K. Butterbach, Q. Cai, H. Cai, T. Caldes, F. Canzian, A. Carracedo, B. D. Carter, J. E. Castelao, T. L. Chan, T. Y. David Cheng, K. Seng Chia, J. Y. Choi, H. Christiansen, C. L. Clarke, M. Collee, D. M. Conroy, E. Cordina-Duverger, S. Cornelissen, D. G. Cox, A. Cox, S. S. Cross, J. M. Cunningham, K. Czene, M. B. Daly, P. Devilee, K. F. Doheny, T. Dork, I. Dos-Santos-Silva, M. Dumont, L. Durcan, M. Dwek, D. M. Eccles, A. B. Ekici, A. H. Eliassen, C. Ellberg, M. Elvira, C. Engel, M. Eriksson, P. A. Fasching, J. Figueroa, D. Flesch-Janys, O. Fletcher, H. Flyger, L. Fritschi, V. Gaborieau, M. Gabrielson, M. Gago-Dominguez, Y. T. Gao, S. M. Gapstur, J. A. Garcia-Saenz, M. M. Gaudet, V. Georgoulias, G. G. Giles, G. Glendon, M. S. Goldberg, D. E. Goldgar, A. Gonzalez-Neira, G. I. Grenaker Alnæs, M. Grip, J. Gronwald, A. Grundy, P. Guenel, L. Haeberle, E. Hahnen, C. A. Haiman, N. Hakansson, U. Hamann, N. Hamel, S. Hankinson, P. Harrington, S. N. Hart, J. M. Hartikainen, M. Hartman, A. Hein, J. Heyworth, B. Hicks, P. Hillemanns, D. N. Ho, A. Hollestelle, M. J. Hooning, R. N. Hoover, J. L. Hopper, M. F. Hou, C. N. Hsiung, G. Huang, K. Humphreys, J. Ishiguro, H. Ito, M. Iwasaki, H. Iwata, A. Jakubowska, W. Janni, E. M. John, N. Johnson, K. Jones, M. Jones, A. Jukkola-Vuorinen, R. Kaaks, M. Kabisch, K. Kaczmarek, D. Kang, Y. Kasuga, M. J. Kerin, S. Khan, E. Khusnutdinova, J. I. Kiiski, S. W. Kim, J. A. Knight, V. M. Kosma, V. N. Kristensen, U. Kruger, A. Kwong, D. Lambrechts, L. Le Marchand, E. Lee, M. H. Lee, J. W. Lee, C. Neng Lee, F. Lejbkowicz, J. Li, J. Lilyquist, A. Lindblom, J. Lissowska, W. Y. Lo, S. Loibl, J. Long, A. Lophatananon, J. Lubinski, C. Luccarini, M. P. Lux, E. S. K. Ma, R. J. MacInnis, T. Maishman, E. Makalic, K. E. Malone, I. M. Kostovska, A. Mannermaa, S. Manoukian, J. E. Manson, S. Margolin, S. Mariapun, M. E. Martinez, K. Matsuo, D. Mavroudis, J. McKay, C. McLean, H. Meijers-Heijboer, A. Meindl, P. Menendez, U. Menon, J. Meyer, H. Miao, N. Miller, N. A. M. Taib, K. Muir, A. M. Mulligan, C. Mulot, S. L. Neuhausen, H. Nevanlinna, P. Neven, S. F. Nielsen, D. Y. Noh, B. G. Nordestgaard, A. Norman, O. I. Olopade, J. E. Olson, H. Olsson, C. Olswold, N. Orr, V. S. Pankratz, S. K. Park, T. W. Park-Simon, R. Lloyd, J. I. A. Perez, P. Peterlongo, J. Peto, K. A. Phillips, M. Pinchev, D. Plaseska-Karanfilska, R. Prentice, N. Presneau, D. Prokofyeva, E. Pugh, K. Pylkas, B. Rack, P. Radice, N. Rahman, G. Rennert, H. S. Rennert, V. Rhenius, A. Romero, J. Romm, K. J. Ruddy, T. Rudiger, A. Rudolph,

M. Ruebner, E. J. T. Rutgers, E. Saloustros, D. P. Sandler, S. Sangrajrang, E. J. Sawyer, D. F. Schmidt, R. K. Schmutzler, A. Schneeweiss, M. J. Schoemaker, F. Schumacher, P. Schurmann, R. J. Scott, C. Scott, S. Seal, C. Seynaeve, M. Shah, P. Sharma, C. Y. Shen, G. Sheng, M. E. Sherman, M. J. Shrubsole, X. O. Shu, A. Smeets, C. Sohn, M. C. Southey, J. J. Spinelli, C. Stegmaier, S. Stewart-Brown, J. Stone, D. O. Stram, H. Surowy, A. Swerdlow, R. Tamimi, J. A. Taylor, M. Tengstrom, S. H. Teo, M. Beth Terry, D. C. Tessier, S. Thanasitthichai, K. Thone, R. A. E. M. Tollenaar, I. Tomlinson, L. Tong, D. Torres, T. Truong, C. C. Tseng, S. Tsugane, H. U. Ulmer, G. Ursin, M. Untch, C. Vachon, C. J. van Asperen, D. Van Den Berg, A. M. W. van den Ouweland, L. van der Kolk, R. B. van der Luijt, D. Vincent, J. Vollenweider, Q. Waisfisz, S. Wang-Gohrke, C. R. Weinberg, C. Wendt, A. S. Whittemore, H. Wildiers, W. Willett, R. Winqvist, A. Wolk, A. H. Wu, L. Xia, T. Yamaji, X. R. Yang, C. Har Yip, K. Y. Yoo, J. C. Yu, W. Zheng, Y. Zheng, B. Zhu, A. Ziogas, E. Ziv, S. R. Lakhani, A. C. Antoniou, A. Droit, I. L. Andrulis, C. I. Amos, F. J. Couch, P. D. P. Pharoah, J. Chang-Claude, P. Hall, D. J. Hunter, R. L. Milne, M. Garcia-Closas, M. K. Schmidt, S. J. Chanock, A. M. Dunning, S. L. Edwards, G. D. Bader, G. Chenevix-Trench, J. Simard, P. Kraft, and D. F. Easton. Association analysis identifies 65 new breast cancer risk loci. *Nature*, 551(7678):92–94, November 2017.

T. H. Morgan. RANDOM SEGREGATION VERSUS COUPLING IN MENDELIAN INHERITANCE. *Science*, 34(873):384, Sep 1911.

J. Morrison, N. Knoblauch, J. H. Marcus, M. Stephens, and X. He. Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics. *Nat. Genet.*, doi: 10.1038/s41588-020-0655-9, May 2020.

N. E. Morton. Sequential tests for the detection of linkage. *Am. J. Hum. Genet.*, 7(3): 277–318, Sep 1955.

M. Muñoz, R. Pong-Wong, O. Canela-Xandri, K. Rawlik, C. S. Haley, and A. Tenesa. Evaluating the contribution of genetics and familial shared environment to common disease using the UK Biobank. *Nat. Genet.*, 48(9):980–983, 09 2016.

D. M. Muzny, M. N. Bainbridge, K. Chang, H. H. Dinh, J. A. Drummond, G. Fowler, C. L. Kovar, L. R. Lewis, M. B. Morgan, I. F. Newsham, J. G. Reid, J. Santibanez, E. Shinbrot, L. R. Trevino, Y. Q. Wu, M. Wang, P. Gunaratne, L. A. Donehower, C. J. Creighton, D. A. Wheeler, R. A. Gibbs, M. S. Lawrence, D. Voet, R. Jing, K. Cibulskis, A. Sivachenko, P. Stojanov, A. McKenna, E. S. Lander, S. Gabriel, G. Getz, L. Ding, R. S. Fulton, D. C. Koboldt, T. Wylie, J. Walker, D. J. Dooling, L. Fulton, K. D. Delehaunty, C. C. Fronick, R. Demeter, E. R. Mardis, R. K. Wilson, A. Chu, H. J. Chun, A. J. Mungall, E. Pleasance, A. Robertson, D. Stoll, M. Balasundaram, I. Birol, Y. S. Butterfield, E. Chuah, R. J. Coope, N. Dhalla, R. Guin, C. Hirst, M. Hirst, R. A. Holt, D. Lee, H. I. Li, M. Mayo, R. A. Moore, J. E. Schein, J. R. Slobodan, A. Tam, N. Thiessen, R. Varhol, T. Zeng, Y. Zhao, S. J. Jones, M. A. Marra, A. J. Bass, A. H. Ramos, G. Saksena, A. D. Cherniack, S. E. Schumacher, B. Tabak, S. L. Carter, N. H. Pho, H. Nguyen, R. C. Onofrio, A. Crenshaw, K. Ardlie, R. Beroukhim, W. Winckler, G. Getz, M. Meyerson, A. Protopopov, J. Zhang,

A. Hadjipanayis, E. Lee, R. Xi, L. Yang, X. Ren, H. Zhang, N. Sathiamoorthy, S. Shukla, P. C. Chen, P. Haseley, Y. Xiao, S. Lee, J. Seidman, L. Chin, P. J. Park, R. Kucherlapati, J. T. Auman, K. A. Hoadley, Y. Du, M. D. Wilkerson, Y. Shi, C. Liquori, S. Meng, L. Li, Y. J. Turman, M. D. Topal, D. Tan, S. Waring, E. Buda, J. Walsh, C. D. Jones, P. A. Mieczkowski, D. Singh, J. Wu, A. Gulabani, P. Dolina, T. Bodenheimer, A. P. Hoyle, J. V. Simons, M. Soloway, L. E. Mose, S. R. Jefferys, S. Balu, B. D. O'Connor, J. F. Prins, D. Y. Chiang, D. Hayes, C. M. Perou, T. Hinoue, D. J. Weisenberger, D. T. Maglinte, F. Pan, B. P. Berman, D. J. Van Den Berg, H. Shen, T. Triche, S. B. Baylin, P. W. Laird, G. Getz, M. Noble, D. Voet, G. Saksena, N. Gehlenborg, D. DiCara, J. Zhang, H. Zhang, C. J. Wu, S. Y. Liu, S. Shukla, M. S. Lawrence, L. Zhou, A. Sivachenko, P. Lin, P. Stojanov, R. Jing, R. W. Park, M. D. Nazaire, J. Robinson, H. Thorvaldsdottir, J. Mesirov, P. J. Park, L. Chin, V. Thorsson, S. M. Reynolds, B. Bernard, R. Kreisberg, J. Lin, L. Iype, R. Bressler, T. Erkkila, M. Gundapuneni, Y. Liu, A. Norberg, T. Robinson, D. Yang, W. Zhang, I. Shmulevich, J. J. de Ronde, N. Schultz, E. Cerami, G. Ciriello, A. P. Goldberg, B. Gross, A. Jacobsen, J. Gao, B. Kaczkowski, R. Sinha, B. Aksoy, Y. Antipin, B. Reva, R. Shen, B. S. Taylor, T. A. Chan, M. Ladanyi, C. Sander, R. Akbani, N. Zhang, B. M. Broom, T. Casasent, A. Unruh, C. Wakefield, S. R. Hamilton, R. Cason, K. A. Baggerly, J. N. Weinstein, D. Haussler, C. C. Benz, J. M. Stuart, S. C. Benz, J. Sanborn, C. J. Vaske, J. Zhu, C. Szeto, G. K. Scott, C. Yau, S. Ng, T. Goldstein, K. Ellrott, E. Collisson, A. E. Cozen, D. Zerbino, C. Wilks, B. Craft, P. Spellman, R. Penny, T. Shelton, M. Hatfield, S. Morris, P. Yena, C. Shelton, M. Sherman, J. Paulauskis, J. M. Gastier-Foster, J. Bowen, N. C. Ramirez, A. Black, R. Pyatt, L. Wise, P. White, M. Bertagnolli, J. Brown, T. A. Chan, G. C. Chu, C. Czerwinski, F. Denstman, R. Dhir, A. Dorner, C. S. Fuchs, J. G. Guillem, M. Iacocca, H. Juhl, A. Kaufman, B. Kohl, X. Van Le, M. C. Mariano, E. N. Medina, M. Meyers, G. M. Nash, P. B. Paty, N. Petrelli, B. Rabeno, W. G. Richards, D. Solit, P. Swanson, L. Temple, J. E. Tepper, R. Thorp, E. Vakiani, M. R. Weiser, J. E. Willis, G. Witkin, Z. Zeng, M. J. Zinner, C. Zornig, M. A. Jensen, R. Sfeir, A. B. Kahn, A. L. Chu, P. Kothiyal, Z. Wang, E. E. Snyder, J. Pontius, T. D. Pihl, B. Ayala, M. Backus, J. Walton, J. Whitmore, J. Baboud, D. L. Berton, M. C. Nicholls, D. Srinivasan, R. Raman, S. Girshik, P. A. Kigonya, S. Alonso, R. N. Sanbhadti, S. P. Barletta, J. M. Greene, D. A. Pot, K. R. Shaw, L. A. Dillon, K. Buetow, T. Davidsen, J. A. Demchok, G. Eley, M. Ferguson, P. Fielding, C. Schaefer, M. Sheth, L. Yang, M. S. Guyer, B. A. Ozenberger, J. D. Palchik, J. Peterson, H. J. Sofia, and E. Thomson. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407):330–337, Jul 2012.

Charles R Nelson and Richard Startz. Some further results on the exact small sample properties of the instrumental variable estimator. *Econometrica (1986-1998)*, 58(4):967, 1990.

D. L. Nicolae, E. Gamazon, W. Zhang, S. Duan, M. E. Dolan, and N. J. Cox. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.*, 6(4):e1000888, Apr 2010.

A. B. Olshen, E. S. Venkatraman, R. Lucito, and M. Wigler. Circular binary segmentation

for the analysis of array-based DNA copy number data. *Biostatistics*, 5(4):557–572, Oct 2004.

H. Ongen, A. Buil, A. A. Brown, E. T. Dermitzakis, and O. Delaneau. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics*, 32(10):1479–1485, 05 2016.

A. Ooe, K. Kato, and S. Noguchi. Possible involvement of CCT5, RGS3, and YKT6 genes up-regulated in p53-mutated tumors in resistance to docetaxel in human breast cancers. *Breast Cancer Res. Treat.*, 101(3):305–315, Mar 2007.

A. A. Pai, J. K. Pritchard, and Y. Gilad. The genetic and mechanistic basis for variation in gene regulation. *PLoS Genet.*, 11(1):e1004857, Jan 2015.

M. Parkes, A. Cortes, D. A. van Heel, and M. A. Brown. Genetic insights into common pathways and complex relationships among immune-mediated diseases. *Nat. Rev. Genet.*, 14(9):661–673, Sep 2013.

B. Pasaniuc and A. L. Price. Dissecting the genetics of complex traits using summary association statistics. *Nat. Rev. Genet.*, 18(2):117–127, 02 2017.

B. Pasaniuc, N. Zaitlen, H. Shi, G. Bhatia, A. Gusev, J. Pickrell, J. Hirschhorn, D. P. Strachan, N. Patterson, and A. L. Price. Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics*, 30(20):2906–2914, Oct 2014.

X. Peng, X. Xu, Y. Wang, D. H. Hawke, S. Yu, L. Han, Z. Zhou, K. Mojumdar, K. J. Jeong, M. Labrie, Y. H. Tsang, M. Zhang, Y. Lu, P. Hwu, K. L. Scott, H. Liang, and G. B. Mills. A-to-I RNA Editing Contributes to Proteomic Diversity in Cancer. *Cancer Cell*, 33(5): 817–828, 05 2018.

N. Petrucelli, M. B. Daly, and G. L. Feldman. Hereditary breast and ovarian cancer due to mutations in BRCA1 and BRCA2. *Genet. Med.*, 12(5):245–259, May 2010.

J. K. Pickrell, T. Berisa, J. Z. Liu, L. Segurel, J. Y. Tung, and D. A. Hinds. Detection and interpretation of shared genetic influences on 42 human traits. *Nat. Genet.*, 48(7): 709–717, 07 2016.

B. L. Pierce, L. Tong, L. S. Chen, R. Rahaman, M. Argos, F. Jasmine, S. Roy, R. Paul-Brutus, H. J. Westra, L. Franke, T. Esko, R. Zaman, T. Islam, M. Rahman, J. A. Baron, M. G. Kibriya, and H. Ahsan. Mediation analysis demonstrates that trans-eQTLs are often explained by cis-mediation: a genome-wide analysis among 1,800 South Asians. *PLoS Genet.*, 10(12):e1004818, Dec 2014.

B. L. Pierce, L. Tong, M. Argos, K. Demanelis, F. Jasmine, M. Rakibuz-Zaman, G. Sarwar, M. T. Islam, H. Shahriar, T. Islam, M. Rahman, M. Yunus, M. G. Kibriya, L. S. Chen, and H. Ahsan. Co-occurring expression and methylation QTLs allow detection of common causal variants and shared biological mechanisms. *Nat Commun*, 9(1):804, 02 2018.

A. Poduri, G. D. Evrony, X. Cai, and C. A. Walsh. Somatic mutation, genomic variation, and neurological disease. *Science*, 341(6141):1237758, Jul 2013.

T. J. Polderman, B. Benyamin, C. A. de Leeuw, P. F. Sullivan, A. van Bochoven, P. M. Visscher, and D. Posthuma. Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nat. Genet.*, 47(7):702–709, Jul 2015.

Shaun Purcell and Christopher Chang. Plink 1.90, 2017. Available from: `http://www.cog-genomics.org/plink/1.9/`.

G. Qi and N. Chatterjee. Mendelian randomization analysis using mixture models for robust and efficient estimation of causal effects. *Nat Commun*, 10(1):1941, 04 2019.

X. Qiu, X. He, Q. Huang, X. Liu, G. Sun, J. Guo, D. Yuan, L. Yang, N. Ban, S. Fan, T. Tao, and D. Wang. Overexpression of CCT8 and its significance for tumor cell proliferation, migration and invasion in glioma. *Pathol. Res. Pract.*, 211(10):717–725, Oct 2015.

J. M. B. Rees, A. M. Wood, F. Dudbridge, and S. Burgess. Robust methods in Mendelian randomization via penalization of heterogeneous causal estimates. *PLoS ONE*, 14(9): e0222362, 2019.

T. G. Richardson, G. Hemani, T. R. Gaunt, C. L. Relton, and G. Davey Smith. A transcriptome-wide Mendelian randomization study to uncover tissue-dependent regulatory mechanisms across the human phenome. *Nat Commun*, 11(1):185, 01 2020.

S. Ripke, B. M. Neale, A. Corvin, J. T. Walters, K. H. Farh, P. A. Holmans, P. Lee, B. Bulik-Sullivan, D. A. Collier, H. Huang, T. H. Pers, I. Agartz, E. Agerbo, M. Albus, M. Alexander, F. Amin, S. A. Bacanu, M. Begemann, R. A. Belliveau, J. Bene, S. E. Bergen, E. Bevilacqua, T. B. Bigdeli, D. W. Black, R. Bruggeman, N. G. Buccola, R. L. Buckner, W. Byerley, W. Cahn, G. Cai, D. Campion, R. M. Cantor, V. J. Carr, N. Carrera, S. V. Catts, K. D. Chambert, R. C. Chan, R. Y. Chen, E. Y. Chen, W. Cheng, E. F. Cheung, S. A. Chong, C. R. Cloninger, D. Cohen, N. Cohen, P. Cormican, N. Craddock, J. J. Crowley, D. Curtis, M. Davidson, K. L. Davis, F. Degenhardt, J. Del Favero, D. Demontis, D. Dikeos, T. Dinan, S. Djurovic, G. Donohoe, E. Drapeau, J. Duan, F. Dudbridge, N. Durmishi, P. Eichhammer, J. Eriksson, V. Escott-Price, L. Essioux, A. H. Fanous, M. S. Farrell, J. Frank, L. Franke, R. Freedman, N. B. Freimer, M. Friedl, J. I. Friedman, M. Fromer, G. Genovese, L. Georgieva, I. Giegling, P. Giusti-Rodriguez, S. Godard, J. I. Goldstein, V. Golimbet, S. Gopal, J. Gratten, L. de Haan, C. Hammer, M. L. Hamshere, M. Hansen, T. Hansen, V. Haroutunian, A. M. Hartmann, F. A. Henskens, S. Herms, J. N. Hirschhorn, P. Hoffmann, A. Hofman, M. V. Hollegaard, D. M. Hougaard, M. Ikeda, I. Joa, A. Julia, R. S. Kahn, L. Kalaydjieva, S. Karachanak-Yankova, J. Karjalainen, D. Kavanagh, M. C. Keller, J. L. Kennedy, A. Khrunin, Y. Kim, J. Klovins, J. A. Knowles, B. Konte, V. Kucinskas, Z. Ausrele Kucinskiene, H. Kuzelova-Ptackova, A. K. Kahler, C. Laurent, J. L. Keong, S. H. Lee, S. E. Legge, B. Lerer, M. Li, T. Li, K. Y. Liang, J. Lieberman, S. Limborska, C. M. Loughland, J. Lubinski, J. Lonnqvist, M. Macek, P. K. Magnusson, B. S. Maher, W. Maier, J. Mallet, S. Marsal, M. Mattheisen,

M. Mattingsdal, R. W. McCarley, C. McDonald, A. M. McIntosh, S. Meier, C. J. Meijer, B. Melegh, I. Melle, R. I. Mesholam-Gately, A. Metspalu, P. T. Michie, L. Milani, V. Milanova, Y. Mokrab, D. W. Morris, O. Mors, K. C. Murphy, R. M. Murray, I. Myin-Germeys, B. Muller-Myhsok, M. Nelis, I. Nenadic, D. A. Nertney, G. Nestadt, K. K. Nicodemus, L. Nikitina-Zake, L. Nisenbaum, A. Nordin, E. O'Callaghan, C. O'Dushlaine, F. A. O'Neill, S. Y. Oh, A. Olincy, L. Olsen, J. Van Os, C. Pantelis, G. N. Papadimitriou, S. Papiol, E. Parkhomenko, M. T. Pato, T. Paunio, M. Pejovic-Milovancevic, D. O. Perkins, O. Pietilainen, J. Pimm, A. J. Pocklington, J. Powell, A. Price, A. E. Pulver, S. M. Purcell, D. Quested, H. B. Rasmussen, A. Reichenberg, M. A. Reimers, A. L. Richards, J. L. Roffman, P. Roussos, D. M. Ruderfer, V. Salomaa, A. R. Sanders, U. Schall, C. R. Schubert, T. G. Schulze, S. G. Schwab, E. M. Scolnick, R. J. Scott, L. J. Seidman, J. Shi, E. Sigurdsson, T. Silagadze, J. M. Silverman, K. Sim, P. Slominsky, J. W. Smoller, H. C. So, C. A. Spencer, E. A. Stahl, H. Stefansson, S. Steinberg, E. Stogmann, R. E. Straub, E. Strengman, J. Strohmaier, T. S. Stroup, M. Subramaniam, J. Suvisaari, D. M. Svrakic, J. P. Szatkiewicz, E. Soderman, S. Thirumalai, D. Toncheva, S. Tosato, J. Veijola, J. Waddington, D. Walsh, D. Wang, Q. Wang, B. T. Webb, M. Weiser, D. B. Wildenauer, N. M. Williams, S. Williams, S. H. Witt, A. R. Wolen, E. H. Wong, B. K. Wormley, H. S. Xi, C. C. Zai, X. Zheng, F. Zimprich, N. R. Wray, K. Stefansson, P. M. Visscher, R. Adolfsson, O. A. Andreassen, D. H. Blackwood, E. Bramon, J. D. Buxbaum, A. D. Børglum, S. Cichon, A. Darvasi, E. Domenici, H. Ehrenreich, T. Esko, P. V. Gejman, M. Gill, H. Gurling, C. M. Hultman, N. Iwata, A. V. Jablensky, E. G. Jonsson, K. S. Kendler, G. Kirov, J. Knight, T. Lencz, D. F. Levinson, Q. S. Li, J. Liu, A. K. Malhotra, S. A. McCarroll, A. McQuillin, J. L. Moran, P. B. Mortensen, B. J. Mowry, M. M. Nothen, R. A. Ophoff, M. J. Owen, A. Palotie, C. N. Pato, T. L. Petryshen, D. Posthuma, M. Rietschel, B. P. Riley, D. Rujescu, P. C. Sham, P. Sklar, D. St Clair, D. R. Weinberger, J. R. Wendland, T. Werge, M. J. Daly, P. F. Sullivan, and M. C. O'Donovan. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, 511(7510):421–427, Jul 2014.

Matthew E Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47, 2015.

M. D. Robinson and A. Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.*, 11(3):R25, 2010.

D. M. Rotroff and A. A. Motsinger-Reif. Embracing Integrative Multiomics Approaches. *Int J Genomics*, 2016:1715985, 2016.

D.B. Rubin. Inference and missing data. *Biometrika*, 63:581–592, 1976.

F.E. Satterthwaite. An Approximate Distribution of Estimates of Variance Components. *Biometrics*, 2:110–114, 1946.

E. E. Schadt, J. Lamb, X. Yang, J. Zhu, S. Edwards, D. Guhathakurta, S. K. Sieberts, S. Monks, M. Reitman, C. Zhang, P. Y. Lum, A. Leonardson, R. Thieringer, J. M. Metzger, L. Yang, J. Castle, H. Zhu, S. F. Kash, T. A. Drake, A. Sachs, and A. J. Lusis. An

integrative genomics approach to infer causal associations between gene expression and disease. *Nat. Genet.*, 37(7):710–717, Jul 2005.

D. J. Schaid, W. Chen, and N. B. Larson. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Review Genetics*, pages 491–504, 2018.

H. W. Schroeder and L. Cavacini. Structure and function of immunoglobulins. *J. Allergy Clin. Immunol.*, 125(2 Suppl 2):41–52, Feb 2010.

H. Semb and G. Christofori. The tumor-suppressor function of E-cadherin. *Am. J. Hum. Genet.*, 63(6):1588–1593, Dec 1998.

A. A. Shabalin. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, 28(10):1353–1358, May 2012.

S. P. Shah, R. D. Morin, J. Khattra, L. Prentice, T. Pugh, A. Burleigh, A. Delaney, K. Gelmon, R. Guliany, J. Senz, C. Steidl, R. A. Holt, S. Jones, M. Sun, G. Leung, R. Moore, T. Severson, G. A. Taylor, A. E. Teschendorff, K. Tse, G. Turashvili, R. Varhol, R. L. Warren, P. Watson, Y. Zhao, C. Caldas, D. Huntsman, M. Hirst, M. A. Marra, and S. Aparicio. Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature*, 461(7265):809–813, Oct 2009.

N. Shan, W. Zhou, S. Zhang, and Y. Zhang. Identification of HSPA8 as a candidate biomarker for endometrial carcinoma by using iTRAQ-based proteomic analysis. *Onco Targets Ther*, 9:2169–2179, 2016.

X. Shao, N. Lv, J. Liao, J. Long, R. Xue, N. Ai, D. Xu, and X. Fan. Copy number variation is highly correlated with differential gene expression: a pan-cancer study. *BMC Med. Genet.*, 20(1):175, Nov 2019.

S. A. Shargh, M. Sakizli, V. Khalaj, A. Movafagh, H. Yazdi, E. Hagigatjou, A. Sayad, N. Mansouri, S. A. Mortazavi-Tabatabaei, and H. R. Khorram Khorshid. Downregulation of E-cadherin expression in breast cancer by promoter hypermethylation and its relation with progression and prognosis of tumor. *Med. Oncol.*, 31(11):250, Nov 2014.

J. Shendure and H. Ji. Next-generation DNA sequencing. *Nat. Biotechnol.*, 26(10):1135–1145, Oct 2008.

X. Shi, S. Cheng, and W. Wang. Suppression of CCT3 inhibits malignant proliferation of human papillary thyroid carcinoma cell. *Oncol Lett*, 15(6):9202–9208, Jun 2018.

H.W. Siemens. *Twin Pathology: Its Importance, Its Methodology, Its Previous Results.* Springer, Berlin, 1924.

D. A. Silvestris, E. Picardi, V. Cesarini, B. Fosso, N. Mangraviti, L. Massimi, M. Martini, G. Pesole, F. Locatelli, and A. Gallo. Dynamic inosinome profiles reveal novel patient stratification and gender-specific differences in glioblastoma. *Genome Biol.*, 20(1):33, 02 2019.

S. Sivakumaran, F. Agakov, E. Theodoratou, J. G. Prendergast, L. Zgaga, T. Manolio, I. Rudan, P. McKeigue, J. F. Wilson, and H. Campbell. Abundant pleiotropy in human complex diseases and traits. *Am. J. Hum. Genet.*, 89(5):607–618, Nov 2011.

A. K. Smith, V. Kilaru, M. Kocak, L. M. Almli, K. B. Mercer, K. J. Ressler, F. A. Tylavsky, and K. N. Conneely. Methylation quantitative trait loci (meQTLs) are consistently detected across ancestry, developmental stage, and tissue type. *BMC Genomics*, 15:145, Feb 2014.

G. D. Smith and S. Ebrahim. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol*, 32(1): 1–22, Feb 2003.

G. K. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, 3:Article3, 2004.

Michael E Sobel. Asymptotic Confidence Intervals for Indirect Effects in Structural Equation Models. *Sociological Methodology*, 13:290–312, 1982.

H. Solomon and M.A. Stephens. Distribution of a Sum of Weighted Chi-Square Variables. *J Amer Statist Assoc*, 72:881–885, 1977.

N. Solovieff, C. Cotsapas, P. H. Lee, S. M. Purcell, and J. W. Smoller. Pleiotropy in complex traits: challenges and strategies. *Nat. Rev. Genet.*, 14(7):483–495, Jul 2013.

S. Srihari, M. Kalimutho, S. Lal, J. Singla, D. Patel, P. T. Simpson, K. K. Khanna, and M. A. Ragan. Understanding the functional impact of copy number alterations in breast cancer using a network modeling approach. *Mol Biosyst*, 12(3):963–972, Mar 2016.

C. Stavraka and S. Blagden. The La-Related Proteins, a Family with Connections to Cancer. *Biomolecules*, 5(4):2701–2722, Oct 2015.

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. *Nat Protoc*, 7(3):500–507, Feb 2012.

James H Stock, Jonathan H Wright, and Motohiro Yogo. A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics*, 20(4):518–529, 2002.

J. D. Storey and R. Tibshirani. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U.S.A.*, 100(16):9440–9445, Aug 2003.

John D. Storey, Andrew J. Bass, Alan Dabney, and David Robinson. qvalue: Q-value estimation for false discovery rate control, 2015. Available from: http://github.com/jdstorey/qvalue.

E.A. Stouffer, S.A.and Suchman, L.C. DeVinney, S.A. Star, and R.M. Jr. Williams. *The American Soldier, Vol.1: Adjustment during Army Life.* Princeton University Press, Princeton, 1949.

M. R. Stratton, P. J. Campbell, and P. A. Futreal. The cancer genome. *Nature*, 458(7239): 719–724, Apr 2009.

A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.*, 102(43):15545–15550, Oct 2005.

K. Suhre, M. Arnold, A. M. Bhagwat, R. J. Cotton, R. Engelke, J. Raffler, H. Sarwath, G. Thareja, A. Wahl, R. K. DeLisle, L. Gold, M. Pezer, G. Lauc, M. A. El-Din Selim, D. O. Mook-Kanamori, E. K. Al-Dous, Y. A. Mohamoud, J. Malek, K. Strauch, H. Grallert, A. Peters, G. Kastenmuller, C. Gieger, and J. Graumann. Connecting genetic risk to disease end points through the human blood plasma proteome. *Nat Commun*, 8:14357, 02 2017.

B. B. Sun, J. C. Maranville, J. E. Peters, D. Stacey, J. R. Staley, J. Blackshaw, S. Burgess, T. Jiang, E. Paige, P. Surendran, C. Oliver-Williams, M. A. Kamat, B. P. Prins, S. K. Wilcox, E. S. Zimmerman, A. Chi, N. Bansal, S. L. Spain, A. M. Wood, N. W. Morrell, J. R. Bradley, N. Janjic, D. J. Roberts, W. H. Ouwehand, J. A. Todd, N. Soranzo, K. Suhre, D. S. Paul, C. S. Fox, R. M. Plenge, J. Danesh, H. Runz, and A. S. Butterworth. Genomic atlas of the human plasma proteome. *Nature*, 558(7708):73–79, 06 2018.

V. Tam, N. Patel, M. Turcotte, Y. Bosse, G. Pare, and D. Meyre. Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.*, 20(8):467–484, 08 2019.

A. B. Taylor and D. P. MacKinnon. Four applications of permutation methods to testing a single-mediator model. *Behav Res Methods*, 44(3):806–844, Sep 2012.

B. S. Taylor, J. Barretina, N. D. Socci, P. Decarolis, M. Ladanyi, M. Meyerson, S. Singer, and C. Sander. Functional copy-number alterations in cancer. *PLoS ONE*, 3(9):e3179, Sep 2008.

The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490:61–70, 2012.

The GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204–213, 10 2017.

D. C. Thomas and D. V. Conti. Commentary: the concept of 'Mendelian Randomization'. *Int J Epidemiol*, 33(1):21–25, Feb 2004.

Sarah Margaret Urbut, Gao Wang, Peter Carbonetto, and Matthew Stephens. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nature Genetics*, 51:187–195, 2019.

J. Vallin and J. Grantham. The role of the molecular chaperone CCT in protein folding and mediation of cytoskeleton-associated processes: implications for cancer cell biology. *Cell Stress Chaperones*, 24(1):17–27, 01 2019.

M. G. P. van der Wijst, H. Brugge, D. H. de Vries, P. Deelen, M. A. Swertz, and L. Franke. Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nat. Genet.*, 50(4):493–497, Apr 2018.

C. Vandiedonck. Genetic association of molecular traits: A help to identify causative variants in complex diseases. *Clin. Genet.*, 93(3):520–532, Mar 2018.

M. Verbanck, C. Y. Chen, B. Neale, and R. Do. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat. Genet.*, 50(5):693–698, 05 2018.

P. M. Visscher, W. G. Hill, and N. R. Wray. Heritability in the genomics era–concepts and misconceptions. *Nat. Rev. Genet.*, 9(4):255–266, Apr 2008.

Peter M. Visscher, Naomi R. Wray, Qian Zhang, Pamela Sklar, Mark I. McCarthy, Matthew A. Brown, and Jian Yang. 10 years of gwas discovery: Biology, function, and translation. *Am J Hum Genet*, 101(1):5–22, 2017.

B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz, and K. W. Kinzler. Cancer genome landscapes. *Science*, 339(6127):1546–1558, Mar 2013.

I. R. Watson, K. Takahashi, P. A. Futreal, and L. Chin. Emerging patterns of somatic mutations in cancer. *Nat. Rev. Genet.*, 14(10):703–718, Oct 2013.

Y. Wee, T. Wang, Y. Liu, X. Li, and M. Zhao. A pan-cancer study of copy number gain and up-regulation in human oncogenes. *Life Sci.*, 211:206–214, Oct 2018.

L. Wei, Z. Jin, S. Yang, Y. Xu, Y. Zhu, and Y. Ji. TCGA-assembler 2: software pipeline for retrieval and processing of TCGA/CPTAC data. *Bioinformatics*, 34(9):1615–1617, May 2018.

Y Wei, Y. Tenzen, and H. Ji. Joint analysis of differential gene expression in multiple studies using correlation motifs. *Biostatistics*, 16(1):31–46, 2015.

X. Wen, R. Pique-Regi, and F. Luca. Integrating molecular qtl data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. *PLoS Genet*, 13(3):e1006646, 2017. ISSN 1553-7404 (Electronic) 1553-7390 (Linking). doi: 10.1371/journal.pgen.1006646. URL `https://www.ncbi.nlm.nih.gov/pubmed/28278150`.

C. J. Willer, E. M. Schmidt, S. Sengupta, G. M. Peloso, S. Gustafsson, S. Kanoni, A. Ganna, J. Chen, M. L. Buchkovich, S. Mora, J. S. Beckmann, J. L. Bragg-Gresham, H. Y. Chang, A. Demirkan, H. M. Den Hertog, R. Do, L. A. Donnelly, G. B. Ehret, T. Esko, M. F. Feitosa, T. Ferreira, K. Fischer, P. Fontanillas, R. M. Fraser, D. F. Freitag, D. Gurdasani, K. Heikkila, E. Hypponen, A. Isaacs, A. U. Jackson, A. Johansson, T. Johnson, M. Kaakinen, J. Kettunen, M. E. Kleber, X. Li, J. Luan, L. P. Lyytikainen, P. K. E. Magnusson, M. Mangino, E. Mihailov, M. E. Montasser, M. Muller-Nurasyid, I. M. Nolte, J. R. O'Connell, C. D. Palmer, M. Perola, A. K. Petersen, S. Sanna, R. Saxena, S. K. Service, S. Shah, D. Shungin, C. Sidore, C. Song, R. J. Strawbridge, I. Surakka, T. Tanaka, T. M.

Teslovich, G. Thorleifsson, E. G. Van den Herik, B. F. Voight, K. A. Volcik, L. L. Waite, A. Wong, Y. Wu, W. Zhang, D. Absher, G. Asiki, I. Barroso, L. F. Been, J. L. Bolton, L. L. Bonnycastle, P. Brambilla, M. S. Burnett, G. Cesana, M. Dimitriou, A. S. F. Doney, A. Doring, P. Elliott, S. E. Epstein, G. Ingi Eyjolfsson, B. Gigante, M. O. Goodarzi, H. Grallert, M. L. Gravito, C. J. Groves, G. Hallmans, A. L. Hartikainen, C. Hayward, D. Hernandez, A. A. Hicks, H. Holm, Y. J. Hung, T. Illig, M. R. Jones, P. Kaleebu, J. J. P. Kastelein, K. T. Khaw, E. Kim, N. Klopp, P. Komulainen, M. Kumari, C. Langenberg, T. Lehtimaki, S. Y. Lin, J. Lindstrom, R. J. F. Loos, F. Mach, W. L. McArdle, C. Meisinger, B. D. Mitchell, G. Muller, R. Nagaraja, N. Narisu, T. V. M. Nieminen, R. N. Nsubuga, I. Olafsson, K. K. Ong, A. Palotie, T. Papamarkou, C. Pomilla, A. Pouta, D. J. Rader, M. P. Reilly, P. M. Ridker, F. Rivadeneira, I. Rudan, A. Ruokonen, N. Samani, H. Scharnagl, J. Seeley, K. Silander, A. Stančaková, K. Stirrups, A. J. Swift, L. Tiret, A. G. Uitterlinden, L. J. van Pelt, S. Vedantam, N. Wainwright, C. Wijmenga, S. H. Wild, G. Willemsen, T. Wilsgaard, J. F. Wilson, E. H. Young, J. H. Zhao, L. S. Adair, D. Arveiler, T. L. Assimes, S. Bandinelli, F. Bennett, M. Bochud, B. O. Boehm, D. I. Boomsma, I. B. Borecki, S. R. Bornstein, P. Bovet, M. Burnier, H. Campbell, A. Chakravarti, J. C. Chambers, Y. I. Chen, F. S. Collins, R. S. Cooper, J. Danesh, G. Dedoussis, U. de Faire, A. B. Feranil, J. Ferrieres, L. Ferrucci, N. B. Freimer, C. Gieger, L. C. Groop, V. Gudnason, U. Gyllensten, A. Hamsten, T. B. Harris, A. Hingorani, J. N. Hirschhorn, A. Hofman, G. K. Hovingh, C. A. Hsiung, S. E. Humphries, S. C. Hunt, K. Hveem, C. Iribarren, M. R. Jarvelin, A. Jula, M. Kahonen, J. Kaprio, A. Kesaniemi, M. Kivimaki, J. S. Kooner, P. J. Koudstaal, R. M. Krauss, D. Kuh, J. Kuusisto, K. O. Kyvik, M. Laakso, T. A. Lakka, L. Lind, C. M. Lindgren, N. G. Martin, W. Marz, M. I. McCarthy, C. A. McKenzie, P. Meneton, A. Metspalu, L. Moilanen, A. D. Morris, P. B. Munroe, I. Njølstad, N. L. Pedersen, C. Power, P. P. Pramstaller, J. F. Price, B. M. Psaty, T. Quertermous, R. Rauramaa, D. Saleheen, V. Salomaa, D. K. Sanghera, J. Saramies, P. E. H. Schwarz, W. H. Sheu, A. R. Shuldiner, A. Siegbahn, T. D. Spector, K. Stefansson, D. P. Strachan, B. O. Tayo, E. Tremoli, J. Tuomilehto, M. Uusitupa, C. M. van Duijn, P. Vollenweider, L. Wallentin, N. J. Wareham, J. B. Whitfield, B. H. R. Wolffenbuttel, J. M. Ordovas, E. Boerwinkle, C. N. A. Palmer, U. Thorsteinsdottir, D. I. Chasman, J. I. Rotter, P. W. Franks, S. Ripatti, L. A. Cupples, M. S. Sandhu, S. S. Rich, M. Boehnke, P. Deloukas, S. Kathiresan, K. L. Mohlke, E. Ingelsson, and G. R. Abecasis. Discovery and refinement of loci associated with lipid levels. *Nat. Genet.*, 45(11):1274–1283, Nov 2013.

F. Windmeijer, H. Farbmacher, N. Davies, and G. Davey Smith. On the Use of the Lasso for Instrumental Variables Estimation with Some Invalid Instruments. *J Am Stat Assoc*, 114(527):1339–1350, 2019.

A. R. Wood, T. Esko, J. Yang, S. Vedantam, T. H. Pers, S. Gustafsson, A. Y. Chu, K. Estrada, J. Luan, Z. Kutalik, N. Amin, M. L. Buchkovich, D. C. Croteau-Chonka, F. R. Day, Y. Duan, T. Fall, R. Fehrmann, T. Ferreira, A. U. Jackson, J. Karjalainen, K. S. Lo, A. E. Locke, R. Magi, E. Mihailov, E. Porcu, J. C. Randall, A. Scherag, A. A. Vinkhuyzen, H. J. Westra, T. W. Winkler, T. Workalemahu, J. H. Zhao, D. Absher, E. Albrecht, D. Anderson, J. Baron, M. Beekman, A. Demirkan, G. B. Ehret, B. Feenstra, M. F.

Feitosa, K. Fischer, R. M. Fraser, A. Goel, J. Gong, A. E. Justice, S. Kanoni, M. E. Kleber, K. Kristiansson, U. Lim, V. Lotay, J. C. Lui, M. Mangino, I. Mateo Leach, C. Medina-Gomez, M. A. Nalls, D. R. Nyholt, C. D. Palmer, D. Pasko, S. Pechlivanis, I. Prokopenko, J. S. Ried, S. Ripke, D. Shungin, A. Stancakova, R. J. Strawbridge, Y. J. Sung, T. Tanaka, A. Teumer, S. Trompet, S. W. van der Laan, J. van Setten, J. V. Van Vliet-Ostaptchouk, Z. Wang, L. Yengo, W. Zhang, U. Afzal, J. Arnlov, G. M. Arscott, S. Bandinelli, A. Barrett, C. Bellis, A. J. Bennett, C. Berne, M. Bluher, J. L. Bolton, Y. Bottcher, H. A. Boyd, M. Bruinenberg, B. M. Buckley, S. Buyske, I. H. Caspersen, P. S. Chines, R. Clarke, S. Claudi-Boehm, M. Cooper, E. W. Daw, P. A. De Jong, J. Deelen, G. Delgado, J. C. Denny, R. Dhonukshe-Rutten, M. Dimitriou, A. S. Doney, M. Dorr, N. Eklund, E. Eury, L. Folkersen, M. E. Garcia, F. Geller, V. Giedraitis, A. S. Go, H. Grallert, T. B. Grammer, J. Grassler, H. Gronberg, L. C. de Groot, C. J. Groves, J. Haessler, P. Hall, T. Haller, G. Hallmans, A. Hannemann, C. A. Hartman, M. Hassinen, C. Hayward, N. L. Heard-Costa, Q. Helmer, G. Hemani, A. K. Henders, H. L. Hillege, M. A. Hlatky, W. Hoffmann, P. Hoffmann, O. Holmen, J. J. Houwing-Duistermaat, T. Illig, A. Isaacs, A. L. James, J. Jeff, B. Johansen, A. Johansson, J. Jolley, T. Juliusdottir, J. Junttila, A. N. Kho, L. Kinnunen, N. Klopp, T. Kocher, W. Kratzer, P. Lichtner, L. Lind, J. Lindstrom, S. Lobbens, M. Lorentzon, Y. Lu, V. Lyssenko, P. K. Magnusson, A. Mahajan, M. Maillard, W. L. McArdle, C. A. McKenzie, S. McLachlan, P. J. McLaren, C. Menni, S. Merger, L. Milani, A. Moayyeri, K. L. Monda, M. A. Morken, G. Muller, M. Muller-Nurasyid, A. W. Musk, N. Narisu, M. Nauck, I. M. Nolte, M. M. Nothen, L. Oozageer, S. Pilz, N. W. Rayner, F. Renstrom, N. R. Robertson, L. M. Rose, R. Roussel, S. Sanna, H. Scharnagl, S. Scholtens, F. R. Schumacher, H. Schunkert, R. A. Scott, J. Sehmi, T. Seufferlein, J. Shi, K. Silventoinen, J. H. Smit, A. V. Smith, J. Smolonska, A. V. Stanton, K. Stirrups, D. J. Stott, H. M. Stringham, J. Sundstrom, M. A. Swertz, A. C. Syvanen, B. O. Tayo, G. Thorleifsson, J. P. Tyrer, S. van Dijk, N. M. van Schoor, N. van der Velde, D. van Heemst, F. V. van Oort, S. H. Vermeulen, N. Verweij, J. M. Vonk, L. L. Waite, M. Waldenberger, R. Wennauer, L. R. Wilkens, C. Willenborg, T. Wilsgaard, M. K. Wojczynski, A. Wong, A. F. Wright, Q. Zhang, D. Arveiler, S. J. Bakker, J. Beilby, R. N. Bergman, S. Bergmann, R. Biffar, J. Blangero, D. I. Boomsma, S. R. Bornstein, P. Bovet, P. Brambilla, M. J. Brown, H. Campbell, M. J. Caulfield, A. Chakravarti, R. Collins, F. S. Collins, D. C. Crawford, L. A. Cupples, J. Danesh, U. de Faire, H. M. den Ruijter, R. Erbel, J. Erdmann, J. G. Eriksson, M. Farrall, E. Ferrannini, J. Ferrieres, I. Ford, N. G. Forouhi, T. Forrester, R. T. Gansevoort, P. V. Gejman, C. Gieger, A. Golay, O. Gottesman, V. Gudnason, U. Gyllensten, D. W. Haas, A. S. Hall, T. B. Harris, A. T. Hattersley, A. C. Heath, C. Hengstenberg, A. A. Hicks, L. A. Hindorff, A. D. Hingorani, A. Hofman, G. K. Hovingh, S. E. Humphries, S. C. Hunt, E. Hypponen, K. B. Jacobs, M. R. Jarvelin, P. Jousilahti, A. M. Jula, J. Kaprio, J. J. Kastelein, M. Kayser, F. Kee, S. M. Keinanen-Kiukaanniemi, L. A. Kiemeney, J. S. Kooner, C. Kooperberg, S. Koskinen, P. Kovacs, A. T. Kraja, M. Kumari, J. Kuusisto, T. A. Lakka, C. Langenberg, L. Le Marchand, T. Lehtimaki, S. Lupoli, P. A. Madden, S. Mannisto, P. Manunta, A. Marette, T. C. Matise, B. McKnight, T. Meitinger, F. L. Moll, G. W. Montgomery, A. D. Morris, A. P. Morris, J. C. Murray, M. Nelis, C. Ohlsson, A. J. Oldehinkel, K. K. Ong, W. H. Ouwehand, G. Pasterkamp, A. Peters, P. P. Pramstaller, J. F. Price, L. Qi, O. T. Raitakari, T. Rankinen, D. C. Rao, T. K. Rice, M. Ritchie, I. Rudan, V. Salomaa, N. J. Samani, J. Saramies, M. A. Sarzyn-

ski, P. E. Schwarz, S. Sebert, P. Sever, A. R. Shuldiner, J. Sinisalo, V. Steinthorsdottir, R. P. Stolk, J. C. Tardif, A. Tonjes, A. Tremblay, E. Tremoli, J. Virtamo, M. C. Vohl, P. Amouyel, F. W. Asselbergs, T. L. Assimes, M. Bochud, B. O. Boehm, E. Boerwinkle, E. P. Bottinger, C. Bouchard, S. Cauchi, J. C. Chambers, S. J. Chanock, R. S. Cooper, P. I. de Bakker, G. Dedoussis, L. Ferrucci, P. W. Franks, P. Froguel, L. C. Groop, C. A. Haiman, A. Hamsten, M. G. Hayes, J. Hui, D. J. Hunter, K. Hveem, J. W. Jukema, R. C. Kaplan, M. Kivimaki, D. Kuh, M. Laakso, Y. Liu, N. G. Martin, W. Marz, M. Melbye, S. Moebus, P. B. Munroe, I. Njølstad, B. A. Oostra, C. N. Palmer, N. L. Pedersen, M. Perola, L. Perusse, U. Peters, J. E. Powell, C. Power, T. Quertermous, R. Rauramaa, E. Reinmaa, P. M. Ridker, F. Rivadeneira, J. I. Rotter, T. E. Saaristo, D. Saleheen, D. Schlessinger, P. E. Slagboom, H. Snieder, T. D. Spector, K. Strauch, M. Stumvoll, J. Tuomilehto, M. Uusitupa, P. van der Harst, H. Volzke, M. Walker, N. J. Wareham, H. Watkins, H. E. Wichmann, J. F. Wilson, P. Zanen, P. Deloukas, I. M. Heid, C. M. Lindgren, K. L. Mohlke, E. K. Speliotes, U. Thorsteinsdottir, I. Barroso, C. S. Fox, K. E. North, D. P. Strachan, J. S. Beckmann, S. I. Berndt, M. Boehnke, I. B. Borecki, M. I. McCarthy, A. Metspalu, K. Stefansson, A. G. Uitterlinden, C. M. van Duijn, L. Franke, C. J. Willer, A. L. Price, G. Lettre, R. J. Loos, M. N. Weedon, E. Ingelsson, J. R. O'Connell, G. R. Abecasis, D. I. Chasman, M. E. Goddard, P. M. Visscher, J. N. Hirschhorn, T. M. Frayling, C. A. McCarty, J. Starren, P. Peissig, R. Berg, L. Rasmussen, J. Linneman, A. Miller, V. Choudary, L. Chen, C. Waudby, T. Kitchner, J. Reeser, N. Fost, M. Ritchie, R. A. Wilke, R. L. Chisholm, P. C. Avila, P. Greenland, M. Hayes, A. Kho, W. A. Kibbe, A. A. Lemke, W. L. Lowe, M. E. Smith, W. A. Wolf, J. A. Pacheco, W. K. Thompson, J. Humowiecki, M. Law, C. Chute, I. Kullo, B. Koenig, M. de Andrade, S. Bielinski, J. Pathak, G. Savova, J. Wu, J. Henriksen, K. Ding, L. Hart, J. Palbicki, E. B. Larson, K. Newton, E. Ludman, L. Spangler, G. Hart, D. Carrell, G. Jarvik, P. Crane, W. Burke, S. M. Fullerton, S. B. Trinidad, C. Carlson, F. Hutchinson, A. McDavid, D. M. Roden, E. Clayton, J. L. Haines, D. R. Masys, L. R. Churchill, D. Cornfield, D. Crawford, D. Darbar, J. C. Denny, B. A. Malin, M. D. Ritchie, J. S. Schildcrout, H. Xu, A. H. Ramirez, M. Basford, J. Pulley, B. Alizadeh, R. A. de Boer, H. M. Boezen, M. Bruinenberg, L. Franke, P. van der Harst, H. L. Hillege, M. M. van der Klauw, G. Navis, J. Ormel, D. S. Postma, J. G. Rosmalen, J. P. Slaets, H. Snieder, R. P. Stolk, B. H. Wolffenbuttel, C. Wijmenga, S. Kathiresan, B. F. Voight, S. Purcell, K. Musunuru, D. Ardissino, P. M. Mannucci, S. Anand, J. C. Engert, N. J. Samani, H. Schunkert, J. Erdmann, M. P. Reilly, D. J. Rader, T. Morgan, J. A. Spertus, M. Stoll, D. Girelli, P. P. McKeown, C. C. Patterson, D. S. Siscovick, C. J. O'Donnell, R. Elosua, L. Peltonen, V. Salomaa, S. M. Schwartz, O. Melander, D. Altshuler, D. Ardissino, P. A. Merlini, C. Berzuini, L. Bernardinelli, F. Peyvandi, M. Tubaro, P. Celli, M. Ferrario, R. Fetiveau, N. Marziliano, G. Casari, M. Galli, F. Ribichini, M. Rossi, F. Bernardi, P. Zonzin, A. Piazza, P. M. Mannucci, S. M. Schwartz, D. S. Siscovick, J. Yee, Y. Friedlander, R. Elosua, J. Marrugat, G. Lucas, I. Subirana, J. Sala, R. Ramos, S. Kathiresan, J. B. Meigs, G. Williams, D. M. Nathan, C. A. MacRae, C. J. O'Donnell, V. Salomaa, A. S. Havulinna, L. Peltonen, O. Melander, G. Berglund, B. Voight, S. Kathiresan, J. N. Hirschhorn, R. Asselta, S. Duga, M. Spreafico, K. Musunuru, M. J. Daly, S. Purcell, B. F. Voight, S. Purcell, J. Nemesh, J. M. Korn, S. A. McCarroll, S. M. Schwartz, J. Yee, S. Kathiresan, G. Lucas, I. Subirana, R. Elosua, A. Surti, C. Guiducci, L. Gianniny, D. Mirel, M. Parkin, N. Burtt, S. B. Gabriel,

N. J. Samani, J. R. Thompson, P. S. Braund, B. J. Wright, A. J. Balmforth, S. G. Ball, A. S. Hall, I. Schunkert, J. Erdmann, P. Linsel-Nitschke, W. Lieb, A. Ziegler, I. R. Konig, C. Hengstenberg, M. Fischer, K. Stark, A. Grosshennig, M. Preuss, H. E. Wichmann, S. Schreiber, H. Schunkert, N. J. Samani, J. Erdmann, W. Ouwehand, C. Hengstenberg, P. Deloukas, M. Scholz, F. Cambien, A. Goodall, M. P. Reilly, M. Li, Z. Chen, R. Wilensky, W. Matthai, A. Qasim, H. H. Hakonarson, J. Devaney, M. S. Burnett, A. D. Pichard, K. M. Kent, L. Satler, J. M. Lindsay, R. Waksman, C. W. Knouff, D. M. Waterworth, M. C. Walker, V. Mooser, S. E. Epstein, D. J. Rader, T. Scheffold, K. Berger, M. Stoll, A. Huge, D. Girelli, N. Martinelli, O. Olivieri, R. Corrocher, T. Morgan, J. A. Spertus, P. P. McKeown, C. C. Patterson, H. Schunkert, J. Erdmann, P. Linsel-Nitschke, W. Lieb, A. Ziegler, I. Konig, C. Hengstenberg, M. Fischer, K. Stark, A. Grosshennig, M. Preuss, H. E. Wichmann, S. Schreiber, H. Holm, G. Thorleifsson, U. Thorsteinsdottir, K. Stefansson, J. C. Engert, R. Do, C. Xie, S. Anand, S. Kathiresan, D. Ardissino, P. M. Mannucci, D. Siscovick, C. J. O'Donnell, N. J. Samani, O. Melander, R. Elosua, L. Peltonen, V. Salomaa, S. M. Schwartz, D. Altshuler, T. Matise, S. Buyske, J. Higashio, R. Williams, A. Nato, J. L. Ambite, E. Deelman, T. Manolio, L. Hindorff, K. E. North, G. Heiss, K. Taylor, N. Franceschini, C. Avery, M. Graff, D. Lin, M. Quibrera, B. Cochran, L. Kao, J. Umans, S. Cole, J. MacCluer, S. Person, J. Pankow, M. Gross, E. Boerwinkle, M. Fornage, P. Durda, N. Jenny, B. Patsy, A. Arnold, P. Buzkova, D. Crawford, J. Haines, D. Murdock, K. Glenn, K. Brown-Gentry, T. Thornton-Wells, L. Dumitrescu, J. Jeff, W. S. Bush, S. L. Mitchell, R. Goodloe, S. Wilson, J. Boston, J. Malinowski, N. Restrepo, M. Oetjens, J. Fowke, W. Zheng, K. Spencer, M. Ritchie, S. Pendergrass, L. Le Marchand, L. Wilkens, L. Park, M. Tiirikainen, L. Kolonel, U. Lim, I. Cheng, H. Wang, R. Shohet, C. Haiman, D. Stram, B. Henderson, K. Monroe, F. Schumacher, C. Kooperberg, U. Peters, G. Anderson, C. Carlson, R. Prentice, A. LaCroix, C. Wu, C. Carty, J. Gong, S. Rosse, A. Young, J. Haessler, J. Kocarnik, Y. Lin, R. Jackson, D. Duggan, and L. Kuller. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.*, 46(11):1173–1186, Nov 2014.

X. Wu, Y. X. Zhuang, C. Q. Hong, J. Y. Chen, Y. J. You, F. Zhang, P. Huang, and M. Y. Wu. Clinical importance and therapeutic implication of E-cadherin gene methylation in human ovarian cancer. *Med. Oncol.*, 31(8):100, Aug 2014.

Y. H. Wu, R. E. Graff, M. N. Passarelli, J. D. Hoffman, E. Ziv, T. J. Hoffmann, and J. S. Witte. Identification of Pleiotropic Cancer Susceptibility Variants from Genome-Wide Association Studies Reveals Functional Characteristics. *Cancer Epidemiol. Biomarkers Prev.*, 27(1):75–85, Jan 2018.

Z. Xu, Q. Duan, S. Yan, W. Chen, M. Li, E. Lange, and Y. Li. DISSCO: direct imputation of summary statistics allowing covariates. *Bioinformatics*, 31(15):2434–2442, Aug 2015.

F. Yang, J. Wang, B. L. Pierce, L. S. Chen, F. Aguet, K. G. Ardlie, B. B. Cummings, E. T. Gelfand, G. Getz, K. Hadley, R. E. Handsaker, K. H. Huang, S. Kashin, K. J. Karczewski, M. Lek, X. Li, D. G. MacArthur, J. L. Nedzel, D. T. Nguyen, M. S. Noble, A. V. Segrè, C. A. Trowbridge, T. Tukiainen, N. S. Abell, B. Balliu, R. Barshir, O. Basha, A. Battle, G. K. Bogu, A. Brown, C. D. Brown, S. E. Castel, L. S. Chen, C. Chiang,

D. F. Conrad, N. J. Cox, F. N. Damani, J. R. Davis, O. Delaneau, E. T. Dermitzakis, B. E. Engelhardt, E. Eskin, P. G. Ferreira, L. Frésard, E. R. Gamazon, D. Garrido-Martìn, A. D. H. Gewirtz, G. Gliner, M. J. Gloudemans, R. Guigo, I. M. Hall, B. Han, Y. He, F. Hormozdiari, C. Howald, H. K. Im, B. Jo, E. Y. Kang, Y. Kim, S. Kim-Hellmuth, T. Lappalainen, G. Li, X. Li, B. Liu, S. Mangul, M. I. McCarthy, I. C. McDowell, P. Mohammadi, J. Monlong, S. B. Montgomery, M. Muñoz-Aguirre, A. W. Ndungu, D. L. Nicolae, A. B. Nobel, M. Oliva, H. Ongen, J. J. Palowitch, N. Panousis, P. Papasaikas, Y. Park, P. Parsana, A. J. Payne, C. B. Peterson, J. Quan, F. Reverter, C. Sabatti, A. Saha, M. Sammeth, A. J. Scott, A. A. Shabalin, R. Sodaei, M. Stephens, B. E. Stranger, B. J. Strober, J. H. Sul, E. K. Tsang, S. Urbut, M. van de Bunt, G. Wang, X. Wen, F. A. Wright, H. S. Xi, E. Yeger-Lotem, Z. Zappala, J. B. Zaugg, Y. H. Zhou, J. M. Akey, D. Bates, J. Chan, L. S. Chen, M. Claussnitzer, K. Demanelis, M. Diegel, J. A. Doherty, A. P. Feinberg, M. S. Fernando, J. Halow, K. D. Hansen, E. Haugen, P. F. Hickey, L. Hou, F. Jasmine, R. Jian, L. Jiang, A. Johnson, R. Kaul, M. Kellis, M. G. Kibriya, K. Lee, J. B. Li, Q. Li, X. Li, J. Lin, S. Lin, S. Linder, C. Linke, Y. Liu, M. T. Maurano, B. Molinie, S. B. Montgomery, J. Nelson, F. J. Neri, M. Oliva, Y. Park, B. L. Pierce, N. J. Rinaldi, L. F. Rizzardi, R. Sandstrom, A. Skol, K. S. Smith, M. P. Snyder, J. Stamatoyannopoulos, B. E. Stranger, H. Tang, E. K. Tsang, L. Wang, M. Wang, N. Van Wittenberghe, F. Wu, R. Zhang, C. R. Nierras, P. A. Branton, L. J. Carithers, P. Guan, H. M. Moore, A. Rao, J. B. Vaught, S. E. Gould, N. C. Lockart, C. Martin, J. P. Struewing, S. Volpi, A. M. Addington, S. E. Koester, A. R. Little, L. E. Brigham, R. Hasz, M. Hunter, C. Johns, M. Johnson, G. Kopen, W. F. Leinweber, J. T. Lonsdale, A. McDonald, B. Mestichelli, K. Myer, B. Roe, M. Salvatore, S. Shad, J. A. Thomas, G. Walters, M. Washington, J. Wheeler, J. Bridge, B. A. Foster, B. M. Gillard, E. Karasik, R. Kumar, M. Miklos, M. T. Moser, S. D. Jewell, R. G. Montroy, D. C. Rohrer, D. R. Valley, D. A. Davis, D. C. Mash, A. H. Undale, A. M. Smith, D. E. Tabor, N. V. Roche, J. A. McLean, N. Vatanian, K. L. Robinson, L. Sobin, M. E. Barcus, K. M. Valentino, L. Qi, S. Hunter, P. Hariharan, S. Singh, K. S. Um, T. Matose, M. M. Tomaszewski, L. K. Barker, M. Mosavel, L. A. Siminoff, H. M. Traino, P. Flicek, T. Juettemann, M. Ruffier, D. Sheppard, K. Taylor, S. J. Trevanion, D. R. Zerbino, B. Craft, M. Goldman, M. Haeussler, W. J. Kent, C. M. Lee, B. Paten, K. R. Rosenbloom, J. Vivian, and J. Zhu. Identifying cis-mediators for trans-eQTLs across many human tissues using genomic mediation analysis. *Genome Res.*, 27(11):1859–1871, 11 2017.

J. Yang, T. Ferreira, A. P. Morris, S. E. Medland, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. N. Weedon, R. J. Loos, T. M. Frayling, M. I. McCarthy, J. N. Hirschhorn, M. E. Goddard, and P. M. Visscher. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.*, 44(4):369–375, Mar 2012.

C. Yao, R. Joehanes, A. D. Johnson, T. Huan, T. Esko, S. Ying, J. E. Freedman, J. Murabito, K. L. Lunetta, A. Metspalu, P. J. Munson, and D. Levy. Sex- and age-interacting eQTLs in human complex diseases. *Hum. Mol. Genet.*, 23(7):1947–1956, Apr 2014.

T. Z. Yi, J. Guo, L. Zhou, X. Chen, R. R. Mi, Q. X. Qu, J. H. Zheng, and L. Zhai.

Prognostic value of E-cadherin expression and CDH1 promoter methylation in patients with endometrial carcinoma. *Cancer Invest.*, 29(1):86–92, Jan 2011.

W. Yin, J. Chen, G. Wang, and D. Zhang. MicroRNA-106b functions as an oncogene and regulates tumor viability and metastasis by targeting LARP4B in prostate cancer. *Mol Med Rep*, 20(2):951–958, Aug 2019.

T. I. Zack, S. E. Schumacher, S. L. Carter, A. D. Cherniack, G. Saksena, B. Tabak, M. S. Lawrence, C. Z. Zhsng, J. Wala, C. H. Mermel, C. Sougnez, S. B. Gabriel, B. Hernandez, H. Shen, P. W. Laird, G. Getz, M. Meyerson, and R. Beroukhim. Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.*, 45(10):1134–1140, Oct 2013.

G. Zeng, J. Wang, Y. Huang, Y. Lian, D. Chen, H. Wei, C. Lin, and Y. Huang. Overexpressing CCT6A Contributes To Cancer Cell Growth By Affecting The G1-To-S Phase Transition And Predicts A Negative Prognosis In Hepatocellular Carcinoma. *Onco Targets Ther*, 12:10427–10439, 2019.

L. Zeng, A. Morinibu, M. Kobayashi, Y. Zhu, X. Wang, Y. Goto, C. J. Yeom, T. Zhao, K. Hirota, K. Shinomiya, S. Itasaka, M. Yoshimura, G. Guo, E. M. Hammond, M. Hiraoka, and H. Harada. Aberrant IDH3α expression promotes malignant tumor growth by inducing HIF-1-mediated metabolic reprogramming and angiogenesis. *Oncogene*, 34(36):4758–4766, Sep 2015.

B. Zhang, J. Wang, X. Wang, J. Zhu, Q. Liu, Z. Shi, M. C. Chambers, L. J. Zimmerman, K. F. Shaddox, S. Kim, S. R. Davies, S. Wang, P. Wang, C. R. Kinsinger, R. C. Rivers, H. Rodriguez, R. R. Townsend, M. J. Ellis, S. A. Carr, D. L. Tabb, R. J. Coffey, R. J. Slebos, D. C. Liebler, S. A. Carr, M. A. Gillette, K. R. Klauser, E. Kuhn, D. R. Mani, P. Mertins, K. A. Ketchum, A. G. Paulovich, J. R. Whiteaker, N. J. Edwards, P. B. McGarvey, S. Madhavan, P. Wang, D. Chan, A. Pandey, I. e. M. Shih, H. Zhang, Z. Zhang, H. Zhu, G. A. Whiteley, S. J. Skates, F. M. White, D. A. Levine, E. S. Boja, C. R. Kinsinger, T. Hiltke, M. Mesri, R. C. Rivers, H. Rodriguez, K. M. Shaw, S. E. Stein, D. Fenyo, T. Liu, J. E. McDermott, S. H. Payne, K. D. Rodland, R. D. Smith, P. Rudnick, M. Snyder, Y. Zhao, X. Chen, D. F. Ransohoff, A. N. Hoofnagle, D. C. Liebler, M. E. Sanders, Z. Shi, R. J. Slebos, D. L. Tabb, B. Zhang, L. J. Zimmerman, Y. Wang, S. R. Davies, L. Ding, M. J. Ellis, and R. R. Townsend. Proteogenomic characterization of human colon and rectal cancer. *Nature*, 513(7518):382–387, Sep 2014.

H. Zhang, T. Liu, Z. Zhang, S. H. Payne, B. Zhang, J. E. McDermott, J. Y. Zhou, V. A. Petyuk, L. Chen, D. Ray, S. Sun, F. Yang, L. Chen, J. Wang, P. Shah, S. W. Cha, P. Aiyetan, S. Woo, Y. Tian, M. A. Gritsenko, T. R. Clauss, C. Choi, M. E. Monroe, S. Thomas, S. Nie, C. Wu, R. J. Moore, K. H. Yu, D. L. Tabb, D. Fenyo, V. Bafna, Y. Wang, H. Rodriguez, E. S. Boja, T. Hiltke, R. C. Rivers, L. Sokoll, H. Zhu, I. M. Shih, L. Cope, A. Pandey, B. Zhang, M. P. Snyder, D. A. Levine, R. D. Smith, D. W. Chan, K. D. Rodland, S. A. Carr, M. A. Gillette, K. R. Klauser, E. Kuhn, D. R. Mani, P. Mertins, K. A. Ketchum, R. Thangudu, S. Cai, M. Oberti, A. G. Paulovich, J. R. Whiteaker, N. J. Edwards, P. B. McGarvey, S. Madhavan, P. Wang, D. W. Chan, A. Pandey, I. M. Shih, H. Zhang, Z. Zhang, H. Zhu, L. Cope, G. A. Whiteley, S. J. Skates, F. M. White, D. A.

Levine, E. S. Boja, C. R. Kinsinger, T. Hiltke, M. Mesri, R. C. Rivers, H. Rodriguez, K. M. Shaw, S. E. Stein, D. Fenyo, T. Liu, J. E. McDermott, S. H. Payne, K. D. Rodland, R. D. Smith, P. Rudnick, M. Snyder, Y. Zhao, X. Chen, D. F. Ransohoff, A. N. Hoofnagle, D. C. Liebler, M. E. Sanders, Z. Shi, R. J. C. Slebos, D. L. Tabb, B. Zhang, L. J. Zimmerman, Y. Wang, S. R. Davies, L. Ding, M. J. C. Ellis, and R. R. Townsend. Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer. *Cell*, 166 (3):755–765, Jul 2016a.

J. M. Zhang and J. An. Cytokines, inflammation, and pain. *Int Anesthesiol Clin*, 45(2): 27–37, 2007.

Y. Zhang, Y. Wang, Y. Wei, J. Wu, P. Zhang, S. Shen, H. Saiyin, R. Wumaier, X. Yang, C. Wang, and L. Yu. Molecular chaperone CCT3 supports proper mitotic progression and cell proliferation in hepatocellular carcinoma cells. *Cancer Lett.*, 372(1):101–109, Mar 2016b.

J. Zhao, J. Ming, X. Hu, G. Chen, J. Liu, and C. Yang. Bayesian weighted Mendelian randomization for causal inference based on summary statistics. *Bioinformatics*, 36(5): 1501–1508, Mar 2020.

M. Zhao and Z. Zhao. Concordance of copy number loss and down-regulation of tumor suppressor genes: a pan-cancer study. *BMC Genomics*, 17 Suppl 7:532, 08 2016.

Qingyuan Zhao, Jingshu Wang, Gibran Hemani, Jack Bowden, and Dylan S. Small. Statistical inference in two-sample summary-data mendelian randomization using robust adjusted profile score. *arXiv*, doi: 1801.09652, 2019.

D. V. Zhernakova, P. Deelen, M. Vermaat, M. van Iterson, M. van Galen, W. Arindrarto, P. van 't Hof, H. Mei, F. van Dijk, H. J. Westra, M. J. Bonder, J. van Rooij, M. Verkerk, P. M. Jhamai, M. Moed, S. M. Kielbasa, J. Bot, I. Nooren, R. Pool, J. van Dongen, J. J. Hottenga, C. D. Stehouwer, C. J. van der Kallen, C. G. Schalkwijk, A. Zhernakova, Y. Li, E. F. Tigchelaar, N. de Klein, M. Beekman, J. Deelen, D. van Heemst, L. H. van den Berg, A. Hofman, A. G. Uitterlinden, M. M. van Greevenbroek, J. H. Veldink, D. I. Boomsma, C. M. van Duijn, C. Wijmenga, P. E. Slagboom, M. A. Swertz, A. Isaacs, J. B. van Meurs, R. Jansen, B. T. Heijmans, P. A. 't Hoen, and L. Franke. Identification of context-dependent expression quantitative trait loci in whole blood. *Nat. Genet.*, 49(1): 139–145, 01 2017.

Z. Zhu, F. Zhang, H. Hu, A. Bakshi, M. R. Robinson, J. E. Powell, G. W. Montgomery, M. E. Goddard, N. R. Wray, P. M. Visscher, and J. Yang. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet*, 48(5): 481–7, 2016. ISSN 1546-1718 (Electronic) 1061-4036 (Linking). doi: 10.1038/ng.3538. URL `https://www.ncbi.nlm.nih.gov/pubmed/27019110`.