Carnegie Mellon University
Dietrich College of Humanities and Social Sciences
Dissertation

Submitted in Partial Fulfillment of the Requirements
For the Degree of Doctor of Philosophy

Title: Selective inference approaches for augmenting genetic association studies with multi-omics metadata
Presented by: Ronald Yurko
Accepted by: Department of Statistics & Data Science
Readers:

KATHRYN ROEDER, ADVISOR	DATE
MAX G'SELL, ADVISOR	DATE
BERNIE DEVLIN	DATE
AADITYA RAMDAS	DATE
VALÉRIE VENTURA	DATE

Approved by the Committee on Graduate Degrees:

RICHARD SCHEINES, DEAN	DATE

Selective inference approaches for augmenting genetic association studies with multi-omics metadata

Ronald Yurko

April 1, 2022

A dissertation submitted in partial fulfillment
of the requirements for the Degree of Doctor of Philosophy

Department of Statistics & Data Science
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA 15213

Thesis Committee:
Kathryn Roeder, Chair
Max G'Sell, Chair
Bernie Devlin
Aaditya Ramdas
Valérie Ventura

Abstract

To correct for a large number of hypothesis tests, most researchers rely on simple multiple testing corrections. Yet, new selective inference methodologies could improve power by enabling exploration of test statistics with meta-data for informative weights while retaining desired statistical guarantees. My thesis revolves around this theme by developing statistical and computational tools to address the challenges especially arising from studying complex, neuropsychiatric disorders. In chapter 2 we explore one such framework, adaptive p-value thresholding (AdaPT), in the context of testing individual single nucleotide polymorphisms (SNPs) for schizophrenia. We demonstrate a substantial increase in power using flexible gradient boosted trees to account for covariates constructed with GWAS statistics from genetically-correlated phenotypes, as well as measures capturing association with gene expression and coexpression subnetwork membership. In chapter 3, we address a popular approach for computing gene-level p-values that is based on an invalid approximation for the combination of two-sided test statistics. Our correction ensures error rate control and alleviates null distribution concerns necessary for selective inference procedures. In chapter 4, we introduce an agglomerative algorithm, based on the dependence induced from linkage disquilibrium (LD), to test the aggregation of SNPs into gene-based test statistics for autism spectrum disorder (ASD). The advantages of our approaches are twofold: increased power and increased interpretability, with the latter expediting our understanding of the etiology of human diseases, disorders, and other phenotypes. Finally, in chapter 5, we demonstrate in simulations an improvement in power in the context of rare variant studies by augmenting testing corrections with annotation information and explore the use of data blurring to explore annotation structure providing ways to address the challenges of multiplicity persistent in whole genome sequencing.

Acknowledgments

It is impossible for me to express my sincere thanks and appreciation to everyone that have contributed to my personal and academic growth over the course of my PhD. I firmly believe that dating back to my time as an undergraduate at Carnegie Mellon (ten years ago!), I have been incredibly fortunate to be at the right place at the right time, surrounded by incredible people. I will likely not be able to express the proper gratitude and respect I feel towards the people below (in no particular order), but I will still try regardless:

- To Kathryn Roeder, who has been my academic mentor ever since she was willing to randomly meet with me one day to discuss potential research projects. She has always supported me through the highs and lows in my academic career, and has helped me grow into the researcher and educator I am today. Her dedication and support to all of the students she advises has created not only an intellectually engaging research group, but also fosters a fun community of people that I am proud to be a member of. I am forever grateful for the patience she has displayed with me over these years, and I am excited to continue working with and learning from her in the years to come.

- To Max G'Sell, who has been a role model for me since my time as an undergrad in his Data Mining class. As an undergrad, he helped inspire me to attend grad school. And throughout my PhD career, his enthusiasm, positive attitude, and ability to generate ideas has been constant motivation for me and is a template for the type of statistician I hope to be one day.

- To Bernie Devlin, who has effectively been a third advisor throughout my PhD. He has been a mentor and has taught me incredible amount about research in genetics and science in general. I have learned so much from our conversations and am excited to continue collaborating together in the future.

- I would also like to thank my other thesis committee members: Aaditya Ramdas and Valerie Ventura. To Aaditya, who has inspired me since his job talk with his incredible

presentations and fantastic course on reproducibility. To Valerie, who taught me how to craft effective presentations with lessons I will follow for the rest of my career.

- To everyone in the lab group, past and present: Bert, Lora, Maria, Jiebiao, Fuchen, Kevin, Minshi, Yixuan, Xuran, Tim, Yue, Jinjin and others whose names I'm forgetting to mention. The bi-weekly lab meetings and our student/postdoc reading groups have been a pleasure to attend and I have learned from so much from all of you.

- To Rebecca Nugent, who has been a role model for me as well as a friend. I remember sitting in 36-401 all those years ago and telling my friends "I think I want to be like her one day" - to which they responded "Ron, you're crazy" and they were probably right... But I am forever grateful for everything she has done for me in my career, from research opportunities as an undergrad, to letting me join the clustering group meetings when I was not a student and "working", to advising my ADA and other projects, to initiating #CMSAC, and for the mentorship she is always somehow able to find time to provide. I am beyond excited to work in a department under her leadership.

- To Peter Freeman, Ann Lee, Andrew Thomas, Cosma Shalizi, Joel Greenhouse, and Howard Seltman, whose classes and advising inspired me to pursue a PhD in Statistics.

- To Sam Ventura, who introduced me to the world of statistics in sports research and has been a great mentor and friend throughout my academic career.

- To Kostas, Lee, Francesca, Taylor, Bmac, and everyone else in the stats in sports group for the thought provoking discussions and fun collaborations.

- To everyone in my cohort, for powering through classes and all the fun times together - I'm excited to see the path you all take in your careers and I was incredibly lucky to be admitted with such an excellent group.

- To every member of the Random Walkers IM teams - thanks for the fun memories together and distractions from everyday life.

- To the SAC "leadership group", Nic, Mikaela, and Alec for the fun golf outings.

- To the CMU Statistics & Data Science department staff, especially Laura, Jess, Heidi, Margie, Beth, Danielle, Kira, and Sam for all the work they do behind the scenes simplifying my life as a PhD student. And a special thanks to Carl Skipper who had to put with I don't know how many emails I sent him with countless tedious computing questions.

- To all of the professors in the CMU Statistics & Data Science department as a whole, for their continued support, encouragement and teaching throughout the PhD years.

- To my parents, brothers, sisters-in-law, nieces, and the rest of my family and friends for their unwavering support throughout this endeavor. I am only able to complete this because of them.
- And finally I want to thank Madeline Marco Scanlon for her love, support, and ability to cheer me up throughout this entire journey. You're stuck with me, now and forever.

Contents

One

Introduction

This introductory chapter provides the background and overview necessary for understanding this thesis, including an introduction to genome-wide association studies (GWAS), commonly used multiple testing corrections, advances in statistical methodology to account for available meta-data to improve multiple testing power, and an introduction to category-wide association studies (CWAS) developed for testing rare variant associations that are now feasible to detect due to advances in technology from whole-genome sequencing (WGS). All of these topics are taken into consideration with respect to detecting associations with complex neuropsychiatric disorders, such as schizophrenia and autism spectrum disorder. Section 1.1 - The introduction to GWAS, single-nucleotide polymorphisms (SNPs), and linkage disequilibrium (LD). Section 1.2 - Commonly used multiple testing corrections for family-wise error rate (FWER) and false-discovery rate (FDR) control and approaches for gene-level testing. Section 1.3 - An introduction to new approaches for flexible multiple testing corrections that account for available meta-data in order to improve power. Section 1.4 - An introduction to rare variant analysis in the context of WGS, de novo mutations, and CWAS methodology.

1.1 Introduction to genome-wide association studies

A major goal for performing human genetics research is to detect genetic risk factors for complex diseases such as schizophrenia and autism spectrum disorder. The first three chapters of this thesis focus on the use of GWAS, which are used to identify common variants in the human population (Figure 1.1) and measure their association with some phenotype, e.g., neuropsychiatric disorders. The use of GWAS to study human genetics has continued to surge over the past fifteen years, with over 5,000 publications and their results available in the GWAS Catalog [Buniello et al., 2019].

The most abundant form of genetic variation in the human genome are single base-pair changes known as single-nucleotide polymorphisms (SNPs). SNPs are the main form of common variation studied in GWAS, with many SNPs present in a large fraction of the human population. Typically SNPs have two alleles, i.e., two common possibilities for

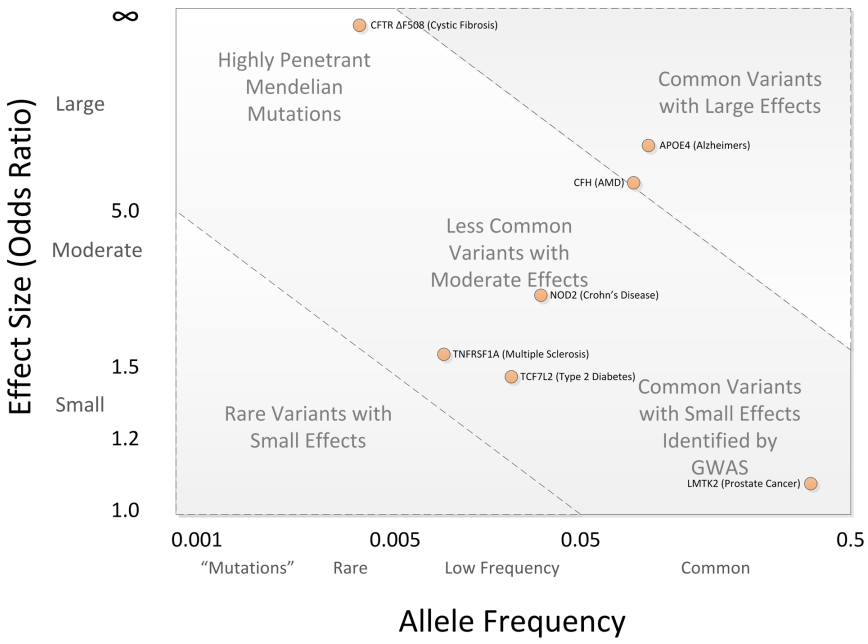

Figure 1.1: GWAS are primarily used to identify common variants with small effect sizes in the lower right portion [Bush and Moore, 2012].

a single base-pair at a specified SNP location. SNPs are often summarized in terms of the *minor allele frequency* (MAF). For example, consider a SNP with two alleles: major allele T and minor allele C. A SNP with MAF $= .35$ implies that 35% of the population has the allele C versus the 65% of the population with the more common allele T. The main type of analysis conducted in GWAS are marginal tests of association for each SNP's minor allele. Thus treating each SNP independently to measure their marginal association with some target phenotype. Throughout this thesis we focus on categorical phenotypes regarding some neuropsychiatric disorder status, relying on the use of case/control studies with measures of association typically from either chi-squared tests or logistic regression. Due to the abundance of studies that are often performed, the main results reported from GWAS are results from meta-analysis pooling information and weak signals together across studies [Willer et al., 2010]. In this common variant setting, GWAS summary statistics are necessarily two-sided tests because which SNP allele confers risk is not known *a priori*.

The presence of linkage disequilibrium (LD) creates a challenge for interpreting GWAS results. LD refers to the degree to which the alleles of two SNPs are inherited together within a population through physical proximity on a chromosome. This leads to LD-induced correlation between the resulting GWAS summary statistics that are assessed for phenotypic

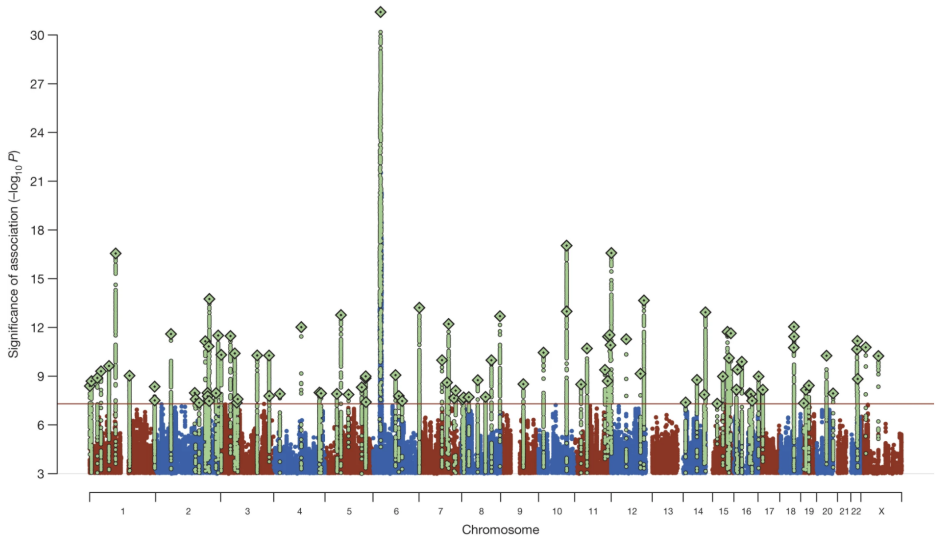

Figure 1.2: Manhattan plot displaying GWAS discoveries for SCZ at genome-wide significance threshold displayed by horizontal line [Ripke et al., 2014].

association. This leads inevitably to the detection of indirect associations, referring to significant SNP associations that are not necessarily causal mechanisms but rather in high LD with the causal SNP. This leads to the use of follow-up studies and techniques in the vein of *fine-mapping* procedures in order to identify causal variants [Schaid et al., 2018].

1.2 Multiple testing corrections for GWAS

1.2.1 Family-Wise Error Rate

Because millions of SNPs are tested in a GWAS, to overcome the multiple testing challenge and limit false positives, researchers typically use a strict multiple testing correction. The most widely accepted GWAS approach corresponds to the *genome-wide* threshold of $< 5 \times 10^{-8}$ for significance. This roughly correspond to to a Bonferroni correction [Bonferroni, 1935] for controlling the Family-Wise Error Rate (FWER), $Pr(V > 0) \leq \alpha$, where V is the number of Type I errors (i.e., false positives) and α is the target Type I error rate. This strict correction has worked well for studies of large sample sizes, e.g., human height, as well as recently obtained results for complex neuropsychiatric disorders (Figure 1.2).

1.2.2 False Discovery Rate

The conservative nature of the classic FWER control presents a challenge for detecting associations for GWAS with less informative sample sizes such as autism spectrum disorder

(ASD). Introduced by [Benjamini and Hochberg, 1995], false discovery rate (FDR) control has become a popular approach to improve power for detecting weak effects by limiting the expected false discovery proportion (FDP) instead of the more classical FWER. Rather than limit the possibility of making any Type 1 error, FDR controlling methods focus on the expected fraction of mistakes out of the total number of rejections R:

$$\text{FDR} = \mathbb{E}[\text{FDP}], \text{ where FDP} = \frac{V}{\max(R, 1)}.$$

The Benjamini-Hochberg (BH) procedure was the first method to control FDR at target level α using a step-up procedure that is *adaptive* to the set of p-values for the hypotheses of interest.

1.2.3 Gene-level testing

A natural strategy for improving GWAS power is gene-based testing: m SNPs are assigned to genes and a global null test is performed for each gene $g \in G$, i.e. all SNPs $i \in S_g$ in gene g are null versus at least one SNP in the gene is non-null,

$$H_{0,g} : H_i = 0 \ \forall i \in S_g \text{ versus } H_{1,g} : \exists i \in S_g \text{ such that } H_i = 1, \tag{1.1}$$

where $H_i = 0$ if SNP i is null, and $H_i = 1$ if non-null. This can improve power to detect weak signal by reducing the multiple testing burden and pooling signal strength, which can be advantageous for settings with weaker signal such as ASD. While there are many approaches for global testing, the presence of LD poses a challenge here: the combination of dependent SNP-level summary statistics at the gene-level must adjust for the LD-induced covariance of SNPs. While there are a number of approaches for global testing in the presence of dependence, such as harmonic means [Wilson, 2019, Tian et al., 2021] or Cauchy combinations [Liu et al., 2019, Liu and Xie, 2020], in this thesis we consider approaches featured in popular gene-level testing software *VEGAS* [Liu et al., 2010, Mishra and Macgregor, 2015] and *MAGMA* [de Leeuw et al., 2015, v1.08]. By focusing tests on genes instead of SNPs dispersed throughout the genome, gene-based testing provides increased interpretability with regards to detecting functional units of interest. Furthermore, it is common to see methods for FDR control applied in gene-level testing [Sey et al., 2020].

1.3 Augmenting multiple testing corrections with meta-data

Other methods for FDR control have led to improvements in power over BH by incorporating prior information, such as by the use of p-value weights [Genovese et al., 2006]. With the realization that multiple omics – genomics, epigenomics, proteomics, etc. – are required for describing phenotypic variation, it is natural to think that accounting for multi-omics metadata in the form of a priori hypothesis weights can improve power. However, until

recently, it was not clear how to choose these weights in an exploratory manner while maintaining valid error rate guarantees. Recent methodologies have been developed enabling the inclusion of metadata in the form of model covariates to improve power while maintaining some form of FDR control [Scott et al., 2015, Ignatiadis et al., 2016, Boca and Leek, 2018, Li and Barber, 2019, Zhang et al., 2019]. A recent review paper [Korthauer et al., 2019] covered the performance of various covariate-informed methods for FDR control, including a selective inference approach, called adaptive p-value thresholding [Lei and Fithian, 2018, AdaPT]. Unlike other considered approaches with asymptotic FDR control, and under similar assumptions, the AdaPT framework guarantees finite-simple FDR control. However, based on simulations and real datasets with one and two-dimensional covariates, [Korthauer et al., 2019] observed a poor results by AdaPT based on off-the-shelf performance with one and two-dimensional covariate examples. Their criticisms of AdaPT include that it: (1) suffers with uninformative covariates, (2) requires careful specification of functional relationships, and (3) displayed low power or failure to reject any tests in many data sets.

1.4 Introduction to rare variant analysis with category-wide association studies

While the majority of this thesis focuses on gains in power for weak, common-variant signals in the GWAS setting, we also explore selective approaches in the context of testing rare variants. Recently, the development of whole-genome sequencing (WGS) has enabled greater exploration into the impact de novo mutations (variants observed in child but not in parents) located in noncoding regions of the genome have on complex disorders. The non-coding portions covers $\approx 98.\%$ of the human genome, and includes elements regulating how protein-coding genes are transcribed. At this point, it is still largely unknown to the extent at which de no novo variation in the noncoding genome contributes to the genetic risk of ASD [An et al., 2018].

Although WGS presents a promising opportunity to reveal insight about noncoding regions, the size and unknown number of tests poses a unique multiple testing challenge. To address the unique multiple burden respecting the scale of the noncoding genome, a category-wide association studies (CWAS) framework has been introduced by defining over fifty-thousand annotation categories to test for association with ASD [Werling et al., 2018]. However, in the analysis of case-control data from a limited number of quartet-families (parents, probands, and siblings for controls), it was unable to detect a single noncoding annotation category that met a category-wide significance threshold (Figure 1.3). Similar null results were observed by with the inclusion of more families, however a de novo risk score analysis implicated the contribution of de novo mutations in promoter regions to ASD [An et al., 2018].

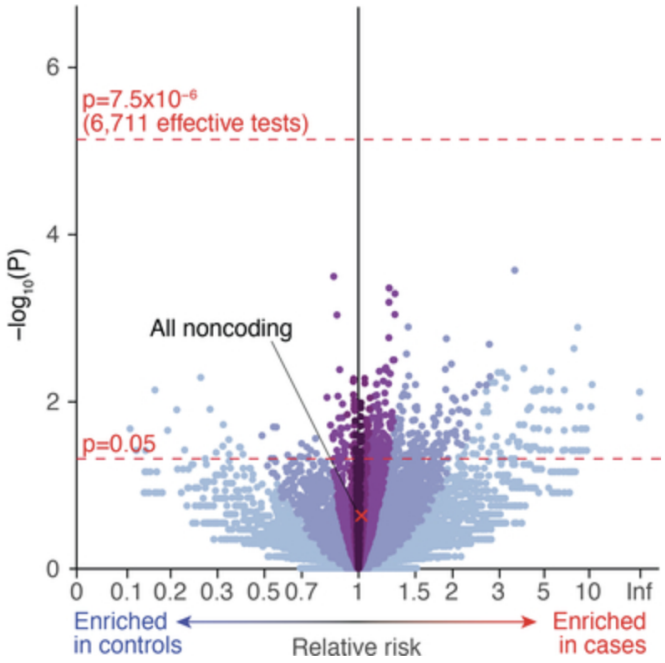

Figure 1.3: CWAS fails to detect any significant enrichment (y-axis) in noncoding variants for cases (right x-axis) or controls (left x-axis) [An et al., 2018].

1.5 Thesis overview

My thesis aims to contribute to addressing these challenging problems in genetics studies to improve power to detect associations for under-powered studies in the context of neuropsychiatric disorders. I have completed these aims in the following ways:

1 Built a pipeline to select a subset of SNPs documented to affect gene expression and then incorporate covariates from independent GWAS and gene expression studies into AdaPT via gradient boosted trees to ulimately improve our power. Our boosting implementation of AdaPT scales with more covariates and addresses the perceived modeling weakness of AdaPT, enabling practitioners to capture interactions and non-linear effects from resources of available multi-omics metadata (chapter 2).

2 When investigating a popular tool for gene-based testing, Multi-marker Analysis of GenoMic Annotation [de Leeuw et al., 2015, MAGMA], we discovered it yielded an unusual distribution of gene-level p-values which would violate necessary assumptions for AdaPT to maintain FDR control. Despite undocumented, ad-hoc corrections in

MAGMA, we observe via simulations and recent applications that it yields incorrect null p-value distribution resulting in inflated error rates. This is due to the inappropriate application of an approximation that is valid for only one-sided tests, while GWAS summary statistics are two-sided (chapter 3).

3 We observed that current gene-based testing approaches do not capture LD of SNPs falling in different nearby genes, which can induce correlation of gene-based test statistics. This compromises the interpretability of gene-based testing, thus obscuring the meaning of error-rate guarantees. We introduce an algorithm to account for this correlation directly, based on the LD-induced correlation of commonly used quadratic gene-level test statistics. (chapter 4).

4 Demonstrate an improvement in CWAS power by augmenting testing corrections with annotation-level information in simulation studies. We also investigate the use of data blurring in the context of exploring annotation structure with the goal of aiding hypothesis testing. While testing power remains stagnant, the use of blurring provides the opportunity for estimation post-selection to reveal greater insight about noncoding associations (chapter 5).

Two

An Implementation of Adaptive p-value Thresholding for GWAS with Gradient Boosted Trees

To correct for a large number of hypothesis tests, most researchers rely on simple multiple testing corrections. Yet, new methodologies of selective inference could potentially improve power while retaining statistical guarantees, especially those that enable exploration of test statistics using auxiliary information (covariates) to weight hypothesis tests for association. We explore one such method, adaptive p-value thresholding [Lei and Fithian, 2018, AdaPT], in the framework of genome-wide association studies (GWAS) and gene expression/coexpression studies, with particular emphasis on schizophrenia (SCZ). Selected SCZ GWAS association p-values play the role of the primary data for AdaPT; SNPs are selected because they are gene expression quantitative trait loci (eQTLs). This natural pairing of SNPs and genes allow us to map the following covariate values to these pairs: GWAS statistics from genetically-correlated bipolar disorder, the effect size of SNP genotypes on gene expression, and gene-gene coexpression, captured by subnetwork (module) membership. In all 24 covariates per SNP/gene pair were included in the AdaPT analysis using flexible gradient boosted trees. We demonstrate a substantial increase in power to detect SCZ associations using gene expression information from the developing human prefontal cortex [Werling et al., 2020b]. We interpret these results in light of recent theories about the polygenic nature of SCZ. Importantly, our entire process for identifying enrichment and creating features with independent complementary data sources can be implemented in many different high-throughput settings to ultimately improve power.

This chapter appears in [Yurko et al., 2020].

2.1 Introduction

Large scale experiments, such as scanning the human genome for variation affecting a phenotype, typically result in a plethora of hypothesis tests. To overcome the multiple

testing challenge, one needs corrections to limit false positives while maximizing power. Introduced by [Benjamini and Hochberg, 1995], false discovery rate (FDR) control has become a popular approach to improve power for detecting weak effects by limiting the expected false discovery proportion (FDP) instead of the more classical Family-Wise Error Rate. The Benjamini-Hochberg (BH) procedure was the first method to control FDR at target level α using a step-up procedure that is *adaptive* to the set of p-values for the hypotheses of interest [Benjamini and Hochberg, 1995]. Other methods for FDR control have led to improvements in power over BH by incorporating prior information, such as by the use of p-value weights [Genovese et al., 2006]. In the "omics" world – genomics, epigenomics, proteomics, and so on – the challenge of multiple testing is burgeoning, in part because our ability to characterize omics features grows continually and in part because of the realization that multiple omics are required for describing phenotypic variation. One might imagine merging complementary omics data and tests using a priori hypothesis weights to improve power; however, until recently, it was not clear how to choose these weights in a data driven manner.

Recent methodologies have been proposed to account for covariates or auxiliary information while maintaining FDR control [Scott et al., 2015, Ignatiadis et al., 2016, Boca and Leek, 2018, Li and Barber, 2019, Zhang et al., 2019]. We implement a selective inference approach, called adaptive p-value thresholding [Lei and Fithian, 2018, AdaPT], to explore prior auxiliary information while maintaining guaranteed finite-sample FDR control. A recent review compared the performance of AdaPT with other covariate-informed methods for FDR control with off-the-shelf one and two-dimensional covariate examples [Korthauer et al., 2019]. One of the weaknesses they ascribe to AdaPT is the unintuitive modeling framework for incorporating covariates; however, AdaPT is not a specific algorithm that one can simply apply to a dataset, but rather a meta-algorithm for marrying machine learning methods to multiple testing problems without compromising FDR control. We fully embrace AdaPT's flexibility via gradient boosted trees in a much richer, high-dimensional setting. Our boosting implementation of AdaPT easily scales with more covariates, enabling practitioners to capture interactions and non-linear effects from the rich resources of prior information available.

In this manuscript, we demonstrate our gradient boosted trees implementation of AdaPT on results from genome-wide association studies (GWAS), incorporating covariates constructed from independent GWAS and gene expression studies. Specifically, we apply AdaPT to GWAS for detecting single nucleotide polymorphisms (SNPs) associated with schizophrenia (SCZ) using bipolar disorder (BD) GWAS results from an independent dataset as a covariate. Additionally, we incorporate results from the recent BrainVar study to identify a set of expression-SNPs (eSNPs) based on 176 neurotypical brains, sampled from pre- and post-natal tissue from the human dorsolateral prefrontal cortex [Werling et al., 2020a]. Along with the genetically correlated BD z-statistics, we create additional features from this

complementary data source by summarizing the associated developmental gene expression quantitative trait loci (eQTL) slopes and membership in gene co-expression networks. We demonstrate that this process of identifying an enriched set of eSNPs and applying AdaPT with covariates summarizing gene expression from the developing human prefrontal cortex yield substantial improvement in power with each additional piece of information from the BrainVar study. Furthermore, we validate the replication of our results using more recent, independent SCZ studies.

This study had two goals, to explore the use of AdaPT in a realistic high-dimensional multi-omics setting and to determine what can be learned about the neurobiology of SCZ by this exploration. Our results revealed the power of incorporating auxiliary information with flexible gradient boosted trees. While each covariate independently provided at best a modest increase in power, our adaptive search discovered a more complex model with far greater power. These discoveries also led to increasing support for the polygenic basis of SCZ, complementing recent findings and suggesting that there are many physiological avenues to its underlying neurobiology. We emphasize that the process and analysis undertaken with this implementation of AdaPT can be extended to a variety of "omics" and other settings to utilize the rich contextual information that is often ignored by standard multiple testing corrections. We highlight this feature by analyzing two other sets of GWAS studies, type 2 diabetes (T2D) and body mass index (BMI), using results from these analyses to interpret findings from SCZ.

2.2 Results

2.2.1 Methodology overview

AdaPT is an iterative search procedure, introduced by [Lei and Fithian, 2018], for determining a set of discoveries/rejections, $\mathcal{R}$, with guaranteed finite-sample FDR control at target level α under conditions outlined below. We apply AdaPT to the collection of p-values and auxiliary information, $(p_i, x_i)_{i \in n}$, testing hypothesis H_i regarding SNP i's association with the phenotype of interest (e.g. SCZ). The covariates from some feature space, $x_i \in \mathcal{X}$, capture information collected independently of p_i, but potentially related to whether or not the null hypothesis for H_i is true and the effect size under the alternative. AdaPT provides a flexible framework to incrementally *learn* these relationships, potentially increasing the power of the testing procedure, while maintaining valid FDR control.

For each step $t = 0, 1, \ldots$ in the AdaPT search, we first determine the rejection set $\mathcal{R}_t = \{i : p_i \leq s_t(x_i)\}$, where $s_t(x_i)$ is the rejection threshold at step t that is *adaptive* to the covariates x_i. This provides us with both the number of discoveries/rejections $R_t = |\mathcal{R}_t|$, as well as a *pseudo*-estimate for the number of false discoveries $A_t = |\{i : p_i \geq 1 - s_t(x_i)\}|$ (i.e. number of p-values above the "mirror estimator" of $s_t(x_i)$). These quantities are used

to estimate the FDP at the current step t,

$$\widehat{\text{FDP}}_t = \frac{1 + A_t}{\max\{R_t, 1\}}. \tag{2.1}$$

If $\widehat{\text{FDP}}_t \leq \alpha$, then the AdaPT search ends and the set of discoveries $\mathcal{R}_t$ is returned. Otherwise, we proceed to update the rejection threshold while satisfying two protocols: (1) the updated threshold must be more stringent $s_{t+1}(x_i) \leq s_t(x_i)$, and (2) p-values determining R_t and A_t are *partially* masked,

$$\tilde{p}_{t,i} = \begin{cases} p_i, \text{if } s_t(x_i) < p_i < 1 - s_t(x_i), \\ \{p_i, 1 - p_i\}, \text{ otherwise.} \end{cases} \tag{2.2}$$

Under these protocols, the rejection threshold can be updated using R_t, A_t, and $(x_i, \tilde{p}_{t,i})_{i \in [n]}$. The flexibility in how this update takes place is one of AdaPT's key strengths and allows it to easily incorporate other approaches from the multiple testing literature, such as a conditional version of the two-groups model [Efron et al., 2001] with estimates for the probability of being non-null, π_1, and the effect size under the alternative, μ.

The algorithm proceeds by sequentially updating the threshold $s_{t+1}(x_i)$ to discard the most likely null element in the current rejection region, as measured by the conditional local false discovery rate (fdr): i.e., $i^* = \arg\max_{i \in \mathcal{R}_t} \text{fdr}_{t,i}$ is removed from $\mathcal{R}_t$. With the threshold updated, the AdaPT search repeats by estimating FDP and updating the rejection threshold until the target FDR level is reached $\widehat{\text{FDP}}_t \leq \alpha$ or $\mathcal{R}_t = 0$.

This procedure guarantees finite-sample FDR control under independence of the null p-values and as long as the null distribution of p-values is *mirror conservative*, i.e. the large "mirror" counterparts $1 - p_i \geq 0.5$ are at least as likely as the small p-values $p_i \leq 0.5$. To address the assumption of independence, we select a subset of weakly correlated SNPs detailed in *Data*, and additionally provide simulations in *Method Appendix* showing that AdaPT appears to maintain FDR control in relevant positive dependence settings. However, one practical limitation we encounter with the FDP estimate in Equation 2.1 is observing p-values *exactly* equal to one. While this can understandably occur with publicly available GWAS summary statistics, p-values equal to one will *always* contribute to the estimated number of false discoveries A_t. This nuance can lead to a failure of obtaining discoveries at a desired target α, such as the reported AdaPT results by [Korthauer et al., 2019] for multiple case-studies. However, we demonstrate in *Method Appendix* an adjustment to the p-values for T2D and BMI GWAS applications that alleviates this problem, although future work should explore modifications to the FDP estimator itself.

The modeling step of AdaPT estimates conditional local fdr with an EM algorithm. In this context, we use gradient boosted trees, which constructs a flexible predictive function as

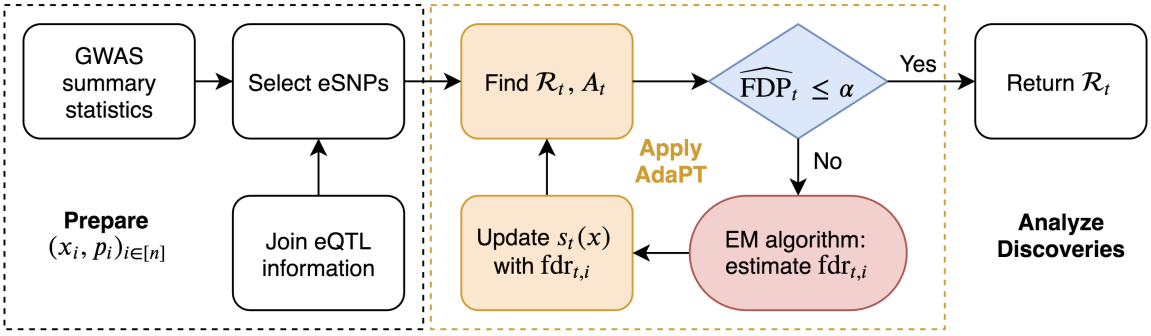

Figure 2.1: Summary of AdaPT GWAS implementation for selected set of SNPs. See Figure S1 for a summary of the AdaPT EM algorithm.

a weighted sum of many simple trees, fit using a gradient descent procedure that minimizes a specified objective function. The two objective functions considered correspond to estimating the probability of a test being non-null and the distribution of the effect size for non-null tests. The advantage of this approach to function fitting is that it is invariant to monotonic variable transformations, automatically incorporates important variable interactions, and is able to handle a large number of covariates without degrading significantly in performance due to the high dimensionality. In contrast, less effective methods might fail to capture useful information because the covariates are poorly transformed for a linear model, because the important information is only revealed through a combination of covariates, or because the important signal is simply swamped by the number of possible predictors to search through. Our choice of method gives the flexibility to include many potentially useful covariates without being overly concerned about the functional form with which they enter or their marginal utility. In our implementation, we employ the XGBoost library [Chen and Guestrin, 2016] to capitalize on its computational advantages. Figure 2.1 displays the full pipeline of our implementation of AdaPT to GWAS summary statistics for SNPs using expression quantitative trait loci (eQTL) to select the SNPs under investigation.

2.2.2 Data

Our investigation includes AdaPT analyses of published GWAS p-values, $\{p_i, i = 1, \dots n\}$, for body mass index [Locke et al., 2015, BMI], type 2 diabetes [Mahajan et al., 2018, T2D], and schizophrenia [Ruderfer et al., 2014, SCZ], but we focus our presentation on SCZ results. SCZ is a highly heritable, severe neuropsychiatric disorder. It is most strongly correlated, genetically, with another severe disorder, bipolar disorder (BD) [Lichtenstein et al., 2009, Cross-Disorder Group of the Psychiatric Genomics Consortium, 2013]. Because of this genetic correlation, reported z-statistics from BD GWAS, z_i^{BD}, can be used as informative covariates for determining the SCZ rejection threshold. As an application of our AdaPT implementation, we use the GWAS summary statistics reported by [Ruderfer et al., 2014],

specifically 19,779 subjects diagnosed with either SCZ or BD with 19,423 control subjects (data are available from the Psychiatric Genomics Consortium, PGC). SCZ and BD subjects were completely independent and independent controls were bulk matched to the sample sizes of the two case samples. Results from more recent studies in [Ruderfer et al., 2018a] are used for replication analysis of our results (combined 53,555 SCZ and BD cases with 54,065 controls). However, the *2014-only* studies from [Ruderfer et al., 2014] are a subset of the *all-2018* studies from [Ruderfer et al., 2018a]. Although we do not have access to the raw genotype data, we use the fact that both papers report inverse variance-weighted fixed effects meta-analysis results [Willer et al., 2010]. We then separate the summary statistics for the *2018-only* studies exclusive to [Ruderfer et al., 2018a], thus independent of the *2014-only* studies, and create an appropriate hold-out for replication analysis.

After matching alleles from both *2014-only* and *all 2018* studies and limiting SNPs to those with imputation score $INFO > 0.6$ for both BD and SCZ in *2014-only* [Ruderfer et al., 2014], we obtained 1,109,226 SNPs. Rather than test all SNPs, we chose to investigate a selected subset of SNPs, eSNPs, whose genotypes are correlated with gene expression; this additional filtering step captures a set of SNPs that are more likely to be functional and not highly correlated [Nicolae et al., 2010]. These eSNPs were identified from two sources. First, we evaluated the BrainVar study of dorsolateral prefrontal cortex samples across a developmental span [Werling et al., 2020a]. BrainVar included cortical tissue from 176 individuals falling into two developmental periods: pre-natal, 112 individuals; and post-natal, 60 individuals. We identified $n_{\text{SCZ}} = 25{,}076$ eSNPs as any eQTL SNP-gene pairs provided by [Werling et al., 2020a] meeting Benjamini-Hochberg $\alpha \leq 0.05$ for at least one of the three sample sets (pre-natal, post-, and complete = all). These eSNPs were used for the SCZ analysis, which is a neurodevelopmental disorder and thus a developmental cohort seemed most appropriate for our analyses.

The second source was the Genotype-Tissue Expression (GTEx) V7 project dataset [GTEx Consortium and others, 2015] with adult samples from fifty-three tissues. As the first winnowing step, we identified the set of GTEx eQTLs for *any* of the available tissues at target FDR level $\alpha = 0.05$. Rather than use all GTEx eQTLs, however, we selected eQTL SNP-gene whose genotypes are most predictive of expression for each gene. The GTEx eSNPs were used for analysis of T2D and BMI, both of which typically onset in adults (for details see *Method Appendix*).

For each eSNP i, we created a vector of covariates x_i to incorporate auxiliary information collected independently of p_i, including p-values from GWAS studies of related phenotypes, and relationships inferred from gene expression studies. First, we utilize the mapping of eSNPs to genes derived from eQTLs assessed in a relevant tissue type r. Although the majority of observed eSNPs have one unique cis-eQTL gene pairing, 14% of SNPs in BrainVar were eQTL for multiple genes. Let $\mathcal{G}_i^r$ denote the set of genes whose expression is associated

with eSNP i and summarize the level of expression as the average absolute eQTL slope for variants in $\mathcal{G}_i^r$ to obtain $\bar{\beta}_i^r$. Additionally, we account for gene co-expression networks as covariates using the $J = 20$ modules reported in the BrainVar study, which were generated using weighted gene co-expression network analysis [Zhang and Horvath, 2005, WGCNA]. For each of the $j = 1, \ldots, J$ WGCNA modules, we create an indicator variable $\ell_{i,j}^r$ denoting whether or not eSNP i has *any* associated cis-eQTL genes in module j.

For the n_{SCZ} eSNPs, we calculate $\bar{\beta}_i^{type}$ where $type \in \{\text{pre, post, complete}\}$ to capture the eSNP's overall expression association across different epochs of the developmental span. Additionally, we use the 20 WGCNA modules (including unassigned *gray*) reported by [Werling et al., 2020a] to create indicator variables $\ell_{i,j}^{\text{SCZ}}$ for $j = 1, \ldots, 20$. This culminates in a vector of twenty-four covariates $x_i^{\text{SCZ}} = (z_i^{\text{BD}}, \bar{\beta}_i^{\text{pre}}, \bar{\beta}_i^{\text{post}}, \bar{\beta}_i^{\text{complete}}, \ell_{i,1}^{\text{SCZ}}, \ldots, \ell_{i,20}^{\text{SCZ}})$. Although we use WGCNA modules to make use of the results from the BrainVar study, future applications could explore other approaches to account for gene set and pathway analysis [Zhu and Stephens, 2018].

2.2.3 AdaPT discoveries

As noted elsewhere [Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014], eSNPs are more likely to be associated with a GWAS phenotype than are randomly chosen SNPs. This is true for the eSNP from BrainVar too, when evaluated in light of the SCZ GWAS p-values (Figure 2.2A). To evaluate the performance of the AdaPT search algorithm using the eSNP data, we compare the fitted full covariate model to results from its *intercept-only* version (Figure 2.2B versus 2.2C). As expected, the intercept-only analysis performs better than BH, with all 269 BH discoveries contained within the intercept-only discoveries, because it incorporates an estimate for the proportion of non-null tests. The full model rejects $R_{\text{SCZ}} = 843$ of the $n_{\text{SCZ}} = 25{,}076$ BrainVar eSNPs versus 361 discoveries for the intercept-only model. For insight into AdaPT's performance, we sequentially include (1) only the BD z-statistics, then (2) include eQTL slope summaries, and then (3) the WGCNA indicators (Figure 2.2D-E).

The largest number of discoveries occurs when all twenty-four covariates are fitted (Figure 2.2 D), highlighting that all three types of information *together* are required. Notably, only 540 associations are discovered using all covariates without interactions, fewer discoveries than only using module-based covariates with interactions. This highlights the improvement in AdaPT's performance from modeling the interactions between covariates via gradient boosted trees. As might be expected from their counts of discoveries (Figure 2.2D), the greatest overlap with the full model occurs by fitting all covariates, but without interactions, or by fitting the module-based covariates (Figure 2.2E).

Additional discoveries are of little interest if they consist primarily of SNPs in LD with SNPs already discovered using a simpler model, such as the logit model typically used for

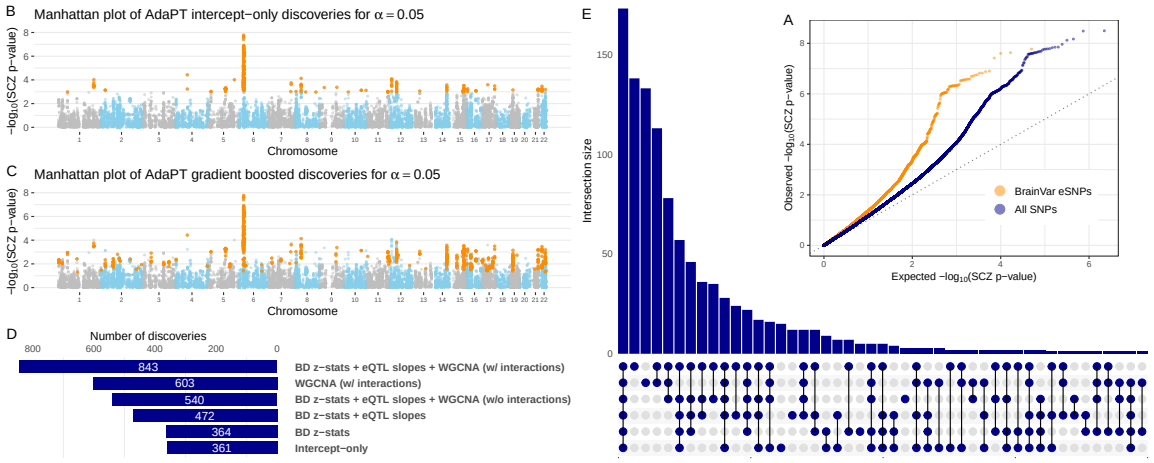

Figure 2.2: AdaPT results from analysis of schizophrenia (SCZ) p-values. *(A)* Comparison of qq-plots revealing SCZ enrichment for both BrainVar eSNPs compared to the full set of SNPs from *2014* studies. (B-C) Manhattan plots of SCZ AdaPT discoveries (in orange) using *(B)* intercept-only model compared to *(C)* covariate-informed model at target $\alpha = 0.05$. *(D-E)* Comparison of the number of discoveries at target $\alpha = 0.05$ for AdaPT with varying levels of covariates (D) and *(E)* their resulting discovery set intersections.

SCZ GWAS. For context, however, of the initial 25,076 eSNPs we analyzed, only four have p-values $< 5 \times 10^{-8}$, the standard GWAS threshold, and all four occur in the discovery sets for the AdaPT full and intercept-only models. To investigate how the Adapt procedure performs using completely independent eSNPs, we identified the "lead" SNP in each LD block using the approach delineated in [Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014] and compared model performance for this set of approximately independent SNPs (*Method Appendix*). This thinning results in roughly 3,960 eSNPs to be analyzed by the different models (Figure 2.2). (Ties in q-values add or subtract a few SNPs to this 3,960 count, depending on the model analyzed.) When AdaPT is fit to these independent SNPs, we obtain analogous improvements in performance compared to the larger set of SNPs (Figures S2 and S3): the full AdaPT model discovers 95 independent loci, while the intercept-only model discovers only 42 loci. Likewise, the full model is the best model and interactions remain important. Finally, no location in the genome exerts unusual influence on the results, which is also the case for the analyses of 25,076 eSNPs.

As described previously, we performed similar analyses of T2D and BMI GWAS p-values. All results for these analyses, as well as more details regarding analyses of SCZ, are available in Dataset S1 and *Method Appendix*.

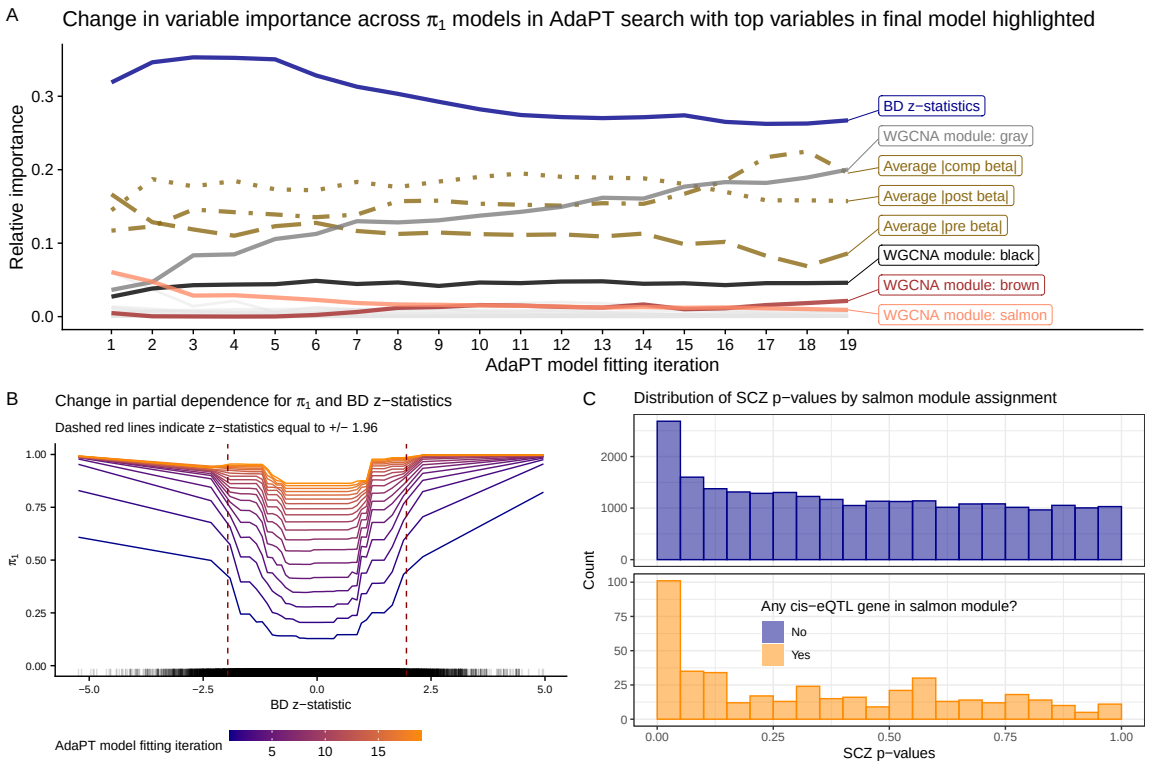

Figure 2.3: Variable importance and relationships. *(A)* Change in variable importance for AdaPT estimated probability of non-null π_1 model across the search, with top variables in final model highlighted. *(B)* Change in partial dependence for estimated probability of being non-null π_1 and BD z-statistics across π_1 models in AdaPT search. *(C)* SCZ enrichment of eSNPs based on *salmon* WGCNA module membership, the most important WGCNA module indicator in the first model fitting step.

2.2.4 Variable importance and relationships

We examine the variable importance and partial dependence plots from the gradient boosted models to provide insight into the relationships between each of the covariates and SCZ associations. Figure 2.3(A) displays the change in variable importance for the probability of being non-null (π_1) at each model fitting iteration, with the top variables in the final model highlighted. We see that the BD z-statistics are estimated as the most important for each π_1 model, but they decrease in importance in the final steps. In contrast, the unassigned *gray* module increases in important throughout the AdaPT search. This change in variable importance across the AdaPT search highlights that the difference in the discriminatory power of covariates depends on the remaining masked p-values.

Figure 2.3(B) displays the partial-dependence plot [Friedman, 2001] at each AdaPT model fitting iteration for the estimated marginal relationship between the BD z-statistics and the probability of being non-null, evaluated at the $0, 2.5\%, 5\%, \ldots, 100\%$ percentiles. Because the goal of the AdaPT two-groups model (detailed in *Methods*) is to order the remaining masked p-values, the π_1 model predicts values relative to the remaining masked p-values: as the rejection threshold $s_t(x_i)$ becomes more stringent, the masked p-values are more likely non-null (assuming there is signal). However, for each model iteration, Figure 2.3(B) reveals an increasing likelihood for non-null results as the BD z-statistics grow in magnitude from zero, as well as a diminished impact of BD z-statistics on the estimated π_1 for later model iterations. Figure 2.3(C) displays the clear enrichment for eSNPs with cis-eQTL genes that are members of the *salmon* WGCNA module reported by [Werling et al., 2020a], which was the most important WGCNA module indicator in the first model fitting step. This differs from the unassigned *gray* module variable: it is predictive of SNPs that are classified as null, rather than associated with the phenotype. Taken together, Figures 2.3(A-C) emphasize the use of all covariates across different steps of the AdaPT search. See *Method Appendix* for more analyses highlighting the advantages of accounting for interactions between covariates.

2.2.5 Replication in independent studies

Next, we examine the replicability of the *2014-only* SCZ AdaPT results using independent *2018-only* studies. We find (Figure 2.4) an increasing smoothing spline relationship between these sets of values, with noticeably increasing evidence indicated by the *2018-only* p-values for the set of AdaPT discoveries at $\alpha = 0.05$. Additionally, of the 843 discoveries from the *2014-only* studies at target FDR level $\alpha = 0.05$, approximately 55.2% (465 eSNPs) were nominal replications for *2018-only* (p-values < 0.05), comparable to the replication fraction expected on the basis of power (see *Method Appendix* for supporting simulations).

2.2.6 Gene ontology comparison

Using the SNP discoveries, which span the genome, we next sought biological insights. We applied gene ontology enrichment analysis [Ashburner et al., 2000, The Gene Ontology Consortium, 2018] to the 136 genes obtained from the eQTL variant-gene pairs associated with the 843 discoveries. This analysis produced no clear signal, yielding only a minor enrichment for biological processes related to peptide antigen assembly. Several explanations are plausible, we explore two: either AdaPT is discovering SNPs of such small effect that the discoveries are not meaningful or SCZ is a highly complex disorder with a large number of biological processes involved. For comparison we applied our full pipeline to GWAS summary statistics for T2D [Mahajan et al., 2018]. This comparison is of interest because T2D is a disease with a well understood functional basis and this is a well powered study (74,124 T2D cases and 824,006 controls). We restricted our analysis to 176,246 eSNPs based on eQTLs obtained using GTEx data. Next, we created eQTL-based covariates using pancreas, liver, and

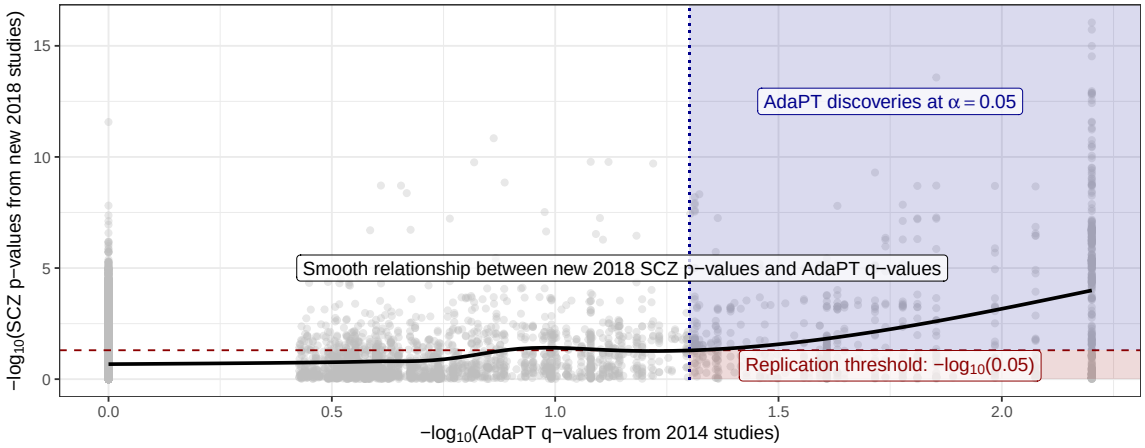

Figure 2.4: Relationship between the *2018-only* p-values and the resulting *2014-only* q-values from the AdaPT search. Black line displays smoothed relationship between SCZ p-values from *2018-only* studies and AdaPT q-values from *2014-only* studies. Blue region indicates AdaPT discoveries at $\alpha = 0.05$ that are nominal replications, p-values from *2018-only* studies $< .05$, while red region denotes discoveries that failed to replicate.

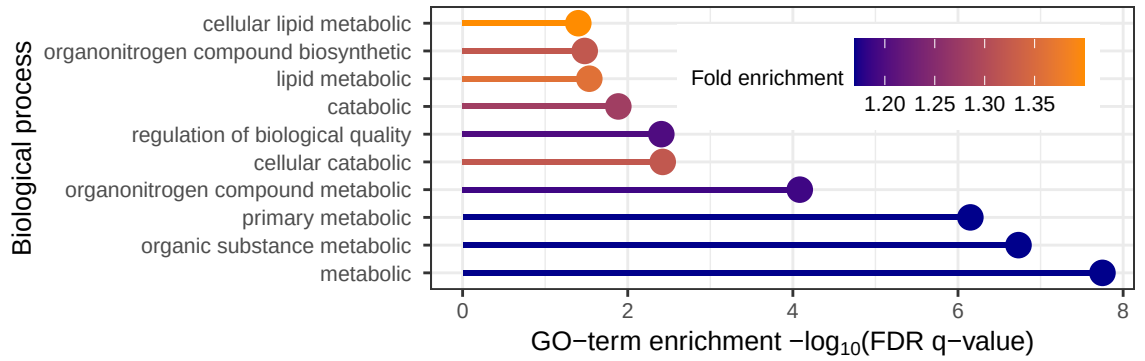

Figure 2.5: T2D gene ontology enrichment analysis results for top ten biological processes based on positive fold enrichment.

adipose tissue samples (see *Method Appendix* for more details). After creating a vector of covariates from GTEx, AdaPT returned 14,920 eSNPs at $\alpha = 0.05$, resulting in 5,970 associated genes. Applying gene ontology enrichment analysis to this gene list, we discovered enrichment for biological processes related to lipid metabolic process (Figure 2.5), consistent with previous literature [Cirillo et al., 2018]. These results provide some reassurance that the lack of specificity in the SCZ results can be attributed to the complex etiology of SCZ. For comparison to the well powered BMI GWAS (339,224 subjects), we found a lack of gene ontology enrichment in our gene discoveries (*Method Appendix*).

2.2.7 Pipeline results for all 2018 studies

In addition to applying the pipeline to SCZ p-values from the *2014-only* studies in [Ruderfer et al., 2014], we also modeled p-values from *all 2018* studies. The latter yields far more discoveries due to smaller standard errors from increased study sizes, even though the covariates were the same: for x_i^{SCZ}, we find 2,228 discoveries at target FDR level $\alpha = 0.05$ when the pipeline was applied to the p-values for most up-to-date set of studies versus 843 for the *2014-only* studies. Notably, the intercept-only version of AdaPT returned 1,865 discoveries at $\alpha = 0.05$, meaning the covariates contributed to $\approx 19\%$ increase in discovery rate for *all 2018* studies versus the $\approx 134\%$ increase (361 to 843 eSNPs) from using the covariates for the *2014-only* studies. This reinforces the value of using auxiliary information in studies with lower power. Complementary to this observation, AdaPT applied to BMI GWAS the covariate informed models did not yield more discoveries than the intercept-only version (details presented in *Method Appendix*). Simply accounting for more auxiliary information does not guarantee an improvement in power and the advantages thereof diminishes as power increases, as witnessed by results for *all 2018* studies for SCZ and the large-scale BMI GWAS. Additionally, the larger number of discoveries for the SCZ *all 2018* studies, 2,228, maps onto 382 genes. Despite this increase, these genes did not reveal any clear signal from the Gene Ontology enrichment analysis, comporting with results from the *2014-only* results.

2.3 Discussion

Our goals in this study were to explore the use of AdaPT for high-dimensional multi-omics settings and investigate the neurobiology of SCZ in the process. AdaPT was used to analyze a selected set of GWAS summary statistics for SNPs, together with numerous covariates. Specifically, SNPs were selected if they were documented to affect gene expression; these SNP-gene pairs were dubbed eSNPs. Covariates for these eSNPs included GWAS test statistics from a genetically correlated phenotype, BD, which were mapped to eSNPs through SNP identity; as well as features of gene expression and co-expression networks, which were mapped to eSNPs through genes. By coupling flexible gradient boosted trees with the AdaPT procedure, relationships among eSNP GWAS test statistics and covariates were uncovered and more SNPs were found to be associated with SCZ, while maintaining

guaranteed finite-sample FDR control. The tree-based handling of covariates addresses a perceived weakness of AdaPT, namely the unintuitive modeling framework for incorporating covariates [Korthauer et al., 2019]. Moreover, it is worth noting that the original approach implemented by [Lei and Fithian, 2018], a generalized linear model with spline bases, yields similar results (361 discoveries at target $\alpha = 0.05$) when applied to the univariate case of only using BD z-statistics. This is an even more straightforward implementation for handling covariates without interactions. The pipeline we built should be simple to mimic for a wide variety of omics and other analyses.

The results shed light on the level of complexity underlying the neurobiology of SCZ. If the origins of SCZ arose by perturbations of one or a few pathways, we would expect to converge on those pathways as we accrue more and more genetic associations. On the other hand, if the ways to generate vulnerability to SCZ were myriad — even if there is an single ultimate cause shared across all cases — then we might expect no such convergence, at least with regards to the common variation assessed through GWAS. Gene ontology analysis of associated discovery genes from either the *2014-only* or *all 2018* studies reveals no enrichment for biological processes for SCZ. There are many possible explanations for these null findings, one of which is simply a lack of power or specificity of our results. However, the result stands in stark contrast to the results for T2D, for which the gene ontology analysis converges nicely on accepted pathways to T2D risk; yet they comport with those for BMI, which is known to have myriad genetic and environmental origins. Therefore our results are consistent with myriad pathways to vulnerability for SCZ, although it is impossible to rule out other explanations: for example, the possibility that we understand so little about brain functions that gene ontology analyses lack specificity. In any case, our results are consistent with two recent theories underlying the genetics of SCZ, namely extreme polygenicity [O'Connor et al., 2019] and "omnigenic" origins [Boyle et al., 2017].

Although the examples considered in this manuscript pertain to omics data, this process can be adapted for a large variety of settings. We demonstrate in *Method Appendix* simulations showing that AdaPT appears to maintain FDR control in positive dependence settings emulating linkage disequilibrium (LD) block structure underlying GWAS results. There is a clear need, however, for future work to explore AdaPT's properties and computational challenges under various dependence regimes. The growing abundance of contextual information available in "omics" settings provides ample opportunity to improve power for detecting associations, using a flexible approach such as AdaPT, when addressing the multiple testing challenge.

2.4 Methods

2.4.1 Two-groups model

The most critical step in the AdaPT algorithm [Lei and Fithian, 2018] involves updating the rejection threshold $s_t(x_i)$. Following [Lei and Fithian, 2018], we use a conditional version of the classical two-groups model [Efron et al., 2001, Scott et al., 2015] where the null p-values are modeled as uniform ($f_0(p|x) \equiv 1$) and we model the non-null p-value density with a beta distribution density parametrized by $\mu_i = \mathbb{E}[-\log(p_i)]$, resulting in a conditional density for a beta mixture model, $f(p|x_i) = \pi_1(x_i)\frac{1}{\mu_i}p^{1/\mu_i - 1} + 1 - \pi_1(x_i)$. In this form, we can model the non-null probability $\pi_1(x_i) = \mathbb{E}[H_i|x_i]$ and the effect size for non-null hypotheses $\mu(x_i) = \mathbb{E}[-\log(p_i)|x_i, H_i = 1]$ with two separate gradient boosted tree-based models. The XGBoost library [Chen and Guestrin, 2016] provides logistic and Gamma regression implementations which we use for $\pi_1(x_i)$ and $\mu(x_i)$ respectively.

There are two categories of missing values in these regression problems: H_i is never observed, and at each step t of the search, the p-values for tests $\{i : p_i \leq s_t(x_i) \text{ or } p_i \geq 1 - s_t(x_i)\}$ are masked as $\tilde{p}_{t,i}$. An expectation-maximization (EM) algorithm can be used to estimate both $\hat{\pi}_1(x_i)$ and $\hat{\mu}(x_i)$ by maximizing the partially observed likelihood. We briefly restate the EM algorithm from (1), and provide details in our supplementary materials that reflect the approach taken in the `R adaptMT` package by the same authors, which differs slightly from (1).

During the E-step of the $d = 0, 1, \ldots$ iteration of the EM algorithm, conditional on the partially observed data fixed at step t, $(x_i, \tilde{p}_{t,i})_{i\in[n]}$, we compute both, $\hat{H}_i^{(d)}$ and $\hat{b}_i^{(d)}$, where $\hat{b}_i^{(d)}$ indicates how likely $p'_{t,i} = \min(\tilde{p}_{t,i})$ equals p_i for non-null hypotheses. The explicit calculations of $\hat{H}_i^{(d)}$ and $\hat{b}_i^{(d)}$ are available in the supplementary materials of [Lei and Fithian, 2018].

The M-step consists of estimating $\hat{\pi}_1^{(d)}$ and $\hat{\mu}^{(d)}$ with separate gradient boosted trees, using *pseudo*-datasets to handle the partially masked data. In order to fit the model for $\pi_1(x_i)$, we construct the response vector $y_\pi^{(d)} = (1, \ldots, 1, 0, \ldots, 0) \in \mathbb{R}^{2n}$ and use weights $w_\pi^{(d)} = (\hat{H}_1^{(d)}, \ldots, \hat{H}_n^{(d)}, 1-\hat{H}_1^{(d)}, \ldots, 1-\hat{H}_n^{(d)}) \in \mathbb{R}^{2n}$. Then we estimate $\hat{\pi}_1^{(d)}(x_i)$ using the first n predictions from a classification model using $y_\pi^{(d)}$ as the response variable with the covariate matrix $(x_i)_{i\in[n]}$ replicated twice and weights $w_\pi^{(d)}$. Similarly, for estimating $\hat{\mu}^{(d)}(x_i)$ we construct a response vector $y_\mu^{(d)} = (-\log(p_1), \ldots, -\log(p_n), -\log(1-p_1), \ldots, -\log(1-p_n)) \in \mathbb{R}^{2n}$ with weights $w_\mu^{(d)} = (\hat{b}_1^{(d)}, \ldots, \hat{b}_n^{(d)}, 1 - \hat{b}_1^{(d)}, \ldots, 1 - \hat{b}_n^{(d)}) \in \mathbb{R}^{2n}$, and again take the first n predicted values using the duplicated covariate matrix.

We follow the procedure detailed in Section 4.3 of [Lei and Fithian, 2018] to estimate the conditional local fdr for each $p'_{t,i}$, and then update the rejection threshold to $s_{t+1}(x_i)$ by removing test $i^* = \arg\max_{i\in\mathcal{R}_t} \text{fdr}_{t,i}$ from $\mathcal{R}_t$.

2.4.2 AdaPT gradient boosted trees with CV steps

As a flexible approach for modeling the conditional local fdr, we use gradient boosted trees [Friedman, 2001] via the open-source XGBoost implementation [Chen and Guestrin, 2016]. Gradient boosted trees are an ensemble of many small tree models that jointly contribute to predictions. Let $f_p \in \mathcal{F}$ be an individual regression tree, then the sum-of-trees model can be written as, $\hat{y}_i = \sum_{p=1}^{P} f_p(x_i)$ to minimize $\sum_i^n L(y_i, \hat{y}_i) + \sum_{p=1}^{P} \Omega(f_p)$ where L is the loss function and Ω measures the complexity of each tree such as the maximum depth, regularization, etc. [Chen and Guestrin, 2016] detail the algorithms for fitting the model in an additive manner as well as determining the splits for each tree.

To tune the variety of parameters for gradient boosted trees within AdaPT, such as the number of trees P and maximum depth of each tree, we use the cross-validation (CV) approach recommended in [Lei and Fithian, 2018]. If we are considering M different options of boosting parameters, then we evaluate each of the M choices during the modeling phase of the AdaPT search. At step t, we divide the data into K folds preserving the relative proportions of masked and unmasked hypotheses. Then for each set of boosting parameters $m = 1, ..., M$, and for each fold $k = 1, ...K$: (1) apply EM-algorithm to estimate $\hat{\pi}_1^{(m)}(x_i)$ and $\hat{\mu}^{(m)}(x_i)$ using parameters m with data from folds $\{1, \ldots, K\} \backslash \{k\}$, and (2) compute expected-loglikelihood $\tilde{l}_k^{(m)}$ on hold-out set k using two-groups model parameters from m following convergence, and compute total across folds as $\tilde{l}_m = \sum_{k=1}^{K} \tilde{l}_k^{(m)}$. Finally we use the set of parameters $m^* = \arg\max_m \tilde{l}^{(m)}$ in another instance of the EM algorithm to estimate $\hat{\pi}_1^{(m^*)}(x_i)$ and $\hat{\mu}^{(m^*)}(x_i)$ on all data.

2.4.3 Computational aspects of AdaPT

Practical decisions are necessary to implement the AdaPT search. In addition to the covariates and p-values $(x_i, p_{t,i})_{i \in [n]}$, an initial rejection threshold $s_0(x_i)$ is required to begin the search. Rather begin the search with a high starting threshold, such as $s_0^* = 0.45$ recommended by [Lei and Fithian, 2018], we instead begin the AdaPT search with $s_0^* = 0.05$. Our decision to lower the starting threshold is advantageous for multiple reasons. First, intuitively, this starts our search in the regime of interest for target level $\alpha = 0.05$, whereas we would not expect to detect discoveries with larger p-values using this flexible multiple testing correction. Additionally, by lowering the starting threshold, more true information is available to the gradient boosted trees at the start of the AdaPT search. For instance, with the set of BrainVar eSNPs, 21,248 true p-values are immediately revealed with $s_0^* = 0.05$ as compared to only 2,290 when $s_0^* = 0.45$. Simulations detailed in *Method Appendix* show that on average our choice for using a lower threshold results in higher power.

The most computationally intensive part of the procedure is updating the rejection threshold via the EM algorithm. Instead of updating the model for estimating $\text{fdr}_{t,i}$ at each step of the search, we re-estimate every $[n/20]$ steps as recommended by [Lei and Fithian, 2018].

However, the inclusion of the previously described K-fold CV procedure (we use $K = 5$) for tuning the gradient boosted trees obviously adds computational complexity to the AdaPT search, and would be expensive to apply every time the model fitting takes place. Rather, we apply the CV step once at the beginning, and then another time half-way through the search based on the similarity of simulation performance with varying number of CV steps in *Method Appendix*. Additionally, one needs to choose the potential M model parameter choices. Technically, unique combinations can be used for both models, π_1 and μ, but for simplicity we only consider matching settings for both models, i.e. both models have the same number of trees and maximum depth. As a reminder, AdaPT guarantees finite-sample FDR control **regardless** of potentially over-fitting to the data when using the CV procedure. Simulations are provided in *Method Appendix* showing how extensively increasing the number of trees P leads to decreasing power, but maintains valid FDR control.

2.4.4 Code availability

We provide a modified version of the `adaptMT R` package to implement the AdaPT-CV tuning steps with XGBoost models at `https://github.com/ryurko/adaptMT`, and provide all code used to generate the manuscript's results at `https://github.com/ryurko/AdaPT-GWAS-manuscript-code`.

2.5 Method Appendix

2.5.1 AdaPT conditional two-groups model

This section provides a more detailed explanation of updating the rejection threshold $s_t(x_i)$ in the AdaPT procedure, expanding on the description from *Methods* in the main manuscript. As in the main text, this is essentially an explanation of the EM approach of Lei18. Note that for coherence some text is repeated from the main manuscript. Lei18 use a conditional version of the classical two-groups model Efron01 yielding the conditional mixture density,

$$f(p|x) = \pi_1(x) f_1(p|x) + 1 - \pi_1(x), \tag{2.3}$$

where the null p-values are modeled as uniform ($f_0(p|x) \equiv 1$). They proceed to use a *conservative* estimate for the conditional local false discovery rate, $\text{fdr}(p|x) = \hat{f}(1|x)/\hat{f}(p|x)$, by setting $1 - \pi_1(x) = f(1|x)$.

We model the non-null p-value density with a beta distribution density parametrized by μ_i,

$$f_1(p|x_i) = h(p; \mu_i) = \frac{1}{\mu_i} p^{1/\mu_i - 1}, \tag{2.4}$$

where $\mu_i = \mathbb{E}[-\log(p_i)]$, resulting in a conditional density for a beta mixture model,

$$f(p|x_i) = \pi_1(x_i) \frac{1}{\mu_i} p^{1/\mu_i - 1} + 1 - \pi_1(x_i). \tag{2.5}$$

In this form, we can model the non-null probability $\pi_1(x_i) = \mathbb{E}[H_i|x_i]$ and the effect size for non-null hypotheses $\mu(x_i) = \mathbb{E}[-\log(p_i)|x_i, H_i = 1]$ with two separate gradient boosted tree-based models. The XGBoost library Chen16 provides logistic and Gamma regression implementations which we use for $\pi_1(x_i)$ and $\mu(x_i)$ respectively.

There are two categories of missing values in these regression problems: H_i is never observed, and at each step t of the search, the p-values for tests $\{i : p_i \leq s_t(x_i) \text{ or } p_i \geq 1 - s_t(x_i)\}$ are masked as $\tilde{p}_{t,i}$. An expectation-maximization (EM) algorithm can be used to estimate both $\hat{\pi}_1(x_i)$ and $\hat{\mu}(x_i)$ by maximizing the partially observed likelihood. The complete log-likelihood for the conditional two-groups model is,

$$l(\pi_1, \mu; p, H, x) = \sum_{i=1}^{n} \{H_i \log(\pi_1(x_i) + (1 - H_i)\log(1 - \pi_1(x_i))\} + \sum_{i=1}^{n} H_i \log\{h(p_i; \mu(x_i))\}. \tag{2.6}$$

During the E-step of the $d = 0, 1, \ldots$ iteration of the EM algorithm, conditional on the partially observed data fixed at step t, $(x_i, \tilde{p}_{t,i})_{i\in[n]}$, we compute both,

$$\hat{H}_i^{(d)} = \mathbb{E}_{\hat{\pi}_1^{(d-1)}, \hat{\mu}^{(d-1)}}[H_i \big| (x_i, \tilde{p}_{t,i})_{i\in[n]}] \tag{2.7}$$

$$\hat{b}_i^{(d)} = \mathbb{E}_{\hat{\pi}_1^{(d-1)}, \hat{\mu}^{(d-1)}}[\mathbf{1}(p'_{t,i} = p_i) \big| (x_i, \tilde{p}_{t,i})_{i\in[n]}, H_i = 1], \tag{2.8}$$

where $\hat{b}_i^{(d)}$ indicates how likely $p'_{t,i} = \min(\tilde{p}_{t,i})$ equals p_i for non-null hypotheses. The explicit calculations of $\hat{H}_i^{(d)}$ and $\hat{b}_i^{(d)}$ for both the revealed, $\tilde{p}_{t,i} = p'_{t,i}$, and masked p-values, $\tilde{p}_{t,i} = \{p_i, 1 - p_i\}$, are available in the supplementary materials of Lei18.

The M-step consists of estimating $\hat{\pi}_1^{(d)}$ and $\hat{\mu}^{(d)}$ with separate gradient boosted trees, using *pseudo*-datasets to handle the partially masked data. In order to fit the model for $\pi_1(x_i)$, we construct the response vector $y_\pi^{(d)} = (1, \ldots, 1, 0, \ldots, 0) \in \mathbb{R}^{2n}$ and use weights $w_\pi^{(d)} = (\hat{H}_1^{(d)}, \ldots, \hat{H}_n^{(d)}, 1-\hat{H}_1^{(d)}, \ldots, 1-\hat{H}_n^{(d)}) \in \mathbb{R}^{2n}$. Then we estimate $\hat{\pi}_1^{(d)}(x_i)$ using the first n predictions from a classification model using $y_\pi^{(d)}$ as the response variable with the covariate matrix $(x_i)_{i\in[n]}$ replicated twice and weights $w_\pi^{(d)}$. Similarly, for estimating $\hat{\mu}^{(d)}(x_i)$ we construct a response vector $y_\mu^{(d)} = (-\log(p_1), \ldots, -\log(p_n), -\log(1-p_1), \ldots, -\log(1-p_n)) \in \mathbb{R}^{2n}$ with weights $w_\mu^{(d)} = (\hat{b}_1^{(d)}, \ldots, \hat{b}_n^{(d)}, 1 - \hat{b}_1^{(d)}, \ldots, 1 - \hat{b}_n^{(d)}) \in \mathbb{R}^{2n}$, and again take the first n predicted values using the duplicated covariate matrix.

The conditional local fdr is estimated for each $p'_{t,i}$,

$$\text{fdr}_{t,i} = \frac{\hat{\pi}_1(x_i) h(1; \hat{\mu}(x_i) + 1 - \hat{\pi}_1(x_i)}{\hat{\pi}_1(x_i) h(p'_{t,i}; \hat{\mu}(x_i) + 1 - \hat{\pi}_1(x_i)}, \tag{2.9}$$

and we follow the procedure detailed in Section 4.3 of Lei18 to update the rejection threshold to $s_{t+1}(x_i)$ by removing test $i^* = \arg\max_{i \in \mathcal{R}_t} \text{fdr}_{t,i}$ from $\mathcal{R}_t$. A summary diagram of the EM algorithm is displayed in Figure 2.6.

2.5.2 SCZ results with independent loci

One potential concern regarding the assessment of performance of AdaPT is the impact of linkage disequilibrium (LD). In the Manhattan plots of Figures 2.2(B-C), the discoveries visually appear to be located close to one another. However, the visual appearance of genomic positions is somewhat misleading because our initial selection of eSNPs greatly reduces the number of SNPs commonly portrayed in Manhattan plots – many of these SNPs are not very close to each other in the genome and not in high LD, although the format of the Manhattan plot makes this feature hard to see. To take this analysis further we follow common practice for GWAS results by identifying the "best" or "lead" SNPs in a LD block/cluster, using a similar approach as scz14, for each of the set of discoveries presented in Figure 2.2:

1. order the SNPs by the AdaPT -$\log_{10}$(q-value) in descending order,
2. starting with the SNP with the largest value for the AdaPT -$\log_{10}$(q-value),
    - remove all SNPs with $r^2 \geq 0.1$ within a 500kb window,
    - move on to next SNP that is still remaining,
3. return the retained SNPs as the LD-independent SNPs in low LD ($r^2 < 0.1$). (Remark: this approach excludes SNPs whose contribution to the GWAS signal is partially independent of the lead SNP, but it has the advantage of simplicity.)

We use the reference European sample genotype data from the 1000 Genomes project 1000G to compute the r^2 values between SNPs. In the GWAS setting this LD clumping procedure is typically applied to the reported SNP p-values, but because the ordering of SNPs varies between the different sets of discoveries (intercept-only versus use of covariates) we perform the operation separately with their respective q-values. For each of the different set of covariates considered, this results in reducing the 25,076 selected eSNPs down to the following number of "independent loci":

- Intercept-only: 3,958
- BD z-stats: 3,966
- BD z-stats + eQTL slopes: 3,962

- BD z-stats + eQTL slopes + WGCNA (w/ interactions): 3,963
- BD z-stats + eQTL slopes + WGCNA (w/o interactions): 3,959
- WGCNA: 3,954

The differences in counts are due to the different number of ties that take place between the resulting q-values for each considered set of covariates. Next, for the identified set of "lead" SNPs we observe how many have q-values less than the target FDR level $\alpha = 0.05$ (i.e. associations detected at $\alpha = 0.05$). The results are displayed in the Figure 2.7, including Manhattan plots Figures 2.7(A-B) of the q-values for the AdaPT intercept-only and BD z-stats + eQTL slopes + WGCNA (w/ interactions) results, rather than using the actual p-values. The lead SNPs in each of the Manhattan plots are denoted by an X shape. In conjunction with Figures 2.7(C-D), the relative improvement in the set of independent loci within the discovery sets from AdaPT is analogous to the results presented in Figure 2.2, emphasizing the advantage of accounting for covariates and their interactions via gradient boosted trees. Additionally, Figure 2.8 further emphasizes that the improvement in power is not restricted to a particular section of the genome. As seen in Figure 2.9, we observe a similar improvement in the number of independent loci when ordering the SNPs with the observed 2014-only studies SCZ p-values.

While we maintain FDR control on the original set of discoveries (see Figure 2.3 in *Results*), we do not retain any guarantees regarding the detected independent loci presented in Figure 2.7. In order to maintain FDR control on the set of discovered independent loci, an alternative approach or adjustment to the AdaPT algorithm is required. A simple alternative is to first apply LD pruning/clumping as initial step prior to applying AdaPT to a reduced set of lead SNPs. However, this encounters the challenge of defining lead SNPs without data "snooping" based on using the observed p-values. Future work will explore modifications for AdaPT, potentially exploring recent developments ren2020knockoffs, to maintain FDR control on an independent subset of SNPs.

2.5.3 SCZ variable importance and partial dependence

We explore further the variable relationships from the gradient boosted trees. First, Figure 2.10 displays the change in variable importance for the non-null effect size (μ) at each model fitting iteration, with the top variables in the final model highlighted. The variable importance measures are relatively stable across all model iterations with the BD z-statistics and eQTL slope measures maintaining the highest level of importance. Figure 2.11 displays the partial-dependence plot at each AdaPT model fitting iteration for the estimated marginal relationship between the BD z-statistics and the non-null effect size μ, evaluated at the $0, 2.5\%, 5\%, \ldots, 100\%$ percentiles. The estimates reveal an increasing effect size as the BD z-statistics grow in magnitude, which is relatively stable across the model iterations. Figures

2.12(A-C) display the relationships for the probability of non-null model, while (D-F) display relationships for the effect size under the alternative. Although the partial dependence plots show considerable variability due to the high dimensional of the model, we can still see general trends consistent with the variable importance plots from Figure 2.3(A) and Figure 2.10.

In Figure 2.13 we display the p-value distributions comparing the enrichment for membership in the different WGCNA modules reported by werling2020whole. While many of the WGCNA modules lack clear evidence or contain too few eSNPs, as denoted by their respective y-axes, the *cyan* and *salmon* modules display noticeable enrichment. Additionally, as mentioned previously, membership in the *gray* module displays a lack of enrichment versus no associated cis-eQTL gene affiliated with the unassigned WGCNA module.

As additional context for the improved performance from using all covariates with interactions, Figures 2.14(A-B) display the change in partial dependence between the BD z-statistics and probability of being non-null π_1 across the AdaPT search for the AdaPT results using (A) BD z-statistics only and (B) all covariates without interactions. When compared to the results using all covariates with interactions in Figure Figure 2.3(B), we see that both versions of these results display relatively flat relationships near the end of the AdaPT search. This provides evidence of the importance of the interactions between other covariates and the BD z-statistics in retaining discriminatory power of the eSNPs near the end of the AdaPT search.

2.5.4 Replication simulations

We use simulations to empirically assess the observed nominal replication rate, percentage of discoveries with p-values less than 0.05 in holdout *2018-only* studies, of 55.2% for the 843 SCZ discoveries from the *2014-only* studies at target FDR level $\alpha = 0.05$. We use the final non-null effect size model returned by the AdaPT, $\hat{\mu}^*$, to generate simulated p-values $\boldsymbol{p}^{sim}$ and nominal replication rates to compare the observed rate against. For the simulations, we assume that all 843 SCZ discoveries from the *2014-only* studies are truly non-null, and we use the actual eSNPs, their observed standard errors σ_{14}, σ_{18} from the *2014-only* and *2018-only* studies respectively, as well as their actual covariates for generating $\boldsymbol{p}^{sim}$. A single iteration of the simulation proceeds as follows:

- For each of the $R_{\text{SCZ}} = 843$ discoveries $i \in \mathcal{R}_{\text{SCZ}}$:
  1. Assume test status is non-null: $H_i = 1$.
  2. Generate effect size using final AdaPT model as truth:
  $$-\log p_i^{sim} | x_i^{\text{SCZ}} \sim \text{Exp}\big(1/\hat{\mu}^*(x_i^{\text{SCZ}})\big). \tag{2.10}$$
  3. Transform effect sizes to p-value p_i^{sim}.

4. Convert simulated p-value to z-statistic $z_i^{sim} = \left|\Phi^{-1}(p_i^{sim}/2)\right|$.
5. Calculate updated z-statistic to reflect observed reduction in standard error for *2018-only* studies relative to *2014-only*,

$$z_i^{*,sim} = z_i^{sim} \cdot \frac{\sigma_{14}}{\sigma_{18}}. \tag{2.11}$$

6. Convert updated z-statistic to p-value:

$$p_i^{*,sim} = 2 \cdot \Phi(-|z_i^{*,sim}|). \tag{2.12}$$

- Calculate nominal replication rate using $\boldsymbol{p}^{sim} = (p_i^{*,sim}, \ldots, p_{R_{\text{SCZ}}}^{*,sim})$,

$$\text{Nominal replication rate } = \frac{|\{i : p_i^{*,sim} \leq .05\}|}{R_{\text{SCZ}}}. \tag{2.13}$$

We repeat this process to generate ten-thousand simulated values for the nominal replication rate. The distribution of the simulated values ranges from approximately 51% to 63%, with an average and median of $\approx 57\%$, close to the observed rate of 55.2%. Obviously, assuming that all of the 843 rejections are truly non-null is an overtly optimistic assumption given the use of FDR error control. Thus, the average simulated nominal replication rate of $\approx 56.6\%$ is reassuringly close to the observed rate and likely higher than what would be expected if false discoveries were accounted for among the 843 considered eSNPs.

2.5.5 SCZ results with *all 2018* studies

We generate the AdaPT results using the SCZ p-values from *all-2018* studies to the same set of $n_{\text{SCZ}} = 25,076$ eSNPs with the same covariates x_i^{SCZ}. As a comparison to the results displayed in Figure 2.2 using the *2014-only* studies, Figures 2.15(A-D) display the same figures but with the results from *all 2018* at target FDR level $\alpha = 0.05$. In contrast to before, we see that due to the increase in power from the study size, the use of modeling the auxiliary information provides a much smaller increase in power with just an approximately 19% increase in discoveries from the intercept-only results (1,865 discoveries) to using all twenty-four covariates with interactions (2,228 discoveries).

For comparison, we additionally examine the change in variable importance and partial dependence plots returned by AdaPT using *all 2018* studies. Similar to before, Figures 2.16(A-B) display the change in variable importance plots for both the probability of being non-null π_1 and effect size under alternative μ models using the SCZ p-values from *all 2018* studies respectively. The results are similar to before, but with the complete sample eQTL slopes possessing the highest importance. The BD z-statistics are again highly important for *all 2018* studies, displaying the similarly increasing relationships across the AdaPT

models as seen in the partial dependence plots in Figures 2.17(C-D). The change in partial dependence plots for the different eQTL slopes summaries are seen in Figures 2.18(A-F). Figure 2.19 displays the levels of SCZ enrichment for *all 2018* studies, revealing modules that are consistent with the *2014-only* studies such as *cyan* and *salmon*.

2.5.6 Type 2 diabetes results

Using GWAS summary statistics for type 2 diabetes (T2D), unadjusted for BMI, available from Diabetes Genetics Replication And Meta-analysis (DIAGRAM) consortium Mahajan18, we applied our full pipeline outlined in Figure 2.1. Of the initial set of over twenty-three million SNPs available, we identified 176,246 eSNPs from eQTL variant-gene pairs from any GTEx tissue sample using the definition of the GTEx eSNPs explained in *Data*. Figure 2.20 displays the enrichment for these GTEx eSNPs compared to the original set of SNPs from the T2D GWAS results.

We create a vector of covariates x_i^{T2D} summarizing expression level information from GTEx for pancreas, liver, and two adipose tissues, *subcutaneous* and *visceral (omentum)*. Specifically, we calculate $\tilde{\beta}_i^{r^{\text{T2D}}}$ for each r^{T2D} in the set of tissues: pancreas, liver, adipose - subcutaneous, adipose - visceral (omentum). Additionally, we generate WGCNA module assignments using protein coding genes for pancreas samples from GTEx. To generate the WGCNA results, we only consider protein coding genes identified using the `grex` package in `R` grex, rlang. Additionally, all genes with expression levels of zero for over half of the provided samples were removed. This resulted in fourteen different module, including the unassigned *gray* module. Unlike the SCZ application, we do not use independent GWAS results from another phenotype.

Using x_i^{T2D} defined above, we applied AdaPT to the 176,246 GTEx eSNPs. However, we encountered an issue for this data where we were unable to discover any hypotheses at target FDR level $\alpha \leq 0.05$. This was due to the fact that 640 eSNPs had p-values *exactly* equal to one. While this can understandably occur with publicly available GWAS summary statistics, p-values equal to one will then *always* contribute to the *pseudo*-estimate for the number of false discoveries A_t during the AdaPT search (see *Methodology overview*). With a relatively high number of p-values equal to one, AdaPT is unable to search through rejection sets for lower α values. To overcome this challenge, we draw random replacement p-values for the 640 eSNPs from a uniform distribution between 0.97 and $1 - 1\text{E}^{-15}$, a value strictly less than one, to allow some leeway. We refer to this set of p-values as *adjusted*, while the original observed p-values are *unadjusted*. For comparison, Figure 2.21 shows the difference in the number of discoveries for the *adjusted* and *unadjusted* p-values across different target α values. Due to the similarity in performance for α values greater than 0.1, we use results for the *adjusted* p-values moving forward.

At target FDR level $\alpha = 0.05$, AdaPT yields 14,920 T2D discoveries using the *adjusted*

p-values with covariates x_i^{T2D} (compared to 14,693 intercept-only discoveries). The change in variable importance for the T2D AdaPT models are displayed in Figure 2.22. This set of eSNPs is associated with 5,970 cis-eQTL genes for which we then applied gene ontology enrichment analysis to Ashburner00, GO18, identifying the gene enrichment for biological processes displayed in Figure 2.5.

2.5.7 BMI results

We also applied our pipeline of analysis to BMI, unadjusted for waist-to-hip ratio (WHR), using GWAS results for individuals of European ancestry available from the GIANT Consortium. Specifically, we approached BMI in the same manner as SCZ: apply AdaPT to GWAS results from earlier studies with a sample size of 322,154 individuals Locke15; then compare the nominal replication results on recently conducted studies with a sample size of approximately 700,000 individuals Yengo18. As before, all of the *2015-only* studies from Locke15 were included as a subset of *all 2018* studies Yengo18. Because both Locke15 and Yengo18 use the inverse variance-weighted fixed effects approach for meta-analysis, we then compute statistics for the studies exclusive to *2018-only* studies in Yengo18. Additionally, to make this example more comparable to the SCZ use, we also use GWAS results for WHR Shungin15 as a covariate (analogous to BD for SCZ). Following pre-processing steps (matching SNPs across studies and effect alleles in both WHR and BMI), we identified 47,690 GTEx eSNPs from a set of nearly two million SNPs, based on the definition explained in *Data*. Figure 2.23 displays the enrichment for the GTEx eSNPs compared to the original set of pre-processed SNPs for the *2015-only* studies.

Based on previous knowledge of BMI tissue expression associations Locke15, we create a vector of covariates x_i^{BMI} summarizing expression level information from GTEx for brain and adipose tissues (both *subcutaneous* and *visceral (omentum)*). Specifically, we calculate $\tilde{\beta}_i^{r^{\text{BMI}}}$ for each $r^{\text{BMI}} \in \{\text{GTEx brain tissues, adipose - subcutaneous, adipose - visceral (omentum)}\}$, where we consider the following brain tissues: (1) *amygdala*, (2) *anterior cingulate cortex BA24*, (3) *caudate basal ganglia*, (4) *cerebellar hemisphere*, (5) *frontal cortex BA9*, (6) *hippocampus*, (7) *hypothalamus*, (8) *nucleus accumbens basal ganglia*, (9) *putamen basal ganglia*, (10) *spinal cord cervical c-1*, and (11) *substantia nigra*. We do not consider the available *cerebellum cortex* tissue samples from GTEx as these are duplicates of *cerebellar hemisphere* and *frontal cortex BA9* respectively. We instead only use the samples taken the same time as the other brain sub-regions at the University of Miami Brain Endowment Bank, preserved by snap freezing (see GTEx FAQs).

We also created an aggregate across $\mathcal{G}_i^{r^{\text{nc}}}$, all cis-eQTL genes associated with eSNP i for each non-cerebellar hemisphere brain tissue region r^{nc},

$$\bar{\beta}_i^{\text{nc}} = \frac{1}{|\mathcal{G}_i^{r^{\text{nc}}}|} \sum_{g \in \mathcal{G}_i^{r^{\text{nc}}}} |\beta_{i,g}^{r^{\text{nc}}}|. \tag{2.14}$$

We did not include the cerebellum tissue samples in this aggregate due to the reported distinctness of the cerebellum relative to other brain tissue samples GTEx2015. Similarly, we computed an average across the two adipose tissues. As before, when calculating the various eQTL slopes summaries, if eSNP i was not an eQTL for a particular region then we impute a value of zero reflecting the lack of associated expression.

Furthermore, WGCNA module assignments were generated using protein coding genes for three different sets of tissues: (1) all non-cerebellar hemisphere brain tissues, (2) cerebellar hemisphere only tissue, and (3) adipose tissues (using same settings described previously in *Type 2 diabetes results*). Together with the WHR z-statistics and covariates accounting for the associations and WGCNA module indicators, x_i^{BMI} contained 110 variables.

For BMI eSPS, 376 have p-value exactly equal to one, leading to the same problem as we encountered in the T2D analysis. Again, we proceed by randomly drawing replacement p-values for these 376 eSNPs from a uniform distribution between 0.97 and $1 - 1\text{E}^{-15}$. Figure 2.24 shows how AdaPT fails to obtain any discoveries across the various α levels without making an adjustment to the p-values. With this limitation recognized, we proceed to focus on the discoveries returned by AdaPT using the adjusted p-values at $\alpha = 0.05$.

Unlike SCZ and T2D, AdaPT using all of the covariates (with the same tuning parameters as SCZ) detected fewer discoveries: 1,383 eSNPs compared to 1,624 eSNPs discovered by the intercept-only AdaPT model at target FDR level $\alpha = 0.05$. With further boosting regularization, beyond what is considered here, one could achieve the intercept-only results with gradient boosted trees. Of these 1,383 discoveries, approximately 83% (1,140 eSNPs) were nominal replications with p-values less than or equal to 0.05 in the independent *2018-only* studies. Figure 2.25 displays the increasing smoothing spline relationship between the *2018-only* p-values and the resulting *2015-only* q-values from the AdaPT search on the $\log_10$ scale. The much higher observed nominal replication rate is not surprising given the well powered size of the BMI studies, as indicated by the y-axis of Figure 2.25, which reflects the level of enrichment for the *2018-only* studies.

Additionally, gene ontology enrichment analysis for the 1,383 discoveries using all covariates revealed no significant biological process enrichment at target FDR level $\alpha = 0.05$. One concern is that a model with 110 variables is excessive, because the variable importance plots for the BMI AdaPT models in Figures 2.26(A-B), along with the partial dependence plots in Figures 2.27(A-B), emphasize the relative importance of the WHR z-statistics compared to other covariates. To test this conjecture, we explored two simpler models using (1) WHR z-statistics only and (2) WHR z-statistics with eQTL slope summaries. These produced 1,324 and 1,351 discoveries at the 0.05 level, respectively. We conclude that the available covariates do not provide sufficient additional information beyond the signal available with this immense sample and consequently including covariates in the AdaPT model does not increase the power of the procedure.

2.5.8 CV tuning for SCZ, T2D, and BMI results

Rather than fixing the parameter settings for the XGBoost gradient boosted trees, we use the CV algorithm (detailed in *Methods*) at two steps of the search to tune the models (see the following section for justification of using two CV steps). For our search space, we evaluate a small range of values for the number of trees P and limit the maximum tree depth D to result in reasonably shallow trees (referred to as `nrounds` and `max_depth` in the `xgboost` package xgboost19).

First, for SCZ analysis, when exploring the improvement in discovery rate for the eSNPs by incrementally including more information, we used the following XGBoost settings:

- BD z-stats: Combinations of $P \in \{100, 150\}$, $D \in \{1, 6\}$,
- BD z-stats + eQTL slopes: Combinations of $P \in \{100, 150\}$, $D \in \{3, 6\}$,
- BD z-stats + eQTL slopes + WGCNA: Combinations of $P \in \{100, 150\}$, $D \in \{2, 3\}$,
- WGCNA only: Combinations of $P \in \{100, 150\}$, $D \in \{1, 2, 3\}$.

We explored different settings for the different possible covariates to address the types of variables included. For instance, when using the BD z-statistics only, we considered both single-split "stumps" as well as more depth with six splits to potentially handle the variable's symmetric relationship. Once we have all three types of covariates (BD z-statistics, eQTL slope summaries, and WGCNA results), we limit the maximum depth to be at least two to ensure possible interactions can be captured.

The selected number of trees P and maximum depth D for each of these sets of covariates is displayed in Table 2.1. When using only the BD z-statistics, as well as only including the eQTL slopes, the single-split settings were selected in the first CV step while the higher depth was selected in the second CV step. When using all covariates, the most complex settings (largest number of trees and largest depth) are selected in both CV steps. This agreement in selection is not surprising given the choice of the low starting threshold $s_0 = 0.05$, which differs from the results displayed in Table 2.3 of the next section using $s_0 = 0.45$. We evaluated the same possible settings for the various *all 2018* results displayed in Figures 2.15(C-D): the same choices for P and D displayed in Table 2.1 were selected in both CV steps.

For the T2D and BMI results with their full set of covariates, we evaluated four combinations: (1) $P = 100$, $D = 2$, (2) $P = 150$, $D = 2$, (3) $P = 100$, $D = 3$, and (4) $P = 150$, $D = 3$. For the BMI results using only WHR z-statistics, we varied over $P \in \{100, 150\}$ and $D \in \{1, 6\}$; for the results using WHR z-statistics with the eQTL slopes, we used combinations of $P \in \{100, 150\}$, $D \in \{3, 6\}$. The selected number of trees P and

maximum depth D for each of these sets of AdaPT results at both CV steps is displayed in Table 2.2.

2.5.9 Selection of s_0 and number of CV steps

To justify the selection of both the starting threshold s_0 and number of CV steps for the AdaPT search, we generated simulations from the first AdaPT models returned from the SCZ *2014-only* results. While these models are based on AdaPT results with a starting threshold of $s_0 = 0.05$ following one CV step, they are only from the first model and are not explicitly parametrized by s_0 and the number of CV steps. We know, however, that these first models are the result of using $P = 150$ trees with a maximum depth of $D = 3$, as indicated in Table 2.1 of the previous section.

Let $\hat{\pi}_1^*$ and $\hat{\mu}^*$ be the first models for the probability of non-null and effect size under the alternative that AdaPT returns for the eSNPs using all covariates x_i^{SCZ}. We use these models as the "truth" for generating data, in which a single iteration of the simulation proceeds as follows:

- For each eSNP $i \in [n^*_{\text{SCZ}}]$
  1. Generate test status: $H_i | x_i^{\text{SCZ}} \sim \text{Bernoulli}\big(\hat{\pi}_1^*(x_i^{\text{SCZ}})\big)$.
  2. Generate simulated effect sizes:
  $$-\log p_i | H_i, x_i^{\text{SCZ}} \sim \begin{cases} \text{Exp}\big(1\big) \text{ if } H_i = 0, \\ \text{Exp}\big(1/\hat{\mu}^*(x_i^{\text{SCZ}})\big) \text{ if } H_i = 1. \end{cases} \tag{2.15}$$
  3. Transform to p-values p_i.
- Apply AdaPT to simulated study p-values with specified s_0 and v CV steps with two candidate settings:
  1. number of trees $P = 100$ and maximum depth $D = 2$,
  2. number of trees $P = 150$ and maximum depth $D = 3$.
- Compute observed power and FDP at range of target FDR α values.

We generate one-hundred simulations this way for each possible threshold $s_0 \in \{0.05, 0.25, 0.45\}$ and $v \in \{1, 2, 5\}$ CV steps. Figure 2.28 displays the average difference in power between the different starting threshold values by the number of CV steps. Although the differences are small, we see that using $s_0 = 0.05$ results in higher power, on average, than both 0.25 and the recommended 0.45 value. Using this low starting threshold of $s_0 = 0.05$, we then directly compute the difference in power between the different number of CV steps displayed in

Figure 2.29. Unsurprisingly, while again the differences are small, only one CV step results in the lowest power, on average. Since the computational cost of AdaPT with CV tuning is reduced by only using two CV steps instead of a higher number, such as five, and the simulations demonstrate on average no difference in power at both α values of 0.05 and 0.10, we use the starting threshold of $s_0 = 0.05$ with two CV steps in our applications of AdaPT.

In the previous section, Table 2.1 displayed the selections in both CV steps with $s_0 = 0.05$. For comparison, Table 2.3 displays the selections using $s_0 = 0.45$. Instead of selecting the same settings in both steps, the higher initial threshold selects the least complex settings (smallest number of trees and minimum depth) in the first CV step before flipping to the most complex settings in the second step. Intuitively, the higher initial threshold means more information is masked from the models, so it is not surprising to see less complex settings chosen. This further reinforces the use of the lower initial threshold $s_0 = 0.05$: it starts with more revealed information and selects model settings corresponding to improved CV performance for tests with lower p-values of interest.

2.5.10 Dependent p-value block simulation

To demonstrate the performance of AdaPT in the presence of dependent tests, we construct simulations with a block-correlation scheme to emulate LD structure for SNPs. We consider a setting with two independent covariates,

$$x_i = (x_{i1}, x_{i2}),$$
$$\text{where } x_{i1}, x_{i2} \sim \text{Uniform}(0, 1).$$

For each test $i \in [n]$, we define a linear relationship for the log-odds of being non-null using these covariates,

$$\text{logit}(\pi_{1,i}(x_i)) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}.$$

Then, the resulting status of the test H_i is a Bernoulli random variable based on the probability $\pi_{1,i}(x_i)$ where $H_i = 1$ indicates the test i is non-null while $H_i = 0$ indicates a true null,

$$H_i \sim \text{Bernoulli}(\pi_{1,i}(x_i)).$$

Given this test status, a vector of true effect sizes $\boldsymbol{\mu} = c(\mu_i, \ldots, \mu_n)$ is also generated as a function of the covariates,

$$\mu_i(x_i) = \begin{cases} \max\{\mu_{floor},\ \gamma_1 x_{i1} + \gamma_2 x_{i2}\} \text{ if } H_i = 1, \\ 0 \text{ otherwise.} \end{cases}$$

To simulate observed effect sizes, we construct an $n \times n$ covariance matrix $\boldsymbol{\Sigma}$ with B blocks of equal size $\frac{n}{B}$. Each block $b \in [B]$ has constant correlation ρ between all tests *within* the block, while each block is independent of each other. This results in constructing

individual block covariance matrices, Σ_b, with ones along the diagonal and ρ for the off-diagonal elements. Each of these individual matrices are placed along the diagonal of $\mathbf{\Sigma}$, with the remaining off-diagonal elements set to zero so blocks are independent of each other. As an example, if each block contained only two tests they would be constructed in the following manner,

$$\Sigma_b = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \quad \Rightarrow \quad \mathbf{\Sigma} = \begin{bmatrix} \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \mathbf{0} & \dots & \dots & \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \end{bmatrix}$$

Using this block-wise construction of the covariance matrix, we then proceed to generate the vector of observed effect sizes $\boldsymbol{z} = (z_i, \dots, z_n)$ from a multivariate Gaussian distribution,

$$\boldsymbol{z} \sim \text{Normal}(\boldsymbol{\mu}, \mathbf{\Sigma}).$$

We compute the resulting two-side p-value $p_i = 2 \cdot \Phi(-|z_i|)$ for each test's observed effect size.

For each dataset generated using this process above, we compute both the observed FDP and power for the classical BH procedure and two different versions of AdaPT:

1. intercept-only,
2. gradient boosted trees with covariates: $x_i = (x_{i1}, x_{i2})$.

We fix both $n = 10{,}000$ and $B = 500$ blocks, resulting in 500 blocks of twenty tests each. Rather than force all non-nulls together in the same blocks, we first calculate the minimum number of blocks required to hold all non-null tests, $B_A^* = \lceil |\{i : H_i = 1\}|/20 \rceil$. The non-null tests are then randomly assigned to $B_A = \lceil (500 + B_A^*)/2 \rceil$ blocks, ensuring that there will be blocks containing both null and non-null tests. The $|\{i : H_i = 0\}|$ tests are randomly assigned to available spots within the B_A blocks as well as the remaining $500 - B_A$ strictly null blocks.

In our simulations, we fix $\beta_0 = -3$ and require that both $\beta_1 = \beta_2$ and $\gamma_1 = \gamma_2$. We vary the following settings in our simulations:

1. block correlation $\rho \in \{0, 0.25, 0.5, 0.75, 1\}$ where each block has the same value for ρ,

2. $\beta_1, \beta_2 \in \{1, 2, 3\}$,
3. $\mu_{floor} \in \{0.5, 1, 1.5\}$,
4. $\gamma_1, \gamma_2 \in \{0.5, .75, 1\}$.

We generate 100 simulations using the data generating process above, computing both the FDP and power for BH and the two different versions of AdaPT. For the covariate-informed version of AdaPT, we use gradient boosted trees via XGBoost with $P = 100$ trees and maximum depth $D = 1$. For both versions of AdaPT results, we start with the initial threshold of $s_0 = 0.45$ and update the model ten times throughout the search (rather than the recommended twenty for computational speed).

Figures 2.30, 2.31, and 2.32 display points for the average observed FDP and power across the 100 simulations with plus/minus two standard errors bars for μ_{floor} =0.5, 1, and 1.5 respectively, with target FDR level $\alpha = 0.05$. The columns in each figure correspond to the different values considered for $\gamma_1 = \gamma_2$, while the rows correspond to $\beta_1 = \beta_2$. The x-axis for the figures displays the increasing block correlation ρ. Regardless of the simulation setting, we see that the AdaPT results when accounting for covariates (x_{i1}, x_{i2}) maintains valid FDR control at 0.05 similar to BH. This holds in the settings with greater effect sizes, as well as when the covariate information displays the best performance in terms of observed power (the bottom right panels of each figure). We can see that the intercept-only approach fails to achieve FDR control under block settings with perfect correlation, while the use of covariate information appears to inhibits such behavior. Our focus on positive correlation values is synonymous with the setting faced in genomics regarding LD structure. Further exploration of AdaPT's performance in settings with arbitrary dependence structure presents an opportunity for future work, as well as accounting for covariate information that predict observed correlated noise.

2.5.11 Simulations demonstrating effects of overfitting

It is possible that flexible methods like gradient boosted trees can be overfit, especially on small data sets. This could potentially lead to concerns about their incorporation in AdaPT. To assess the effects of overfitting the gradient boosted trees in AdaPT, we constructed simulated datasets using the first models returned by AdaPT on the SCZ GWAS results, $\hat{\pi}_1^*$ and $\hat{\mu}^*$, with the actual covariates x_i^{SCZ} for each of the $n_{\text{SCZ}}^* = 25{,}076$ eSNPs. We then simulated data using these models in the same manner previously explained for choosing s_0 and the number of CV steps, and computed the observed power and FDP over a range of number of trees $P \in \{100, 300, 500, 700, 900\}$.

Figure 2.33(A) displays the distributions for fifty simulations of the observed FDP as the number of trees in the gradient boosted model increases. Regardless of the number of trees, we still maintain valid FDR control. However, Figure 2.33(B) shows as the number of trees

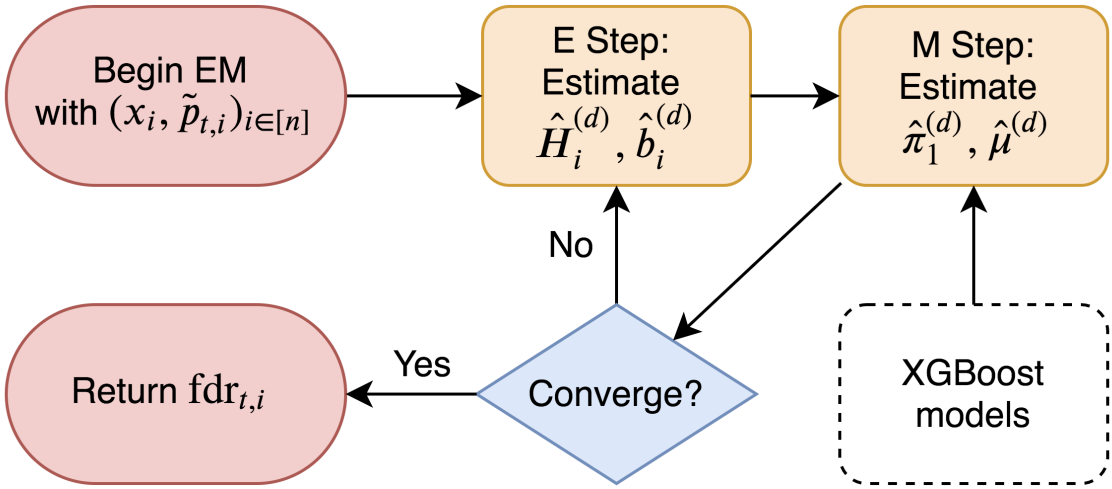

Figure 2.6: Summary of AdaPT EM algorithm.

increases, the method will overfit, resulting in a reduction in power. This reinforces that, although good model tuning can be important for power, the AdaPT method continues to maintain FDR control even as the model breaks down.

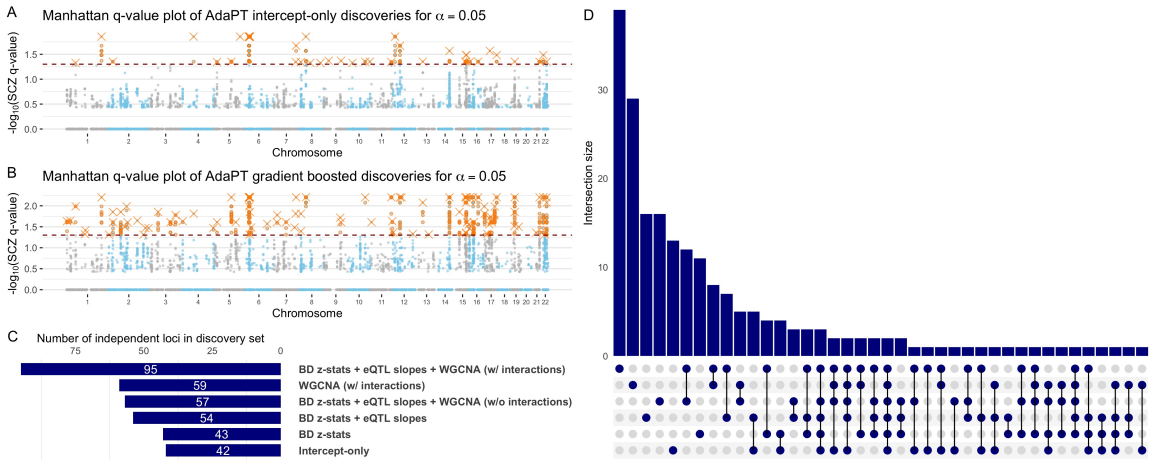

Figure 2.7: Manhattan q-value plots of SCZ AdaPT discoveries (orange) using *(A)* intercept-only model compared to *(B)* covariate informed model at target $\alpha = 0.05$, with lead SNPs for independent loci denoted by Xs. *(C)* Comparison of the number of independent loci for each discovery set at target $\alpha = 0.05$ based on LD pruning with the respective AdaPT q-values and *(D)* their resulting discovery set intersections.

Table 2.1: Selected boosting settings for number of trees P and maximum depth D with AdaPT CV algorithm by covariates for eSNPs in each CV step.

Covariates	m_1^*	m_2^*
BD z-stats	$P = 150,\ D = 1$	$P = 150,\ D = 6$
BD z-stats + eQTL slopes	$P = 150,\ D = 3$	$P = 150,\ D = 6$
BD z-stats + eQTL slopes + WGCNA	$P = 150,\ D = 3$	$P = 150,\ D = 3$
WGCNA only	$P = 150,\ D = 3$	$P = 150,\ D = 3$

Table 2.2: Selected boosting settings for number of trees P and maximum depth D with AdaPT CV algorithm by GWAS results in each CV step.

GWAS results	m_1^*	m_2^*
T2D	$P = 100,\ D = 2$	$P = 150,\ D = 3$
BMI (all covariates)	$P = 100,\ D = 2$	$P = 150,\ D = 3$
BMI (WHR z-stats only)	$P = 150,\ D = 1$	$P = 150,\ D = 1$
BMI (WHR z-stats + eQTL slopes)	$P = 100,\ D = 3$	$P = 150,\ D = 3$

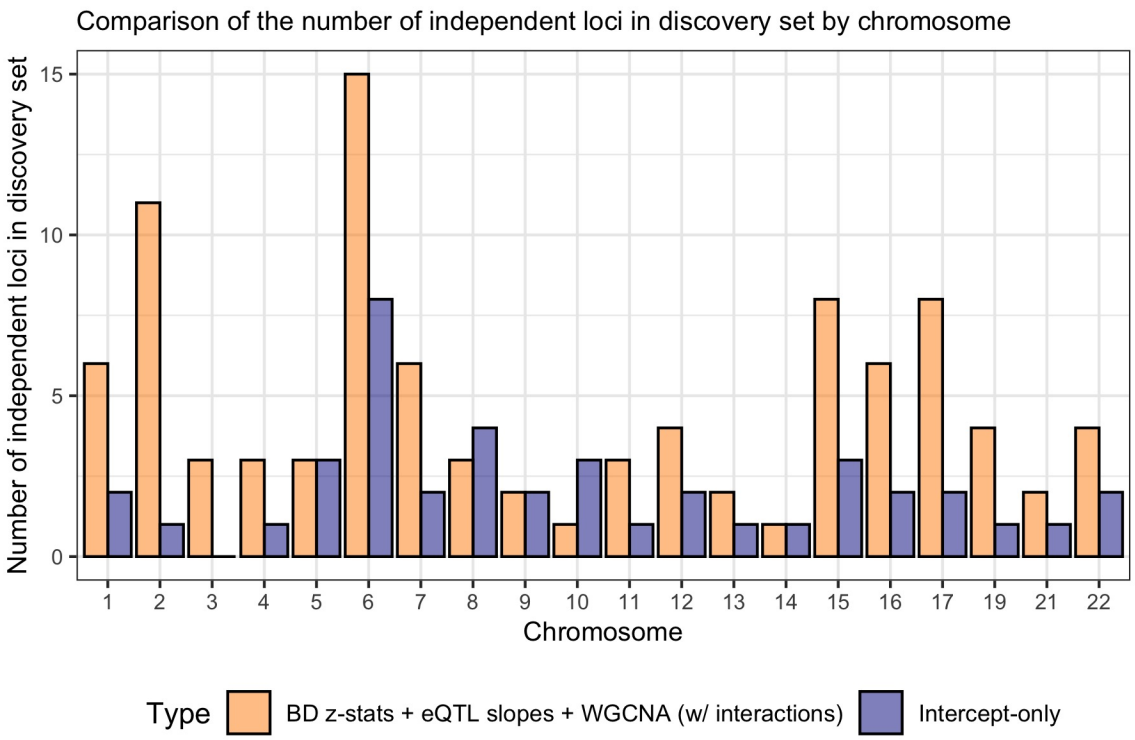

Figure 2.8: Comparison of the number of independent loci in the AdaPT discovery sets by type for each chromosome.

Table 2.3: Selected boosting settings for number of trees P and maximum depth D with AdaPT CV algorithm by covariates for eSNPs with $s_0 = 0.45$.

Covariates	m_1^*	m_2^*
BD z-stats	$P = 50,\ D = 1$	$P = 150,\ D = 1$
BD z-stats + eQTL slopes	$P = 100,\ D = 1$	$P = 150,\ D = 2$
BD z-stats + eQTL slopes + WGCNA	$P = 100,\ D = 2$	$P = 150,\ D = 3$
WGCNA only	$P = 150,\ D = 3$	$P = 150,\ D = 3$

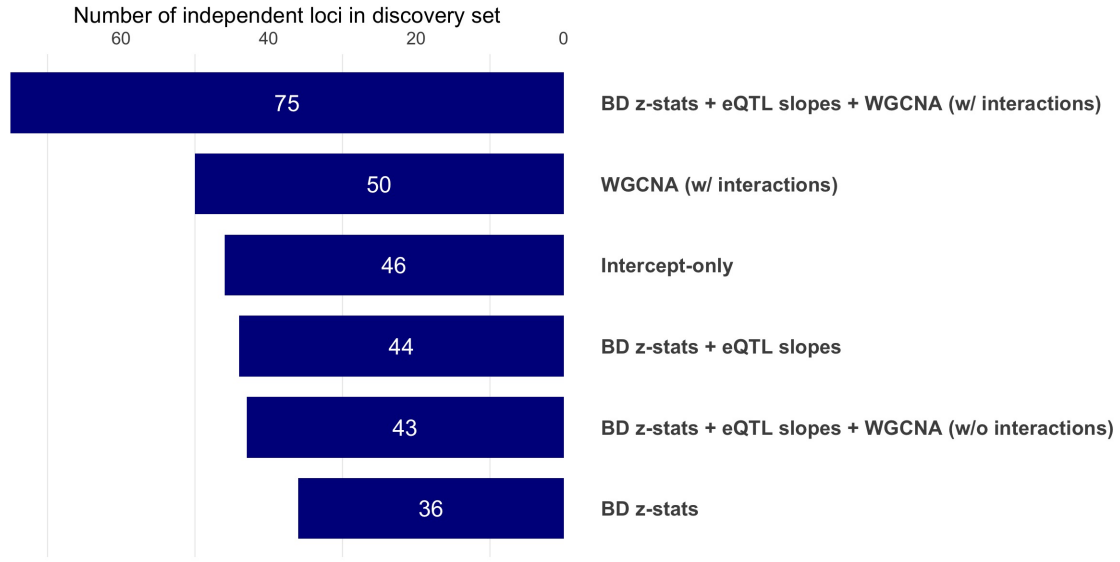

Figure 2.9: Comparison of the number of independent loci for each discovery set at target $\alpha = 0.05$, based on LD pruning with the with *2014-only* SCZ p-values.

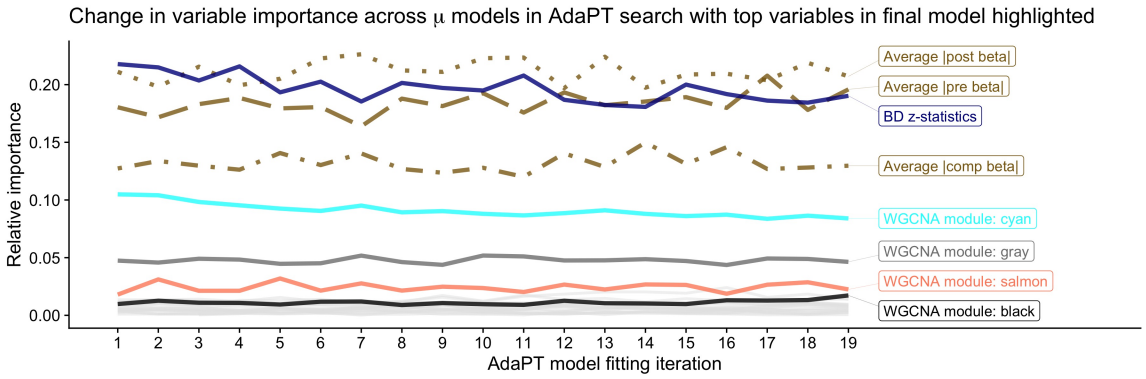

Figure 2.10: Change in variable importance for AdaPT non-null effect size μ model across search, with top variables in final model highlighted.

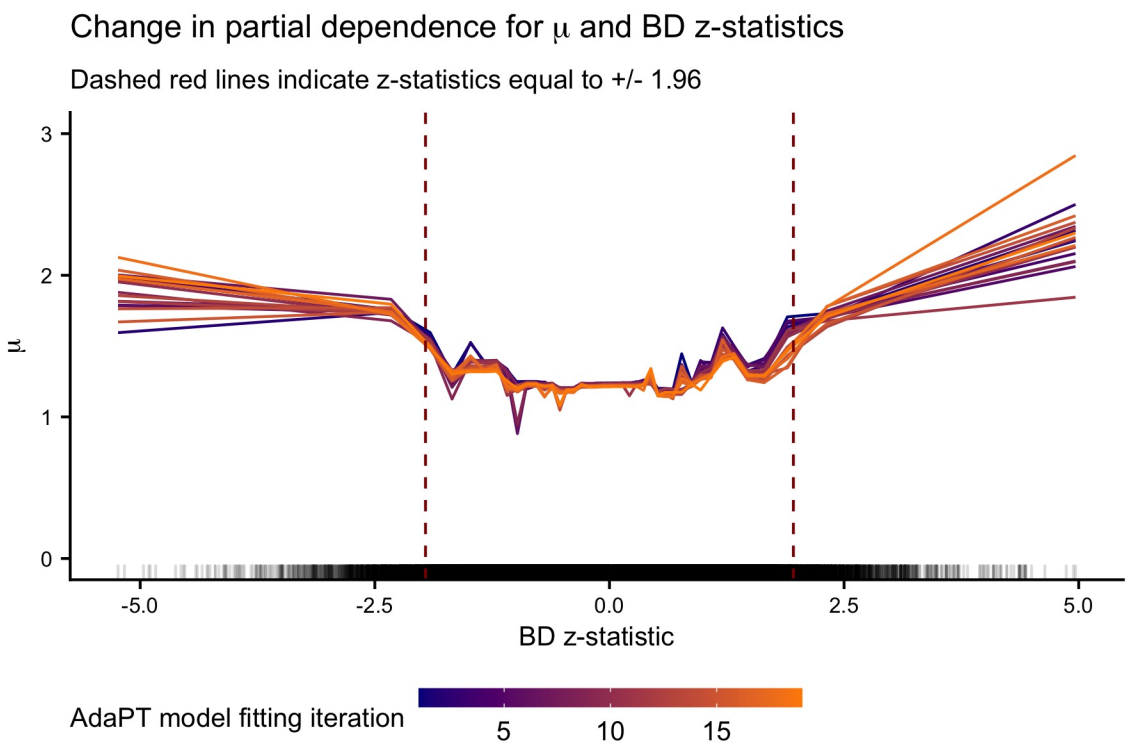

Figure 2.11: Change in partial dependence for non-null effect size μ and BD z-statistics across μ models in AdaPT search.

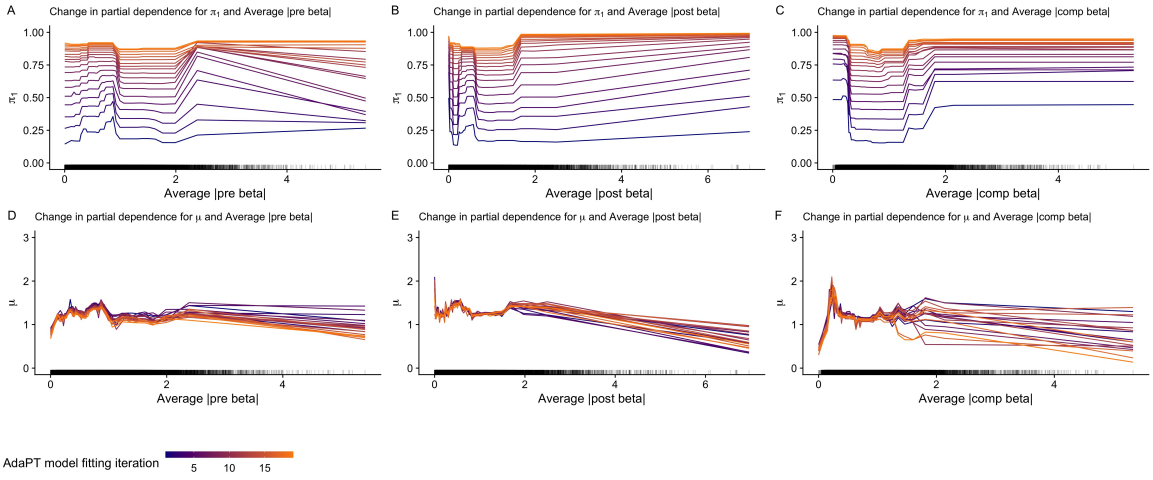

Figure 2.12: Change in partial dependence plots for probability of being non-null π_1 in *(A-C)*, and the effect size under alternative μ in *(D-F)*, for each type of eQTL slope. Rugs along x-axis denote distribution of values for each variable.

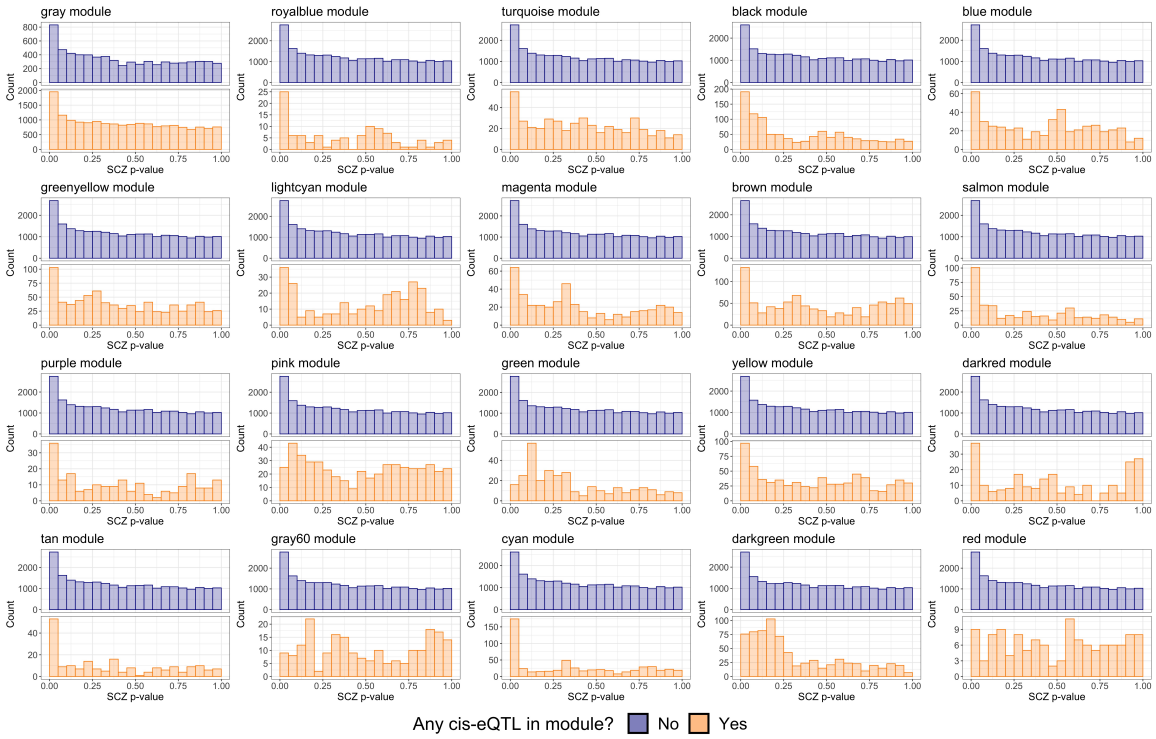

Figure 2.13: Comparison of SCZ p-value distributions from *2014* studies by whether or not the eSNP had an associated cis-eQTL gene in the module.

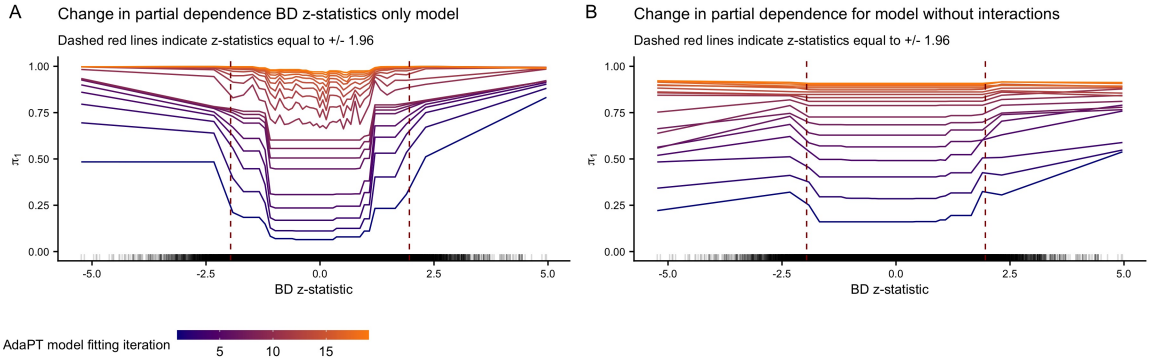

Figure 2.14: Change in partial dependence for BD z-statistics and probability of being non-null π_1 for the AdaPT results using *(A)* only BD z-statistics and *(B)* all covariates without any interactions.

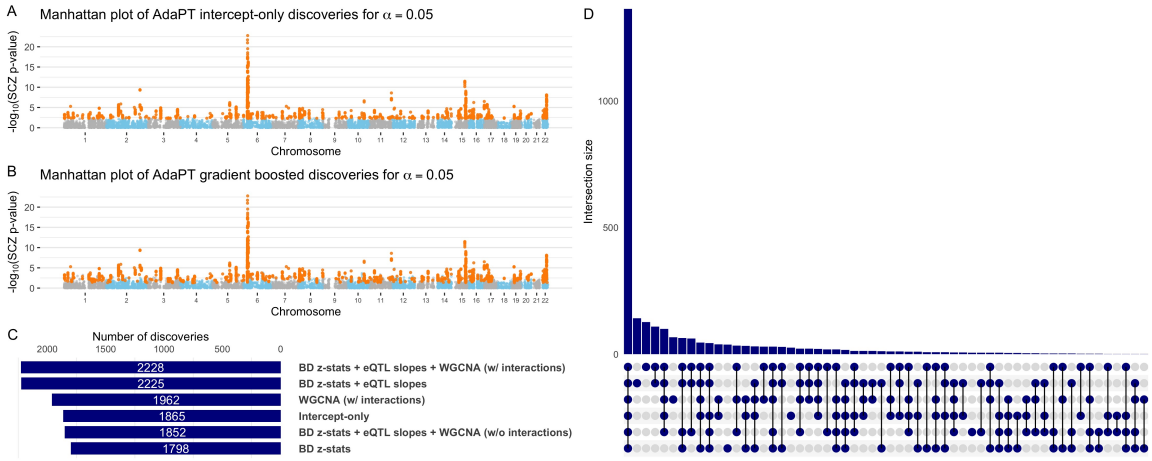

Figure 2.15: Manhattan plots of SCZ AdaPT discoveries (in orange) with *all 2018* studies using *(A)* intercept-only model compared to *(B)* covariate informed model at target $\alpha = 0.05$. *(C)* Comparison of the number of discoveries at target $\alpha = 0.05$ for AdaPT with varying levels of covariates and *(D)* their resulting discovery set intersections.

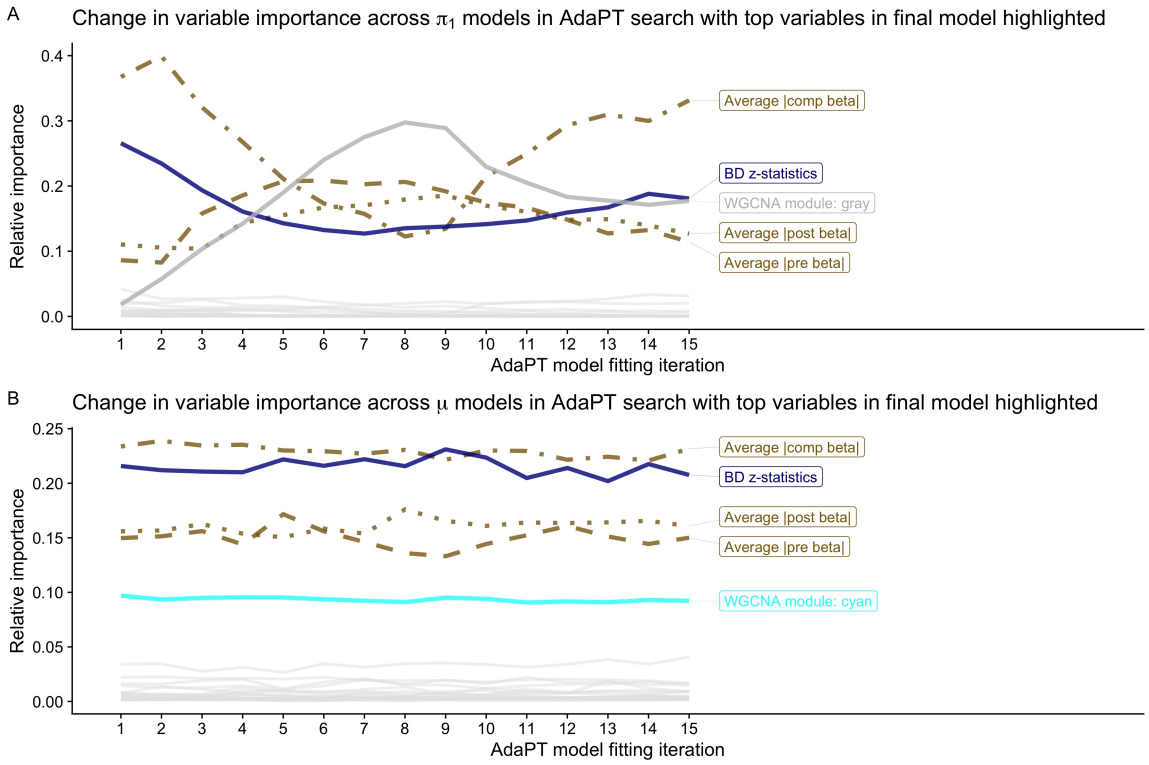

Figure 2.16: Using *all 2018* studies: change in variable importance for AdaPT *(A)* probability of being non-null π_1 and *(B)* effect size under alternative μ models across search, with top variables in final model highlighted.

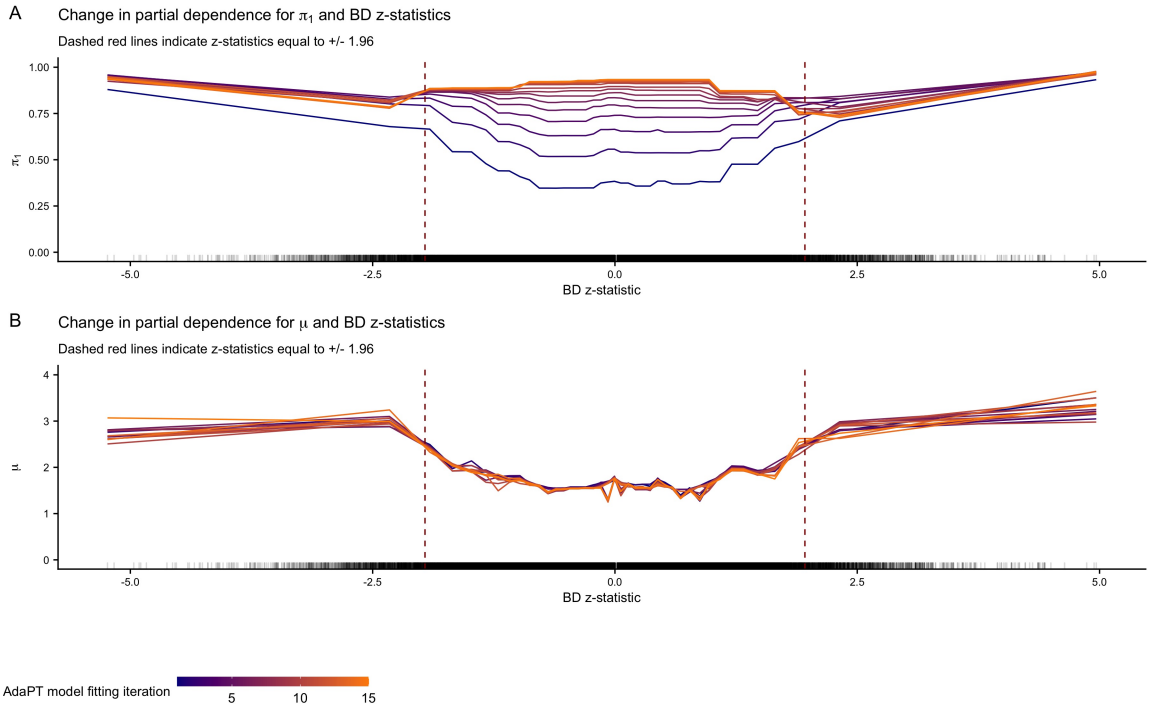

Figure 2.17: Using *all 2018* studies: change in partial dependence for BD z-statistics and AdaPT *(A)* probability of being non-null π_1 and *(B)* effect size under alternative μ models across search.

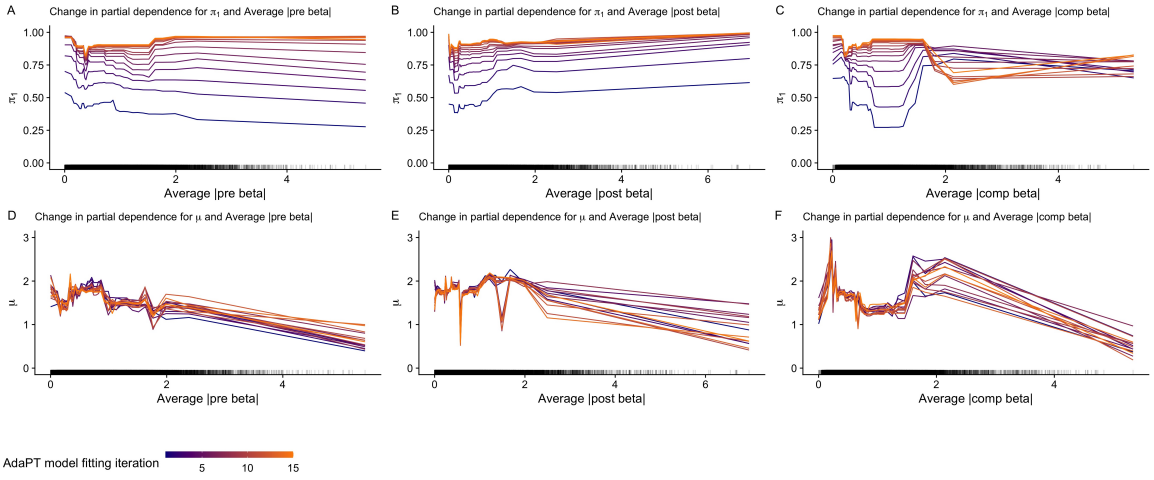

Figure 2.18: Using *all 2018* studies: change in partial dependence plots for probability of being non-null π_1 in *(A-C)*, and the effect size under alternative μ in *(D-F)*, for each type of BrainVar eQTL slope. Rugs along x-axis denote distribution of values for each variable.

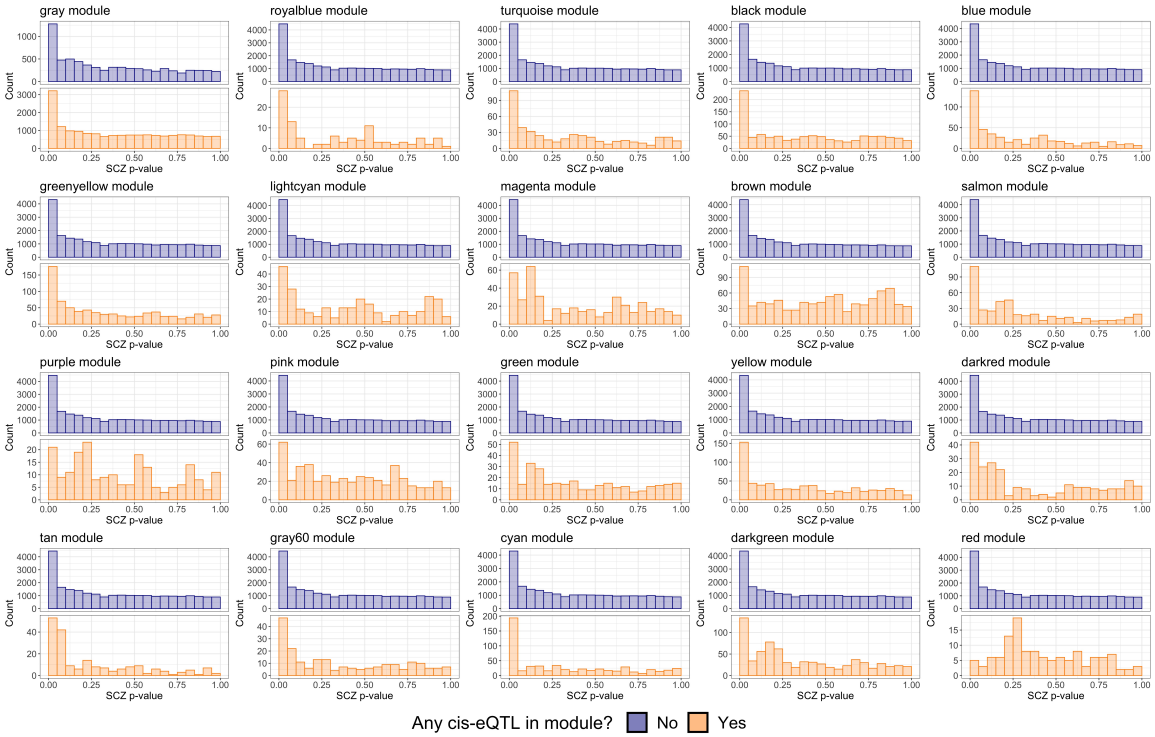

Figure 2.19: Using *all 2018* studies: comparison of SCZ p-value distributions from *2014* studies by whether or not the eSNP had an associated cis-eQTL gene in the module.

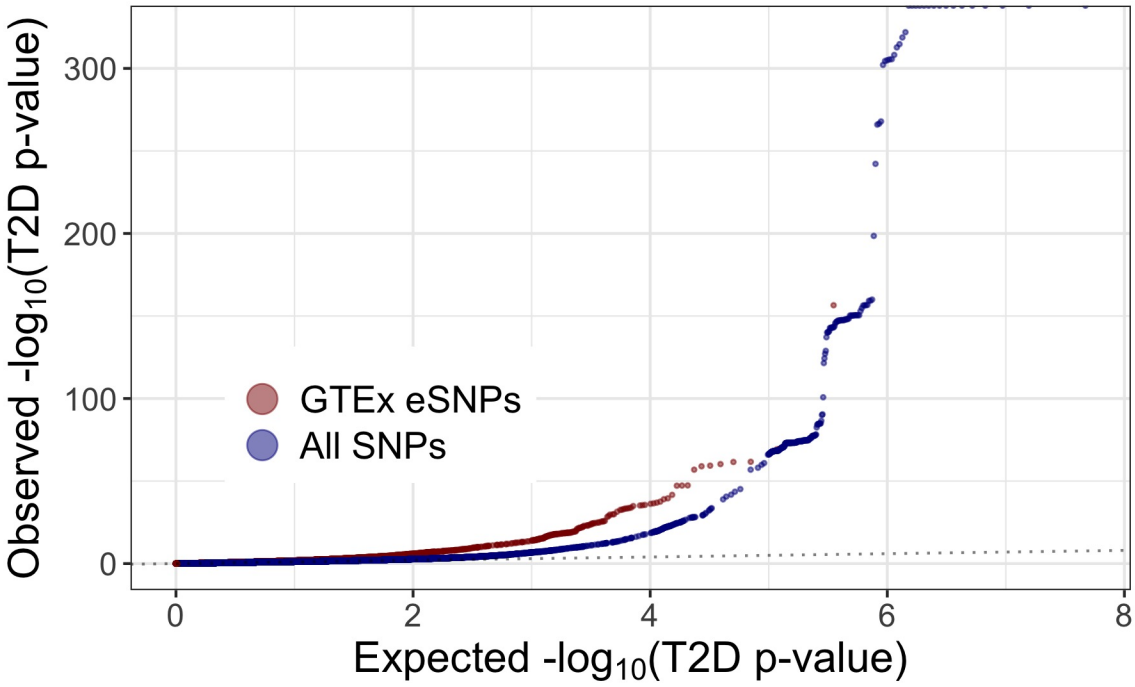

Figure 2.20: A comparison of qq-plots revealing T2D enrichment for GTEx eSNPs compared to full set of SNPs.

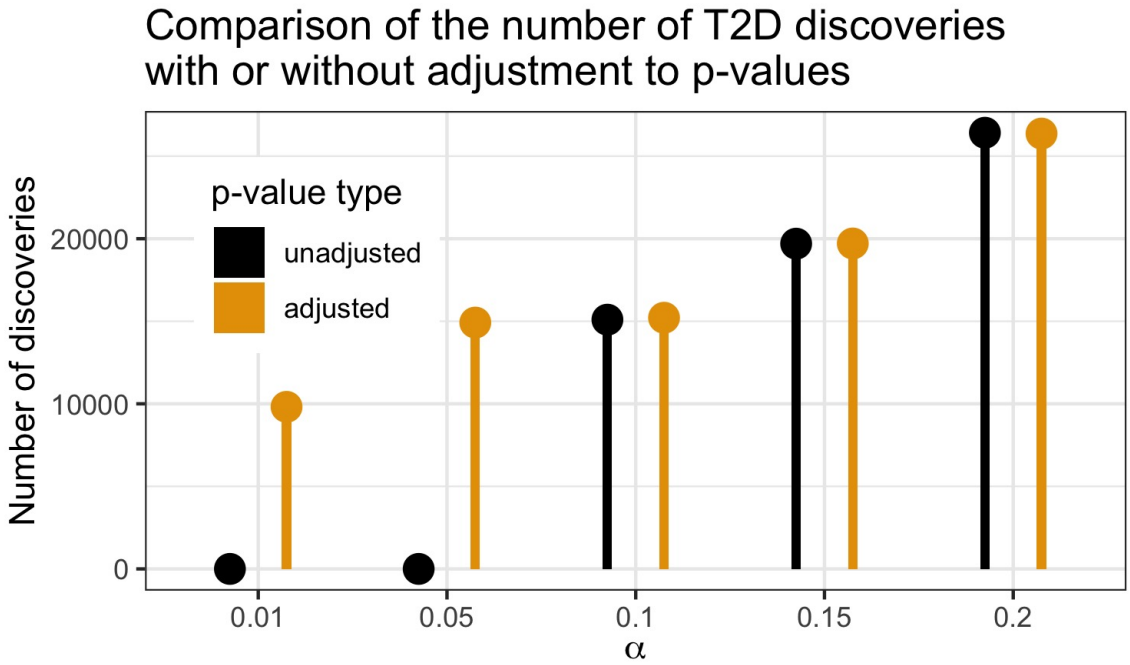

Figure 2.21: Comparison of the number of discoveries by AdaPT for T2D by whether or not the adjusted or unadjusted p-values were used.

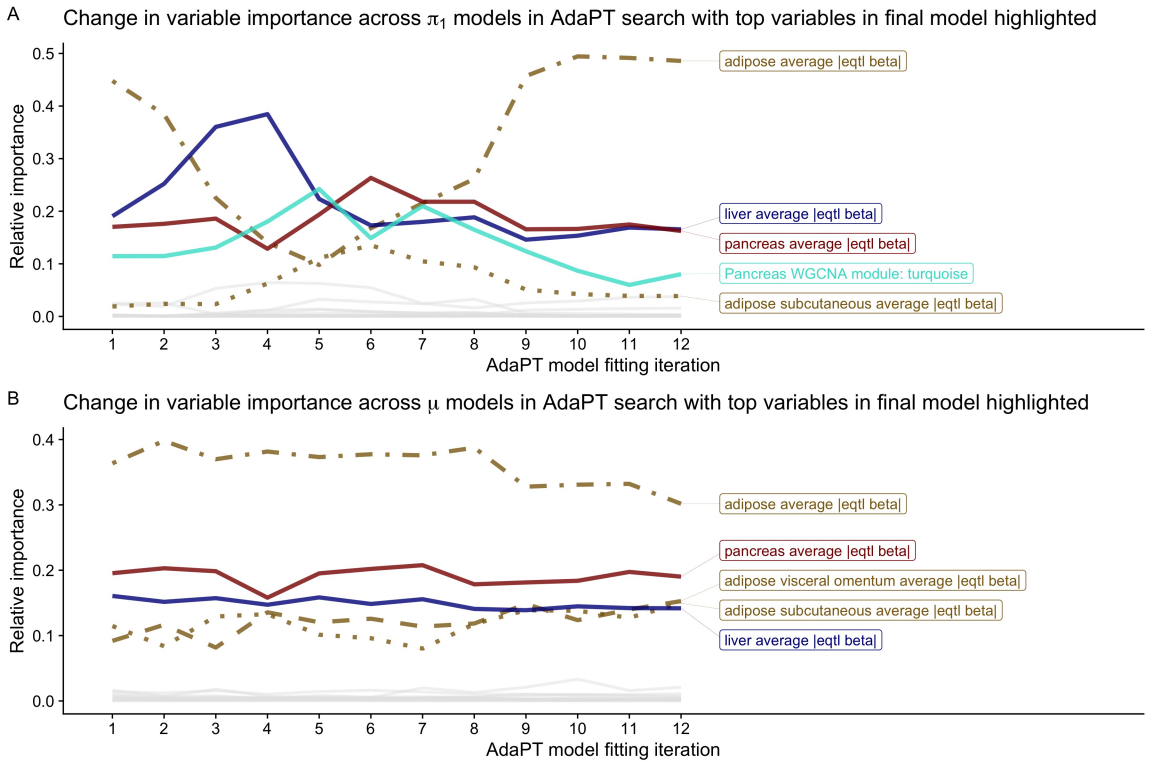

Figure 2.22: Change in T2D variable importance for AdaPT *(A)* probability of being non-null π_1 and *(B)* effect size under alternative μ models across search, with top variables in final model highlighted.

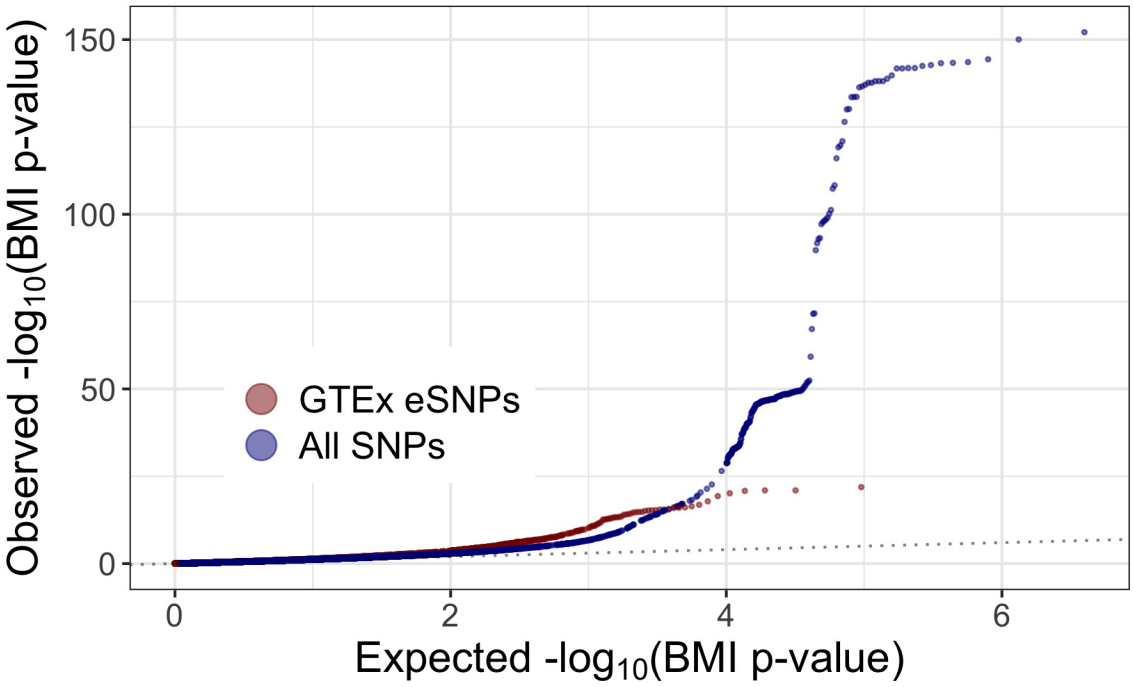

Figure 2.23: Comparison of qq-plots revealing BMI enrichment for GTEx eSNPs compared to full set of SNPs.

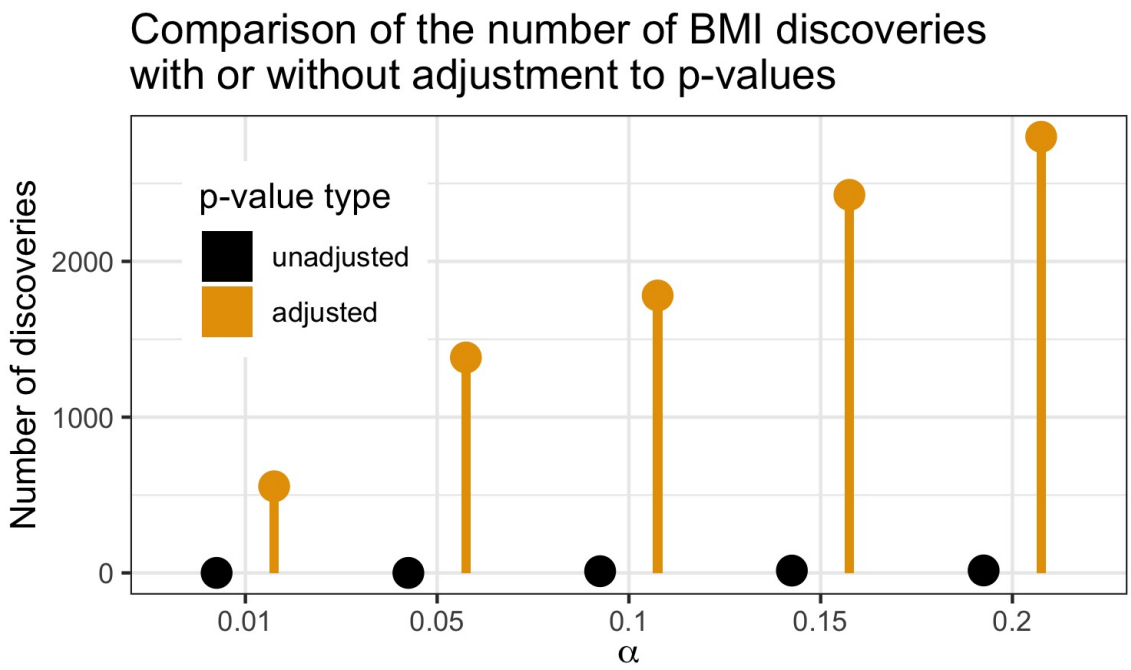

Figure 2.24: Comparison of the number of discoveries by AdaPT for BMI by whether or not the adjusted or unadjusted p-values were used.

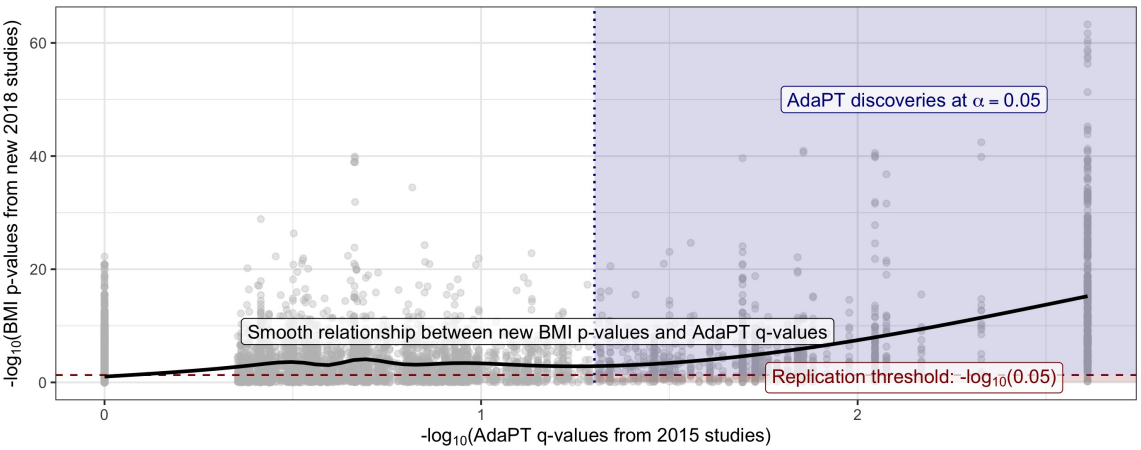

Figure 2.25: Black line displays smooth relationship between BMI p-values from *2018-only* studies and the AdaPT q-values from the *2015-only* studies. Blue-shaded region indicates AdaPT discoveries at $\alpha = 0.05$ that are nominal replications, p-values from the *2018-only* studies < 0.05 while red denotes discoveries which failed to replicate.

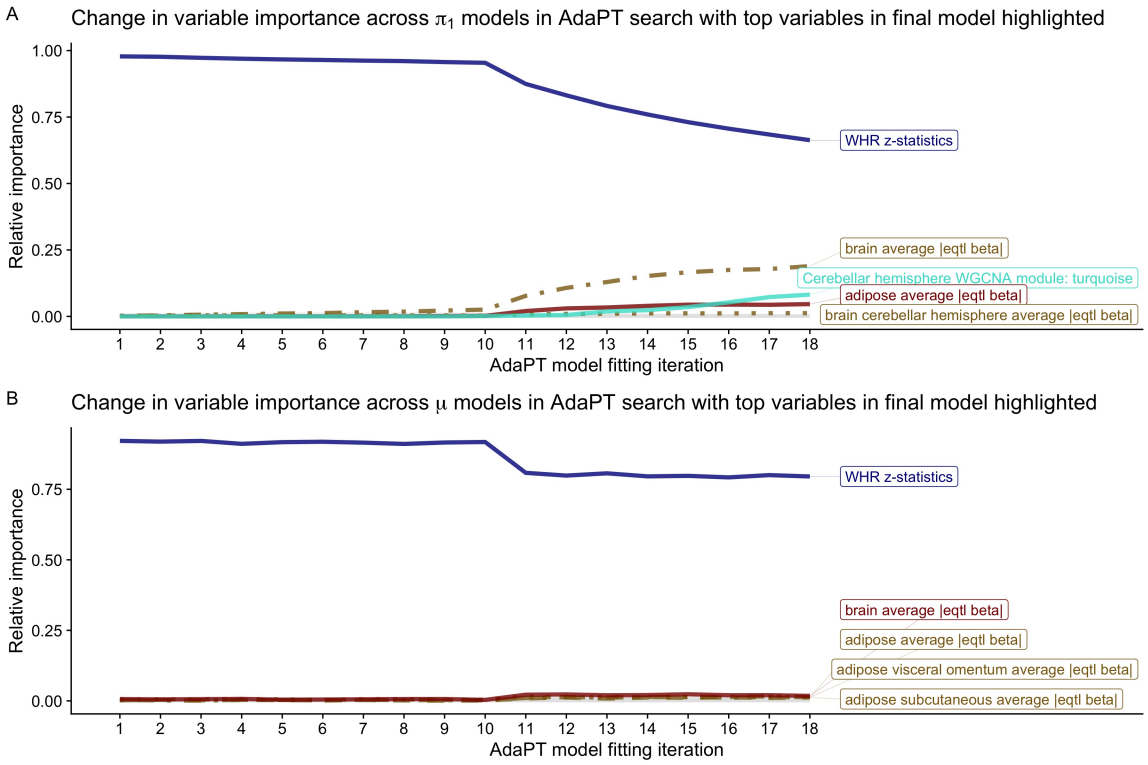

Figure 2.26: Change in BMI variable importance for AdaPT *(A)* probability of being non-null π_1 and *(B)* effect size under alternative μ models across search, with top variables in final model highlighted.

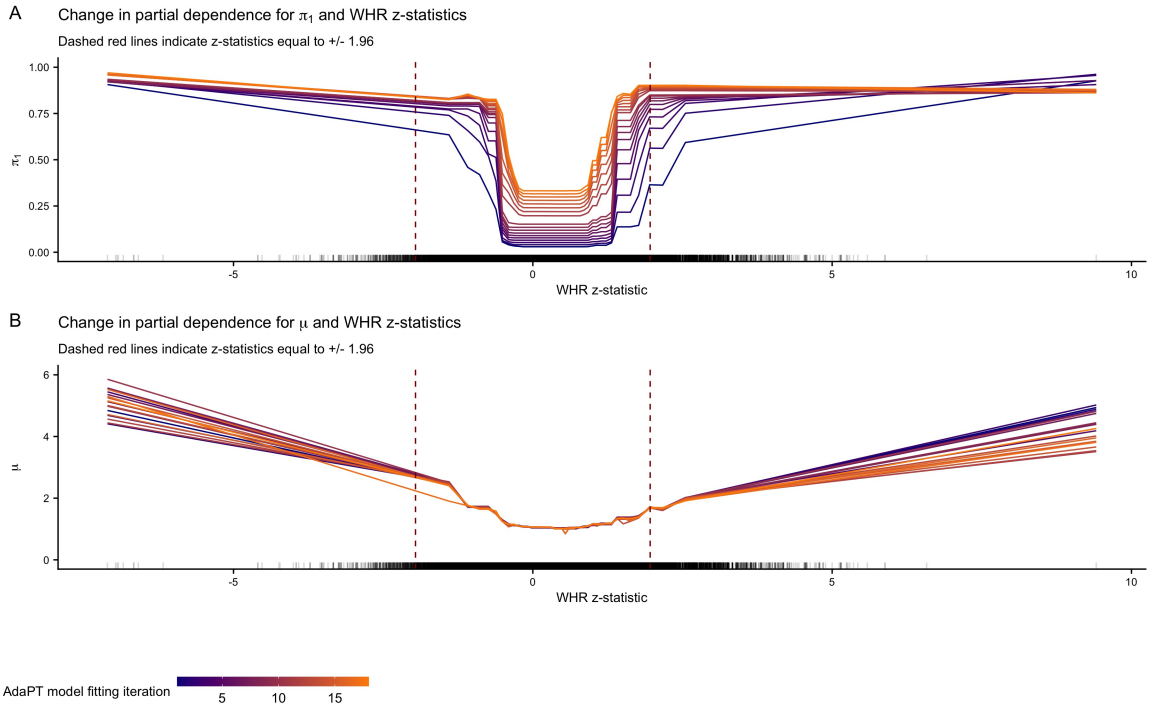

Figure 2.27: Change in BMI partial dependence for WHR z-statistics and AdaPT *(A)* probability of being non-null π_1 and *(B)* effect size under alternative μ models across search.

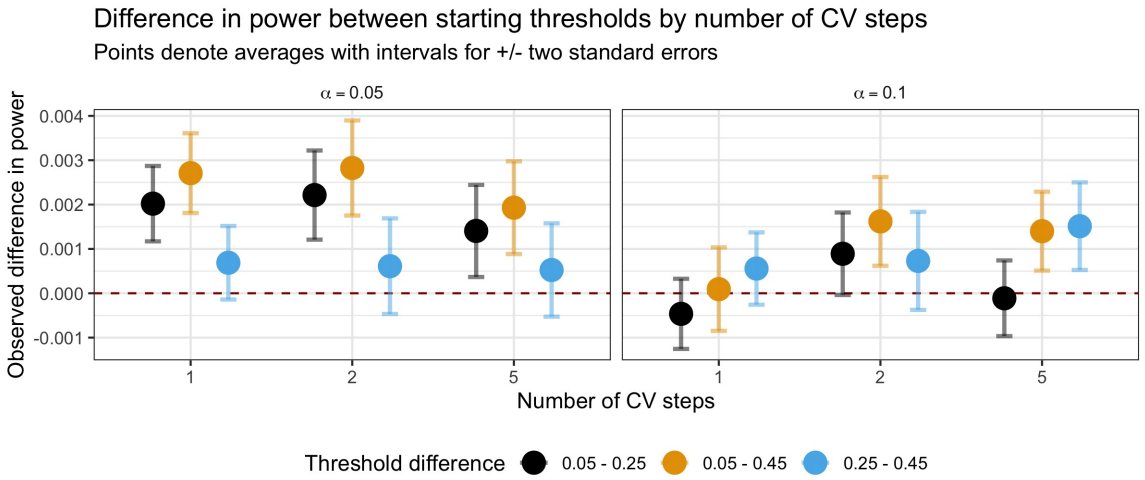

Figure 2.28: Difference in simulation power between different initial thresholds s_0 for AdaPT search by number of CV steps. Points denote averages with plus/minus two standard error bars.

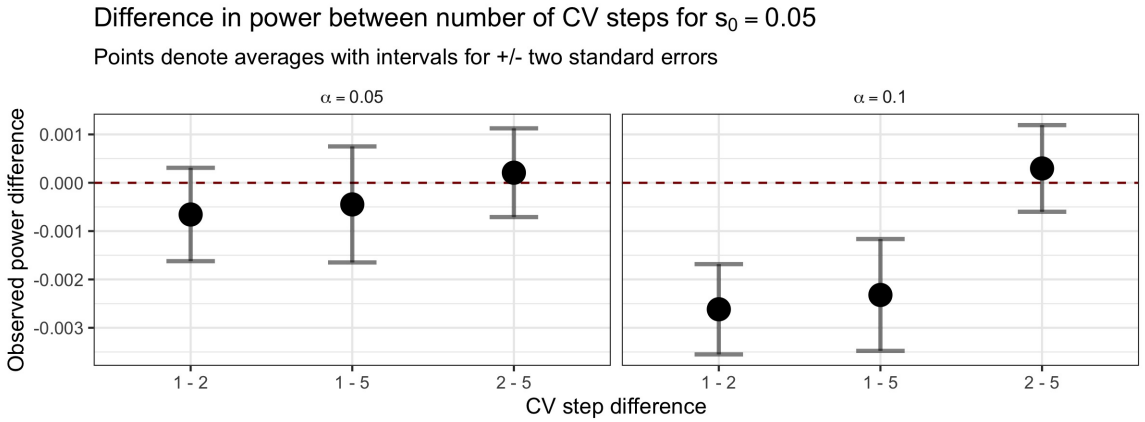

Figure 2.29: Difference in simulation power between the number of CV steps with $s_0 = 0.05$. Points denote averages with plus/minus two standard error bars.

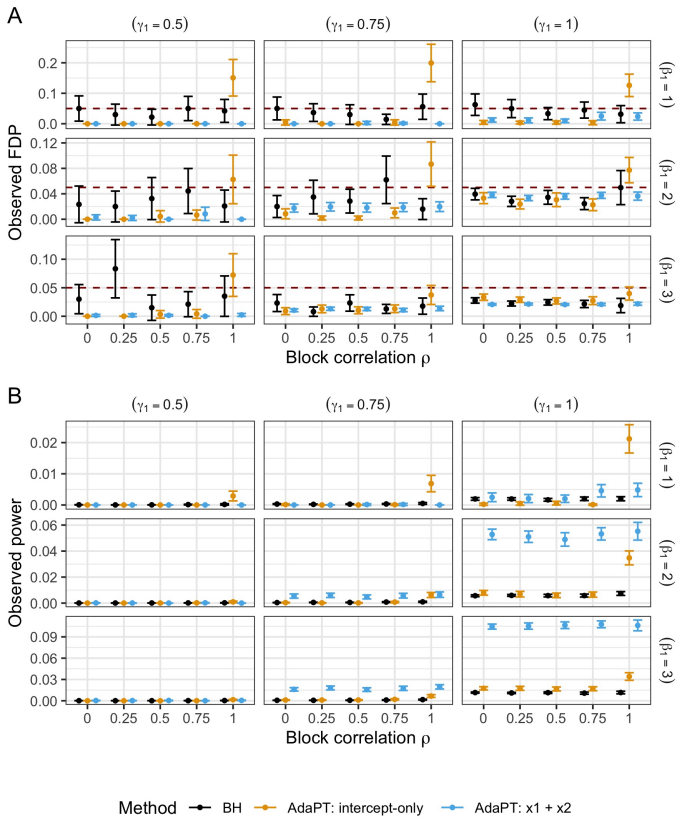

Figure 2.30: Comparison of average (A) FDP and (B) power with plus/minus two standard error bars for 100 simulations with $\mu_{floor} = 0.5$, and varying values for β_1 (rows) and γ_1 (columns) and block correlation ρ.

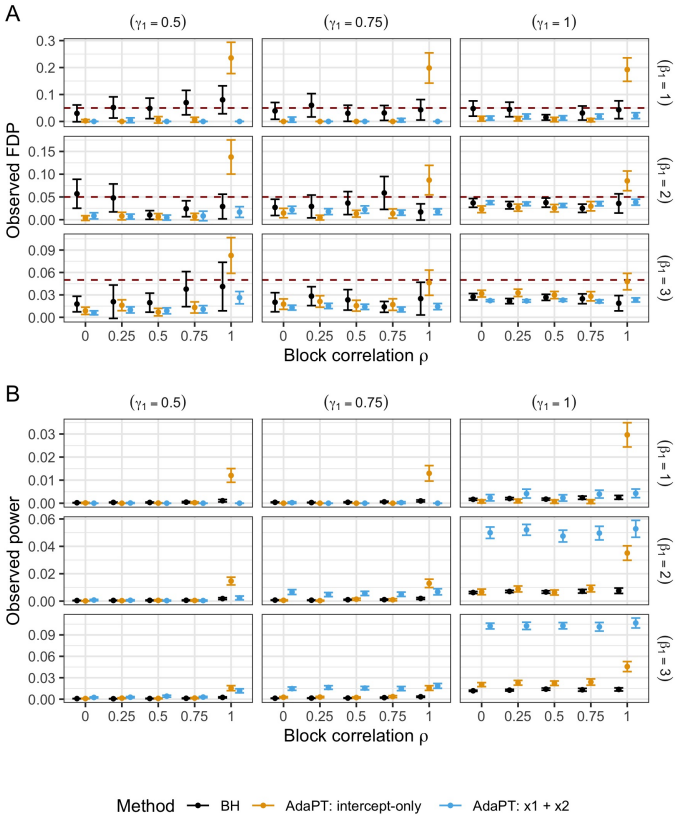

Figure 2.31: Comparison of average (A) FDP and (B) power with plus/minus two standard error bars for 100 simulations with $\mu_{floor} = 1$, and varying values for β_1 (rows) and γ_1 (columns) and block correlation ρ.

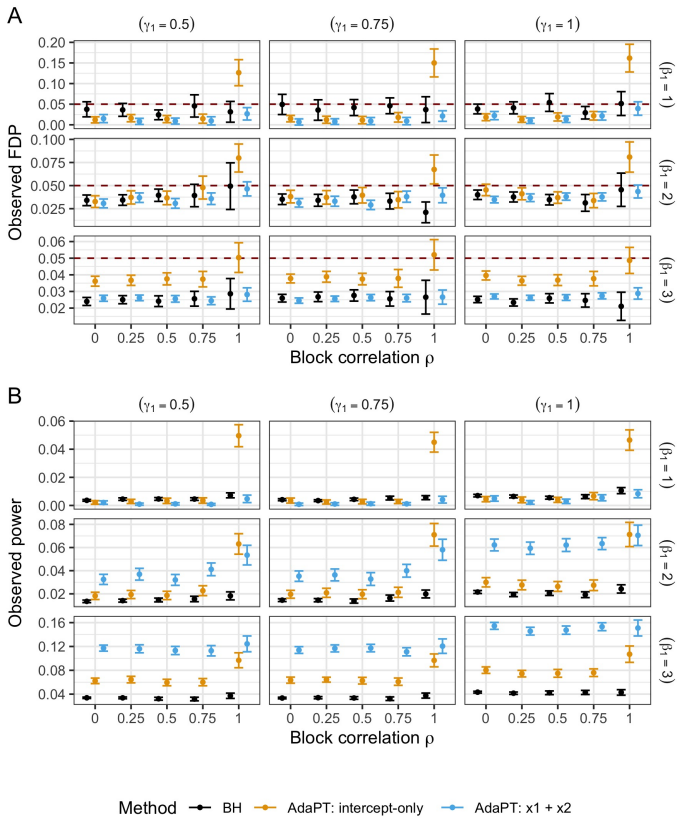

Figure 2.32: Comparison of average (A) FDP and (B) power with plus/minus two standard error bars for 100 simulations with $\mu_{floor} = 1.5$, and varying values for β_1 (rows) and γ_1 (columns) and block correlation ρ.

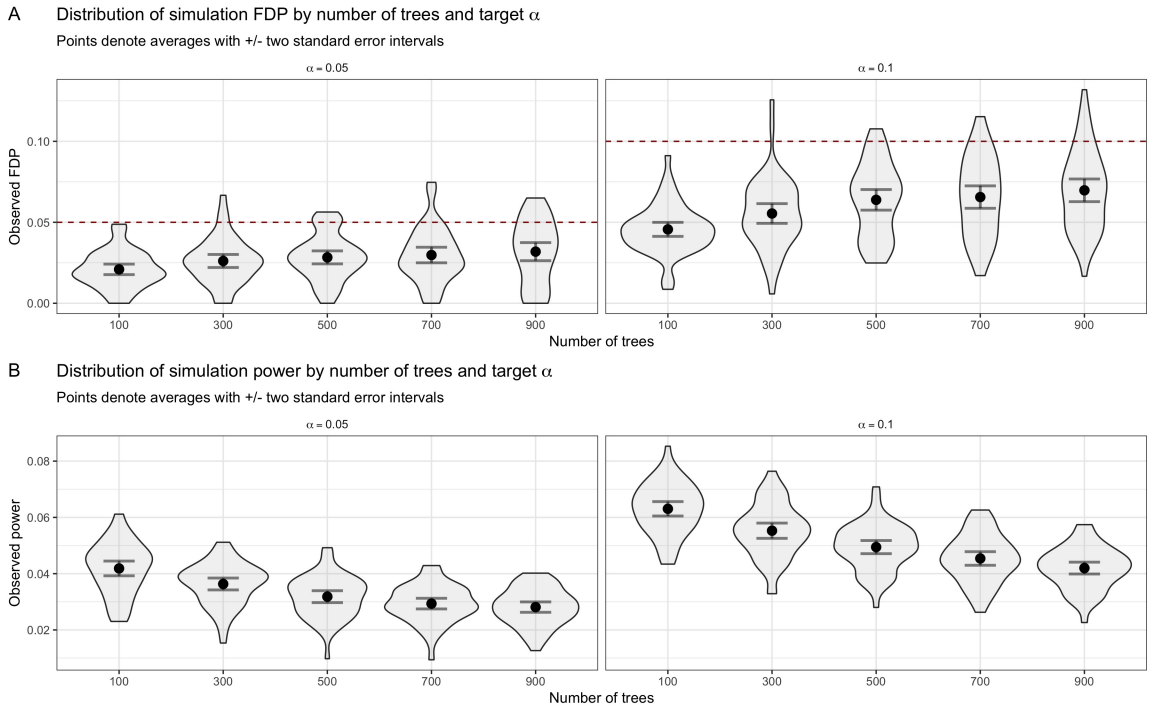

Figure 2.33: Distributions of observed *(A)* FDP and *(B)* power for simulations as the number of AdaPT gradient boosted trees increases by target FDR level α. Points denote averages with plus/minus two standard error intervals.

Three

Identifying and Correcting Type I Error Rate Inflation in Gene-level Testing

The 'snp-wise mean model' of Multi-marker Analysis of GenoMic Annotation is often used to perform gene-level testing for association with disease and other phenotypes. This methodology, in turn, forms the foundation for H-MAGMA. Unfortunately, that foundation is unsound, with implications publications including recent H-MAGMA results published in *Nature Neuroscience* regarding genes associated with psychiatric disorders: e.g., only 125 of H-MAGMA's 275 reported discoveries for autism replicate when the foundation's flaws are corrected.

This chapter appears in [Yurko et al., 2021a].

3.1 Introduction

The 'snp-wise mean model' of Multi-marker Analysis of GenoMic Annotation [de Leeuw et al., 2015] (hereafter MAGMA) is often used to perform gene-level testing for association with disease or other phenotypes, taking as input genomewide association study (GWAS) summary statistics and reference linkage disequilibrium (LD) data. The success of this methodology (MAGMA has 826 Google Scholar citations as of 18 August, 2020) has led to its incorporation into a variety of tools, including an atlas of over 4,700 GWAS results [Watanabe et al., 2019] and various testing methods. For example, it is the foundation for H-MAGMA[Sey et al., 2020], which also incorporates Hi-C data and uses the Benjamini-Hochberg (BH)[Benjamini and Hochberg, 1995] procedure for false discovery rate (FDR) control. When applying MAGMA, however, we noted that its distributional properties did not comport with statistical expectation.

When investigating the basis of MAGMA's distributional properties, we also discovered a critical departure of the accepted MAGMA implementation from the manuscript details[1].

[1]Our findings are pertinent to version 1.07b of MAGMA, since the latest version (1.08) was released with

MAGMA, as described in the manuscript, builds on Brown's approximation of Fisher's method for combining dependent SNP-level p-values [Brown, 1975], adjusting for the LD-induced covariance of SNP p-values. This statistical approximation, however, is *valid only for one-sided tests*. GWAS summary statistics are necessarily two-sided tests because which SNP allele confers risk is not known, a priori [Willer et al., 2010]. When applied to two-sided tests, as in the analysis of GWAS summary statistics by MAGMA, the assumed null distribution is incorrect in both its distributional form and its covariance. The correct null distribution for simulated multivariate normal two-sided test statistics (Fig. 3.1a) does not follow the re-scaled chi-square distribution implied by Brown's approximation (denoted by *MAGMA: paper*). Furthermore, comparison to a known correction to the covariance for two-sided tests [Yang et al., 2016] leads to a stark difference with the one-sided test approximation (Supplementary Fig. 3.3; see Supplement for details of this and other calculations).

This incorrect null distribution should lead to invalid and non-uniform null p-values with a severely inflated error rate. Curiously, this does not comport with the observed performance of MAGMA when tests are conducted at small significance level α. How could this be? The software embodies two undocumented, ad-hoc corrections: replacing the correlation coefficient ρ in Brown's approximation with its square, which acts as a rough correction for one- versus two-sidedness (denoted as *MAGMA:* ρ^2 in Fig. 3.1), followed by an empirically-motivated warping of the p-values to reduce the false positive rate (*MAGMA: code*). While these corrections together result in improved error rate control at small α, ultimately, this extension of Brown's approximation is invalid for two-sided tests (Fig. 3.1a) and it yields an inflated error rate that worsens for larger genes (Fig. 3.1b and Supplementary Fig. 3.4).

MAGMA's incorrect null p-value distributions (Fig. 3.1c) are particularly relevant for procedures that correct for multiple testing, such as family-wise error rates (Supplementary Fig. 3.5 and 3.6). We demonstrate this impact with simulations using real genotype data [1000 Genomes Project Consortium and others, 2012] to show its failure to maintain FDR control using the BH procedure (Fig. 3.1d). Additionally, we observe that the concern for MAGMA in the setting of gene-set enrichment analysis is a loss of power due to the properties of the procedures (see *Supplement* and Supplementary Fig. 3.7-3.9). In comparison, Monte Carlo-based approaches to computing the null guarantee appropriate error-rate control and uniform p-value distributions under the assumed model (Fig. 3.1b-d), even when using the same test statistic as MAGMA (referred to as *Corrected*).

The use of MAGMA impacts the results and development of new procedures that inherit its statistical flaws. We examine the results of one such paper, published in *Nature Neuroscience*, which proposes H-MAGMA[Sey et al., 2020]. After introducing the concept

a different but corrected testing strategy in response to a pre-print of our manuscript. For details regarding the MAGMA v1.08 update: `https://www.biorxiv.org/content/10.1101/2020.09.25.310722v1`

of H-MAGMA, which relies on MAGMA for computing gene-level p-values and BH for FDR control, the authors apply it to data from five psychiatric disorders. Reanalyzing these data, we observe that the H-MAGMA p-value distributions are improper (Fig. 3.2a) and testing using corrected p-values yields substantially smaller subsets of the reported results (Fig. 3.2b and Supplementary Table 3.1). This reduction is more prevalent in the weaker signal setting for autism spectrum disorder, for which only 125 of the 275 genes H-MAGMA associates with autism replicate with the corrected p-values.

Correcting MAGMA's underlying gene-level p-value computation is essential to ensure that novel extensions, such as H-MAGMA, do not yield excess false positives. As our results suggest, a simple solution is to replace Brown's approximation in MAGMA with Monte Carlo-based procedures, similar to *VEGAS*[Liu et al., 2010, Mishra and Macgregor, 2015]. We believe these results are critical for researchers wishing to interpret gene-based testing or for those wishing to build new methods in this challenging area.

3.2 Background

3.2.1 Combining p-values under dependence

Fisher's method [Fisher, R.A., 1925] is a classical approach to combine p-values from m tests. Given observed p-values p_j for each test $j \in [m]$, it computes a summary test statistic,

$$T = \sum_{j}^{m} -2 \log p_j. \tag{3.1}$$

Under the assumption of independent p-values, Fisher's test statistic $T \sim \chi^2_{2m}$. Fisher's method is used to test the global null, i.e. all SNPs $j \in [m]$ in gene g are null versus at least one SNP in the gene is non-null,

$$H_{0,g} : H_j = 0 \ \forall j \in [m] \text{ versus } H_{1,g} : \exists j \in [m] \text{ such that } H_j = 1, \tag{3.2}$$

where $H_j = 0$ if SNP j is null, and $H_j = 1$ if non-null.

However, p-values for nearby SNPs are often dependent due to LD. Brown introduced an extension of Fisher's method in the case of dependence [Brown, 1975], for the setting where m tests are based on multivariate Gaussian random variables,

$$\boldsymbol{z} = (z_1, \ldots, z_m) \sim \text{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{m \times m}), \tag{3.3}$$

where $\boldsymbol{\mu}$ is the vector of true centers for the m tests and $\boldsymbol{\Sigma}_{m \times m}$ is the corresponding covariance matrix. Brown's test uses the same Fisher's test statistics as in Equation 3.1, but assumes that the null distribution is $T \sim c\chi^2_{2v}$, a re-scaled χ^2 distribution. The two

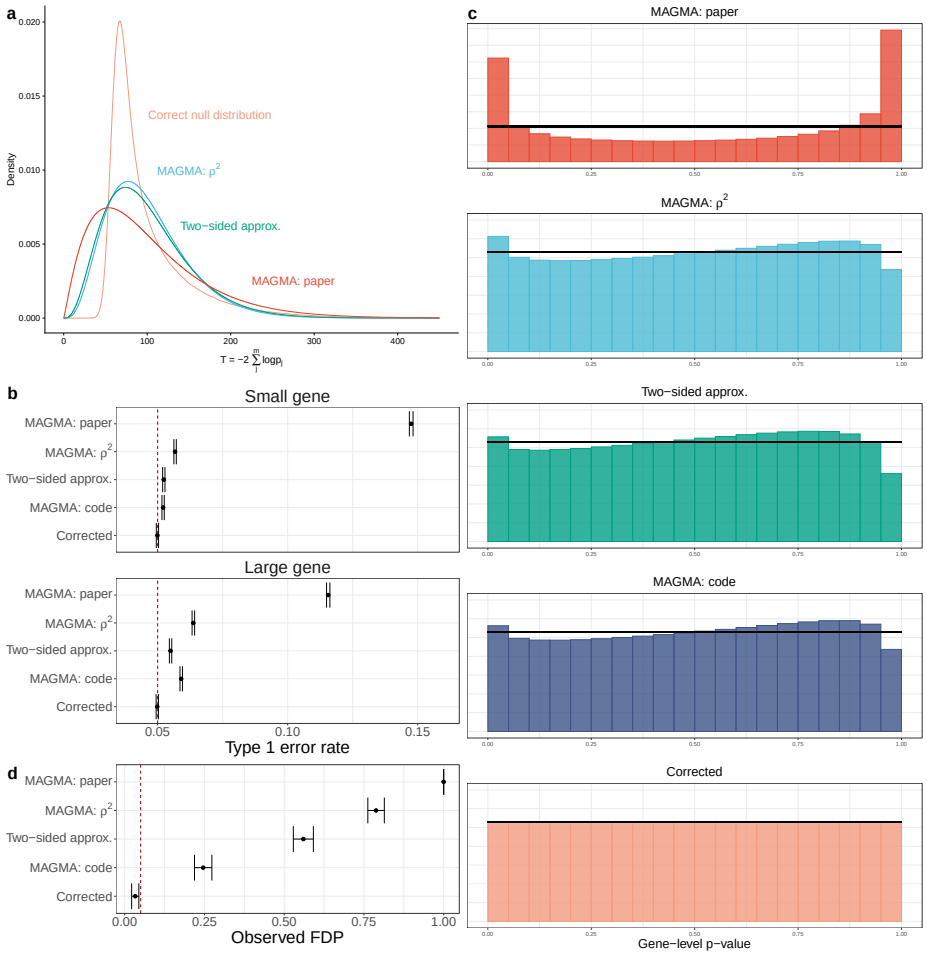

Figure 3.1: **a**, Comparison of the different covariance approximations for fitting the test statistic distribution with Brown's rescaled χ^2 using two-sided summary statistics. Test statistic distribution is generated using fifty-dimensional multivariate Gaussian distribution centered with mean zero and block correlation $\rho = 0.50$. **b**, Comparison of the type 1 error rate control, plus/minus two standard errors at target $\alpha = 0.05$ (denoted by dashed red line) by gene-level method for small (17 SNPs) and large (1,068 SNPs) example genes. **c**, Comparison of null histograms by method for ≈ 53000 genes of different sizes averaged over 1,000 simulations (standard errors are too small to be visible). Horizontal black lines represents ideal uniform distribution for null p-values. **d**, Comparison of average BH false discovery proportions (FDP) across 1000 simulations, plus/minus two standard errors, at target FDR level $\alpha = 0.05$ (denoted with dashed red line) by gene-level method.

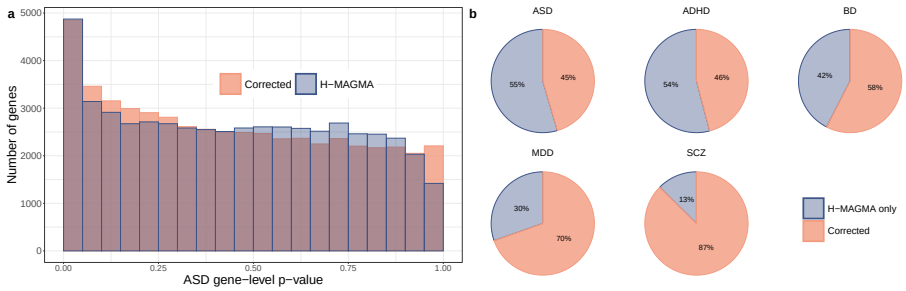

Figure 3.2: **a**, Comparison of autism spectrum disorder (ASD) gene-level p-values, based on the adult Hi-C data annotations, computed with H-MAGMA versus corrected Monte Carlo-based approach. **b**, Percentage of H-MAMGA reported discoveries using BH at target FDR level $\alpha = 0.05$ for five psychiatric disorders due to MAGMA inflation versus the remaining Monte Carlo-corrected set.

constants c and v are calculated by matching the first two moments of the re-scaled χ^2 to the moments of Fisher's test statistic T induced by the multivariate Gaussian,

$$v = \frac{\mathbb{E}[T]^2}{\text{Var}[T]} \text{ and } c = \frac{\text{Var}[T]}{2\mathbb{E}[T]}, \tag{3.4}$$

$$\text{where } \mathbb{E}[T] = 2m \text{ and } \text{Var}[T] = 4m + 2\sum_{j<k} \text{Cov}(-2\log p_j, -2\log p_k), \text{ for } j,k \in [m]. \tag{3.5}$$

Instead of computing the covariance directly with numerical integration, which can be computationally intensive in -omics settings such as GWAS, Brown approximated the covariance between tests j and k as a function of the correlation of the corresponding Gaussians, ρ_{jk},

$$\text{Cov}(-2\log p_j, -2\log p_k) \approx \begin{cases} \rho_{jk} \cdot (3.25 + 0.75\rho_{jk}) \text{ if } \rho_{jk} \geq 0, \\ \rho_{jk} \cdot (3.27 + 0.71\rho_{jk}) \text{ if } 0.5 \leq \rho_{jk} < 0. \end{cases} \tag{3.6}$$

Since Brown's initial publication, there have been further refinements to this approximation [Kost and McDermott, 2002, Yang, 2010]; for example, Kost and McDermott[Kost and McDermott, 2002] used polynomial regression over a grid of values for $-0.98 \leq \rho_{jk} \leq 0.98$ by increments of 0.02,

$$\text{Cov}(-2\log p_j, -2\log p_k) \approx 3.263\rho_{jk} + 0.710\rho_{jk}^2 + 0.027\rho_{jk}^3. \tag{3.7}$$

However, Brown's covariance approximation and its various refinements are based on the usage of one-sided tests $p_j = 1 - \phi(z_j)$, where z_j is the corresponding z-statistic for test j and ϕ is the Gaussian cumulative distribution function. [Yang et al., 2016] introduced an

approximation for two-sided tests, $p_j = 2\phi(-|z_j|)$, in the context of detecting association between a SNP and multivariate phenotypic traits, using a tenth order polynomial,

$$\text{Cov}(-2\log p_j, -2\log p_k) \approx 3.9081\rho_{jk}^2 + 0.0313\rho_{jk}^4 + 0.1022\rho_{jk}^6 - 0.1378\rho_{jk}^8 + 0.0941\rho_{jk}^{10}. \quad (3.8)$$

This yields a drastically different covariance approximation from one-sided tests as displayed in Supplementary Fig. 3.3. We provide a simpler two-sided approximation using polynomial regression over a grid of values for $-1 \leq \rho_{jk} \leq 1$ by increments of 0.01,

$$\text{Cov}(-2\log p_j, -2\log p_k) \approx 3.902364\rho_{jk}^2 + 0.051520\rho_{jk}^4 + 0.032832\rho_{jk}^6. \quad (3.9)$$

The median difference between the approximations in Equations 3.8 and 3.9 is 0.0006465, with a maximum observed difference of 0.011184. Equation 3.9 also displays a smaller maximum absolute difference of 0.0009217044 with a recent calculation using Hermite polynomials [Zhang and Wu, 2020] than Equation 3.8, which has a maximum absolute difference of 0.0104.

3.2.2 MAGMA 'snp-wise-mean model'

The MAGMA 'snp-wise-mean model' is used to compute a gene-level test statistic from GWAS summary statistics. In the original publication [de Leeuw et al., 2015], the authors describe the use of Brown's covariance approximation, which is inappropriate for two-sided tests (Supplementary Fig. 3.3). However, the maintainers of the MAGMA software have made several adjustments to this approximation not described in the manuscript. First, prior to approximating the covariance based on the sign of the correlation, they square the correlation values, ρ_{jk}^2, resulting in,

$$\text{Cov}(-2\log p_j, -2\log p_k) \approx \rho_{jk}^2 \cdot (3.25 + 0.75\rho_{jk}^2). \quad (3.10)$$

This covariance approximation (Supplementary Fig. 3.3) alleviates the initial stark difference in covariance values in the presence of negative correlation from using Brown's one-sided approximation and is much closer to, but still under-estimating, the appropriate approximation for two-sided tests. The software includes an additional adjustment to the resulting gene-level p-value p_g,

$$p_g^* = p_g^c \text{ where } c = (1.025)^{\log_{10}(p_g)}. \quad (3.11)$$

Based on correspondence with the maintainers of MAGMA, they determined the use of an adjusted p-value p_g^* from the power c after viewing simulations to reveal Brown's approximation would yield smaller p-values than the truth. By using the power adjustment above, smaller values p_g will receive a *stronger* adjustment. It is unfortunate that these adjustments were not presented in the manuscript, which thus misleads readers who are exploring the usage of similar methodology for combining p-values from GWAS summary statistics.

3.3 Methodology

3.3.1 Methods for computing gene-level p-values

To evaluate the impact of MAGMA's implementation, we compare these approaches:

1. *MAGMA: paper* - the approach presented in the original MAGMA manuscript following Brown's covariance approximation for one-sided tests, red-dashed line in Supplementary Fig. 3.3,

2. *MAGMA:* ρ^2 - replacement of Brown's covariance approximation with squared correlation values, cyan-dashed line in Supplementary Fig. 3.3,

3. *MAGMA: code* - includes the use of both ρ^2 and the adjustment power c,

4. *Two-sided approximation* - replacement of Brown's covariance approximation with the appropriate two-sided covariance [Yang et al., 2016], green-solid line in Supplementary Fig. 3.3.

5. *Corrected* - Monte Carlo simulation using the Fisher test statistic, $T_i^* = -2 \cdot \sum_j^m \log p_j^*$. Details below.

The *Corrected* Monte Carlo-based approach for computing empirical p-values differs from MAGMA in that it does not rely on Brown's original assumption of a re-scaled χ^2 distribution. Instead, we generate N draws of m-dimensional Gaussian random variables,

$$\boldsymbol{z}^* = (z_1^*, \ldots, z_m^*) \sim \text{Normal}(\mathbf{0}, \boldsymbol{\Sigma}_{m \times m}), \tag{3.12}$$

where $\mathbf{0}$ is a vector of zeroes representing the null distribution with $\boldsymbol{\Sigma}_{m \times m}$ as the LD structure of the variants within a gene. A test statistic T_i^* is calculated for each of the $i \in [N]$ draws and, using the observed test statistic T, an empirical gene-level p-value is calculated as:

$$p = \frac{\sum_i^n 1(T_i^* > T) + 1}{N + 1}. \tag{3.13}$$

To be consistent with MAGMA, we use the same test statistic $T_i^* = -2 \cdot \sum_j^m \log p_j^*$, where $p_j^* = 2 \cdot \phi(-|z_j^*|)$ is a two-sided p-value. While the *Corrected* results in this manuscript refers matching the MAGMA test statistic, we observed equivalent results when using *VEGAS*[Liu et al., 2010, Mishra and Macgregor, 2015], a similar Monte Carlo-based approach based on a different test statistic, $T_i^* = \boldsymbol{z}^{*T} \boldsymbol{z}^*$.

Computational considerations

We use the Cholesky decomposition of $\mathbf{\Sigma}_{m\times m} = \mathbf{L}\mathbf{L}^{\mathrm{T}}$, where $\mathbf{L}$ is a lower triangular $m \times m$ matrix, to simulate N draws from a m-dimensional multivariate Gaussian distribution. A single draw $\boldsymbol{z}^* \sim \text{Normal}(\mathbf{0}, \mathbf{\Sigma}_{m\times m})$ is generated by multiplying $\mathbf{L}$ with a vector of m independent, standard Gaussian random variables. The combined computational cost of the Cholesky decomposition with all N draws is $O(m^3 + Nm^2)$, but blocks of draws can be produced in parallel to reduce time. We implement these steps in the `R`[R Core Team, 2020] package `snpcombineR`[Yurko, 2020] (available at `https://github.com/ryurko/snpcombineR`) using `RcppArmadillo`[Eddelbuettel and Sanderson, 2014] for efficient computation. Using this package we can generate 1,000,000 null simulations in ≈ 10 seconds for a gene with $m = 101$ SNPs and ≈ 575 seconds with $m = 1{,}068$ SNPs (see below for details on example genes), with a 3.6 GHz Intel processor without using parallel computing. While we consider a large fixed N in this manuscript (e.g., 1,000,000 is sufficient for testing $\approx 20,000$ genes), in practice N can be determined adaptively, as described in detail by Liu et. al.[Liu et al., 2010] in their paper describing *VEGAS*..

3.3.2 Multivariate Gaussian simulation

As a 'toy' example for demonstrating the flaw of Brown's approximation for two-sided test statistics, we generate null multivariate Gaussian random variables as in Equation 3.12, which matches the initial assumption made by Brown in Equation 3.3. For simplicity, we a $m = 50$-dimensional covariance matrix with block correlation $\rho = 0.5$, where the diagonal elements $\mathbf{\Sigma}_{j,j} = 1$ and all off-diagonal elements $\mathbf{\Sigma}_{j,k} = 0.5$ for $j \neq k$, and $j, k = 1, \ldots, m$. We use 1,000,000 simulations under this model to generate the correct null distribution for the test statistic $T_i^* = -2 \cdot \sum_j^m \log p_j^*$ in Fig 1a, where $p_j^* = 2 \cdot \phi(-|z_j^*|)$ is a two-sided p-value, as compared to Brown's re-scaled χ^2 distribution with three different covariance approximations: *MAGMA: paper* (red), *MAGMA:* ρ^2 (cyan), and the *Two-sided approximation* (green).

3.3.3 Example genotype data

To evaluate the performance of the gene-level methods with real data, we randomly select genes using the sample of 503 individuals from the 1000 Genomes project [1000 Genomes Project Consortium of European ancestry (build 37), with NCBI gene-locations accessed from the MAGMA site: `https://ctg.cncr.nl/software/magma`. After standard pre-processing steps, excluding SNPs with minor allele frequency (MAF) less than 0.05 and call rate less than 0.95, the remaining SNPs were assigned to 19,326 genes using the MAGMA 'annotate' command with 10kb padding (upstream and downstream). We randomly select fifteen candidate genes from each autosomal chromosome, with stratified sampling of five genes from each of the following size groups: (1) $[10, 100]$, (2) $(100, 500]$, and (3) $(500, \infty)$ SNPs. To remove SNPs displaying high LD values, these candidate genes were pruned in the following manner:

1. order the SNPs in the gene by minor allele frequencies (MAFs) in descending order,

2. starting with the SNP with the highest MAF,
    - remove all SNPs with $r^2 \geq 0.95$ within the gene,
    - move on to the next SNP that is still remaining,
3. return the retained SNPs that are not in high LD for simulations.

We then randomly selected nine example genes, with stratified sampling of three genes from the same size groups as above: (1) $[10, 100]$, (2) $(100, 500]$, and (3) $(500, \infty)$ SNPs remaining after pruning. The nine randomly selected genes (with m SNPs after pruning) are: KRT1 ($m = 17$), FAM13B (24), BMPR2 (38), IDO2 (101), PREX2 (218), RAI14 (220), RYR3 (772), PTPRT (1,026), and AGBL1 (1,068). For each gene of these nine, we resample the reference genotype matrix 5,000 times to increase the sample size, from the original 503 individuals, to yield more stable and realistic GWAS simulations.

3.3.4 Gene simulation

To simulate GWAS SNP-level statistics, we use a gene's $n \times m$ reference genotype matrix $\mathbf{X}$ where $n = 5{,}000$ individuals, $m =$ number of SNPs, and $X_{i,j} \in \{0, 1, 2\}$ is the number of effect alleles for individual i at SNP j. We generate simulations for null genes in the following way:

1. Compute the gene's correlation matrix $\mathbf{R}_{m \times m}$.
2. Determine the phenotype status $Y_i \in \{0, 1\}$ for each individual $i \in [n]$ depending on the expected case rate η:
$$Y_i \sim \text{Bernoulli}(\eta), \tag{3.14}$$
3. For each individual SNP $j \in [m]$, fit a logistic regression model $\text{logit}(\text{P}(Y_i = 1 | X_j)) = \beta_0 + \beta_j \cdot X_j$ using all 5,000 individuals, returning the SNP's two-sided p-value p_j.
4. Compute gene-level p-values using the SNP-level statistics and gene correlation matrix $\mathbf{R}_{m \times m}$ with each of the considered methods: (1) *MAGMA: paper*, (2) *MAGMA: ρ^2*, (3) *MAGMA: code*, (4) *Two-sided approximation*, and (5) *Corrected*.

To measure the type 1 error rate at target $\alpha = 0.05$ for the five different methods, we generate 1,000,000 independent null simulations for each of the nine example genes with the case-rate $\eta = 0.5$.

3.3.5 Multiple testing simulation

We assess each method's multiple testing performance by simulating sets of null genes that are comparable in size to the number of genes tested in H-MAGMA. To make the simulations more realistic, we generate each set of genes based on the distribution of the number of SNPs assigned to each gene for the H-MAGMA autism spectrum disorder (ASD) results using the adult Hi-C data annotations. This corresponds to roughly 82.6% from $[10, 100]$, 16.4% $(100, 500]$, and 1.0% from $(500, \infty)$. Using the nine randomly picked genes from each of the size buckets, we construct each set of null genes by assigning:

- 14,663 from each gene with $m \in [10, 100]$,
- 2,827 from each gene with $m \in (100, 500]$,
- 177 from each gene with $m \in (500, \infty)$.

We repeat this process to generate 1,000 sets of $G = 53,001$ null genes to assess each method's impact on multiple testing corrections. First, we use the Bonferroni correction, rejecting the gene's null hypothesis if its p-value is $\leq \frac{\alpha}{G}$, to control the target family-wise error rate (FWER), the probability of making at least one type 1 error, at $\alpha = 0.05$. We then apply the Benjamini-Hochberg procedure[Benjamini and Hochberg, 1995] (BH) to control the false discovery rate (FDR) at target $\alpha = 0.05$, and compute the average false discovery proportion (FDP) across the simulations.

3.3.6 Gene-set analysis simulation

We measure the downstream impact of the different approaches for computing gene-level p-values on self-contained gene-set analysis, i.e., test whether a set of genes is associated with a phenotype. There are several ways for performing gene-set analysis[De Leeuw et al., 2016], but we consider three approaches for computing a test statistic T_s for gene-set s with G genes:

1. MAGMA[de Leeuw et al., 2015] gene-set analysis: each gene g's p-value p_g is converted to a one-sided z-statistic, $z_g = \phi^{-1}(1 - p_g)$, then the test statistic is

$$T_s = \frac{\sqrt{G}}{SD_s} \cdot \frac{1}{G} \sum_{g}^{G} z_g \sim t_{(G-1)}, \tag{3.15}$$

   where SD_s is the sample standard deviation of the gene-set's one-sided z-statistics. The one-sided gene-set p-value is computed as $p_s = 1 - F_{t_{(G-1)}}(T_s)$, where $F_{t_{(G-1)}}$ is the cumulative distribution function for the t-distribution with $G - 1$ degrees of freedom. *Note: this is a separate test from the MAGMA gene-level test described earlier in this manuscript.*

2. Fisher's combination test: as presented in Equation 3.1,

$$T_s = \sum_{g}^{G} -2 \log p_g \sim \chi^2_{2G} \tag{3.16}$$

3. Stouffer's z-test: similar to MAGMA, one-sided z-statistics, $z_g = \phi^{-1}(1 - p_g)$, are computed for the test statistic,

$$T_s = \frac{1}{\sqrt{G}} \sum_{g}^{G} z_g \sim N(0, 1). \tag{3.17}$$

   The gene-set p-value is then computed as $p_s = 1 - \phi(T_s)$.

Both Fisher's and Stouffer's are testing the global null of the gene-set,

$$H_{0,s} : H_g = 0 \; \forall g \in [G] \text{ versus } H_{1,s} : \exists g \in [G] \text{ such that } H_g = 1, \tag{3.18}$$

i.e. all of the individual genes, $g \in [G]$ in gene-set s, are null versus at least one gene in the set is non-null. In comparison, the MAGMA approach is a one-sided test for whether the genes in gene-set s are jointly associated with the phenotype based on a measure of the set's effect size μ_s,

$$H_{0,s} : \mu_s \leq 0 \text{ versus } H_{1,s} : \mu_s > 0. \tag{3.19}$$

We simulate 200,000 gene-sets of size $G = 45$, constructed with five independent null genes from each example gene, to compare the gene-set analysis type 1 error rate that follows from using the different methods for computing gene-level p-values.

3.3.7 Replicating H-MAGMA Analysis

We assess the impact of correcting the usage of MAGMA for computing gene-level p-values with a Monte Carlo-based approach when applied to the SNP annotations used in H-MAGMA by Sey et al.[Sey et al., 2020]. Specifically, we recreate the results presented in Extended Data Fig. 1b of the H-MAGMA manuscript for the GWAS results of five psychiatric disorders: attention-deficit/hyperactivity disorder[Demontis et al., 2019] (ADHD), autism spectrum disorder[Grove et al., 2019] (ASD), schizophrenia[Pardiñas et al., 2018] (SCZ), bipolar disorder[Stahl et al., 2019] (BD) and major depressive disorder[Howard et al., 2018] (MDD). For each of the five GWAS results, Sey et al.[Sey et al., 2020] assign SNPs to genes based on annotations derived from two different sources of Hi-C data: (1) adult and (2) fetal brain tissue samples (available on GitHub: `https://github.com/thewonlab/H-MAGMA`). H-MAGMA relies on MAGMA to compute the GWAS gene-level p-values for both adult and fetal Hi-C annotations, then proceeds to identify which genes are associated with the

respective GWAS phenotype using BH to control FDR at target $\alpha = 0.05$. The counts reported by Sey et al. in Extended Data Fig. 1b correspond to taking the union of the adult and fetal sets of BH discoveries.

We use the same set of Hi-C derived annotations to compute the *Corrected* Monte Carlo-based gene-level p-values with $N = 1{,}200{,}000$ null simulations for each gene. Each gene's assumed covariance matrix is based on the sample of 503 individuals of European ancestry from the 1000 Genomes project, which is the same reference data used by Sey et al. in H-MAGMA. We proceed to identify the union of adult and fetal BH identified genes at target FDR level $\alpha = 0.05$ for each of the five psychiatric disorders' GWAS and compare to the reported H-MAGMA results. We encountered minor discrepancies in the total number of genes with p-values in a subset of the GWAS results as compared to H-MAGMA, since we were unable to compute gene-level p-values for the following: forty-one adult and forty-five fetal for ADHD, two adult and one fetal for ASD, and one fetal for SCZ. These differences are likely due to pre-processing steps that are not publicly available since our analysis matches the set of genes with p-values returned by using the MAGMA software directly, including the same set of SNPs by accounting for synonymous identifiers. However, these differences are negligible and do not impact our results as only one of these missing genes were reported as an association for ASD by H-MAGMA.

3.4 Results

3.4.1 Comparison of type 1 error rate control

Fig. 1b and Supplementary Fig. 3.4 display the type 1 error rates at target $\alpha = 0.05$, plus/minus two standard errors, for the example genes with each of the considered methods to compute gene-level p-values. The same pattern holds for each gene: the different variants of MAGMA display inflated type 1 error rates while the *Corrected* Monte Carlo-based approach maintains valid control. Unsurprisingly, the incorrect usage of Brown's covariance approximation (*MAGMA: paper*) displays the greatest error rate inflation. *MAGMA: code* yields comparable results to the *Two-sided approximation*, however, it appears to display higher inflated error rates with more severity for larger genes.

These results are consistent with recent analysis regarding the behavior of Brown's approximation when applied to two-sided tests generated from multivariate normal data[Zhang and Wu, 2020], whereas our simulations use real genotype data rather than rely on distributional assumptions. The reason for this poor performance stems from the failure of Brown's re-scaled χ^2 distribution to properly fit the actual null test statistic distribution, regardless of the covariance approximation, as seen in Fig. 1a for the 'toy' multivariate Gaussian example with $m = 50$.

3.4.2 Impact on multiple testing

Supplementary Fig. 3.5 displays the observed FWER across 1,000 simulations of applying the Bonferroni correction to sets of 53,001 null genes, by each gene-level method. Only the *Corrected* Monte Carlo-based approach controls FWER at the target rate $\alpha = 0.05$ (indicated by the dashed red line), while all of the other MAGMA variants do not maintain valid control within plus/minus two standard errors. The *MAGMA: code* approach appears to be more conservative than using the *Two sided approximation* approach, likely from its adjustment power c. For context, Supplementary Fig. 3.6a displays the proportion of simulations with zero to six false positives with the Bonferroni correction for each method (excluding *MAGMA: paper*) while Supplementary Fig. 3.6b displays the distribution of the number of false positives represented by boxplots, with points denoting single simulation results, for each method (dashed red line indicates one observed false positive). The *Corrected* Monte Carlo-based approach is the only method that yield less than one false positive in the vast majority of simulations (as implied by the FWER control), while the other methods display a substantial number of simulations resulting in one to six false positives or, in the case of using the *MAGMA: paper* approach, between two to three hundred false positives. We observe similar patterns of inflation when applying BH for FDR control as displayed in Fig. 1d.

We investigate this further by examining the p-value distributions for the simulated sets of 53,001 genes. Fig. 1c displays histograms for the null gene-level p-values by method averaged over 1,000 simulations (standard errors are too small to be visible), with the black horizontal line denoting the ideal uniform distribution for null p-values. However, only the *Corrected* null p-value distribution displays such behavior. The different MAGMA variants, including the *Two-sided approximation*, display an inflation in smaller p-values along with non-uniform p-value distributions, such as fewer than expected number of genes for the bin of largest p-values (≥ 0.95). The observed inflated FWER and FDR in Supplementary Fig. 3.5 and Fig. 3.1d emphasize the failings of using these improper, non-uniform null distribution in typical multiple testing procedures.

3.4.3 Impact on gene-set analysis type 1 error control

The type 1 error rate for the gene-set simulations are displayed in Supplementary Fig. 3.7. Regardless of the gene-set analysis method, only the *Corrected* Monte Carlo-based approach displays control at the at the target type 1 error rate $\alpha = 0.05$ (as indicated by the dashed red line). The behavior for the other methods varies depending on the type of gene-set analysis. The different types of MAGMA gene-level p-values result in overly conservative results for the MAGMA gene-set analysis approach, while they display inflated error rates for both Fisher and Stouffer's method to varying degrees (with the notable exception that *MAGMA: code* is conservative with Stouffer's method).

The difference in behavior of the gene-set error rate control is driven by the non-uniform

p-value distribution. Supplementary Fig. 3.8 displays a comparison of the resulting gene-set null p-value distributions by gene-set analysis method (columns) and method for computing gene-level p-values (rows). We clearly see that for both the MAGMA and Stouffer's gene-set analysis methods, the *MAGMA: code* approach yields conservative p-value distributions. To further investigate this behavior, Supplementary Fig. 3.9 displays the resulting MAGMA gene-set analysis z_s distributions by the gene-level method, with the expected $t_{(G-1)}$ (where $G = 45$) distribution curve overlaid in red. We see that both the *MAGMA: paper* and *MAGMA: code* approaches are shifted to the left with more negative values. The use of the adjustment power c in the *MAGMA: code* approach, which applies a more conservative adjustment as the p-value decreases, has a noticeable downstream effect on the MAGMA gene-set analysis results. The Monte Carlo-based approaches are the only gene-level methods yielding appropriate null distributions and results regardless of the gene-set analysis method.

3.4.4 Results for H-MAGMA Replication

Supplementary Table 3.1 and Fig. 2b display the substantial reduction in the number of discoveries reported by H-MAGMA after using the *Corrected* Monte Carlo-based approach. Fig. 2a highlights the improper H-MAGMA ASD p-value distribution (based on adult Hi-C data annotations) consistent with behavior observed in our simulations (Fig. 1c).

3.5 Code availability

All code and data used in the chapter are available at `https://github.com/ryurko/HMAGMA-comment`.

Table 3.1: Comparison of number of BH discoveries (union across adult and fetal Hi-C data annotations) with H-MAGMA versus corrected Monte Carlo-based approach.

Phenotype	H-MAGMA	Corrected
ADHD	486	223
ASD	275	125
BD	1,931	1,111
SCZ	9,217	8,066
MDD	3,167	2,201

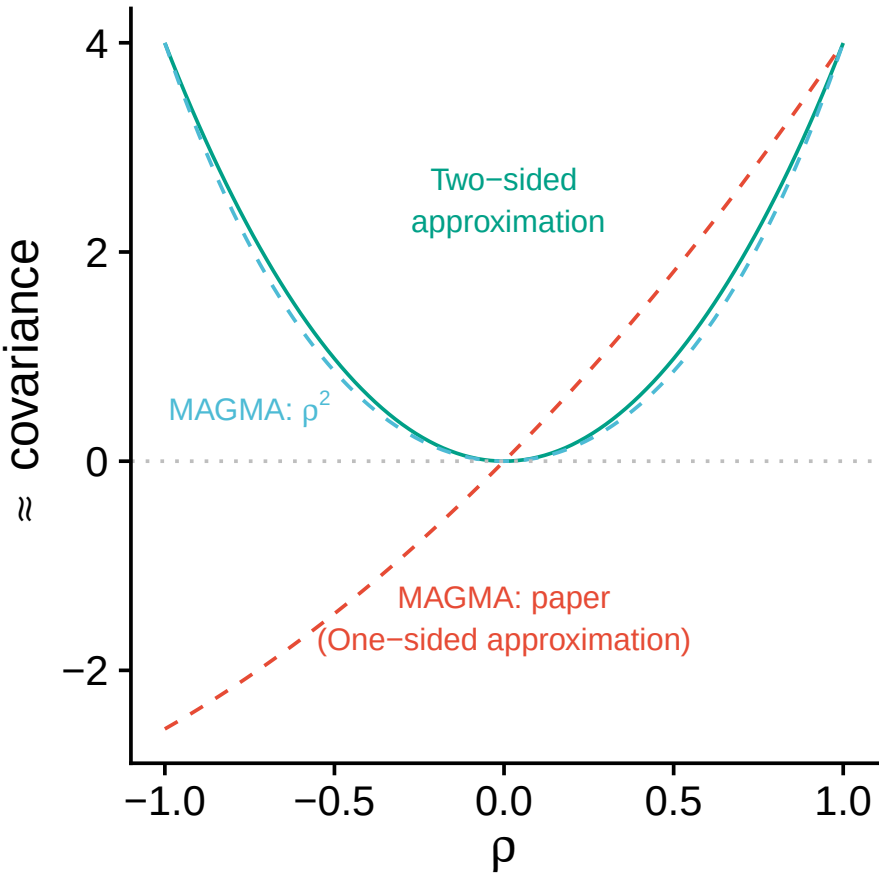

Figure 3.3: Comparison of covariance approximations for two-sided tests (green-solid) versus Brown's one-sided approximation described in MAGMA paper (red-dashed) and the code implemented in MAGMA with ρ^2 (cyan-dashed).

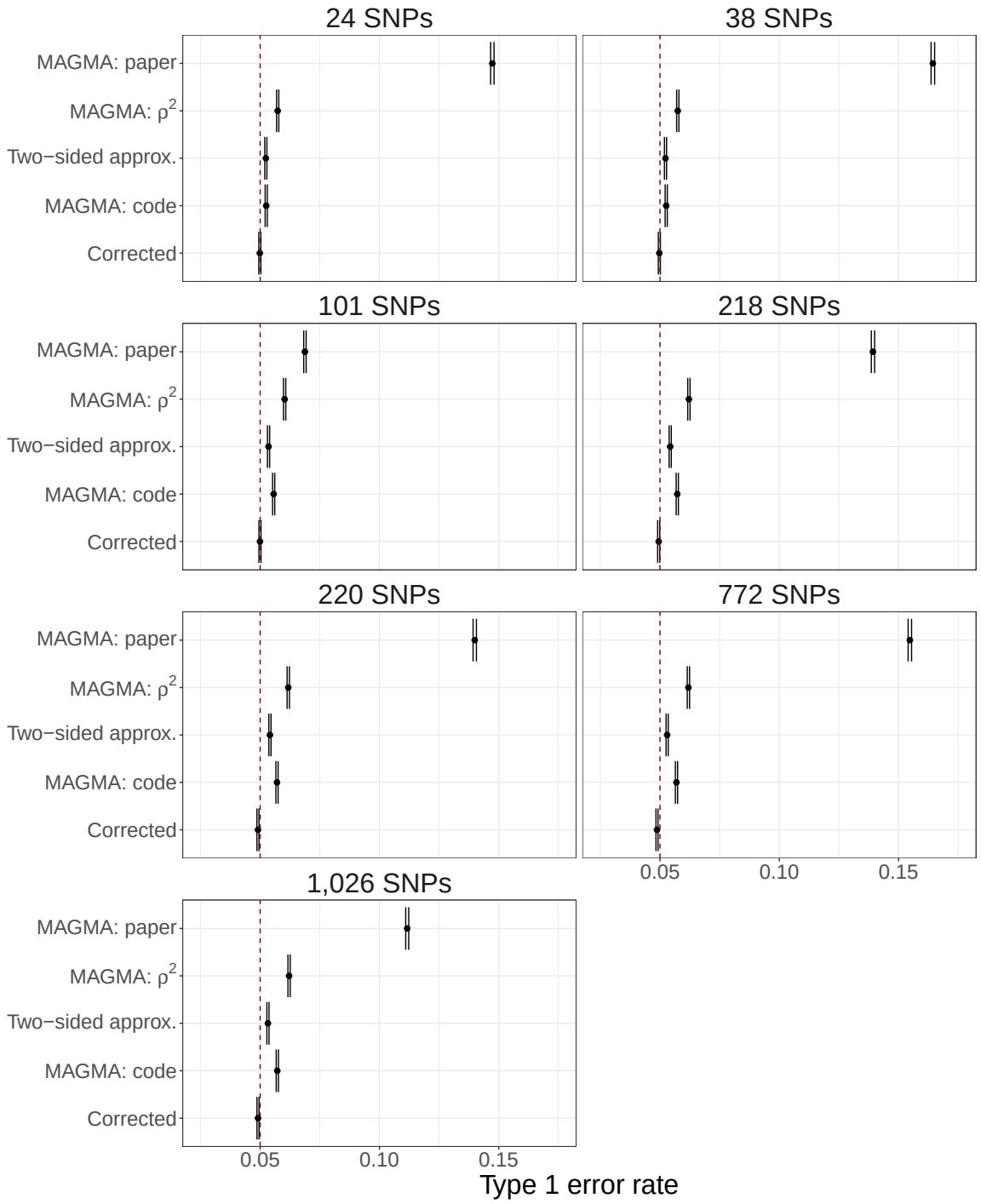

Figure 3.4: Comparison of the type 1 error rate control, plus/minus two standard errors, at target $\alpha = 0.05$ (denoted by dashed red line) by gene-level method for seven example genes of different sizes.

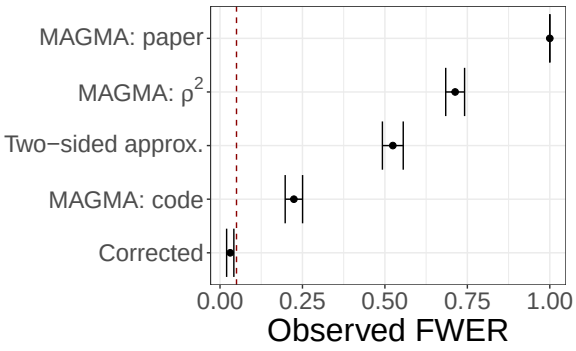

Figure 3.5: Comparison of multiple testing family-wise error rate (FWER), plus/minus two standard errors, at target $\alpha = 0.05$ (denoted with dashed red line) by gene-level method.

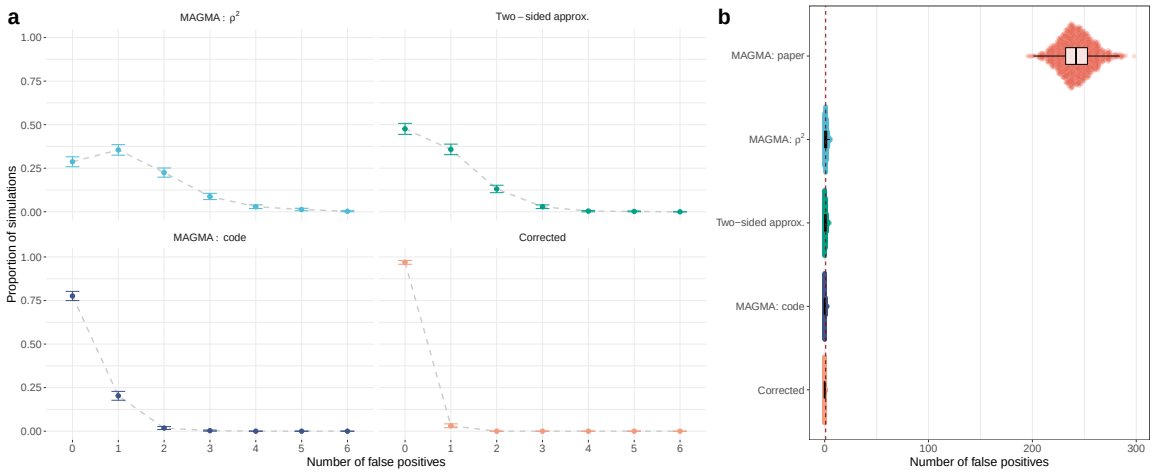

Figure 3.6: **a**, Comparison of the proportion of simulations with zero to six false positives for each method, excluding *MAGMA: paper*, using the Bonferroni correction at target $\alpha = 0.05$ (plus/minus two standard errors). **b**, Distribution of the number of false positives represented by boxplots, with points denoting single simulation results, for each method (dashed red line indicates one observed false positive).

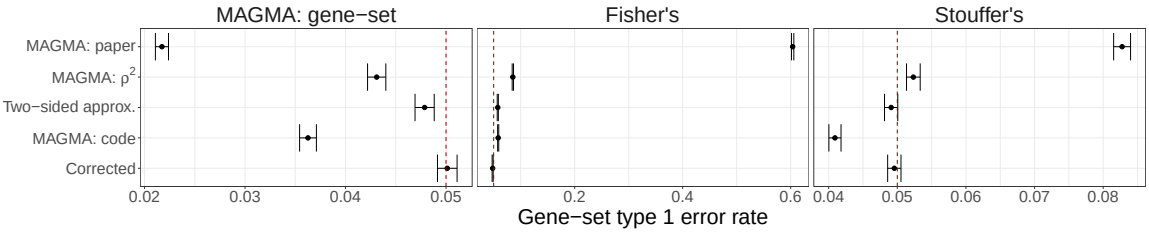

Figure 3.7: Comparison of gene-set analysis type 1 error, plus/minus two-standard errors, at target $\alpha = 0.05$ (denoted with dashed red line) by gene-level method for each considered gene-set analysis method (from left to right) MAGMA: gene-set, Fisher's combination test, and Stouffer's z-test.

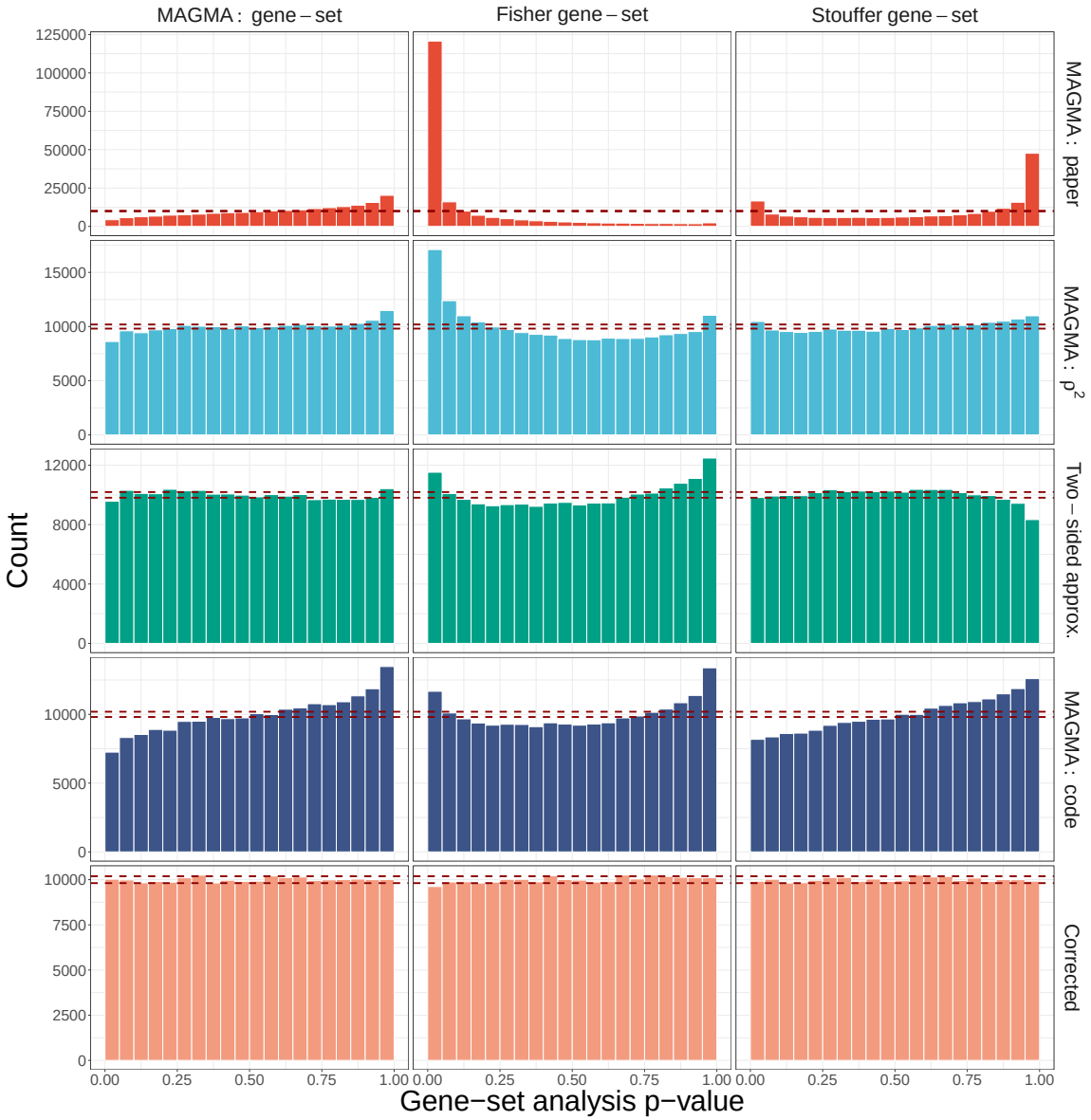

Figure 3.8: Comparison of null histograms by gene-set analysis method (columns) and methods for computing gene-level p-values (rows) for gene-sets of size $G = 45$ genes. Red dashed lines indicate expected 2.5% and 97.5% quantiles for uniform p-values across 200,000 simulations.

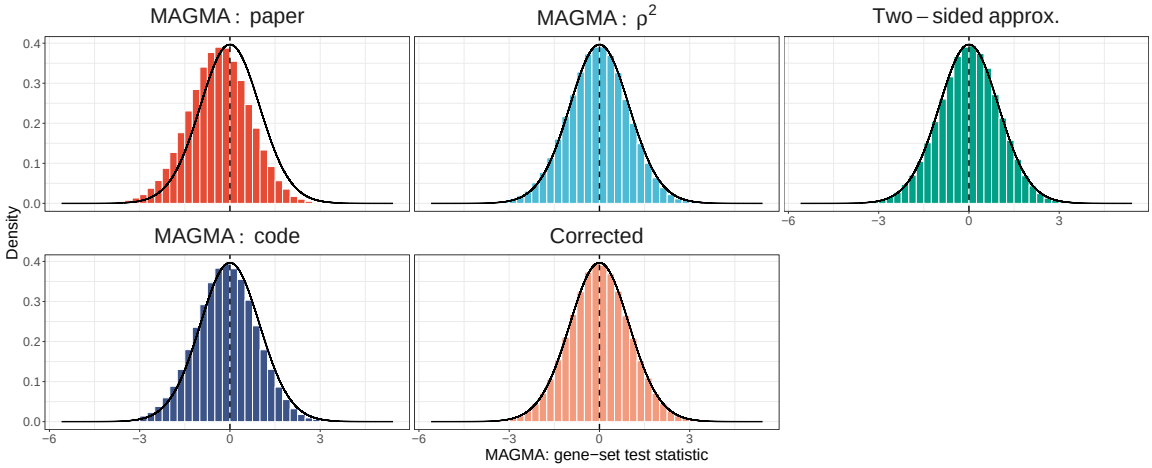

Figure 3.9: Comparison of distributions for MAGMA gene-set analysis standardized test statistic by methods for computing gene-level p-values. Red density curve displays the expected $t_{(G-1)}$ distribution, with the red vertical dashed line denoting the center at zero.

Four

An approach to gene-based testing accounting for dependence of tests among nearby genes

In genome-wide association studies (GWAS), it has become commonplace to test millions of SNPs for phenotypic association. Gene-based testing can improve power to detect weak signal by reducing multiple testing and pooling signal strength. While such tests account for linkage disequilibrium (LD) structure of SNP alleles within each gene, current approaches do not capture LD of SNPs falling in different nearby genes, which can induce correlation of gene-based test statistics. We introduce an algorithm to account for this correlation. When a gene's test statistic is independent of others, it is assessed separately; when test statistics for nearby genes are strongly correlated, their SNPs are agglomerated and tested as a locus. To provide insight into SNPs and genes driving association within loci, we develop an interactive visualization tool to explore localized signal. We demonstrate our approach in the context of weakly powered GWAS for autism spectrum disorder, which is contrasted to more highly powered GWAS for schizophrenia and educational attainment. To increase power for these analyses, especially those for autism, we use adaptive p-value thresholding (AdaPT), guided by high-dimensional metadata modeled with gradient boosted trees, highlighting when and how it can be most useful. Notably our workflow is based on summary statistics. This chapter appears in [Yurko et al., 2021b].

4.1 Introduction

More than 3,000 human GWAS have examined over 1,800 diseases and traits, with uneven success in discovering associations [MacArthur et al., 2017]. For schizophrenia, for example, 280 discoveries were recently announced, while, for genetically correlated autism spectrum disorder, a handful of loci have been discovered [Grove et al., 2019]. The difference largely is due to statistical power. To increase power, one might decrease the number of hypotheses tested and thus reduce the threshold for significance. A natural strategy is gene-based

testing: SNPs are assigned to genes they occur in or nearby [de Leeuw et al., 2015]; within this unit, test statistics for SNPs are aggregated; and, finally, significance is judged by the number of genes tested. By focusing tests on genes instead of SNPs dispersed throughout the genome, gene-based testing also has interpretability as an appealing feature. Power can also be enhanced by choosing false discovery rate (FDR) control for significance testing. These two options, gene-based testing and FDR control, are not mutually exclusive. H-MAGMA [Sey et al., 2020] combines them and also incorporates Hi-C data into its testing scheme. Likewise, when SNPs affect gene expression, these functional SNP-to-gene assignments can be modeled [Gerring et al., 2019].

A related approach is to increase power by incorporating metadata about SNPs or genes in the targeting of multiple testing procedures; selective inference provides approaches to incorporating this information while maintaining valid FDR control. An early approach incorporated metadata directly through the use of p-value weights [Genovese et al., 2006]. More recently, in the setting of SNP-based GWAS, we [Yurko et al., 2020] implemented a data-driven approach to determine weights via the adaptive p-value thresholding (AdaPT) framework [Lei and Fithian, 2018]. In brief [Yurko et al., 2020], we improved power for detecting a subset of weakly correlated SNPs by using gradient boosted trees to model potentially-informative metadata, such as known effects of SNPs on gene expression and on genetically correlated traits.

Here we explore the use of AdaPT in the context of gene-based tests for autism (ASD), schizophrenia (SCZ), and educational attainment (EA), placing special emphasis on how it enhances power to detect associations of genes with ASD. To do so, we utilize gene-based testing methods introduced to account for linkage disequilibrium (LD) among SNPs in a gene [Liu et al., 2010, Mishra and Macgregor, 2015, Yurko et al., 2021a]. LD is not limited by gene boundaries, however. To the contrary, LD among SNPs falling in different genes is commonplace. This compromises the interpretability of current gene-based tests, obscuring the meaning of error guarantees with family-wise error rate and FDR controlling procedures. Because of the extensive and heterogeneous LD in the genome, in our prior SNP-based GWAS using AdaPT, we purposely selected quasi-independent SNPs for analysis. By contrast, for gene-based testing, we introduce an agglomerative algorithm to account for LD-induced correlation of test statistics. This algorithm directly groups genes into 'loci' for which between-loci test statistic correlation—based on LD—is bounded above. This reduces the set of tests to a collection of weakly correlated genes and loci. Importantly, this agglomerative algorithm can be used in any gene-based testing framework to highlight gene-based tests that are dependent.

We analyze results from three GWAS: ASD[Grove et al., 2019], SCZ[Ruderfer et al., 2018b] and EA[Lee et al., 2018]. Using AdaPT guided by loci metadata as our example, we are able to improve power to select ASD-associated genes and LD-defined loci with multiple genes

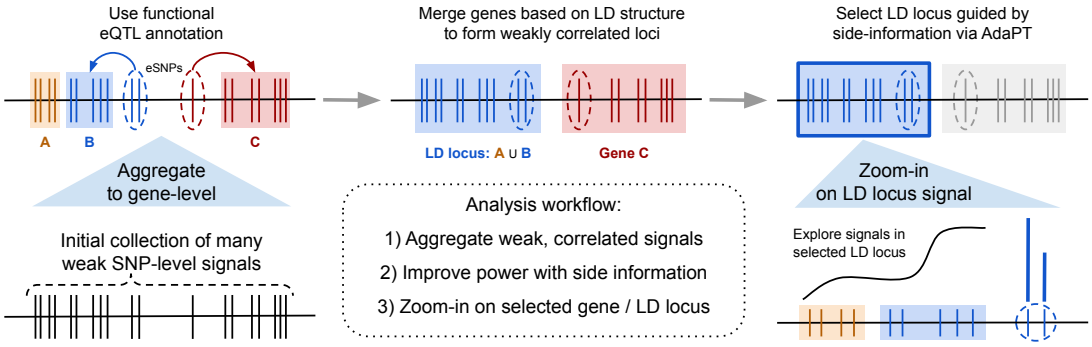

Figure 4.1: Schematic of our workflow to improve power with metadata and highlight interesting signals while adjusting for LD structure.

while maintaining finite-sample FDR control. Improvements are more modest for the other phenotypes, due to the high power of their original GWAS. One novel feature of our analyses is that it groups genes into loci when their test statistics are expected to be highly dependent, due to LD. We complement this feature with graphical tools to examine the distribution of association signal within each such locus. The interactive visualization tool we develop uses `R` Shiny [R Core Team, 2020, Chang et al., 2020] and `plotly` [Sievert, 2020] for exploring and highlighting biological signals therein.

Our workflow for improving power to select associated genes and LD-defined loci follows (Figure 4.1): We first introduce the agglomerative algorithm. Then we demonstrate our approach for detecting associations with in the AdaPT framework. For metadata, we use eQTL data from cortical tissue samples, gene co-expression networks [Zhang and Horvath, 2005], and GWAS results for the other phenotypes, which is motivated by the observation that all three are genetically correlated[Weiner et al., 2017]. Relationships of metadata to gene-based statistic are uncovered by use of gradient boosted trees. We separate our results into two categories, technical features related to gene-based testing and implementation of AdaPT in this setting, and association results for these phenotypes and their implications, with special emphasis on ASD.

4.2 Methods

4.2.1 SNP-to-gene assignment and correlation between gene-level tests

We consider two different approaches for assigning n SNPs to a set of genes G. Following common practice, we use genomic location only for "Positional" SNP assignment: SNP i is assigned to gene g if its genomic position is within gene g's start and end positions (S_g^{Pos}).

In some published analyses, the start and end positions are expanded slightly to include presumed regulatory regions, although we do not do so here. As an alternative approach, recognizing that some SNPs have documented effects on genes, such as eQTL effects, we use Positional SNP assignment and include in the set cis-eQTL SNP-gene pairs, which we dub eSNPs. We denote the collection of eSNPs for gene g as S_g^{eQTL} and call this "Positional + eSNPs" SNP assignment.

After assigning SNPs to genes, each gene's vector of SNP-level z statistics $\boldsymbol{z}_g$ is modeled as multivariate normal (Gaussian) with mean 0 and LD-induced covariance $\boldsymbol{\Sigma}_g$. Following common practice [Liu et al., 2010], we assume $\boldsymbol{\Sigma}_g$ is known using correlations from reference genotype data, specifically 503 individuals from the 1000 Genomes EUR sample as the reference data [1000 Genomes Project Consortium and others, 2012]. Gene-level testing frameworks [de Leeuw et al., 2015, Liu et al., 2010, Mishra and Macgregor, 2015] combine SNP-level signals into gene-level test statistics T_g while accounting for $\boldsymbol{\Sigma}_g$, the LD among SNPs in a gene.

Consider the quadratic gene-level test statistics, $T_g = \boldsymbol{z_g}^T \boldsymbol{z_g}$, featured in *VEGAS* [Liu et al., 2010, Mishra and Macgregor, 2015] and MAGMA (v1.08). Constructed in this way, the quadratic test statistic is merely the sum of the individual z_g^2's at every SNP in the gene. Under the null model, if the $|S_g|$ SNPs had been independent, then the test statistic T_g would have been approximately χ^2 distributed with $|S_g|$ degrees of freedom, and thus have an expectation of $|S_g|$ with variance $2|S_g|$. Here we are concerned with the case where the individual test statistics are dependent. The expectation is unchanged by dependence, but the variance in the sum now becomes $\sum_{i \in S_g} \mathrm{Var}(z_i^2) + 2 \sum_{i<j \in S_g} \sum \mathrm{Cov}(z_i^2, z_j^2)$, which gives equation

$$\mathrm{Var}(T_g) = 2|S_g| + 2 \sum_{i<j \in S_g} \sum (2\rho_{ij}^2). \tag{4.1}$$

Let S_g and $S_{g'}$ denote the sets of SNPs for g and g', respectively. We can compute the induced correlation between the quadratic test statistics for two sets of SNPs, $\mathrm{Cor}(T_g, T_{g'}) = \mathrm{Cov}(T_g, T_{g'}) / \sqrt{\mathrm{Var}(T_g) \cdot \mathrm{Var}(T_{g'})}$, using the induced covariance between the test statistics,

$$\mathrm{Cov}(T_g, T_{g'}) = \sum_{i \in S_g} \sum_{j \in S_{g'}} (2\rho_{ij}^2). \tag{4.2}$$

For nearby gene sets g and g', these correlations can be quite strong. This can confound the interpretability—and even meaning—of the guarantees from multiple testing procedures. To avoid these issues, we will modify the construction of our gene sets to bound the pairwise correlation between the resulting gene sets.

4.2.2 Agglomerative LD loci testing

To account for substantially correlated test statistics, we introduce an agglomerative procedure to group highly correlated genes that are within w megabases (Mb) into sets of genes we refer to as LD loci or simply loci. Given an LD threshold r^2, we apply the following procedure to a set of genes G within a chromosome:

1. Compute $\text{Cor}(T_g, T_{g'})$ for all pairs of genes, $g, g' \in G$, within w Mb of each other (using Equations 4.2 and 4.1).

2. Repeat the following until $(\text{Cor}(T_g, T_{g'}))^2 < r^2$ for all remaining pairs of genes/loci in G:

   - Find genes/loci $\{g_*, g'_*\} = \underset{g,g' \in G, g \neq g'}{\arg\max} \text{Cor}(T_g, T_{g'})$,
   - Merge $\{g_*, g'_*\}$ into locus g_{LD},
   - Update $G = G \setminus \{g_*, g'_*\} \cup \{g_{LD}\}$, and compute $\text{Cor}(T_{g_{LD}}, T_{g'})$ for all $g' \in G$ within w Mb of g_{LD}.

This is essentially agglomerative hierarchical clustering, but with a linkage determined by the LD-based correlation structure of the test statistics. We compute the quadratic test statistic, T_g, for each remaining gene/locus $g \in G$.

Because the resulting distribution T_g does not have a known closed-form solution, we use a Monte Carlo based approach for computing the p-value p_g for the gene/locus to test the null hypothesis that its n_g-dimensional vector of SNP-level z statistics are not associated with trait status. We generate B draws of null, n_g-dimensional Gaussian random variables, $\boldsymbol{z}_g^* \sim \text{Normal}(\mathbf{0}_g, \boldsymbol{\Sigma}_g)$. A quadratic test statistic T_b^* is calculated for each of the $b \in [B]$ draws, resulting in an empirical p-value:

$$p_g = \frac{\sum_b^B 1(T_b^* > T_g) + 1}{B + 1}. \tag{4.3}$$

To generate the B samples, we use the Cholesky decomposition of $\boldsymbol{\Sigma}_g = \mathbf{L}\mathbf{L}^{\mathrm{T}}$, where $\mathbf{L}$ is a lower triangular $n_g \times n_g$ matrix. A single sample is generated by multiplying $\mathbf{L}$ with a vector of n_g independent, standard Gaussian random variables. Across B samples, the combined computational cost is $O(n_g^3 + Bn_g^2)$. Using an efficient implementation of these steps [Yurko et al., 2021a], for a gene/locus with $n_g \approx$ 1,000 SNPs, we can generate one million draws in less than ten minutes with a 3.6 GHz Intel processor. Parallelization can be used to further increase speed.

4.2.3 Overview of GWAS data and eQTL sources

Our investigation focuses on reported GWAS z statistics, $\{z_i, i = 1, \ldots, n\}$ measuring SNP-level association with ASD [Grove et al., 2019], SCZ [Ruderfer et al., 2018b] and EA [Lee et al., 2018]. For GWAS results of one phenotype, we explore SNP-level association statistics from the other two GWAS as potential sources of metadata due to previous evidence of their genetic correlation [Weiner et al., 2017]. We consider $n = 5,238,256$ SNPs whose alleles could be aligned across all three phenotypes and with minor allele frequency (MAF) > 0.05 based on the 1000 Genomes EUR sample reference data [1000 Genomes Project Consortium and others, 2012]. Also, for these SNPs, their hg19 variant locations could be converted to GRCh38 using the LiftOver utility from the UCSC Genome Browser (`http://genome.ucsc.edu/`). Probably due to a smaller sample size, ASD has lower power: 18,381 cases and 27,969 controls, in comparison to SCZ with 33,426 cases and 32,541 controls, and EA with ≈ 1.1 million subjects (Figure 4.7). Because we focus on detecting associations for ASD, a neurodevelopmental disorder, we leverage two different sources of cortical tissue to identify eSNPs for functional SNP-to-gene assignment. The first source of eSNPs was obtained from the BrainVar study of dorsolateral prefrontal cortex from 176 individuals sampled across a developmental span [Werling et al., 2020b]. We identified 151,491 cis-eQTL SNP-gene pairs meeting BH $\alpha \leq 0.05$ for at least one of the three sample sets: prenatal (112 individuals), postnatal (60 individuals), as well as across the complete study. This corresponds to 123,664 eSNPs associated with 6,660 genes, with 85% of the eSNPs associated with one unique cis-eQTL gene pairing.

The second source is adult cortical tissue cis-eQTLs from the Genotype-Tissue Expression (GTEx) V7 project dataset [GTEx Consortium and others, 2015]. Instead of using eQTLs as reported by GTEx, to be consistent with the BrainVar eQTL definition, we identified 414,405 cis-eQTL SNP-gene pairs meeting BH $\alpha \leq 0.05$ for either *Frontal Cortex BA9* or *Anterior cingulate cortex BA24* samples based on the tissue specific files for all SNP-gene associations available at `gtexportal.org`. This resulted in 313,316 GTEx eSNPs associated with 9,012 genes, where 78% of the eSNPs are associated with one gene. However we observe an overlap of 55,313 cis-eQTL SNP-gene pairs with BrainVar, culminating in 510,583 unique cis-eQTL SNP-gene pairs with 370,749 eSNPs associated with 12,854 genes across the union of BrainVar and GTEx sources.

4.2.4 GENCODE version

We use GENCODE v21 [Harrow et al., 2012] for our list of genes with their respective start and end positions based on genome assembly version GRCh38. This matches the version used in the BrainVar study, but differs from GTEx, which is based on v19. When identifying GTEx eQTLs, we removed 187 genes from GENCODE v19 that do not match Ensembl IDs in v21. This provides us with an initial list of $G = 57,005$ candidate genes to potentially assign SNPs to, based on either positional or functional eSNP status.

4.2.5 Metadata

For each gene/locus g, we create a vector of metadata x_g collected independently of p_g. This process is completed in the same manner for both the *Positional* and *Positional + eSNPs* collections. First, we consider the number of SNPs assigned to a gene/locus, $n_g = |S_g|$, which can be viewed as statistical information relevant to the power of the quadratic test statistic. Additionally, we include one-sided z statistics, i.e., $z_g = \Phi^{-1}(1 - p_g)$, constructed using the gene/locus-level p-values from independent GWAS results. For our target phenotype ASD we use SCZ and EA GWAS z statistics, z_g^{SCZ} and z_g^{EA}, while for SCZ (EA) we use z_g^{EA} (z_g^{SCZ}) and z_g^{ASD} as metadata.

Given the set of eSNPs S_g^{eQTL} associated with single genes or genes in LD locus g, we summarize the expression level as the average absolute eQTL slope in a relevant source to obtain $\bar{\beta}_g^{source}$ for five sources: three BrainVar developmental periods (pre-, post-, and complete) and two adult GTEx cortical regions. Furthermore, we account for weighted gene co-expression network analysis (WGCNA) [Zhang and Horvath, 2005] modules by creating two sets of indicators, one set for the twenty modules reported in the BrainVar study and another for eight modules constructed using the GTEx cortical tissue samples. For simplicity, we also construct indicators denoting if gene/loci is not included in any of the modules.

We also include additional context about the gene/loci. Indicator variables determine four GENCODE biotypes: protein coding, antisense, long non-coding RNA, and other. Using gnomAD v2.1.1 [Karczewski et al., 2020], we associate with each gene its loss-of-function observed / expected upper fraction (LOEUF) value, which indicates the gene's tolerance to loss-of-function. Because a lower LOEUF scores indicate strong selection against loss-of-function, we include the minimum LOEUF across all genes in an LD locus in our vector of metadata x_g.

4.2.6 AdaPT implementation

Given a collection of gene/locus-level p-values and metadata, $(p_g, x_g)_{g \in G}$, we apply AdaPT to select a subset of discoveries with FDR control at target level $\alpha = 0.05$. AdaPT is guaranteed finite-sample FDR control under the assumption of independent null p-values, and was demonstrated to maintain control in weak, positive correlated scenarios [Yurko et al., 2020], such as the one considered here.

We incorporate metadata from the feature space $x_g \in \mathcal{X}$ using XGBoost [Chen and Guestrin, 2016] (see *Method Appendix* for details). XGBoost is a popular implementation of gradient boosted trees, which constructs a flexible predictive function as a weighted sum of many simple trees, fit using a gradient descent procedure that minimizes a specified objective function. The two objective functions considered in our AdaPT context correspond to estimating the probability that a hypothesis is non-null, and the distribution of effect size for non-null hypotheses. XGBoost gives us flexibility to include many potentially useful covariates

without being overly concerned about the functional form with which they enter the model or their marginal utility. However, overfitting in this context will lead to a loss of power. To find appropriate settings for the gradient boosted trees (number of trees, learning rate, and maximum depth), we first "tune" AdaPT's performance with synthetic SCZ p-values that are aligned with the ASD p-values in the following manner:

1. Sort SCZ and ASD p-values: $(p^{SCZ}_{(1)}, \ldots, p^{SCZ}_{(G)})$ and $(p^{ASD}_{(1)}, \ldots, p^{ASD}_{(G)})$
2. Replace SCZ with ASD p-values by matching order, i.e., replace $p^{SCZ}_{(1)}$ with $p^{ASD}_{(1)}$, $p^{SCZ}_{(2)}$ with $p^{ASD}_{(2)}$, ...

This transforms the SCZ signal to match the weaker signal in ASD. We then proceed to apply AdaPT using these synthetic SCZ p-values to find candidate settings which yield the highest number of synthetic SCZ discoveries at FDR level $\alpha = 0.05$. Finally, for each phenotype and positional assignment, we use two cross-validation steps within AdaPT [Yurko et al., 2020] to select from among these candidate settings to generate our AdaPT: XGBoost results using the `adapt_xgboost_cv()` function in the `adaptMT` `R` package (available at `https://github.com/ryurko/adaptMT`).

4.2.7 Kernel smoothing localization

Following the selection of interesting genes/loci, researchers may be interested in "zooming" in on localized signals at the gene and/or SNP-level. For a selected gene/locus g^* and its corresponding set of SNPs S_{g^*}, we smooth over the squared z statistics of the locus' positional SNPs $S^{Pos}_{g^*} \subseteq S_{g^*}$, given their genomic positions (BP) using the Nadaraya–Watson estimator:

$$\hat{z}^2_{g^*}(BP) = \sum_{i \in S^{Pos}_{g^*}} z_i^2 \frac{K_h(BP_i, BP)}{\sum_{j \in S^{Pos}_{g^*}} K_h(BP_j, BP)}, \tag{4.4}$$

in which K_h is a one-dimensional Gaussian kernel with bandwidth h selected separately for each gene/locus using generalized cross-validation (as implemented in the `np` package [Hayfield and Racine, 2008]). We then provide the option to display any subset of eSNPs $S^{eQTL}_{g^*} \subseteq S_{g^*}$ separately as bars with their heights indicating individual SNP-level signal. This separation is due to the presence of intergenic eSNPs, however any eSNPs that are also positionally assigned to genes, $S^{Pos}_{g^*} \cap S^{eQTL}_{g^*}$, are included in the positional smoothing in Equation 4.4.

4.3 Results

4.3.1 Assigning SNPs to genes and generating LD loci

We assign SNPs to genes using the two approaches: *Positional*, which assigns 2,779,780 SNPs to 40,581 genes; and *Positional + eSNPs*, which includes an additional 109,042 intergenic

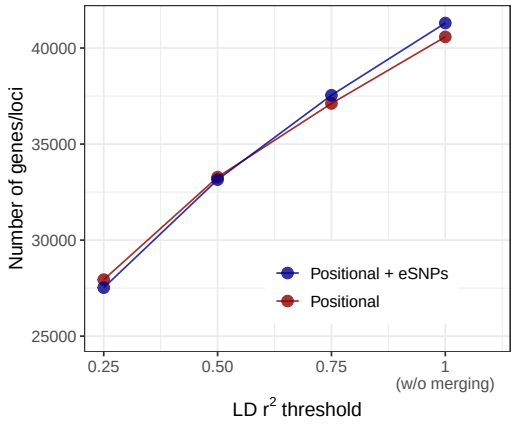

Figure 4.2: Comparison of the number of genes/loci following our agglomerative algorithm over a range of values for the induced LD threshold $r^2 \in \{0.25, 0.50, 0.75\}$ by positional type (in color). The initial number of genes is provided for reference (corresponding to $r^2 = 1$).

cortical tissue eSNPs resulting in 2,888,822 SNPs assigned to 41,301 genes. Next, we generate genes/loci based on the LD-induced correlation of gene-based test statistics using the agglomerative algorithm with window size $w = 6$ Mb and with one of three thresholds for r^2, 0.25, 0.50, 0.75. The number of independent genes/loci decreases substantially as the threshold becomes more strict for both SNP assignment types (Figure 4.2). Even a relatively high threshold of $r^2 = 0.75$ reduces the number of *Positional + eSNPs* (*Positional*) gene/locus tests from 41,301 (40,581) to 37,522 (37,114). We report the conservative threshold $r^2 = 0.25$ in the body of the manuscript (see *Method Appendix* for results with $r^2 \in \{0.50, 0.75\}$). Due to the conservative threshold we combine 17,915 genes to form 4,136 LD loci for *Positional + eSNPs* and 16,625 genes to form 3,985 LD loci for the *Positional* approaches. Over 75% of these loci contain five or fewer genes while the largest is a chromosome 11 locus that groups over sixty genes, most of which encode olfactory receptors. Combined with the 23,386 and 23,956 individual genes that were not merged, this results in 27,522 and 27,941 genes/loci for testing. The reduction of independent testing units highlights the correlation among genes that is often ignored in gene-based testing. We then compute the gene/locus quadratic test statistics and p-values for each phenotype using the Monte Carlo-based approach in Equation 4.3 with B equal to two million simulations.

4.3.2 AdaPT models and results

To generate AdaPT: XGBoost results, we first tune the procedure based on synthetic SCZ p-values, which mimic the distribution of ASD p-values, to find optimal XGBoost settings. To avoid over-fitting, we consider shallow trees with maximum depth $\in \{1, 2\}$, while searching over the number of trees $P \in 100, \ldots, 450$ by increments of fifty and the learning rate

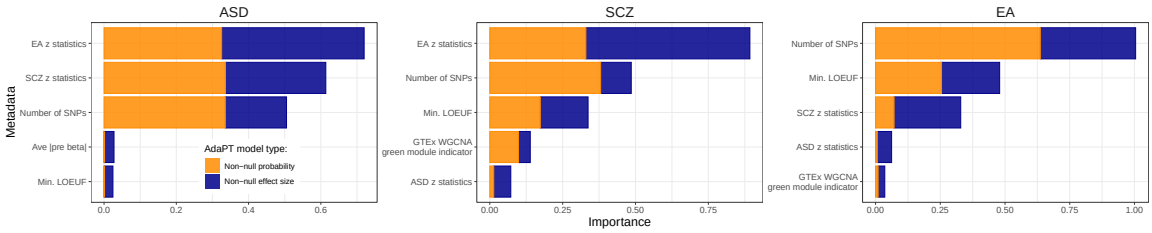

Figure 4.3: The most important variables (top five) predicting association with phenotype as ranked by XGBoost for the *Positional + eSNPs* assignment of SNPs. Variables are sorted in order of importance, while the color of the bars denote the separate parameters for the AdaPT implementation: probability of association (orange) and of non-zero effect size (blue).

$\eta \in 0.03, \ldots, 0.06$ by increments of 0.01. To analyze the real p-values for each phenotype and thereby select associated genes/loci, then, these top setting combinations (Table 4.1) were used in the AdaPT cross-validation steps and while targeting FDR control at $\alpha = 0.05$.

To find genes/loci associated with each phenotype, our AdaPT: XGBoost implementation fits a mixture model that requires two parameters – the probability of association and the effect of each gene/locus on the phenotype – and these are estimated separately for each step of the algorithm. Variables in the metadata inform on each of these parameters to different degrees; see *Method Appendix* and Table 4.2 for details on measuring variable importance. For the *Positional + eSNPs* results, the number of SNPs per gene/locus and z-statistics for at least one genetically correlated trait are important predictors for all three phenotypes (Figure 4.3). SCZ and EA, in contrast to ASD, display increased importance for LOEUF and membership in a WGCNA module constructed from the GTEx cortical tissue samples (Figure 4.3) Lower LOEUF values, which indicate lower tolerance to loss-of-function, were more likely to be associated with SCZ and EA. We observe similar patterns of variable importance for the *Positional* results (Figure 4.8).

Comparing the number of genes/loci selected by AdaPT to baseline results of intercept-only versions of AdaPT and BH, there is a clear gain in gene/locus discovery by accounting for metadata through the AdaPT: XGBoost implementation, regardless of phenotype and SNP-to-gene assignment approach (Figure 4.4A, Table 4.3, see also *Method Appendix*, Figures 4.9 and 4.10 for results with LD threshold of $r^2 \in \{0.50, 0.75\}$). Unsurprisingly, the number of associated genes/loci is much larger for SCZ and EA than for ASD, likely due to the lower power of the original ASD GWAS. For *Positional* ASD, we see that the intercept-only version of AdaPT fails to select any genes/loci due to the weak signal without inclusion of metadata (Figure 4.4A).

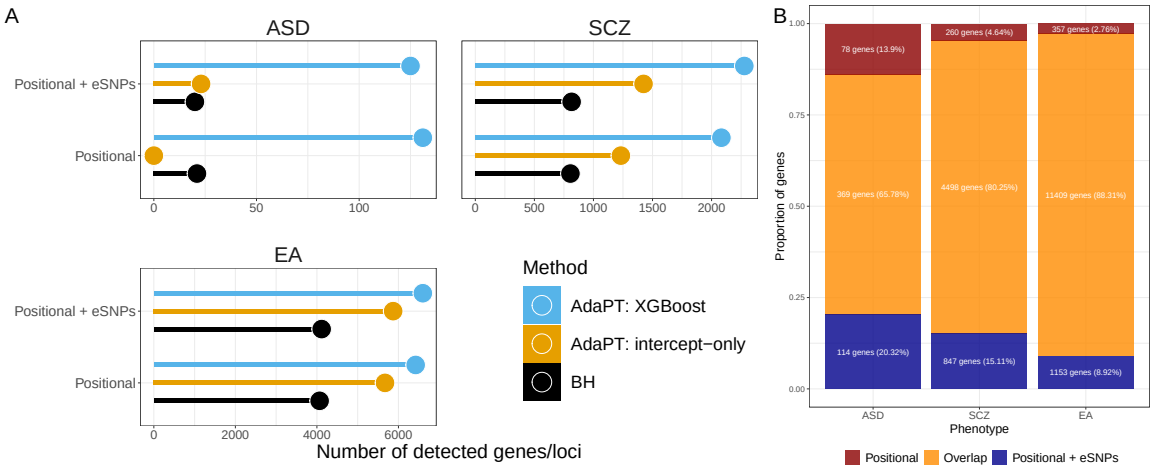

Figure 4.4: (A) Comparison of the number of selected genes/loci at FDR level $\alpha = 0.05$ for each phenotype by positional assignment. AdaPT: XGBoost results are presented in comparison to BH and AdaPT: intercept-only baselines which do not account for metadata. *(B)* Comparison of the proportion of implicated genes that overlap between the two types of positional assignment results, based on the AdaPT: XGBoost results for each phenotype.

4.3.3 Comparison of phenotypic results

For ASD, analysis of *Positional + eSNPs* identifies 483 genes, of which 405 cluster in 47 loci and 78 are unclustered, whereas analysis of *Positional* SNPs alone yields 447 genes, of which 370 cluster in 54 loci (Table 4.3). A substantial portion of these genes overlap (Figure 4.4B). While similar patterns emerge for SCZ and EA, the ratio of unclustered to clustered genes increases with increasing number of genes/loci detected: 0.193 for ASD, 0.414 for SCZ, and 0.681 for EA. This presumably reflects greater power to detect small effects of a SNP on phenotype with larger sample size: decay of this signal tends to cause it to fall below the threshold of detection for SNPs in nearby genes. In contrast, the proportion of genes uniquely identified when eSNP information is included is substantially higher for ASD than it is for SCZ or EA (Figure 4.4B), again likely due to lower power for the ASD sample.

As expected, for all three phenotypes, the number of unclustered genes increases with increasing threshold r^2 (Table 4). If, however, we assume that signal should be sparsely distributed across the genome, then the sum of LD loci and unclustered genes, for $r^2 = 0.25$, should be a reasonable estimate of genes associated given the current data. This translates into 125 genes for ASD, 2,277 for SCZ, and 6,598 for EA, and correspondingly 0.30, 5.51, and 15.98% of the total 41,301 genes analyzed (including protein-coding and non-coding genes). For SCZ and EA, the total number of associated genes per chromosome declines strongly and linearly with chromosome size (Figures 4.11-4.13), which itself is correlated

with the total number of genes per chromosome. This pattern, while more variable, is also seen in a breakdown of genes into protein-coding and other non-coding types (Figure 4.12). For ASD, however, an unusually large number of genes are associated on chromosomes 3, 8, 15, and 17, relative to chromosome size, and the associated gene counts show only a modest relationship with chromosome size (Figures 4.11-4.13), presumably due to lower power and resulting lower number of associated genes per chromosome.

4.3.4 Exploring signal in selected genes/loci

LD loci, as we define them, are expected to exhibit correlated association signal. Nonetheless, the signal is unlikely to be distributed evenly across the locus, instead in many instances it will be concentrated near one or more SNPs generating the signal, depending on the LD pattern in the locus. The same is true for signal in non-clustered genes. To interactively explore localized signal within genes/loci, we developed an LD locus zoom application using `R` Shiny [R Core Team, 2020, Chang et al., 2020] and `plotly` [Sievert, 2020] (available here: `https://ron-yurko.shinyapps.io/ld_locus_zoom/`). This tool displays the gene/locus of interest, represents genes by their location therein, and highlights the association signal by kernel smoothing for positional SNP signals, including interpolation.

3.7 Mb deletion region in chromosome 17:. One of the associated loci, roughly 500 Kb, falls in the well-known 3.7 Mb 17p11.2 deletion/duplication region associated with Smith-Magenis syndrome (deletion, OMIM:182290) [Seranski et al., 1999] and Potocki-Lupski syndrome (duplication, OMIM:610883) [Neira-Fresneda and Potocki, 2015]. The associated LD locus displays overlapping genes and signals from eSNPS and positional SNPs for ASD, SCZ, and EA (Figure 4.5A). For reference, the gray dotted line denotes a point-wise 95^{th} percentile of 1,000 simulations for squared null Gaussian random variables, $z^2_{g^*}$ where $z_{g^*} \sim \text{Normal}(\mathbf{0}_{g^*}, \mathbf{\Sigma}_{g^*})$, given the LD structure $\mathbf{\Sigma}_{g^*}$ of the selected LD locus g^*. Notably, a single gene in the locus has been associated with all three phenotypes to various degrees, *RAI1* [Carmona-Mora and Walz, 2010]. Known to be dosage-sensitive, both deletion and duplication of a single copy of the gene is sufficient to elevate the likelihood of ASD substantially and to diminish cognitive function strongly [Carmona-Mora and Walz, 2010]. Curiously, only the smoothed signal for EA association peaks over *RAI1*, whereas it peaks over *TOM1L1* for ASD and further proximal for SCZ. The peak association for EA over *RAI1* also coincides with the SNP having the lowest p-value for association with this phenotype: this index SNP rs11655029 has p-value 2.84×10^{-9} and falls in an intron of *RAI1*. While *RAI1* is a prime candidate for the target gene in this locus for ASD, due to prior evidence, the peak signal located over *TOM1L2* is intriguing, especially because its index SNP rs4244599, while not strongly associated (p-value = 0.0018), is an eQTL for *TOM1L2*. The index SNP for SCZ, rs8082590, has p-value 6.02×10^{-8}, close to the traditional threshold for GWAS, and coincides roughly with its smoothed signal for association over *GID4*.

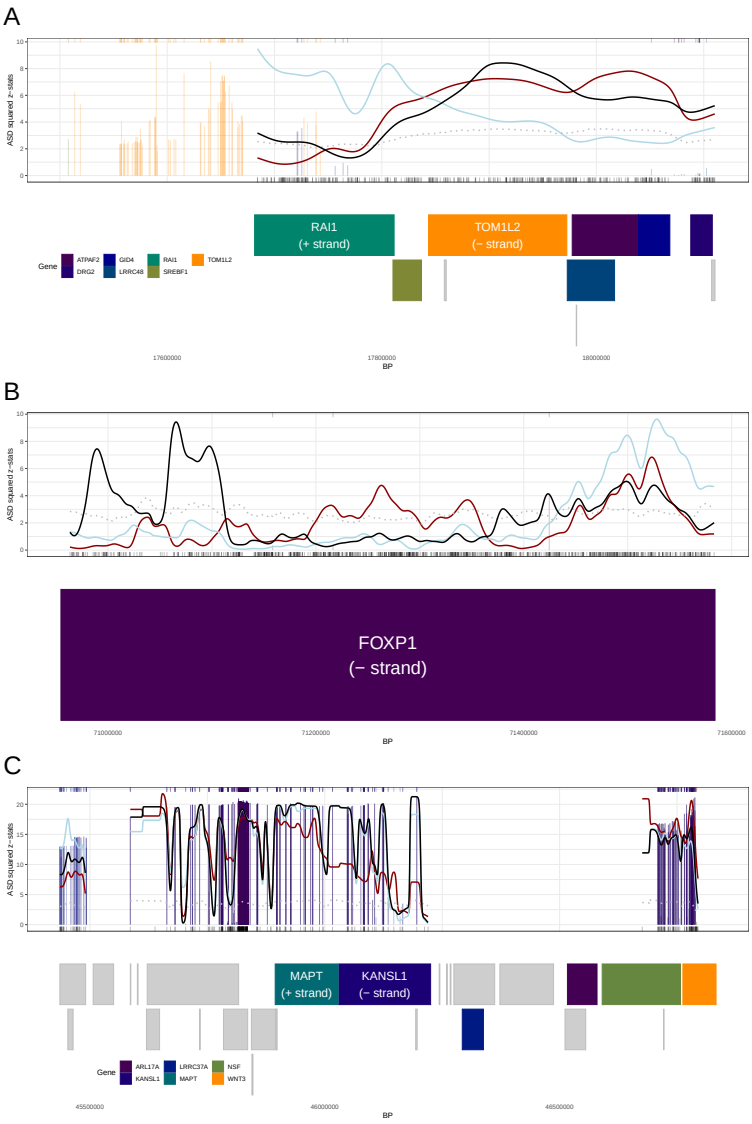

Figure 4.5: Gene/locus zoom displays for *(A)* locus falling in the chromosome 17 3.7 Mb Smith-Magenis syndrome region, *(B) FOXP1* in chromosome 3, and *(C)* locus in chromosome 17 1 Mb inversion region. *(A)-(C)* The genes located within the locus are represented in rectangles denoting their respective start and end positions below the smooth display, arranged by position and size to prevent overlapping. The gene/locus level kernel smoothing of ASD signal (black line) is displayed along with SCZ (red line) and EA (blue line) kernel smoothing signal (both are normalized to appear on the same signal scale as ASD). The gray dotted line denotes a point-wise 95^{th} percentile of 1,000 simulations for null simulations, and vertical bars correspond to eSNP signals (colored by their associated genes). A subset of genes are highlighted with colors in *(A)* and *(B)*, with other genes represented by gray rectangles. Additionally, individual genes are labeled in *(A)*: *RAI1* and *TOM1L2*, *(B)*: *FOXP1*, and *(C)*: *MAPT* and *KANSL1*, with gene direction (+ indicates left to right, - indicates right to left).

FOXP1 in chromosome 3:. An unclustered protein-coding gene on chromosome 3 highlights a set of associated SNPs in *FOXP1*, which has previously been strongly associated with intellectual disability, language impairment, and ASD [Hamdan et al., 2010] (Figure 4.5B). The smoothed association signal for ASD falls near the gene's 3' end whereas, for SCZ and EA, it falls close to the gene's 5' region. It is possible that variation associated with ASD is regulating *FOXP1*'s expression quite differently than that for SCZ and EA. The index SNP for EA is GWAS-significant and the one for SCZ approaches it (rs55736314 and rs4677597 with p-values 1.63×10^{-16} and 3.87×10^{-7} respectively) and both fall closer to the 5' region of *FOXP1*. For ASD, its index SNP rs7616330 carries a more modest signal (p-value 2.26×10^{-4}). Compared to the complete gene, it falls toward the 3' end, but it also falls quite close to the 5' start site of certain transcripts, such as ENST00000650387.1. Given the prominent role that *FOXP1* has in ASD and cognitive function, we speculate that the differential location of signal for SCZ/EA versus ASD could trace to different transcripts and regulation of their expression.

1 Mb inversion region in chromosome 17:. Another locus of interest is a 1.5 Mb region of chromosome 17, namely 17q21 (Figure 4.5C). This region of the genome is well known in human genetics because it comprises a 1.5 Mb inversion polymorphism [Stefansson et al., 2005, Steinberg et al., 2012] and the inversion alleles, actually haplotypes, have been associated with a wide variety of neurodegenerative disorders, including Progressive Supranuclear Palsy [Höglinger et al., 2011], corticobasal degeneration [Kouri et al., 2015], frontotemporal dementia [Furukawa et al., 2003], and other tauopathies [Silva and Haggarty, 2020]. In this locus, altered *MAPT* is well known to affect risk for late-life neurodegenerative disorders. Moreover, the inversion itself inhibits recombination, rendering alleles at SNPs across this region in high LD. Indeed, the complexity of this region inspires the interactive features of our application (conveyed in Figures 4.14 and 4.15) with subsets of genes and their associated eSNPs. Our results suggest variation in the region is associated with all three phenotypes. The index SNP for EA exceeds the standard threshold for single SNP significance (rs74998289, p-value 1.31×10^{-17}), while for ASD it approaches it (rs12942300, p-value 2.06×10^{-6}). Variation in the region has been previously implicated in ASD susceptibility [Cantor et al., 2005]; more recently, *KANSL1* expression has been implicated in cognitive function and ASD [Arbogast et al., 2017]. While the smoothed association signals for ASD and EA show a peak over this gene, signal is distributed across many genes in this locus and gene-based analysis is unlikely to pinpoint any gene therein. A clue to the driver or drivers of association comes by comparison of panels A-C in Figure 4.5. In the inversion region (Figure 4.5C), eSNP signals are noticeably more enriched compared to the other two loci and indeed the index SNP for EA is an eSNP for *KANSL1*. As highlighted by the display (Figure 4.5C), however, careful statistical and molecular analyses will be required to pinpoint what variation and what genes influence each of the three phenotypes at this locus.

4.3.5 Enrichment analysis

A primary motivation for gene-based analyses is to garner insight into the biological mechanisms underlying the phenotype by evaluating the set of genes associated with it. A standard approach infers these mechanisms by gene-set enrichment analysis, which we will implement in two ways. First, we will use the FUMA GENE2FUNC tool [Watanabe et al., 2017] for gene ontology (GO) enrichment analysis of the AdaPT: XGBoost genes/loci. This analysis has the advantage of searching through myriad functional sets of genes in an agnostic fashion. Still, this could also be viewed as a disadvantage if, *a priori*, certain biological functions are likely to affect the phenotype. For example, for all three phenotypes analyzed here, rare variation has already provided evidence linking synaptic, epigenetic, and transcription factor genes to them. Moreover, for ASD there exists a substantial set of genes implicated in risk by studies of rare variation. For these reasons, we implement a second gene-set enrichment analysis, specifically GSEA [Subramanian et al., 2005], which is a tool for testing if different sets of genes are enriched in a ranked gene list. We perform GSEA at the gene/locus-level ranked by their one-sided z statistics, using five different sets of genes/loci to test for enrichment at the top of the phenotype-specific ranked list: (1) brain expressed, (2) synaptic, (3) epigenetic, (4) transcription factors, and (5) 102 ASD risk genes identified based on *de novo* and case-control variation [Satterstrom et al., 2020] (see *Method Appendix* and Table 4.4 for details on compilation of gene lists).

Another reason to use both approaches is that they allow us to handle a confounder, gene size, in different ways. In our analyses, gene/locus size (and the number of SNPs therein) is a predictor of association: larger genes/loci are more likely to be associated (Figure 4.16). So enrichment analysis should control for gene size in some way. One approach is to contrast the associated set of genes with control genes after matching on gene size. We will use this approach in the FUMA analyses (see *Method Appendix* for details). However, a substantial portion of brain-expressed genes are synaptic, which are known to be among the largest genes in the genome. We reasoned that if these phenotypes were influenced by variation altering function of synaptic genes—which rare variant studies suggest they are [De Rubeis et al., 2014, Satterstrom et al., 2020, Fromer et al., 2014, Kurki et al., 2019]—then matching on gene size will tend to match associated synaptic genes with other synaptic genes, thereby over-matching and lowering the power to detect synaptic association. For this reason, in the GSEA analysis, we first remove the effect of gene size on the association z statistics by regressing them on gene size, then entering the residual z' values into the GSEA analysis.

Rather than include all genes in the associated loci for the FUMA enrichment analysis, we used the kernel smoothing results to identify *signal* genes (see *Method Appendix*). For example, this reduces the number of ASD *Positional + eSNPs* genes from 483 to 464 *signal* genes. After matching for gene size, no GO terms show enrichment for associated

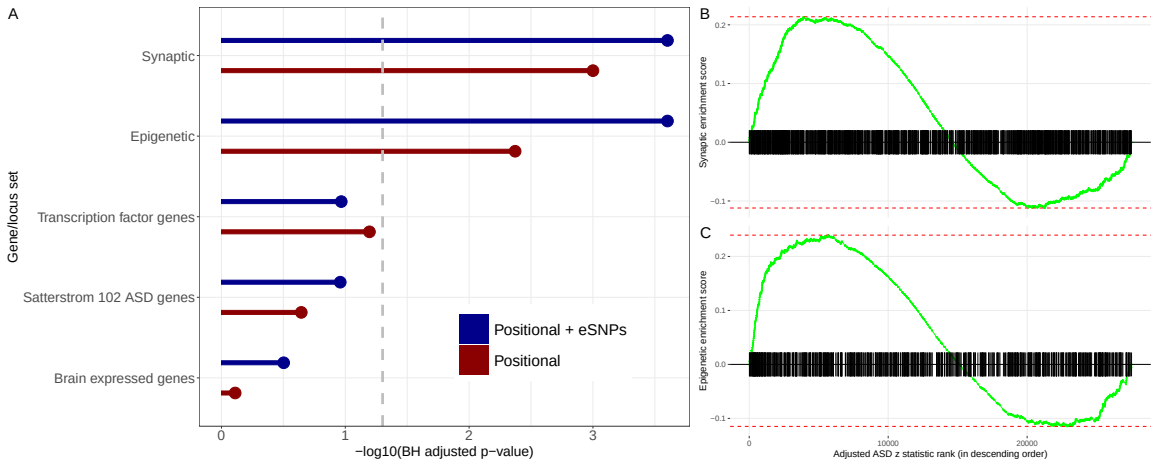

Figure 4.6: Enrichment analysis of genes/loci ranked by one-sided ASD z statistics, adjusted for size, by positional assignment. *(A)* BH-adjusted GSEA p-values are displayed on the -$\log_10$ scale, with FDR target level $\alpha = 0.05$ indicated by dashed gray line. Enrichment score for *(B)* synaptic and *(C)* epigenetic genes/loci with tick marks denoting gene/locus rank based on *Positional + eSNPs* ASD z statistics, adjusted for size, in descending order.

ASD genes. Genes associated with SCZ display GO enrichment: 698 terms in biological processes, 163 terms in cellular components, and 108 in molecular function. The EA genes display enrichment for 114 terms in biological processes, 64 in cellular components, and 27 in molecular function (Table 4.3). Using REVIGO [Supek et al., 2011] to summarize these terms, they highlight neuron projection, synaptic function, cell adhesion, cell cycle, chromosome organization, and many more for both SCZ and EA (Figures 4.17 and 4.18).

GSEA analysis of five gene sets (brain expressed, synaptic, epigenetic, transcription factor, and ASD risk) finds ASD implicated genes enriched for two gene sets, synaptic and epigenetic genes (Figure 4.6, Figure 4.19), whereas for SCZ and EA, all five gene sets are enriched (Table 4.3).

4.4 Conclusion

GWAS studies have identified myriad associations between SNPs and human phenotypes. With few exceptions, these associations are weak. One of the goals in this study is to improve power to detect weak associations through gene-based tests, while also accounting for LD among SNPs in different genes, which can induce correlations among tests for different genes. Although such correlation can elevate the false positive rate and obscure the interpretation of gene-based test results, most published studies have ignored it. To account for LD, we introduce a workflow (Figure 4.1) to aggregate genes into loci if the expected dependence

of their test statistics exceeds a pre-specified threshold. This approach produces a notable reduction in the number of genes tested: nearly 4,000 are aggregated even for a tolerant threshold for correlation of test statistics (aggregate if $r^2 > 0.75$; Figure 4.2).

We recommend practitioners use a stringent LD r^2 threshold, such as the one we use to analyze the data reported here ($r^2 = .25$). This ensures test statistics for genes/loci are largely independent, it will be appropriately conservative regarding reporting of the number of discoveries, and, for loci, it will highlight the complexity of causal inference therein. Practitioners will also need to set the window size to search for correlation among test statistics. For our analyses, we chose a large one, 6Mb. This is computationally expensive, compared to a smaller window. A smaller window will cover most loci, with a few known exceptions including the chromosome 6p MHC region. One should also note that, like all gene-based test approaches that account for LD structure, our approach is sensitive to mismatch of the LD reference panel. Such a mismatch could impact our results in two ways: (1) the individual test-level statistics could have the inappropriate null distribution, (2) the induced correlations between gene-level test statistics will be incorrect, potentially leading to over/under-clustering of the genes. We believe that (1) is the larger concern, and it is a concern that is shared by any testing approach that aggregates SNP-level test statistics (e.g. MAGMA or VEGAS). Properly capturing this sensitivity to correlation in gene-level testing is a topic for future research.

With a collection of uncorrelated or only weakly correlated genes/loci, any gene-based test can be applied. To improve power, we adopt the FDR control framework for hypothesis testing and a particular implementation of it, adaptive p-value thresholding (AdaPT). AdaPT has the potential to increase power over the classical Benjamini–Hochberg (BH) procedure for FDR control by accounting for covariates, which collectively we call metadata, to inform on which genes are likely to be true positives even though their test statistics failed to cross the BH threshold for significance. Importantly, like BH, AdaPT also maintains finite-sample FDR control.

We applied AdaPT to data from three phenotypes, ASD, SCZ, and EA, all of which are genetically correlated. Notably, the ASD GWAS was weakly-powered compared to the well-powered GWAS for SCZ and EA. Although AdaPT increased power for every phenotype and setting, relative to BH, the largest improvement was achieved for ASD (Figure 4.4A, Figures 4.9 and 4.10). We believe the most likely explanation for these contrasting results is the differential information content of the original GWAS. Even without the additional power from AdaPT, the BH-corrected GWAS for SCZ and EA identifies many genes/loci: there is little to be gained by analyses with more power. By contrast, there is far more to be gained by additional power for ASD analyses. Gene-based tests that incorporated the impact of SNPs on gene expression also increased power of gene discovery (Figure 4.4B, Figure 4.11). Reflecting the genetic correlation, summary statistics for other phenotypes were always

useful predictors for the AdaPT model (Figure 4.3, Figure 4.8). Gene conservation also played an important role, as did the size of the gene/locus (Figure 4.3, Figure 4.16).

That gene/locus size was a useful predictor for associated genes is of limited biological and genetic interest; however, it generates an interpretative challenge when the set of associated genes are evaluated for functional relevance, such as by GO enrichment analysis. One approach to enrichment analysis would match associated genes with control genes by size and we did this in our GO analyses. Such an approach, however, could be conservative. For example, a substantial portion of brain-expressed genes are synaptic, they tend to be large, and synaptic genes likely play a role in all three phenotypes analyzed here [De Rubeis et al., 2014, Satterstrom et al., 2020, Fromer et al., 2014, Kurki et al., 2019]. Yet, matching on gene size will tend to contrast associated synaptic genes with other synaptic genes, lowering power to detect their enrichment. For this reason, we also took a different approach to enrichment analysis. Specifically, we regressed the association statistics on size, then entered the residual value into a GSEA analysis. The end result, for ASD, is that synaptic genes were not enriched when genes were matched before GO analysis, while they were enriched in the GSEA analysis (Figure 4.6, Figure 4.19). The same is true for a set of epigenetic genes, which include chromatin readers, remodelers, and so on. Both results agree well with previous rare variant studies [De Rubeis et al., 2014, Satterstrom et al., 2020]. On the other hand, for the well powered SCZ and EA studies, such enrichment emerges regardless of the way gene size is controlled (Figures 4.17 and 4.18).

Here we demonstrated the gain in power of evaluating gene-based association statistics using AdaPT, guided by metadata, to detect genes affecting three specific phenotypes, ASD, SCZ, and EA. As our results show, the greatest gain in power is achieved when the underlying study has intermediate power, neither too high nor too low. Power gains will be modest for highly powered studies and absent for very weakly powered studies. We believe that AdaPT, guided by metadata, can be applied to a wide variety of omics problems, although it will undoubtedly require some adaptation for the specific problem to be solved. The advantage of doing so is twofold, increased power and increased interpretability. We are especially interested in the latter, as a means of expediting our understanding of the etiology of human diseases, disorders, and other phenotypes.

Although gene-based association draws attention to an important functional unit (or units for locus-based association), it does not inform on what variation drives the association. We developed an interactive visualization tool (`https://ron-yurko.shinyapps.io/ld_locus_zoom/`) for exploring the localization of association signal within associated genes/loci and generating hypotheses about mechanisms of action (Figure 4.5, Figures 4.14 and 4.15). For example, using this tool to explore the association of *FOXP1* [Hamdan et al., 2010] reveals that the pattern of association in the gene is different for ASD versus EA and SCZ (Figure 4.5B) and suggests that different patterns of gene expression could be important

for these phenotypes. Another example is the 500Kb associated locus within the 17p11.2 deletion/duplication (Smith-Magenis syndrome) region (Figure 4.5A) [Seranski et al., 1999, Neira-Fresneda and Potocki, 2015]. In this locus, all three phenotypes have been associated to varying degrees to one gene, *RAI1* [Carmona-Mora and Walz, 2010]. Curiously, however, only the association signal for EA peaks within this gene, whereas for ASD it unexpectedly maximizes over an adjacent gene. While we believe this tool can be a useful guide to researchers, this example also underscores a limitation of this approach: we cannot provide error rate guarantees at the localized level. Such an analysis will be a target of our future work in this challenging area.

4.5 Data availability statement

The data and code used in this chapter are available at `https://github.com/ryurko/Agglomerative-LD-loci-testing`.

4.6 Method Appendix

4.6.1 Comparison of GWAS enrichment

We observe that autism spectrum disorder (ASD) GWAS results [Grove et al., 2019] suffer from weaker power in comparison to schizophrenia (SCZ) [Ruderfer et al., 2018b] and educational attainment (EA) [Lee et al., 2018] (Figure 4.7). This is likely due to differences in study size: 18,381 cases and 27,969 controls for ASD in comparison to 33,426 cases and 32,541 controls for SCZ and $\approx$ 1.1 million individuals for EA.

4.6.2 AdaPT overview

AdaPT is an iterative search procedure for selecting R discoveries with guaranteed finite-sample FDR control at target level α, under the assumption of independent null p-values [Lei and Fithian, 2018]. We apply AdaPT to a collection of weakly correlated gene/locus-level p-values and metadata, $(p_g, x_g)_{g \in G}$, testing hypothesis H_g regarding gene/locus' g's association with the phenotype of interest (e.g. ASD). For each step $t = 0, 1, \ldots$ in the AdaPT search, we first determine the rejection set $\mathcal{R}_t = \{g : p_g \leq s_t(x_g)\}$, where $s_t(x_g)$ is the rejection threshold at step t that is *adaptive* to the metadata x_g (except for the starting threshold $s_0 = 0.05$). This provides us with both the number of discoveries/rejections $R_t = |\mathcal{R}_t|$, as well as a *pseudo*-estimate for the number of false discoveries $A_t = |\{g : p_g \geq 1 - s_t(x_g)\}|$. These quantities are used to estimate the FDP at the current step t,

$$\widehat{\text{FDP}}_t = \frac{1 + A_t}{\max\{R_t, 1\}}.$$

If $\widehat{\text{FDP}}_t \leq \alpha$, then the AdaPT search ends and the set of discoveries $\mathcal{R}_t$ is returned. Otherwise, the rejection threshold is updated by discarding the most likely null element in

the current rejection region, as measured by the conditional local false discovery rate (fdr) estimated with an expectation-maximization (EM) algorithm. With the threshold updated, the AdaPT search repeats by estimating FDP and updating the rejection threshold until the target FDR level is reached.

The most critical step in AdaPT involves updating the rejection threshold $s_t(x_i)$ through an EM algorithm to fit a conditional version of the two-groups model [Efron et al., 2001] to estimate local fdr. We consider the same model form as our previous work [Yurko et al., 2020], where the null p-values are modeled as uniform ($f_0(p|x) \equiv 1$) while we model the non-null p-value density with a beta distribution density parametrized by $\mu_g = \mathbb{E}[-\log(p_g)]$. This results in a conditional density for a beta mixture model,

$$f(p|x_g) = \pi_1(x_g)\frac{1}{\mu_g}p^{1/\mu_g - 1} + 1 - \pi_1(x_g).$$

In this form, we can model the non-null probability $\pi_1(x_g) = \mathbb{E}[H_i|x_g]$ and the non-null effect size $\mu(x_g) = \mathbb{E}[-\log(p_i)|x_g, H_g = 1]$ with two separate gradient boosted tree-based models. The XGBoost library [Chen and Guestrin, 2016] provides logistic and Gamma regression implementations which we use for $\pi_1(x_g)$ and $\mu(x_g)$ respectively. See our previous work [Yurko et al., 2020] for details about the implementation of the EM algorithm, and [Lei and Fithian, 2018] for details about updating the AdaPT rejection threshold using $\pi_1(x_g)$ and $\mu(x_g)$.

4.6.3 AdaPT tuning results

In order to avoid overfitting our the non-null probability and effect size models, we first find appropriate settings for the gradient boosted trees (number of trees, learning rate, and maximum depth) using synthetic SCZ p-values which mimic ASD p-values in the following manner:

1. Sort SCZ and ASD p-values: $(p^{SCZ}_{(1)}, \ldots, p^{SCZ}_{(M)})$ and $(p^{ASD}_{(1)}, \ldots, p^{ASD}_{(M)})$
2. Replace SCZ with ASD p-values by matching order,
    - e.g., replace $p^{SCZ}_{(1)}$ with $p^{ASD}_{(1)}$, $p^{SCZ}_{(2)}$ with $p^{ASD}_{(2)}$, ...

We then proceed to apply AdaPT using these ASD-aligned SCZ p-values, finding appropriate settings in our gradient boosted trees. To avoid overfitting, we consider shallow trees with maximum depth $\in \{1, 2\}$, while searching over the number of trees $P \in 100, \ldots, 450$ by increments of fifty and the learning rate $\eta \in 0.03, \ldots, 0.06$ by increments of 0.01. We find the combination P and η yielding the highest number of synthetic SCZ discoveries for both depth values at FDR level $\alpha = 0.05$. The top setting combinations across the considered SNP-to-gene assignment types and r^2 threshold (Table 4.1) were then considered in the

AdaPT cross-validation algorithm [Yurko et al., 2020] when applied to the actual ASD, SCZ, and EA p-values.

4.6.4 Measuring AdaPT metadata importance

We examine the variable importance from the gradient boosted trees to provide insight into the relationships between the metadata x_g and measures of phenotypic association, non-null probability $\pi_1(x_g)$ and non-null effect size $\mu(x_g)$. However, because AdaPT is an iterative search with several modeling steps, we summarize metadata importance by computing the average importance across the steps in the search. This allows us to compare metadata importance across phenotypes and positional assignment since the AdaPT searches vary in terms of the number of modeling steps required to reach the target $\alpha = 0.05$ (Table 4.2).

We rank the top five sources of metadata for each phenotype and positional assignment based on the sum of the average importance for the two types of AdaPT models, non-null probability $\pi_1(x_g)$ and non-null effect size $\mu(x_g)$. We observe similar rankings in metadata importance between the *Positional + eSNPs* (Figure 4.3) and *Positional* results (Figure 4.8), with differences between phenotypes such as the increased importance of LOEUF for SCZ and EA in comparison to ASD.

4.6.5 Results with LD threshold $r^2 \in \{0.50, 0.75\}$

We additionally generate the AdaPT: XGBoost and baseline results using genes/loci formed with LD thresholds of $r^2 \in \{0.50, 0.75\}$, which are not as conservative as the threshold of $r^2 = 0.25$. Both of these higher threshold values lead to a greater number of genes/loci for testing for both positional types, *Positional* and *Positional + eSNPs* (Figure 4.2). Ultimately, in terms of the number of genes/loci selected by AdaPT at target FDR level $\alpha = 0.05$, we observe similar results as before (Figure 4.4A) with increased power from using AdaPT: XGBoost regardless of phenotype with indication that including eSNPs provides an advantage in detecting more signals (Figures 4.9 and 4.10). As expected, for all three phenotypes, the percentage of genes/loci selected corresponding to unclustered genes increases with the r^2 threshold (Table 4.3).

4.6.6 Results per chromosome breakdown

We compare the chromosome breakdown of the selected genes/loci using AdaPT: XGBoost for each phenotype, indicating a greater boost in detecting associations by including eSNPs for ASD in comparison to well-powered SCZ and EA (Figure 4.11). Additionally, for SCZ and EA, the total number of associated genes per chromosome declines strongly and linearly with chromosome size (Figures 4.12 and 4.13).

4.6.7 LD locus zoom application

We developed our LD locus zoom application using `R` Shiny [R Core Team, 2020, Chang et al., 2020] and `plotly` [Sievert, 2020] to interactively explore localized signal within interesting genes/loci.

The genes located within the locus are represented in rectangles denoting their respective start and end positions below the smooth display, arranged by position and size to prevent overlapping. The solid black line denotes the gene/locus-level kernel smoothing for positional SNP signals, including interpolation. For reference, the gray dotted line denotes a point-wise 95^{th} percentile of 1,000 simulations for squared null Gaussian random variables, $z^2_{g^*}$ where $z_{g^*} \sim \text{Normal}(\mathbf{0}_{g^*}, \mathbf{\Sigma}_{g^*})$, given the LD structure $\mathbf{\Sigma}_{g^*}$ of the selected LD locus g^*.

Additionally, we provide several interactive features such as the option to display background kernel smoothing results for SCZ (in red) and EA (in blue), normalized to appear on the same signal scale as ASD, as well as the option to display eSNPs separately as bars (colored to match their associated genes) with bar heights denoting individual eSNP-level signal. Furthermore, one has the option to display gene-level smoothing (Figure 4.14) as well as highlight particular genes and their corresponding eSNP signals with the `plotly` highlighting tool (Figure 4.15). To simplify the visualization of larger loci, we perform the interpolation within subgroup of SNPs that are formed if they are separated by more than five percent of the loci size (using single-linkage clustering), e.g. three subgroups of SNPs with interpolation performed separately for LD locus in chromosome 17 1 Mb inversion region (Figure 4.15).

Within the application, there are three tabs: (1) *ASD results*, (2) *Upload results*, and (3) *Description*. The *ASD results* tab includes the LD locus visualization for our *Positional* and *Positional + eSNPs* results, with the ability to select different genes/loci to display as well as export the image as an SVG file. Furthermore, the tables of the selected LD locus' corresponding genes and SNPs can be downloaded as CSV or Excel files (via the *Genes* and *SNPs* tabs respectively). The gene and SNP tables also include urls to their respective pages in the GWAS catalog [Buniello et al., 2019].

Additionally, users can use the *Upload results* tab to import datasets to generate the same type of kernel smoothing visualization. Additional information about the interactive features of the application are available in the *Description* tab. The LD locus zoom application can be accessed here `https://ron-yurko.shinyapps.io/ld_locus_zoom/`.

4.6.8 Enrichment analysis

A primary motivation for gene-based analyses is to garner insight into the biological mechanisms underlying the phenotype by evaluating the set of genes associated with it. We implement two approaches: (1) FUMA GENE2FUNC tool [Watanabe et al., 2017] for gene ontology (GO) enrichment analysis of the AdaPT: XGBoost genes/loci, and (2) gene-set enrichment analysis (GSEA) [Subramanian et al., 2005] to test if different sets of genes are enriched in a ranked gene list. In both approaches, we address the confounding effect of gene/locus size as larger genes are more likely to be associated (Figure 4.16).

First, we use the FUMA GENE2FUNC tool for GO enrichment analysis of the implicated

genes in the AdaPT: XGBoost genes/loci. Rather than include all implicated genes in the associated loci for the FUMA enrichment analysis, we used the kernel smoothing results to identify *signal* genes. Specifically, we only use genes with kernel smoothing signals above the point-wise 95th percentile of null simulations and any gene with at least one intergenic eSNP displaying a large marginal effect size (i.e., z statistic ≥ 1.96). Then to address the confounder of gene size, we find non-implicated genes to match the *signal* genes with respect to gene size using the `optmatch` [Hansen and Klopfer, 2006] package in `R` [R Core Team, 2020]. Due to the difference in number of *signal* genes for the phenotypes, we find the twenty, two, and the single closest matching non-implicated genes for each *Positional + eSNPs signal* gene with ASD, SCZ, and EA respectively. Using the list of matched non-implicated background genes, we then do not observe any GO enrichment terms for the ASD genes. In comparison, the SCZ (EA) genes display enrichment for 698 (114) terms in biological processes, 163 (64) terms in cellular components, and 103 (27) in molecular function. Using REVIGO [Supek et al., 2011] to summarize these terms, they highlight neuron project, synaptic function, cell adhesion, cell cycle, chromosome organization, and and many more (Figures 4.17 and 4.18).

Next, we performed GSEA for five different sets of genes/loci to test for enrichment at the top of the list of genes/loci ranked by their one-sided z statistics:

1. 18,008 brain expressed genes from SynGO [Koopmans et al., 2019],
2. 1,232 synaptic genes from SynGO [Koopmans et al., 2019],
3. 815 epigenetic genes from EpiFactors [Medvedeva et al., 2015],
4. 1,819 transcription factors compiled from the SeqQC project [de Santiago, 2020], and
5. 102 ASD risk genes identified based on *de novo* and case-control variation [Satterstrom et al., 2020].

We collapsed these five lists of genes into their respective genes/loci based on LD induced correlation for each positional type, indicating that a locus is a member of a list if at least one of its genes is a member of the list (Table 4.4).

To address the confounder of gene/locus size, we compute versions of the z statistics that are adjusted for the gene/locus' size. Specifically, we regress out the effect of the log(size) on the z statistics for each phenotype (Figure 4.16) and use the adjusted z statistics for GSEA. We use the `fgsea` implementation in `R` [Korotkevich et al., 2019], with 10,000 permutations to compute GSEA p-values, to test if any of the five lists are enriched at the top of the genes/loci ordered by the size-adjusted z statistics for all three phenotypes. We observe that both synaptic and epigenetic genes/loci are enriched for both positional types using the size-adjusted ASD z statistics based on Benjamini-Hochberg (BH)

[Benjamini and Hochberg, 1995] adjusted p-values at FDR level $\alpha = 0.05$ (Figure 4.6 and Figure 4.19). In comparison, we observe that all five gene sets, for both positional types, are enriched for both SCZ and EA size-adjusted z statistics (all of their respective BH adjusted p-values were equal to 1 / 10,001, i.e., their observed enrichment scores were more extreme than the 10,000 simulations).

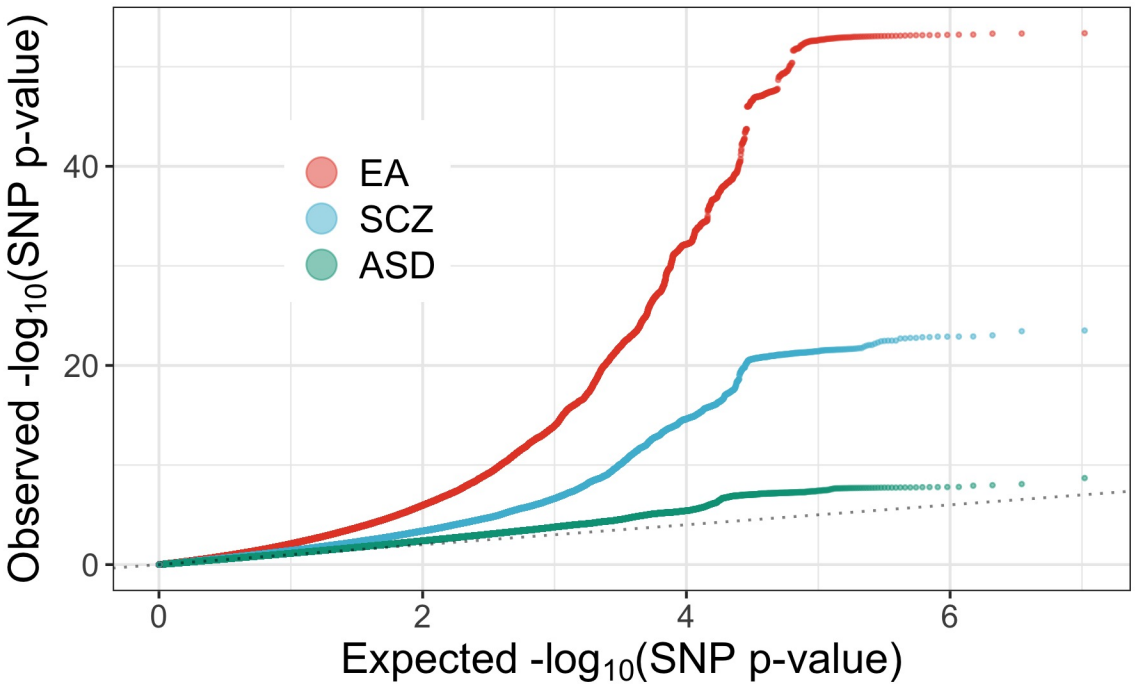

Figure 4.7: Comparison of SNP-level quantile-quantile plots revealing greater enrichment for SCZ and EA in comparison to ASD. Dotted reference line indicates null, uniform distribution.

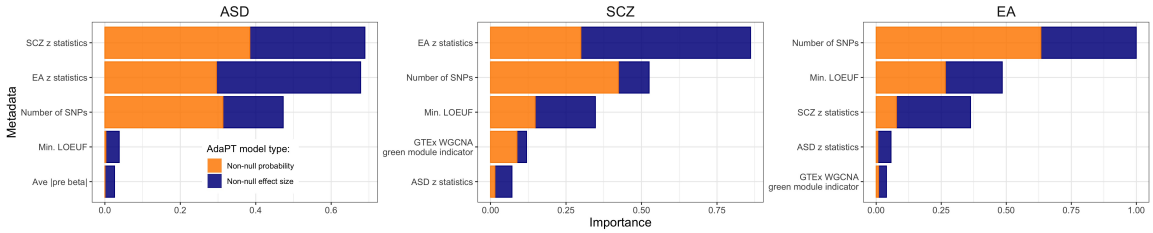

Figure 4.8: The top five sources of metadata as ranked by XGBoost importance for the *Positional* results by phenotype. Metadata are sorted in order of combined importance, while the color of the bars denote the separate measures of importance for the two AdaPT models: non-null probability (orange) and non-null effect size (blue).

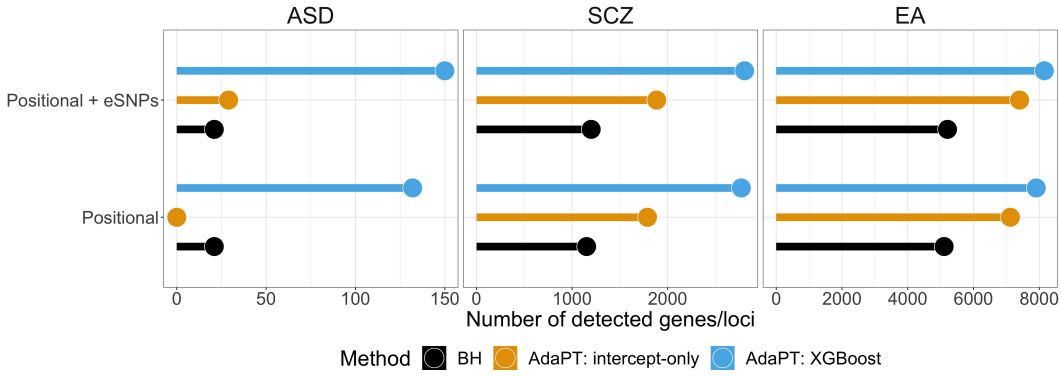

Figure 4.9: Comparison of the number of discoveries at FDR level $\alpha = 0.05$ for each phenotype (by column), comparing the number of genes/loci returned by the type of SNP-to-gene assignment with LD threshold $r^2 = 0.50$. AdaPT: XGBoost results are presented in comparison to BH and AdaPT: intercept-only baselines.

Figure 4.10: Comparison of the number of discoveries at FDR level $\alpha = 0.05$ for each phenotype (by column), comparing the number of genes/loci returned by the type of SNP-to-gene assignment with LD threshold $r^2 = 0.75$. AdaPT: XGBoost results are presented in comparison to BH and AdaPT: intercept-only baselines.

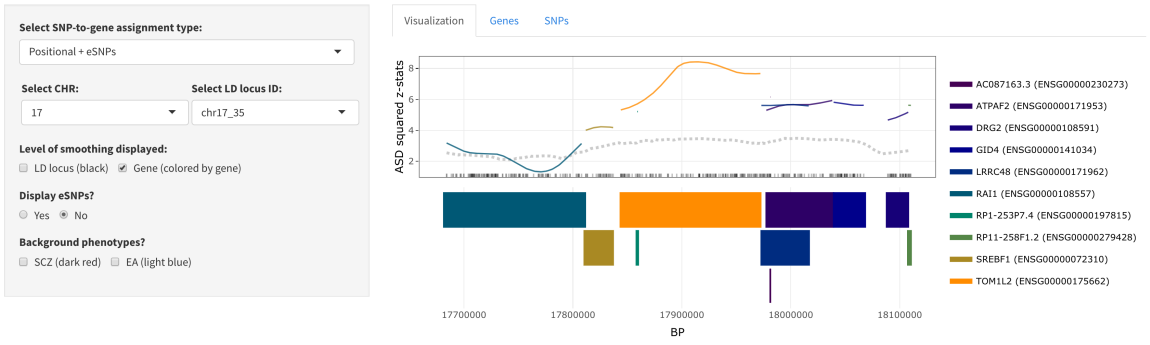

Figure 4.14: Gene-level ASD kernel smoothing for the chromosome 17 LD locus in the 3.7 Mb deletion region. Line colors match the associated genes displayed below the x-axis. The gray dotted line indicates point-wise 95^{th} percentile for null simulations.

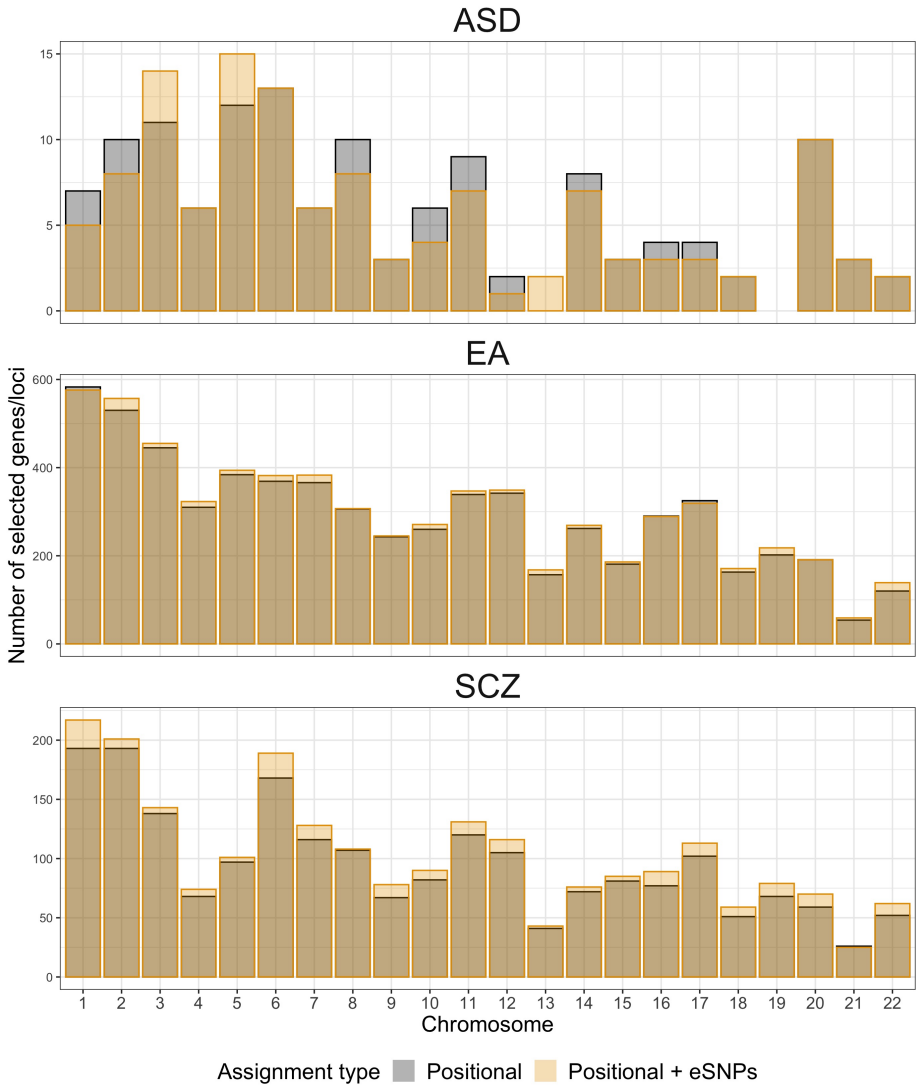

Figure 4.11: Comparison of the number of selected genes/loci by chromosome in the AdaPT: XGBoost by positional assignment type for each phenotype. The bars are overlaid on top of each other, indicating a higher number of genes/loci are selected at a chromosome using the *Positional + eSNPs* approach when the orange bar height is above the dark gray area, e.g., ASD results for chromosome six.

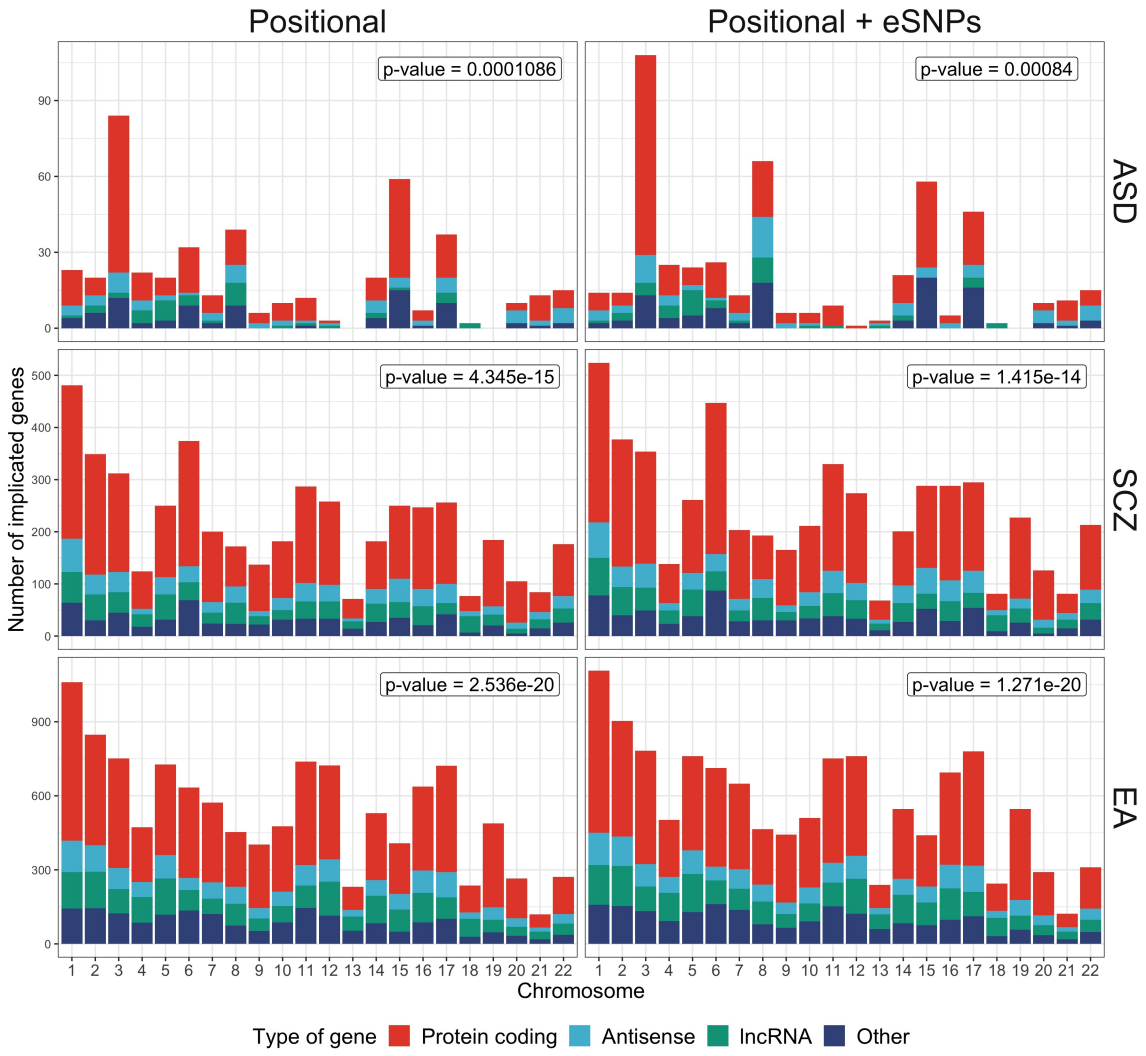

Figure 4.12: Comparison of the number of implicated genes by chromosome in the AdaPT: XGBoost results for both positional assignment types for each phenotype (colored by type of gene).

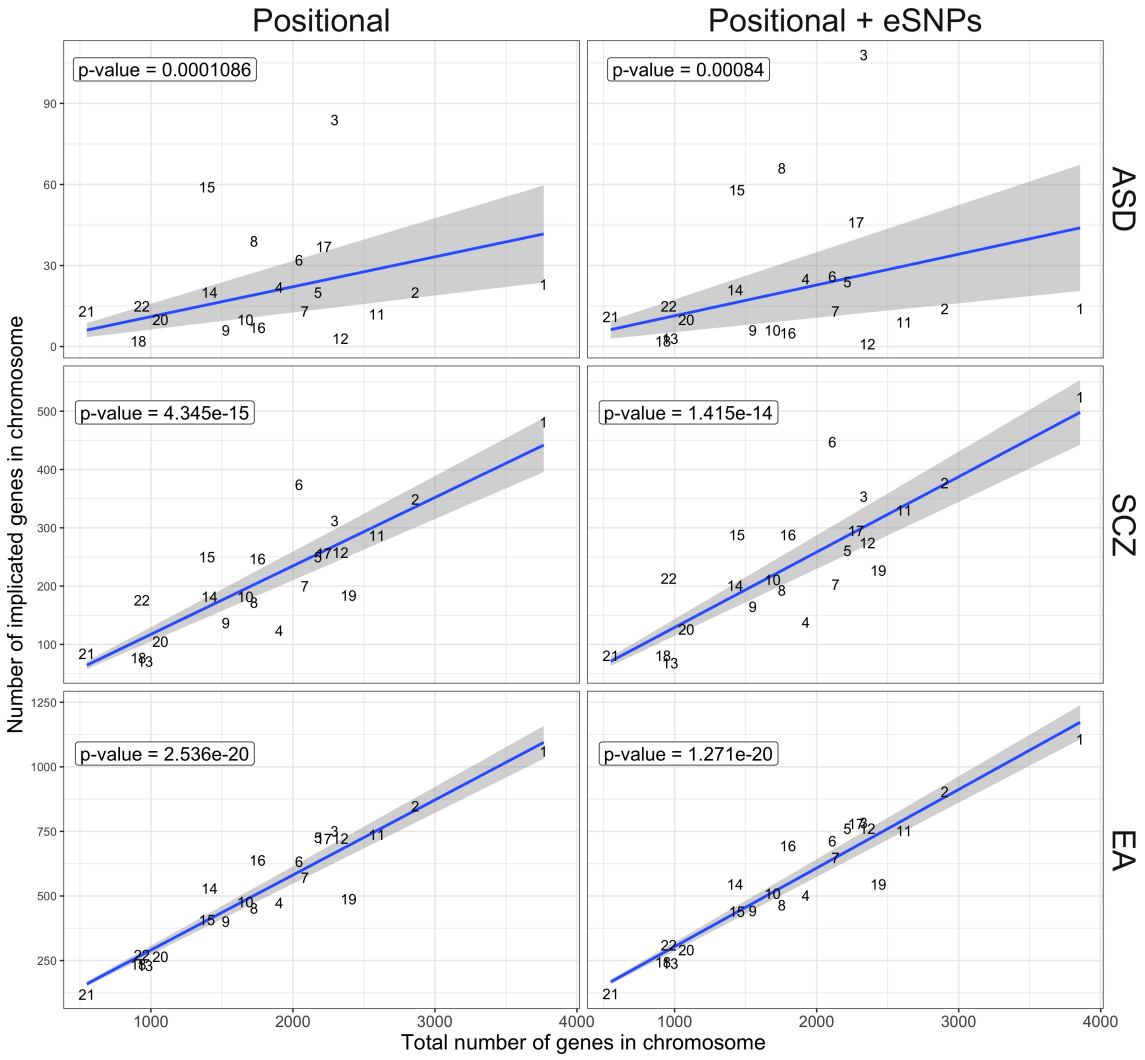

Figure 4.13: Relationship between the number of implicated genes and total number of genes per chromosome by phenotype and positional assignment. Chromosome points are labeled by their respective number. Blue line indicates regression fit with p-values of coefficients labeled in the top-left corners.

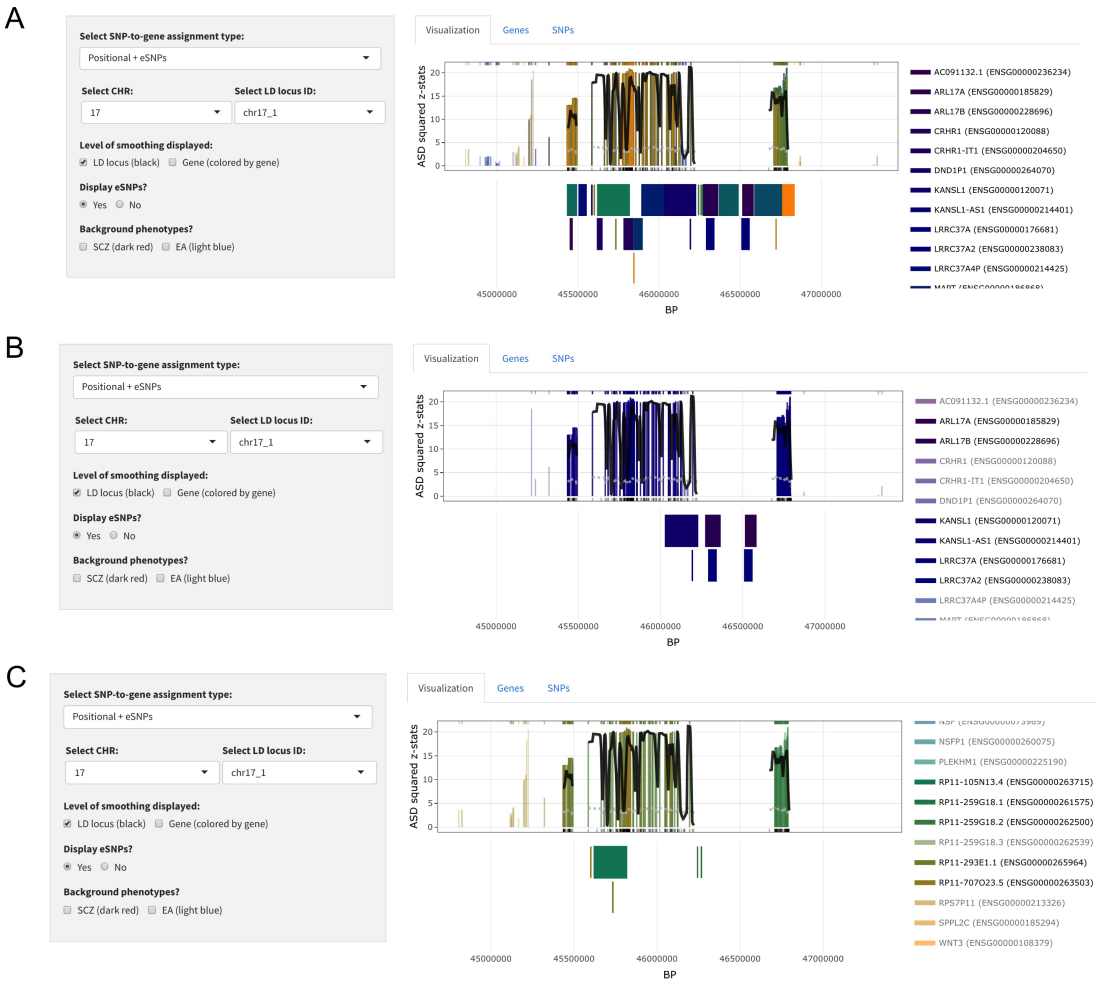

Figure 4.15: Zoom display for LD locus in chromosome 17 1 Mb inversion region. *(A)* Display with all genes and their associated eSNPs. *(B)* and *(C)* Display two separate subsets of genes and their associated eSNPs.

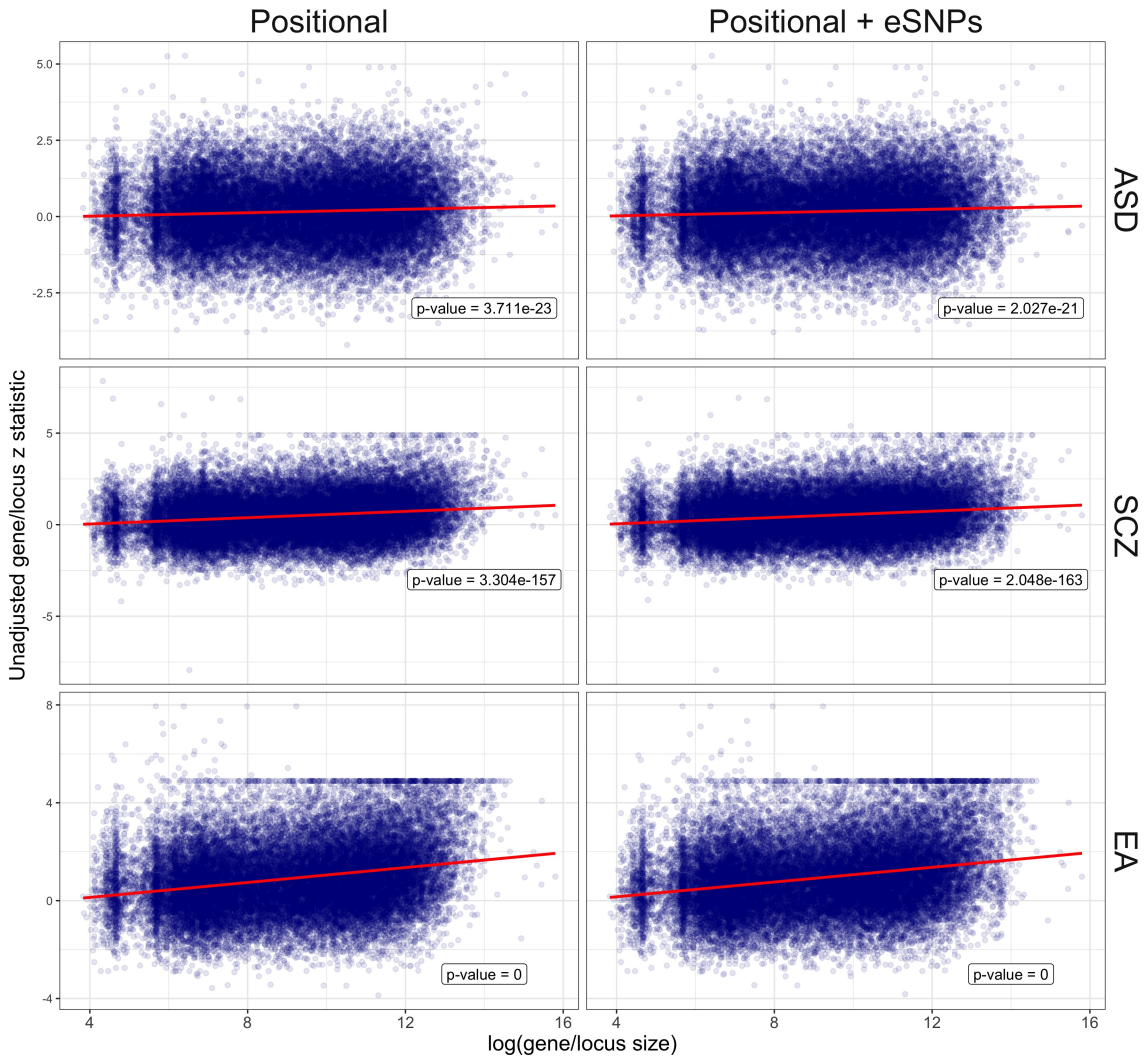

Figure 4.16: Relationship between unadjusted z statistics and log(gene/locus size) by phenotype and positional assignment. Red line indicates regression fit with p-values of coefficients labeled in the bottom-right corners.

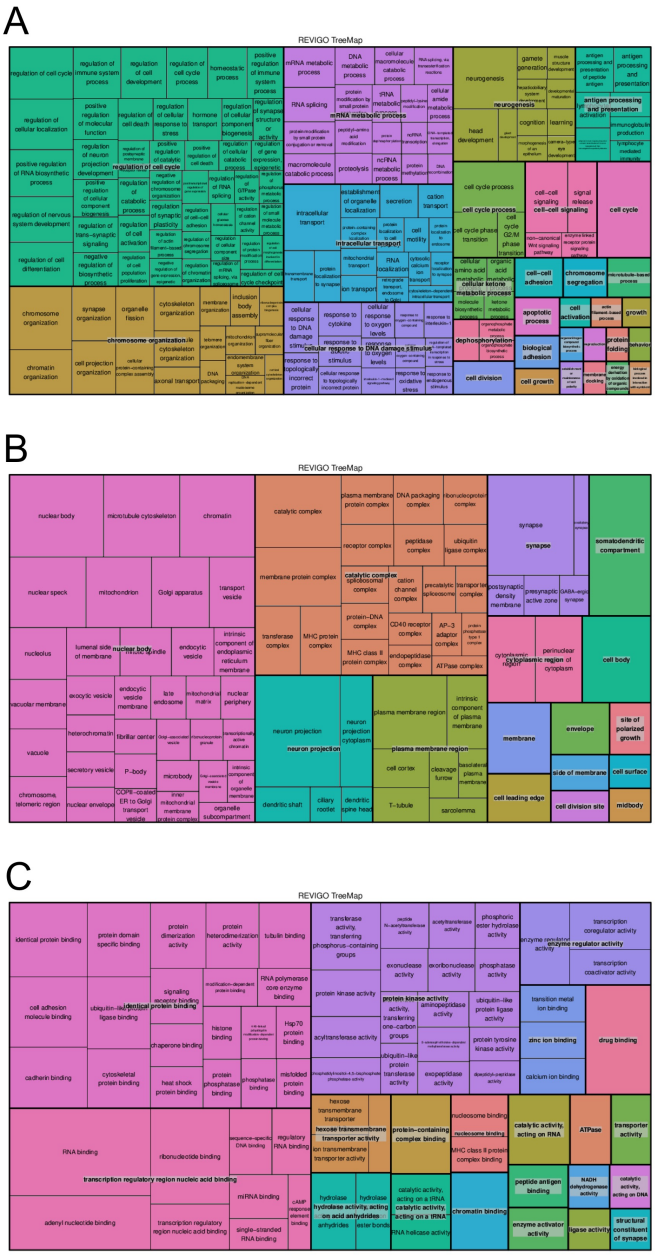

Figure 4.17: Treemap display of GO *(A)* biological processes, *(B)* cellular components, and *(C)* molecular function for SCZ *Positional + eSNPs* results using *signal* genes with size-matched list of background genes.

Figure 4.18: Treemap display of GO *(A)* biological processes, *(B)* cellular components, and *(C)* molecular function for EA *Positional + eSNPs* results using *signal* genes with size-matched list of background genes.

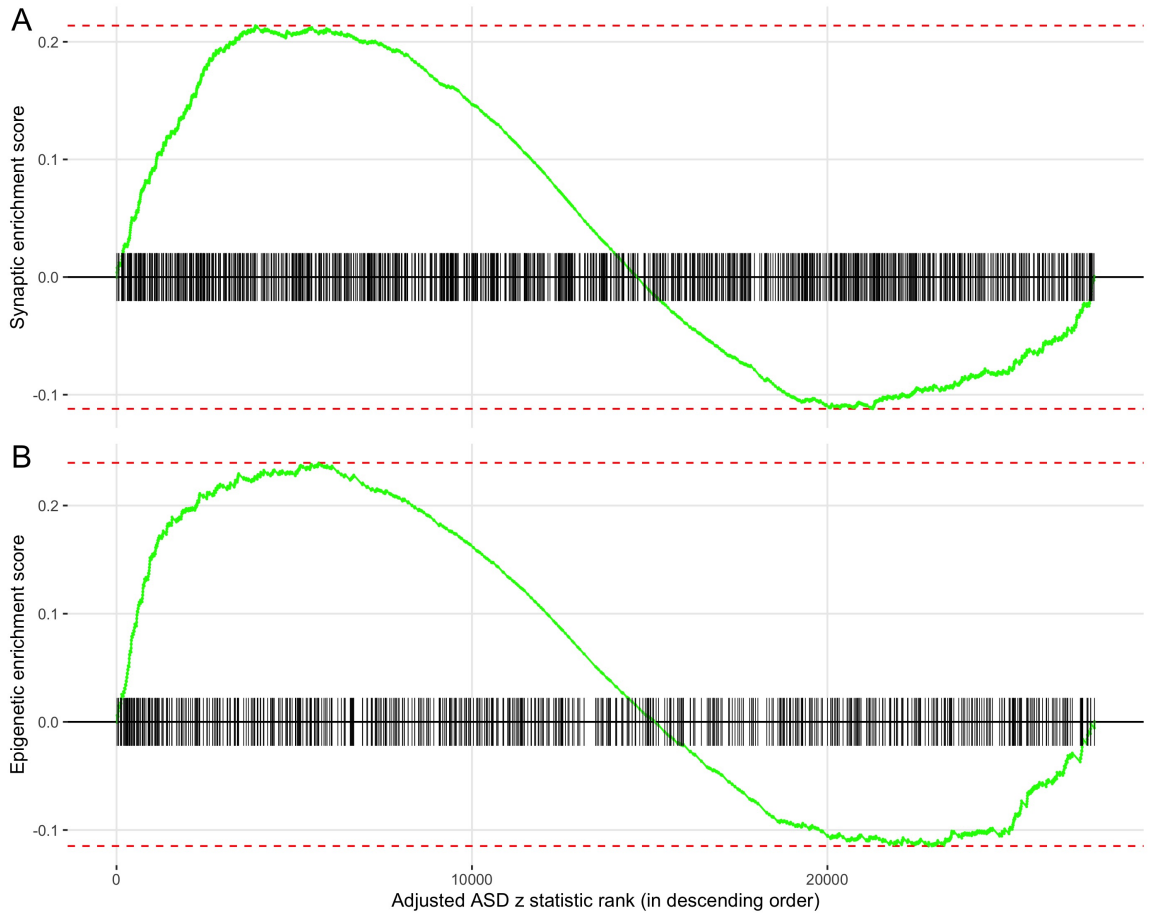

Figure 4.19: Enrichment score for *(A)* synaptic and *(B)* epigenetic genes/loci with tick marks denoting gene/locus rank based on *Positional* ASD z statistics, adjusted for size, in descending order.

Table 4.1: Top boosting settings to consider for number of trees P and learning rate η by maximum depth and SNP-to-gene assignment type from tuning with synthetic SCZ p-values.

SNP-to-gene assignment	r^2 threshold	Depth = 1	Depth = 2
Positional	0.25	$P = 450,\ \eta = 0.06$	$P = 100,\ \eta = 0.04$
	0.50	$P = 150,\ \eta = 0.06$	$P = 250,\ \eta = 0.06$
	0.75	$P = 450,\ \eta = 0.05$	$P = 200,\ \eta = 0.05$
Positional + eSNPs	0.25	$P = 400,\ \eta = 0.06$	$P = 250,\ \eta = 0.05$
	0.50	$P = 450,\ \eta = 0.06$	$P = 300,\ \eta = 0.05$
	0.75	$P = 450,\ \eta = 0.05$	$P = 250,\ \eta = 0.04$

Table 4.2: Number of model fitting steps in AdaPT search to reach target FDR level $\alpha = 0.05$, by phenotype and positional assignment.

Phenotype	*Positional* steps	*Positional + eSNPs* steps
ASD	18	20
SCZ	16	15
EA	8	8

Table 4.3: Comparison of the number (%) of unclustered genes in the AdaPT: XGBoost selected genes/loci as a function of r^2 for each phenotype by positional assignment.

Positional assignment	r^2	ASD	SCZ	EA
Positional	0.25	77 (17.2%)	1,439 (30.2%)	4,991 (42.4%)
	0.50	98 (31.6%)	2,318 (54.6%)	6,882 (63.2%)
	0.75	150 (62.2%)	3,098 (76.5%)	8,683 (81.9%)
Positional + eSNPs	0.25	78 (16.1%)	1,566 (29.3%)	5,091 (40.5%)
	0.50	112 (28.9%)	2,306 (50.6%)	7,054 (61.1%)
	0.75	179 (60.1%)	3,297 (74.5%)	9,071 (80.8%)

Table 4.4: Comparison of the number of genes/loci in each of the five considered list of genes for GSEA by positional assignment.

Gene list	*Positional*	*Positional + eSNPs*
Brain expressed	11,878	11,558
Synaptic	1,109	1,114
Epigenetic	654	664
Transcription Factors	1,453	1,454
102 ASD risk genes	102	102

Five

Augmenting rare variant studies with annotations to improve power

5.1 Introduction

Recently, the development of whole-genome sequencing (WGS) has enabled greater exploration into the impact de novo mutations (variants observed in child but not in parents) located in noncoding regions of the genome have on complex disorders. To address the unique multiple burden respecting the scale of the noncoding genome, [Werling et al., 2018] introduced a category-wide association studies (CWAS) framework defining over fifty-thousand annotation categories to test for association with ASD. However, in the analysis of case-control data from a limited number of quartet-families (parents, probands, and siblings for controls), they did not observe any noncoding annotation categories meeting the category-wide significance threshold. Similar null results were observed by [An et al., 2018] with the inclusion of more families, however a de novo risk score analysis implicated the contribution of de novo mutations in promoter regions to ASD. Our goal is to improve power to detect noncoding, regulatory elements associated ASD by building upon the foundation of the CWAS framework.

The role of annotation categories in the CWAS framework provides a natural setting for selective inference approaches to improve power. First, we provide an overview of the mutation rate model and example data considered in this chapter. We then propose a testing approach analogous to our previous work in gene-level testing [Yurko et al., 2021b], by clustering correlated tests together based on their annotation structure and then incorporating annotations as covariates in AdaPT. We then investigate the use of a data blurring approach to enable exploration of lower-level annotations and compare performances between the considered methods in simulation studies.

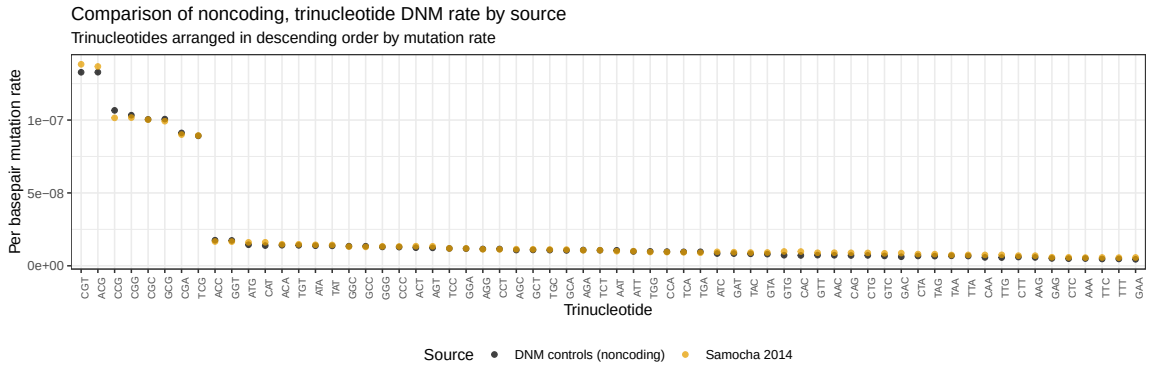

Figure 5.1: Comparison of trinucleotide DNM rate in noncoding regions between values computed using controls to approach based on human-chimp comparison.

5.2 Background and data

5.2.1 Mutation rate model

Following common practice [He et al., 2013, Liu et al., 2018], we model the null *de novo* mutation (DNM) counts Y_w within some noncoding region w following a Poisson distribution,

$$Y_w \sim \text{Poisson}\Big(\lambda_w\Big), \tag{5.1}$$

where $\lambda_w = 2N\mu_w$ is the DNM rate given N individuals in the study and the baseline mutation rate μ_w for region w. We assume the mutation rates μ_w are known from an external source. In this chapter, we use controls from previous studies [An et al., 2018] to compute the baseline mutation rates, which are based on summing across the trinucleotide-specific mutation rates for each base in a region. This yields comparable rates to previous approaches [Samocha et al., 2014] (Figure 5.1).

The aforementioned genomic regions are constructed based on annotations, essentially dividing the genome into non-overlapping regions defined by changes in the status of the considered annotations (Figure 5.2). Rather than work with these regions directly, the CWAS framework instead accumulates these regions based on the annotation status. For example, in the example schematic (Figure 5.2) there are nine regions constructed based on two annotations A and B. This leads to four independent *combinations* of annotations: (1) $A = 0$; $B = 0$, (2) $A = 1$; $B = 0$, (3) $A = 0$; $B = 1$, or (4) $A = 1$; $B = 1$. By this construction, a genomic region can only exist in a single annotation *combination*. For this reason, we then treat each combination $c \in C$ as the base unit of our analysis. We assume the DNM counts for each combination are independent and follow a Poisson distribution with DNM rate λ_c.

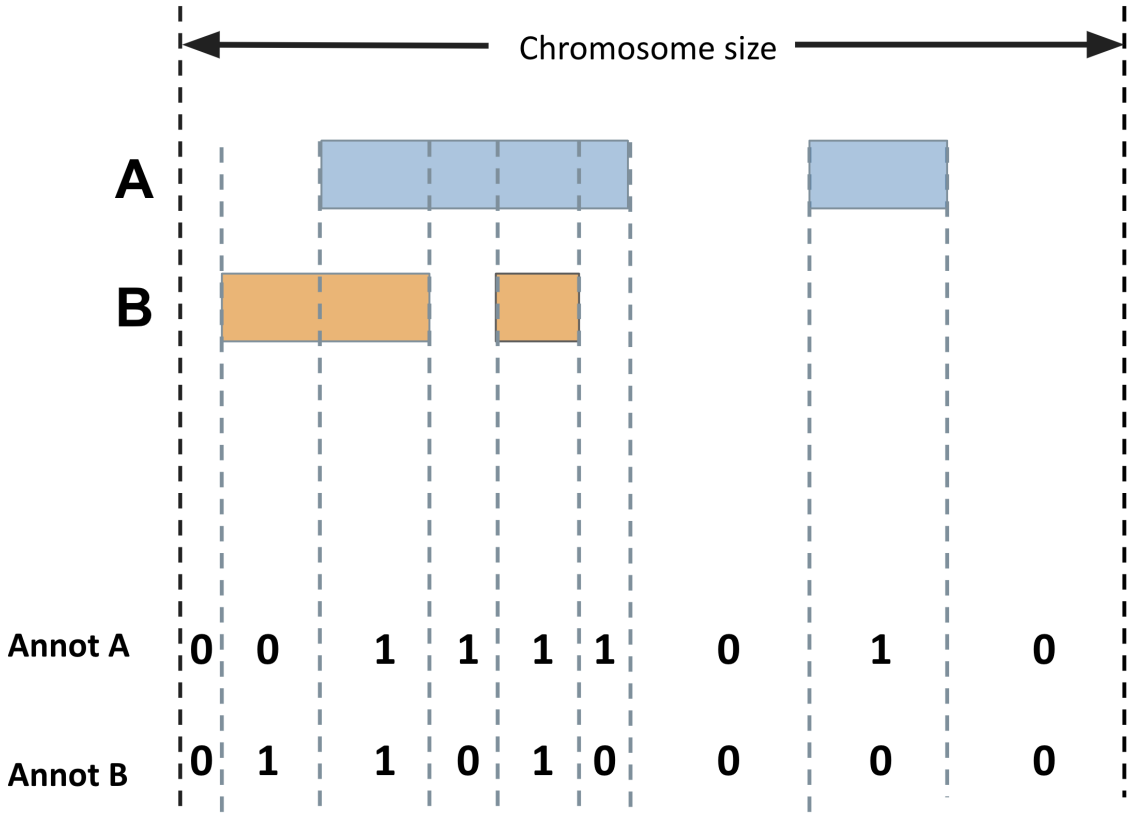

Figure 5.2: Schematic indicating how genomic regions are constructed based on annotations. Each rectangle denotes a particular annotation, while the dashed lines indicate where non-overlapping regions are constructed.

However, we are not interested in testing the association status of the $|C|$ annotation combinations directly. We are instead interested in testing the association status for *intersections* of annotations, i.e., when a particular set of annotation indicators are each equal to one. For example, in the example schematic (Figure 5.2) there are three possible intersections: (1) $A = 1$, (2) $A = 1$; $B = 1$, or (3) $B = 1$. Given this structure, each *intersection* hypothesis $i \in I$ is comprised of a set independent, annotation combinations C_i. We then compute the DNM rate for each intersection hypothesis by summing the rates for its independent combination affiliates, $\lambda_i = \sum_{c \in C_i} \lambda_c$, such that the DNM counts at the intersection follow a Poisson distribution, $Y_i \sim \text{Poisson}(\lambda_i)$.

5.2.2 Example data

Throughout this chapter, we work with data from previous studies [An et al., 2018] to ensure our methods are working on realistic examples. We consider fourteen annotations indicating GENCODE definitions, conservation, and functional status. Since there are annotations that do not overlap with each other, we observe $|C| = 371$ independent combinations of the fourteen annotations, i.e., there are 371 unique observations of fourteen indicator variables. We then observe $\mathcal{I} = 420$ different annotation intersections. Unlike the combinations which always consider all fourteen annotations, we designate intersections by the *order* of their intersection, i.e., the number of annotation indicators equal to one. For instance, in the previously mentioned example, $A = 1$ is considered an order-1 intersection, while $A = 1$; $B = 1$ is an order-2 intersection. Because there are certain annotations that do not overlap with each, for our example data we observe up to order-6 intersections. These different order of intersections could be interpreted as a DAG-structure, where order-1 intersections represent the top of the DAG while the combinations (which are technically order-14 by our construction) are the leaves. For the remainder of this chapter, we decide to remove the order-1 and order-2 intersections from our analysis due to their overwhelming size and the potential lack of insight gained from detecting associations at such a high level. This leaves us with $\mathcal{I} = 344$ remaining intersection hypotheses for testing.

5.3 Methods

5.3.1 Agglomerative testing approach

By construction, we can observe overlap between the considered annotation intersections leading to test statistics with substantial levels of correlation, thus confounding the interpretability of our error rate guarantees from various multiple testing procedures. Because of the structure of the independent annotation combinations, we can compute the covariance between the DNM counts for two intersections i and i' based on the overlap in their independent combination members:

$$\text{Cov}(Y_i, Y_{i'}) = \begin{cases} \sum_{c \in C_i \bigcap C_{i'}} \lambda_c \text{ if } |C_i \bigcap C_{i'}| > 0 \\ 0, \text{ else} \end{cases} \tag{5.2}$$

To account for substantially correlated test statistics, we use an agglomerative procedure that was implemented in the context of gene-level testing [Yurko et al., 2021b] to cluster highly correlated intersections together for testing. Given a correlation threshold r, we apply the following procedure to a set of annotation intersections $\mathcal{I}$:

1. Let each intersection be its own test i with $|\mathcal{I}|$ initial hypotheses in total.

2. Compute $\text{Cor}(Y_i, Y_{i'})$ for all pairs of intersections, $i, i' \in \mathcal{I}$, (using Equation 5.2 and each intersection's assumed Poisson variance λ_i and $\lambda_{i'}$).

3. Repeat the following until $\text{Cor}(Y_i, Y_{i'}) < r$ for all remaining pairs of annotation intersections in $\mathcal{I}$:

   - Find intersection $\{i_*, i'_*\} = \underset{i,i' \in \mathcal{I}, i \neq i'}{\arg\max} \text{Cor}(Y_i, Y_{i'})$,
   - Merge $\{i_*, i'_*\}$ into group i_{merge},
   - Update $\mathcal{I} = \mathcal{I} \setminus \{i_*, i'_*\} \cup \{i_{merge}\}$, and compute $\text{Cor}(Y_{i_{merge}}, Y_{i'})$ for all remaining $i' \in \mathcal{I}$.

This is an agglomerative clustering algorithm with a linkage function determined by the assumed correlation structure of the Poisson test statistics. We can use the observed DNM counts y_i as the test statistic for each clustered annotation intersection $i \in \mathcal{I}$, and then compute the p-values based on one-sided enrichment tests, i.e., $p_i = \Pr(Y_i \geq y | \lambda_i)$. For our example data, we observe $\mathcal{I} = 96$ clustered intersections at threshold $r = 0.50$, thus greatly reducing the number of tests but also indicating the prevalent levels of correlation between the various intersections.

5.3.2 AdaPT implementation with annotation features

Given a collection of intersection-level p-values and metadata, $(p_i, x_i)_{i \in \mathcal{I}}$, we apply AdaPT [Lei and Fithian, 2018] to select a subset of discoveries with FDR control at target level $\alpha = 0.05$. However, due discreteness of the p-values for one-sided Poisson tests and the potentially conservative inflation of p-values equal to one exactly, we cannot use the original masking function in AdaPT. Instead, we implement recent developments in masking functions [Duan et al., 2020b, Chao and Fithian, 2021] to handle the smaller number of tests and also remove the conservative p-values equal to one that would result in AdaPT failing to reject anything. These general masking functions take the form of,

$$g(p_i) = \begin{cases} \frac{\nu - p_i}{\zeta} & p_i \in [\eta, \nu] \\ p_i & \text{otherwise} \end{cases} \tag{5.3}$$

where ζ is a "stretch factor" based on inputs for the thresholds α_m, ν, and η:

$$\zeta = \frac{\nu - \eta}{\alpha_m}. \tag{5.4}$$

This leads to an updated FDP estimate at step t of the AdaPT search,

$$\text{FDP}_t = \frac{A_t + 1}{\zeta R_t}. \tag{5.5}$$

We then choose values for these inputs such that AdaPT only needs a minimal number of rejections to be made in order to work, as well as avoid the inflation of p-values equal to one. In the simulation studies below, we set the values to be $\alpha_m = 0.125$, $\eta = 0.125$, and set $\zeta = 6.6$ such that the upper bound for p-values is $\nu \approx 0.95$. At target FDR level $\alpha = 0.05$, this means we need at least three rejections for AdaPT to be able to return any discoveries.

Similar to [Yurko et al., 2021b], we incorporate metadata about these annotation hypotheses from feature space $x_i \in \mathcal{X}$ using XGBoost [Chen and Guestrin, 2016]. For each annotation intersection i, we create a vector of information summarizing which annotations comprise each intersection. For each annotation $a \in \mathcal{A}$, we let x_{ia} denote the fraction of the intersection's independent combination where that annotation indicator is equal to one:

$$x_{ia} = \frac{1}{|C_i|} \sum_{c \in C_i} x_{ca}, \tag{5.6}$$

where x_{ca} is an indicator variable denoting whether or not annotation a is equal to one for combination c. Obviously, this is just a single approach to encoding the annotation information at this intersection-level structure. We leave further encoding for future work.

5.3.3 Data blurring augmentation

While we can create features summarizing the annotations at the intersection-level, we are interested in leveraging the unique structure of the independent annotation combinations. We want to somehow explore the independent-level data to help us effectively order the intersections we are interested in testing. In order to explore the annotations, we consider a *data blurring* approach which has been primarily used in the context of estimation after selection [Leiner et al., 2021]. We consider the following steps:

Step 1 - Blur data. Given the use of independent Poissons, we observe y_c DNMs for each independent annotation combination. We then create a blurred version $\tilde{y}_c$ which is defined as,

$$\tilde{y}_c = y_c + z_c \text{ where } z_c \sim \text{Poisson}(\tau \lambda_c), \tag{5.7}$$

and $\tau > 0$ controls the amount of blurring that takes place. Under this form, we assume $\tilde{Y}_c \sim \text{Poisson}(\lambda_c \cdot (1 + \tau))$.

Step 2 - Model blurred data. Next, we can model the blurred counts $\tilde{y}_c$ at the independent combination level to derive new features based on exploring the blurred data. First, we consider a Poisson model of the blurred counts accounting for the offset term from the blurred data:

$$\log \lambda_c^*(x_c) = \log(\lambda_{0c} \cdot (1 + \tau)) + \beta(x_c), \tag{5.8}$$

where λ_{0c} is our assumed null DNM rate and $\beta(x_c)$ is the effect size as a function of the annotation indicators x_c, which we learn with gradient boosted trees. Additionally, we

also implement a version of the classical two-groups model [Efron et al., 2001] yielding the conditional mixture probability mass function,

$$f(\tilde{y_c}|x_c) = \pi_1(x_c)f(\tilde{y_c}; \lambda_{0c}e^{\beta(x_c)} \cdot (1+\tau)) + (1-\pi_1(x_c))f_0(\tilde{y_c}; \lambda_{0c} \cdot (1+\tau)), \tag{5.9}$$

where $f(y;\lambda)$ is the Poisson probability mass function for a Poisson random variable. In this form, we model the non-null probability $\pi_1(x)$ and the non-null effect size $\beta(x_c)$. In this model, there is one missing variable: the hypothesis status H_c regarding the combination is never observed. An expectation-maximization (EM) algorithm can be used to estimate both $\hat{\pi}_1(x_c)$ and $\hat{\beta}(x_c)$ by maximizing the partially observed likelihood. The complete data log-likelihood for the conditional two-groups model in this form is,

$$l(\pi_1, \beta; y_c, H_c, x_c) = \sum_{c \in C}\{H_c \text{log}\pi_1(x_c) + (1-H_c)\text{log}(1-\pi_1(x_c))\} + \tag{5.10}$$

$$\sum_{c \in C}\{H_c \text{log}f(y_c; \lambda_{0c}e^{\beta(x_c)} \cdot (1+\tau)) + (1-H_c)\text{log}f(y_c; \lambda_{0c} \cdot (1+\tau))\}. \tag{5.11}$$

During the E-step of the iteration of the EM algorithm, given estimates $\hat{\pi}_1$ and $\hat{\beta}$, we compute:

$$\hat{H}_c = \frac{\pi_{1c}f(Y_c; \lambda_{0c}e^{\beta(x_c)} \cdot (1+\tau))}{\pi_{1c}f(y_c; \lambda_{0c}e^{\mu(x_c)} \cdot (1+\tau)) + 1 - \pi_{1c}f(y_c; \lambda_{0c} \cdot (1+\tau))}. \tag{5.12}$$

The M-step consists of estimating $\hat{\pi}_1$ and $\hat{\mu}$ with separate gradient boosted trees. In order to fit the model for $\pi_1(x_i)$, we construct the response vector $v_\pi^{(d)} = (1, \ldots, 1, 0, \ldots, 0) \in \mathbb{R}^{2|C|}$ and use weights $w_\pi = (\hat{H}_1, \ldots, \hat{H}_{|C|}, 1-\hat{H}_1, \ldots, 1-\hat{H}_{|C|}) \in \mathbb{R}^{2|C|}$. Then we estimate $\hat{\pi}_1(x_c)$ using the first $|C|$ predictions from a classification model using v_π as the response variable with the covariate matrix $(x_c)_{c\in[C]}$ replicated twice and weights w_π. Then, for estimating $\hat{\mu}(x_c)$ we construct a response vector with the blurred counts $v_\mu = (\tilde{y}_1, \ldots, \tilde{y}_{|C|}) \in \mathbb{Z}^{+|C|}$ with weights $w_\mu = (\hat{H}_1, \ldots, \hat{H}_{|C|}) \in \mathbb{R}^n$.

We use logistic and Poisson regression implementations in XGBoost [Chen and Guestrin, 2016] to model π_1 and β respectively. For initialization of β, we use an unweighted model of the blurred counts $\tilde{y}_c$. For initialization of π_1, we use the same approach by [Boca and Leek, 2018] and [Ignatiadis and Huber, 2021].

Step 3 - Accumulate blurred data to intersection level. Because we are interested in testing the merged groups of intersections, we first accumulate the blurred data to the intersection-level. Based on the mapping of the combinations to intersections, denoted by the set C_i, we compute the blurred counts as $\tilde{y}_i = \sum_{c \in C_i} \tilde{y}_c$. We then accumulate the conditionally independent non-null probability and effect size estimates based on the mapping of combinations to intersections. To compute the intersection-level effect size estimate $\hat{\beta}_i$, we simply sum up

the estimated non-null rates for each corresponding member combination and divide by the assumed null rate under the blurred model:

$$\log \hat{\beta}_i = \log \big(\sum_{c \in C_i} \lambda_{0c} e^{\hat{\beta}_c} \cdot (1+\tau) \big) - \log \big(\lambda_i \cdot (1+\tau) \big). \tag{5.13}$$

Similarly, we compute the non-null probability based on the product of the conditionally independent non-null probability estimates at the combination level,

$$\hat{\pi}_{1i} = 1 - \prod_{c \in C_i} (1 - \hat{\pi}_{1c}). \tag{5.14}$$

We then compute the conditional local fdr [Efron et al., 2001] for each test as,

$$\text{lfdr}_i = \frac{(1-\hat{\pi}_i) f\big(\tilde{y}_i; \lambda_{0i} \cdot (1+\tau)\big)}{\hat{\pi}_i f\big(\tilde{y}_i; \lambda_{0i} e^{\hat{\beta}_i} \cdot (1+\tau)\big) + (1-\hat{\pi}_i) f\big(\tilde{y}_i; \lambda_{0i} \cdot (1+\tau)\big)}. \tag{5.15}$$

Step 4 - Perform conditional binomial test. Since we used the blurred counts $\tilde{y}_i$ (from modeling $\tilde{y}_c$) to explore the hypotheses, we then must test $Y_i | \tilde{Y}_i$ to ensure our hypothesis testing is independent of the exploration step (*Step 2* above) [Leiner et al., 2021]. This conditional test becomes a one-sided binomial test because,

$$Y_i | \tilde{Y}_i \sim \text{Binomial}(\tilde{Y}_i, \frac{1}{1+\tau}). \tag{5.16}$$

By design, the blurring parameter τ controls the tradeoff we have between hypothesis exploration in *Step 2* and the testing step in *Step 4*. As τ increases, the power in the conditional binomial test will improve - but the exploration step in modeling the blurred counts will likely suffer.

Step 5 - Guide conditional test with blurred exploration. To overcome the power lost by introducing variance from blurring the data, we use the estimates constructed in *Step 2* as new metadata for the intersection-level hypotheses. Specifically, in the simulations below we demonstrate the performance using the local fdr estimates as features in AdaPT, coupled with the annotation proportion values that were constructed prior to blurring. The goal is to see if additional information constructed from modeling the blurred data can ultimately lead to improved power over the results without blurring.

5.4 Simulation studies

In order to evaluate our potential approaches for augmenting CWAS, we consider simulation iterations with the following steps:

- Choose a set of annotation interactions to be non-null, i.e., where a particular choice of annotations are all equal to one.

- Determine which of the annotation combinations $c \in C$ are non-null (C_1) based on if they contain the selected non-null interaction, e.g., if the triplet (a_1, a_2, a_3) denotes a non-null annotation interaction then $H_c = 1$ if $x_{ca_1} \cdot x_{ca_2} \cdot x_{ca_3} = 1$ regardless of the status of the other eleven annotations.

- Under the assumption we have access to null mutation rates λ_{0c} for each independent annotation combination, we multiply it's mutation rate by a factor of e^{β} if it's non-null:

$$\lambda_c = \begin{cases} \lambda_{0c} \cdot e^{\beta}, \text{ if } c \in C_1 \\ \lambda_{0c} \text{ otherwise} \end{cases} \tag{5.17}$$

- Generate the observed number of DNMs y_c for each of the independent combinations, given the desired total number of DNMs Y^*:

$$Y_c \sim \text{Multinomial}\Big(Y^*, \frac{\lambda_c}{\sum_{d \in C} \lambda_d}\Big) \tag{5.18}$$

- Proceed to compute both number of DNMs y_i and null rates λ_{0i} for each of the considered intersection/group hypotheses $i \in I$, by summing across the their respective member combinations y_c and λ_{0c} for $c \in C_i$. The null status of the intersection/group hypotheses H_i is determined by the structure of which independent combinations are accumulated for that intersection,

$$H_i = \begin{cases} 1, \text{ if } \exists H_c = 1, \text{ for } c \in C_i \\ 0, \text{ otherwise} \end{cases} \tag{5.19}$$

In the simulation studies below we set $Y^* = 63492$ to match the number of non-coding DNMs observed in new datasets, providing us with realistic scenarios for our simulations.

5.4.1 Agglomerative AdaPT results without blurring

We first compare the performance of AdaPT with the annotation features considered above relative to an AdaPT intercept-only approach (without accounting for metadata) and BH [Benjamini and Hochberg, 1995] on the clustered intersection hypotheses with threshold $\rho = 0.5$. These tests are based on using the p-values from one-sided Poisson tests given the simulated case counts y_i. We generate 1,000 simulations following the above steps values of β ranging from 0 (null) to $\log 4$, over a grid of ten values, and observe both the intersection-level power and FDR control at target level $\alpha = 0.5$. We observe clear gains in

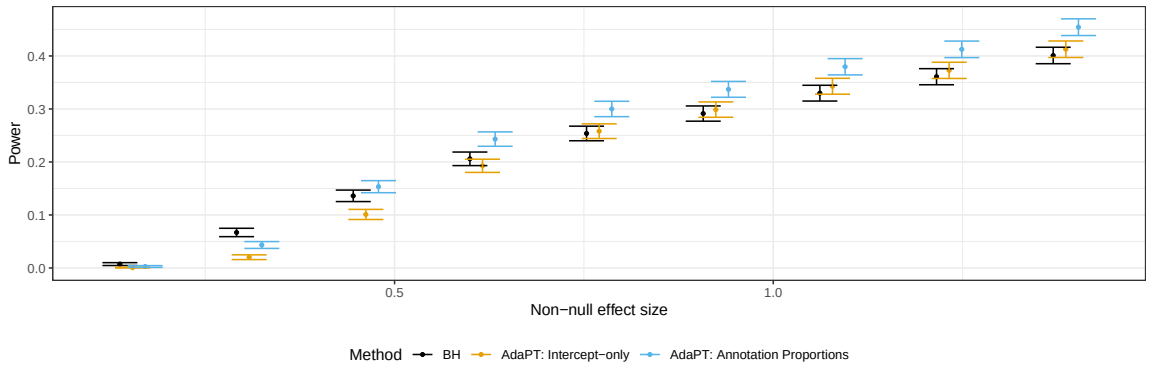

Figure 5.3: Comparison of power (+/- one standard error) across 1,000 simulations between AdaPT with annotation proportions (blue) to BH (black) and the AdaPT intercept-only approach (orange), as a function of the non-null effect size.

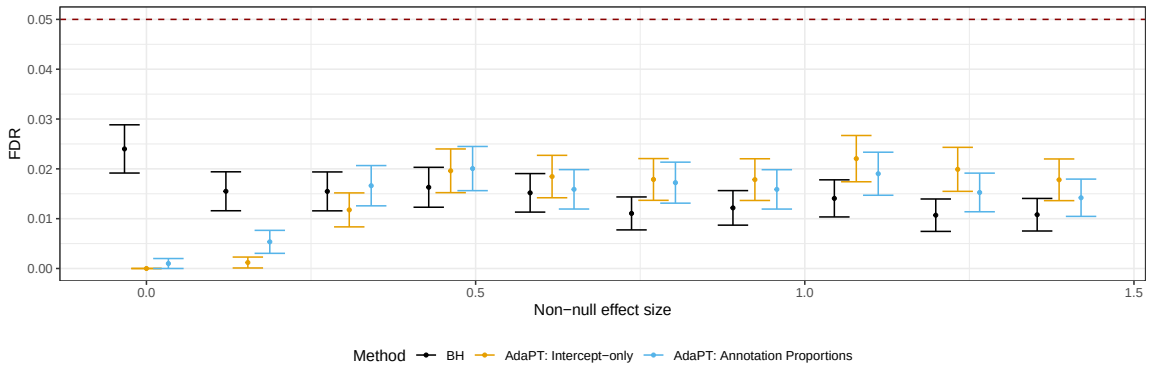

Figure 5.4: Comparison of FDR control (+/- one standard error) across 1,000 simulations between AdaPT with annotation proportions (blue) to BH (black) and the AdaPT intercept-only approach (orange), as a function of the non-null effect size.

power from using AdaPT with the annotation proportions (AdaPT: Annotation Proportions in blue) compared to both BH (black) and AdaPT without any side information (orange) - with the only drawback seen in the initial weak settings where we suffer from power to detect anything (Figure 5.3). All three approaches suffer from conservative control likely due to the positive dependence structure remaining post-merging intersection hypotheses at the chosen threshold of $\rho = 0.5$ (Figure 5.4). These results indicate the clear advantages of including the annotation indicators in some form to demonstrably improve power.

5.4.2 Comparison of results with blurring

Next, we compare the performance of AdaPT using features created from exploring data via the blurring steps based on 100 simulations across the same grid of non-null effect sizes μ but now varying the amount of blurring τ, ranging from 0.5 to 10 in increments of 0.5. To evaluate the effectiveness of the performance with blurring, we compute the power for BH on the conditional binomial tests post-blurring along with AdaPT using as a feature the local fdr estimates defined in the previous section. For reference, we consider the performance of AdaPT using only the local fdr features from blurring versus using only the annotation proportions (again post-blurring), then finally using both the local fdr and annotation proportions together. The goal is to see if any information modeling in the blurring step can be used to improve our power over the performance with the annotation features constructed without blurring.

While we can see the gains in power from using our features constructed from the blurred exploration step over pre-blurring BH (black dashed line), we fail to improve our power over the AdaPT results using the annotation proportions without blurring (blue dashed line, Figure 5.5). We can see that the inclusion of the blurred estimates for local fdr as features in AdaPT, whether it is exclusive or included with the annotation proportions, leads to the same results post-blurring as the AdaPT results with annotation proportions. While we have evidence indicating the annotation proportions are useful features alone (Figure 5.3), our current approach for using features derived from exploring blurred data does not lead to any additional improvement.

5.5 Conclusions

In this study, we proposed the use of an agglomerative algorithm to cluster overlapping annotation intersections and then demonstrated in simulations an improvement in power by augmenting CWAS hypotheses with annotation summaries. This approach is analogous to our previous work in the common variant setting [Yurko et al., 2021b] but adapted for the context of rare variant studies. We only considered annotation proportions as metadata in this chapter, but additional summary statistics and external data can be accounted for to further improve our power. We then explored the use of data blurring in order to see if an additional exploration step of the lower-level annotation combinations could lead to further improvements in power. Our current approach based on generating additional features for AdaPT from the blurred step, in the form of conditional local fdr estimates, failed to demonstrate any meaningful gains in power beyond available annotation summaries.

For next steps, data blurring enables estimation of CWAS effect sizes post-selection [Leiner et al., 2021]. For instance, we can use AdaPT with summaries about annotations and additional metadata to improve our power on blurred DNM counts, with the amount of blurring chosen such that AdaPT displays the same level of power as BH without

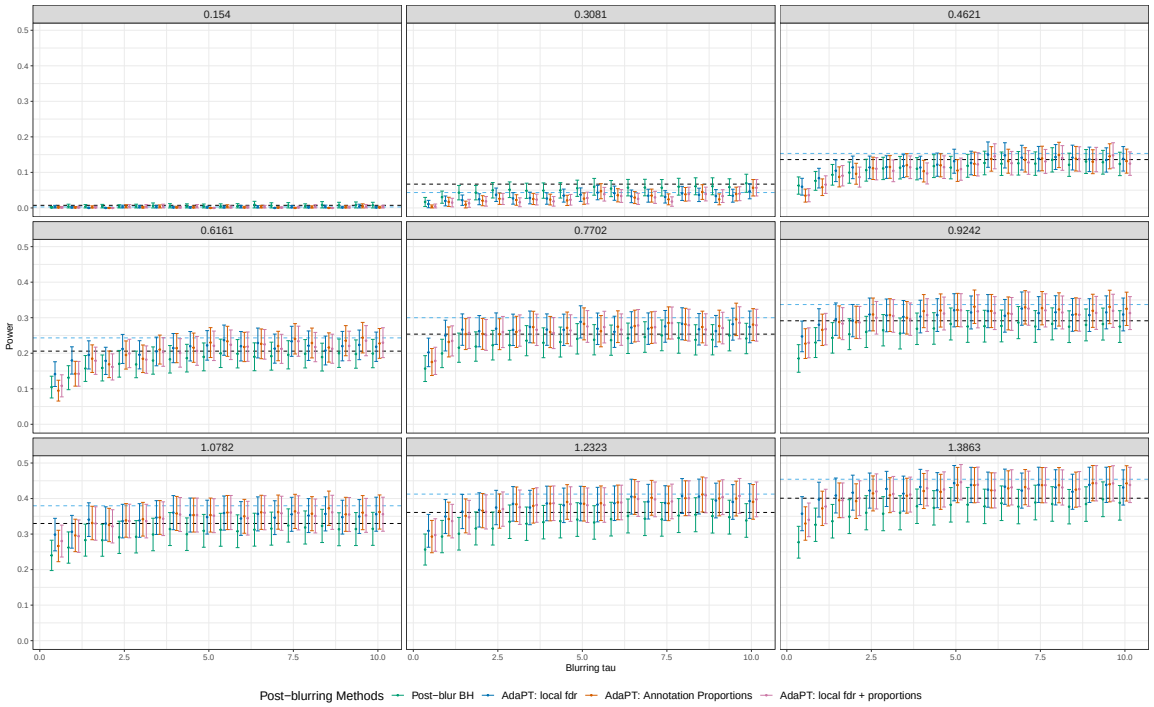

Figure 5.5: Comparison of power (+/- one standard error) across 100 simulations between AdaPT with different sets of features in comparison to the results before blurring with annotation proportions (dashed blue line) and BH (dashed black line) as a function of the blurring parameter τ. Each panel corresponds to an increasing non-null effect size β.

blurring (Figure 5.6). Then we could construct valid post-selection confidence intervals on the conditional binomial test statistics [Leiner et al., 2021], or perform conditional testing [Heller et al., 2019] within the clustered hypotheses, while retaining greater power than considering these same steps post-selection with BH. While we did not observe improvements in overall selection power via blurring in our current approach, the use of blurring may prove optimal when considering the optimal approach for aggregating tests in this context. For example, different global hypotheses may vary in terms of the distribution of signal, e.g., sparse versus dense signal regimes. Rather than computing a single test statistic in the manner we have in this chapter, we could instead explore blurred data to figure out the optimal testing strategy for a particular global hypothesis. Coupled with multiple testing corrections guided by metadata, determining the best way to aggregate lower-level test statistics in a valid manner with blurred data could yield fruitful results in both the CWAS context as well as analogous problems in gene-level testing with common variants.

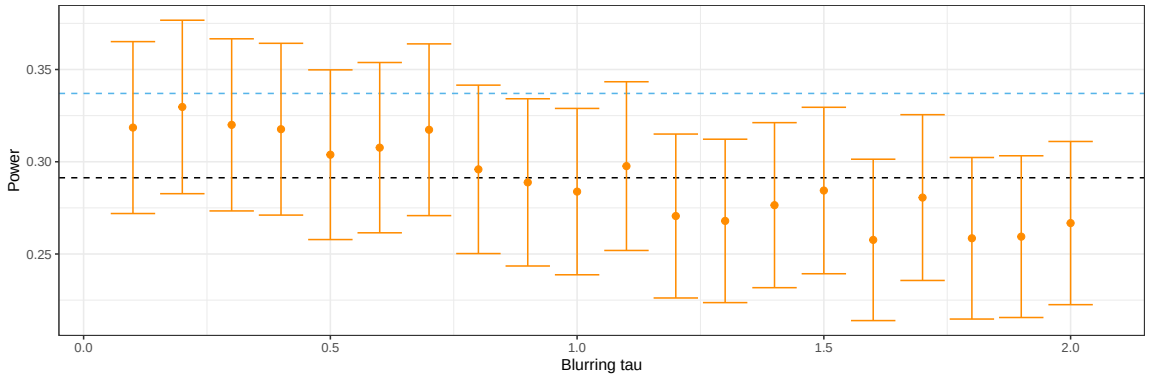

Figure 5.6: Comparison of power (+/- one standard error) across 100 simulations for AdaPT with annotation proportions based on one-sided Poisson tests with blurred DNM counts as a function of blurring parameter τ for a particular choice of non-null effect size β. For reference, the pre-blurring results for AdaPT with annotation proportions (blue) and BH (black) are displayed as horizontal dashed lines.

Six

Conclusions and future work

My thesis explores the challenging problem of improving power to detect association in under-powered genetics studies with the use of selective inference methods. While there has been great progress in the development of selective inference methodology in recent years, there has been limited applied work in the setting of genetic studies taking advantage of these approaches for handling the burden of multiple testing in the presence of weak effects. This slow adoption can be partially attributed to implementation challenges for these approaches, both statistical and computational. My Ph.D. has focused on addressing such implementation difficulties in an effort to encourage adoption of these methods in genetics.

The thesis contains four related projects aimed to augmenting genetic association studies for neuropsychiatric disorders with metadata across both common and rare variant analysis.

In chapter 2, we boost the performance of adaptive AdaPT by recognizing it is not a specific algorithm that one can simply apply to a dataset, but rather a meta-algorithm for coupling machine learning methods to multiple testing problems without compromising FDR control. We embrace AdaPT's flexibility via gradient boosted trees in a rich, high-dimensional multi-omics settings and investigate the neurobiology of schizophrenia in the process. Specifically, we built a pipeline to select a subset of single-nucleotide polymorphisms (SNPs) documented to affect gene expression and then incorporate covariates from independent GWAS and gene expression studies into AdaPT to ulimately improve our power. Our boosting implementation of AdaPT scales with more covariates and addresses the perceived modeling weakness of AdaPT, enabling practitioners to capture interactions and non-linear effects from resources of available multi-omics metadata.

Our investigation motivates several questions that could lead to interesting extensions that are worthy of future work. First, although we demonstrate in simulation AdaPT appears to maintain FDR control in relevant dependence settings to LD structure underlying GWAS results, there is a need to explore in greater detail AdaPT's properties under various dependence regimes. Furthermore, there are opportunities to address adaptive testing schemes in the presence of dependence directly [Fithian and Lei, 2020]. We also envision our GWAS analysis with AdaPT to assist in identifying variants likely driving GWAS association

signal with functional roles regarding gene expression, assisting in colocalization analysis [Foley et al., 2021, Giambartolomei et al., 2018]. Additionally, this work could be used to potentially improve the performance of polygenic risk scores, which are numerical summaries indicating an individual's risk for a disease / phenotype based on identified genetic risk factors [Ni et al., 2021]. The use of adaptive thresholding approaches guided by metadata could help inform to indentify and weight potential risk factors [Amariuta et al., 2020].

In chapter 3, we transition to gene-based testing, which can improve power to detect weak signal by reducing multiple testing and pooling signal strength. This can be advantageous for settings with weaker signal, such as those observed in studies for neuropsychiatric disorders. While there are many approaches for global testing, the presence of LD poses a challenge here: the combination of dependent SNP-level summary statistics at the gene-level must adjust for the LD-induced covariance of SNPs. When investigating MAGMA, a popular tool for this problem, we discovered it yielded an unusual distribution of gene-level p-values, which would violate necessary assumptions for AdaPT to maintain FDR control. Despite undocumented, ad-hoc corrections in MAGMA, we observe via simulations and recent applications that it yields incorrect null p-value distribution resulting in inflated error rates. This is due to the inappropriate application of an approximation that is valid for only one-sided tests, while GWAS summary statistics are two-sided.

In chapter 4, we observed that current gene-based testing approaches do not capture LD of SNPs falling in nearby genes, which can induce correlation of gene-based test statistics. This compromises the interpretability of a gene-based test, obscuring the meaning of error-rate guarantees. We introduce an algorithm to account for this correlation directly, based on the LD-induced correlation of commonly used quadratic gene-level test statistics. When a gene's test statistic is independent of others, it is assessed separately, but when test statistics for genes are strongly correlated, their SNPs are agglomerated and tested as a locus. Using our implementation of AdaPT guided by gradient boosted trees, we are able to improve power to select ASD-associated genes and LD-defined loci while maintaining finite-sample FDR control. We observe how improvements are modest for other, well-powered phenotypes in comparison to ASD. We complement this algorithm with an interactive visualization tool to explore localized signal and shed light on biological signals therein.

Our work in the common variant setting demonstrates clear advantages of augmenting association studies with metadata to improve power. But can we make more general statements about which sources of side information are relevant for improving power? Or is it specific to the context of the association study of interest? These are questions relevant to practitioners desiring to use these approaches in their research. While we provided variable importance summaries of gradient boosted trees, there is ample opportunity for deeper investigation. One problem pertinent to gene-level testing is the assumption of availability of using a LD reference panel. However, this means our gene-based approaches are sensitive to

potential mismatches with the assumed LD reference structure. A mismatch could lead to an inappopriate null distribution, as well as potentially over/under-clustering genes. Properly accounting for sensitivity to LD mismatch is an important area for future research it also pertains to the portability of our approach across ancestries.

In chapter 5, we focus on problems in the context of rare variant analysis to improve power for detecting noncoding associations in CWAS. Via simulations, we demonstrated improvements in power by accounting for annotation-level summaries using AdaPT, in an analogous manner with our gene-level testing approach by also clustering strongly correlated hypotheses together. We also investigated a data blurring approach to separately model and explore lower-level annotation combinations to guide our multiple testing corrections. However, our current approach failed to yield power improvements over the application of AdaPT with features constructed without blurring. We emphasize, however, that the role blurring still leads to advantages in terms of estimation post-selection with an adaptive threshold providing further insight into understanding rare variant associations in noncoding regions of the genome.

Our preliminary simulation studies for CWAS indicate similar findings with our work in common variants: clear gains in power from augmenting testing procedures with AdaPT and available side information via gradient boosted trees. But there are several opportunities for future work in this space, such as exploiting the hierarchical structure of annotation categories in CWAS [Werling et al., 2018]. The STAR framework [Lei et al., 2021] is a potential way to address this problem, as a more careful testing scheme could be designed to incorporate the DAG structure of the annotation categories. Additionally, while the Cauchy combination test provides one way to aggregate several tests together [Liu et al., 2019, Liu and Xie, 2020], there may be ways to exploit data blurring to improve selection power in the context of determining the optimal approach for aggregating test statistics for global testing and coupling it with interactive procedures [Duan et al., 2020a]. While this thesis has focused on approaches for controlling FDR, there is an opportunity to meld approaches for adaptively controlling FWER in genetics studies with high-dimensional metadata [Duan et al., 2020b, Ignatiadis and Huber, 2021]. I believe the contribution of my thesis helps bridge the gap between selective inference methodology and practical applications in genetic studies, and motivates the need for the development of new methods and future studies in this challenging area.

Bibliography

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

[Amariuta et al., 2020] Amariuta, T., Ishigaki, K., Sugishita, H., Ohta, T., Koido, M., Dey, K. K., Matsuda, K., Murakami, Y., Price, A. L., Kawakami, E., et al. (2020). Improving the trans-ancestry portability of polygenic risk scores by prioritizing variants in predicted cell-type-specific regulatory elements. *Nature genetics*, 52(12):1346–1354.

[An et al., 2018] An, J.-Y., Lin, K., Zhu, L., Werling, D. M., Dong, S., Brand, H., Wang, H. Z., Zhao, X., Schwartz, G. B., Collins, R. L., et al. (2018). Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. *Science*, 362(6420).

[Arbogast et al., 2017] Arbogast, T., Iacono, G., Chevalier, C., Afinowi, N. O., Houbaert, X., van Eede, M. C., Laliberte, C., Birling, M.-C., Linda, K., Meziane, H., et al. (2017). Mouse models of 17q21. 31 microdeletion and microduplication syndromes highlight the importance of kansl1 for cognition. *PLoS genetics*, 13(7):e1006886.

[Ashburner et al., 2000] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29.

[Benjamini and Hochberg, 1995] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

[Boca and Leek, 2018] Boca, S. M. and Leek, J. T. (2018). A direct approach to estimating false discovery rates conditional on covariates. *PeerJ*, 6:e6035.

[Bonferroni, 1935] Bonferroni, C. E. (1935). Il calcolo delle assicurazioni su gruppi di teste. *Studi in onore del professore salvatore ortu carboni*, pages 13–60.

[Boyle et al., 2017] Boyle, E. A., Li, Y. I., and Pritchard, J. K. (2017). An expanded view of complex traits: From polygenic to omnigenic. *Cell*, 169(7):1177–1186.

[Brown, 1975] Brown, M. B. (1975). 400: A method for combining non-independent, one-sided tests of significance. *Biometrics*, 31(4):987–992.

[Buniello et al., 2019] Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2019). The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research*, 47(D1):D1005–D1012.

[Bush and Moore, 2012] Bush, W. S. and Moore, J. H. (2012). Chapter 11: Genome-wide association studies. *PLoS computational biology*, 8(12):e1002822.

[Cantor et al., 2005] Cantor, R. M., Kono, N., Duvall, J. A., Alvarez-Retuerto, A., Stone, J. L., Alarcón, M., Nelson, S. F., and Geschwind, D. H. (2005). Replication of autism linkage: fine-mapping peak at 17q21. *The American Journal of Human Genetics*, 76(6):1050–1056.

[Carmona-Mora and Walz, 2010] Carmona-Mora, P. and Walz, K. (2010). Retinoic acid induced 1, rai1: a dosage sensitive gene related to neurobehavioral alterations including autistic behavior. *Current genomics*, 11(8):607–617.

[Chang et al., 2020] Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J. (2020). *shiny: Web Application Framework for R*. R package version 1.4.0.2.

[Chao and Fithian, 2021] Chao, P. and Fithian, W. (2021). Adapt-gmm: Powerful and robust covariate-assisted multiple testing. *arXiv preprint arXiv:2106.15812*.

[Chen and Guestrin, 2016] Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794. ACM.

[Cirillo et al., 2018] Cirillo, E., Kutmon, M., Gonzalez Hernandez, M., Hooimeijer, T., Adriaens, M. E., Eijssen, L. M. T., Parnell, L. D., Coort, S. L., and Evelo, C. T. (2018). From snps to pathways: Biological interpretation of type 2 diabetes (t2dm) genome wide association study (gwas) results. *PLOS ONE*, 13(4):1–19.

[Cross-Disorder Group of the Psychiatric Genomics Consortium, 2013] Cross-Disorder Group of the Psychiatric Genomics Consortium (2013). Genetic relationship between five psychiatric disorders estimated from genome-wide snps. *Nature Genetics*, 45:984 EP –.

[de Leeuw et al., 2015] de Leeuw, C. A., Mooij, J. M., Heskes, T., and Posthuma, D. (2015). Magma: Generalized gene-set analysis of gwas data. *PLOS Computational Biology*, 11(4):1–19.

[De Leeuw et al., 2016] De Leeuw, C. A., Neale, B. M., Heskes, T., and Posthuma, D. (2016). The statistical properties of gene-set analysis. *Nature Reviews Genetics*, 17(6):353.

[De Rubeis et al., 2014] De Rubeis, S., He, X., Goldberg, A. P., Poultney, C. S., Samocha, K., Cicek, A. E., Kou, Y., Liu, L., Fromer, M., Walker, S., et al. (2014). Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature*, 515(7526):209–215.

[de Santiago, 2020] de Santiago, I. (2020). Seqqc. `http://inesdesantiago.github.io/SeqQC.blog/`.

[Demontis et al., 2019] Demontis, D., Walters, R. K., Martin, J., Mattheisen, M., Als, T. D., Agerbo, E., Baldursson, G., Belliveau, R., Bybjerg-Grauholm, J., Bækvad-Hansen, M., et al. (2019). Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nature genetics*, 51(1):63–75.

[Duan et al., 2020a] Duan, B., Ramdas, A., Balakrishnan, S., and Wasserman, L. (2020a). Interactive martingale tests for the global null. *Electronic Journal of Statistics*, 14(2):4489–4551.

[Duan et al., 2020b] Duan, B., Ramdas, A., and Wasserman, L. (2020b). Familywise error rate control by interactive unmasking. In *International Conference on Machine Learning*, pages 2720–2729. PMLR.

[Eddelbuettel and Sanderson, 2014] Eddelbuettel, D. and Sanderson, C. (2014). Rcpparmadillo: Accelerating r with high-performance c++ linear algebra. *Computational Statistics and Data Analysis*, 71:1054–1063.

[Efron et al., 2001] Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). Empirical bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96(456):1151–1160.

[Fisher, R.A., 1925] Fisher, R.A. (1925). *Statistical Methods for Research Workers*. Oliver & Boyd, Edinburgh.

[Fithian and Lei, 2020] Fithian, W. and Lei, L. (2020). Conditional calibration for false discovery rate control under dependence. *arXiv preprint arXiv:2007.10438*.

[Foley et al., 2021] Foley, C. N., Staley, J. R., Breen, P. G., Sun, B. B., Kirk, P. D., Burgess, S., and Howson, J. M. (2021). A fast and efficient colocalization algorithm for identifying shared genetic risk factors across multiple traits. *Nature communications*, 12(1):1–18.

[Friedman, 2001] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232.

[Fromer et al., 2014] Fromer, M., Pocklington, A. J., Kavanagh, D. H., Williams, H. J., Dwyer, S., Gormley, P., Georgieva, L., Rees, E., Palta, P., Ruderfer, D. M., et al. (2014). De novo mutations in schizophrenia implicate synaptic networks. *Nature*, 506(7487):179–184.

[Furukawa et al., 2003] Furukawa, K., Wang, Y., Yao, P. J., Fu, W., Mattson, M. P., Itoyama, Y., Onodera, H., D'Souza, I., Poorkaj, P. H., Bird, T. D., et al. (2003). Alteration in calcium channel properties is responsible for the neurotoxic action of a familial frontotemporal dementia tau mutation. *Journal of neurochemistry*, 87(2):427–436.

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

[Gerring et al., 2019] Gerring, Z. F., Gamazon, E. R., Derks, E. M., et al. (2019). A gene co-expression network-based analysis of multiple brain tissues reveals novel genes and molecular pathways underlying major depression. *PLoS genetics*, 15(7):e1008245.

[Giambartolomei et al., 2018] Giambartolomei, C., Zhenli Liu, J., Zhang, W., Hauberg, M., Shi, H., Boocock, J., Pickrell, J., Jaffe, A. E., Consortium, C., Pasaniuc, B., et al. (2018). A bayesian framework for multiple trait colocalization from summary association statistics. *Bioinformatics*, 34(15):2538–2545.

[Grove et al., 2019] Grove, J., Ripke, S., Als, T. D., Mattheisen, M., Walters, R. K., Won, H., Pallesen, J., Agerbo, E., Andreassen, O. A., Anney, R., et al. (2019). Identification of common genetic risk variants for autism spectrum disorder. *Nature genetics*, 51(3):431–444.

[GTEx Consortium and others, 2015] GTEx Consortium and others (2015). The genotype-tissue expression (gtex) pilot analysis: multitissue gene regulation in humans. *Science*, 348(6235):648–660.

[Hamdan et al., 2010] Hamdan, F. F., Daoud, H., Rochefort, D., Piton, A., Gauthier, J., Langlois, M., Foomani, G., Dobrzeniecka, S., Krebs, M.-O., Joober, R., et al. (2010). De novo mutations in foxp1 in cases with intellectual disability, autism, and language impairment. *The American Journal of Human Genetics*, 87(5):671–678.

[Hansen and Klopfer, 2006] Hansen, B. B. and Klopfer, S. O. (2006). Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics*, 15(3):609–627.

[Harrow et al., 2012] Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B. L., Barrell, D., Zadissa, A., Searle, S., et al. (2012). Gencode: the

reference human genome annotation for the encode project. *Genome research*, 22(9):1760–1774.

[Hayfield and Racine, 2008] Hayfield, T. and Racine, J. S. (2008). Nonparametric econometrics: The np package. *Journal of Statistical Software*, 27(5).

[He et al., 2013] He, X., Sanders, S. J., Liu, L., De Rubeis, S., Lim, E. T., Sutcliffe, J. S., Schellenberg, G. D., Gibbs, R. A., Daly, M. J., Buxbaum, J. D., et al. (2013). Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. *PLoS Genet*, 9(8):e1003671.

[Heller et al., 2019] Heller, R., Meir, A., and Chatterjee, N. (2019). Post-selection estimation and testing following aggregate association tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(3):547–573.

[Höglinger et al., 2011] Höglinger, G. U., Melhem, N. M., Dickson, D. W., Sleiman, P. M., Wang, L.-S., Klei, L., Rademakers, R., De Silva, R., Litvan, I., Riley, D. E., et al. (2011). Identification of common variants influencing risk of the tauopathy progressive supranuclear palsy. *Nature genetics*, 43(7):699.

[Howard et al., 2018] Howard, D. M., Adams, M. J., Shirali, M., Clarke, T.-K., Marioni, R. E., Davies, G., Coleman, J. R., Alloza, C., Shen, X., Barbu, M. C., et al. (2018). Genome-wide association study of depression phenotypes in uk biobank identifies variants in excitatory synaptic pathways. *Nature communications*, 9(1):1–10.

[Ignatiadis and Huber, 2021] Ignatiadis, N. and Huber, W. (2021). Covariate powered cross-weighted multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(4):720–751.

[Ignatiadis et al., 2016] Ignatiadis, N., Klaus, B., Zaugg, J. B., and Huber, W. (2016). Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature methods*, 13(7):577–580.

[Karczewski et al., 2020] Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581(7809):434–443.

[Koopmans et al., 2019] Koopmans, F., van Nierop, P., Andres-Alonso, M., Byrnes, A., Cijsouw, T., Coba, M. P., Cornelisse, L. N., Farrell, R. J., Goldschmidt, H. L., Howrigan, D. P., et al. (2019). SynGO: an evidence-based, expert-curated knowledge base for the synapse. *Neuron*, 103(2):217–234.

[Korotkevich et al., 2019] Korotkevich, G., Sukhov, V., and Sergushichev, A. (2019). Fast gene set enrichment analysis. *bioRxiv*.

[Korthauer et al., 2019] Korthauer, K., Kimes, P. K., Duvallet, C., Reyes, A., Subramanian, A., Teng, M., Shukla, C., Alm, E. J., and Hicks, S. C. (2019). A practical guide to methods controlling false discoveries in computational biology. *Genome biology*, 20(1):1–21.

[Kost and McDermott, 2002] Kost, J. T. and McDermott, M. P. (2002). Combining dependent p-values. *Statistics Probability Letters*, 60(2):183 – 190.

[Kouri et al., 2015] Kouri, N., Ross, O. A., Dombroski, B., Younkin, C. S., Serie, D. J., Soto-Ortolaza, A., Baker, M., Finch, N. C. A., Yoon, H., Kim, J., et al. (2015). Genome-wide association study of corticobasal degeneration identifies risk variants shared with progressive supranuclear palsy. *Nature communications*, 6(1):1–7.

[Kurki et al., 2019] Kurki, M. I., Saarentaus, E., Pietiläinen, O., Gormley, P., Lal, D., Kerminen, S., Torniainen-Holm, M., Hämäläinen, E., Rahikkala, E., Keski-Filppula, R., et al. (2019). Contribution of rare and common variants to intellectual disability in a sub-isolate of northern finland. *Nature communications*, 10(1):1–15.

[Lee et al., 2018] Lee, J. J., Wedow, R., Okbay, A., Kong, E., Maghzian, O., Zacher, M., Nguyen-Viet, T. A., Bowers, P., Sidorenko, J., Linnér, R. K., et al. (2018). Gene discovery and polygenic prediction from a 1.1-million-person gwas of educational attainment. *Nature Genetics*, 50(8):1112.

[Lei and Fithian, 2018] Lei, L. and Fithian, W. (2018). Adapt: an interactive procedure for multiple testing with side information. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):649–679.

[Lei et al., 2021] Lei, L., Ramdas, A., and Fithian, W. (2021). A general interactive framework for false discovery rate control under structural constraints. *Biometrika*, 108(2):253–267.

[Leiner et al., 2021] Leiner, J., Duan, B., Wasserman, L., and Ramdas, A. (2021). Data blurring: sample splitting a single sample. *arXiv preprint arXiv:2112.11079*.

[Li and Barber, 2019] Li, A. and Barber, R. F. (2019). Multiple testing with the structure-adaptive benjamini–hochberg algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(1):45–74.

[Lichtenstein et al., 2009] Lichtenstein, P., Yip, B. H., Björk, C., Pawitan, Y., Cannon, T. D., Sullivan, P. F., and Hultman, C. M. (2009). Common genetic determinants of schizophrenia and bipolar disorder in swedish families: a population-based study. *The Lancet*, 373(9659):234–239.

[Liu et al., 2010] Liu, J. Z., McRae, A. F., Nyholt, D. R., Medland, S. E., Wray, N. R., Brown, K. M., Investigators, A., Hayward, N. K., Montgomery, G. W., Visscher, P. M., Martin, N. G., and Macgregor, S. (2010). A versatile gene-based test for genome-wide association studies. *American journal of human genetics*, 87(1):139–145.

[Liu et al., 2019] Liu, Y., Chen, S., Li, Z., Morrison, A. C., Boerwinkle, E., and Lin, X. (2019). Acat: a fast and powerful p value combination method for rare-variant analysis in sequencing studies. *The American Journal of Human Genetics*, 104(3):410–421.

[Liu et al., 2018] Liu, Y., Liang, Y., Cicek, A. E., Li, Z., Li, J., Muhle, R. A., Krenzer, M., Mei, Y., Wang, Y., Knoblauch, N., et al. (2018). A statistical framework for mapping risk genes from de novo mutations in whole-genome-sequencing studies. *The American Journal of Human Genetics*, 102(6):1031–1047.

[Liu and Xie, 2020] Liu, Y. and Xie, J. (2020). Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. *Journal of the American Statistical Association*, 115(529):393–402.

[Locke et al., 2015] Locke, A. E., Kahali, B., Berndt, S. I., Justice, A. E., Pers, T. H., Day, F. R., Powell, C., Vedantam, S., Buchkovich, M. L., Yang, J., Croteau-Chonka, D. C., Esko, T., Fall, T., Ferreira, T., Gustafsson, S., Kutalik, Z., Luan, J., Mägi, R., Randall, J. C., Winkler, T. W., Wood, A. R., Workalemahu, T., Faul, J. D., Smith, J. A., Hua Zhao, J., Zhao, W., Chen, J., Fehrmann, R., Hedman, Å. K., Karjalainen, J., Schmidt, E. M., Absher, D., Amin, N., Anderson, D., Beekman, M., Bolton, J. L., Bragg-Gresham, J. L., Buyske, S., Demirkan, A., Deng, G., Ehret, G. B., Feenstra, B., Feitosa, M. F., Fischer, K., Goel, A., Gong, J., Jackson, A. U., Kanoni, S., Kleber, M. E., Kristiansson, K., Lim, U., Lotay, V., Mangino, M., Mateo Leach, I., Medina-Gomez, C., Medland, S. E., Nalls, M. A., Palmer, C. D., Pasko, D., Pechlivanis, S., Peters, M. J., Prokopenko, I., Shungin, D., Stančáková, A., Strawbridge, R. J., Ju Sung, Y., Tanaka, T., Teumer, A., Trompet, S., van der Laan, S. W., van Setten, J., Van Vliet-Ostaptchouk, J. V., Wang, Z., Yengo, L., Zhang, W., Isaacs, A., Albrecht, E., Ärnlöv, J., Arscott, G. M., Attwood, A. P., Bandinelli, S., Barrett, A., Bas, I. N., Bellis, C., Bennett, A. J., Berne, C., Blagieva, R., Blüher, M., Böhringer, S., Bonnycastle, L. L., Böttcher, Y., Boyd, H. A., Bruinenberg, M., Caspersen, I. H., Ida Chen, Y.-D., Clarke, R., Warwick Daw, E., de Craen, A. J. M., Delgado, G., Dimitriou, M., Doney, A. S. F., Eklund, N., Estrada, K., Eury, E., Folkersen, L., Fraser, R. M., Garcia, M. E., Geller, F., Giedraitis, V., Gigante, B., Go, A. S., Golay, A., Goodall, A. H., Gordon, S. D., Gorski, M., Grabe, H.-J., Grallert, H., Grammer, T. B., Gräßler, J., Grönberg, H., Groves, C. J., Gusto, G., Haessler, J., Hall, P., Haller, T., Hallmans, G., Hartman, C. A., Hassinen, M., Hayward, C., Heard-Costa, N. L., Helmer, Q., Hengstenberg, C., Holmen, O., Hottenga, J.-J., James, A. L., Jeff, J. M., Johansson, Å., Jolley, J., Juliusdottir, T., Kinnunen, L., Koenig, W., Koskenvuo, M., Kratzer, W.,

Laitinen, J., Lamina, C., Leander, K., Lee, N. R., Lichtner, P., Lind, L., Lindström, J., Sin Lo, K., Lobbens, S., Lorbeer, R., Lu, Y., Mach, F., Magnusson, P. K. E., Mahajan, A., McArdle, W. L., McLachlan, S., Menni, C., Merger, S., Mihailov, E., Milani, L., Moayyeri, A., Monda, K. L., Morken, M. A., Mulas, A., Müller, G., Müller-Nurasyid, M., Musk, A. W., Nagaraja, R., Nöthen, M. M., Nolte, I. M., Pilz, S., Rayner, N. W., Renstrom, F., Rettig, R., Ried, J. S., Ripke, S., Robertson, N. R., Rose, L. M., Sanna, S., Scharnagl, H., Scholtens, S., Schumacher, F. R., Scott, W. R., Seufferlein, T., Shi, J., Vernon Smith, A., Smolonska, J., Stanton, A. V., Steinthorsdottir, V., Stirrups, K., Stringham, H. M., Sundström, J., Swertz, M. A., Swift, A. J., Syvänen, A.-C., Tan, S.-T., Tayo, B. O., Thorand, B., Thorleifsson, G., Tyrer, J. P., Uh, H.-W., Vandenput, L., Verhulst, F. C., Vermeulen, S. H., Verweij, N., Vonk, J. M., Waite, L. L., Warren, H. R., Waterworth, D., Weedon, M. N., Wilkens, L. R., Willenborg, C., Wilsgaard, T., Wojczynski, M. K., Wong, A., Wright, A. F., Zhang, Q., Study, T. L. C., Brennan, E. P., Choi, M., Dastani, Z., Drong, A. W., Eriksson, P., Franco-Cereceda, A., Gådin, J. R., Gharavi, A. G., Goddard, M. E., Handsaker, R. E., Huang, J., Karpe, F., Kathiresan, S., Keildson, S., Kiryluk, K., Kubo, M., Lee, J.-Y., Liang, L., Lifton, R. P., Ma, B., McCarroll, S. A., McKnight, A. J., Min, J. L., Moffatt, M. F., Montgomery, G. W., Murabito, J. M., Nicholson, G., Nyholt, D. R., Okada, Y., Perry, J. R. B., Dorajoo, R., Reinmaa, E., Salem, R. M., Sandholm, N., Scott, R. A., Stolk, L., Takahashi, A., Tanaka, T., van't Hooft, F. M., Vinkhuyzen, A. A. E., Westra, H.-J., Zheng, W., Zondervan, K. T., Consortium, T. A., Group, T. A.-B. W., Consortium, T. C., Consortium, T. C., GLGC, T., ICBP, T., Investigators, T. M., Consortium, T. M., Consortium, T. M., Consortium, T. P., Consortium, T. R., Consortium, T. G., Consortium, T. I. E., Heath, A. C., Arveiler, D., Bakker, S. J. L., Beilby, J., Bergman, R. N., Blangero, J., Bovet, P., Campbell, H., Caulfield, M. J., Cesana, G., Chakravarti, A., Chasman, D. I., Chines, P. S., Collins, F. S., Crawford, D. C., Adrienne Cupples, L., Cusi, D., Danesh, J., de Faire, U., den Ruijter, H. M., Dominiczak, A. F., Erbel, R., Erdmann, J., Eriksson, J. G., Farrall, M., Felix, S. B., Ferrannini, E., Ferrières, J., Ford, I., Forouhi, N. G., Forrester, T., Franco, O. H., Gansevoort, R. T., Gejman, P. V., Gieger, C., Gottesman, O., Gudnason, V., Gyllensten, U., Hall, A. S., Harris, T. B., Hattersley, A. T., Hicks, A. A., Hindorff, L. A., Hingorani, A. D., Hofman, A., Homuth, G., Kees Hovingh, G., Humphries, S. E., Hunt, S. C., Hyppönen, E., Illig, T., Jacobs, K. B., Jarvelin, M.-R., Jöckel, K.-H., Johansen, B., Jousilahti, P., Wouter Jukema, J., Jula, A. M., Kaprio, J., Kastelein, J. J. P., Keinanen-Kiukaanniemi, S. M., Kiemeney, L. A., Knekt, P., Kooner, J. S., Kooperberg, C., Kovacs, P., Kraja, A. T., Kumari, M., Kuusisto, J., Lakka, T. A., Langenberg, C., Le Marchand, L., Lehtimäki, T., Lyssenko, V., Männistö, S., Marette, A., Matise, T. C., McKenzie, C. A., McKnight, B., Moll, F. L., Morris, A. D., Morris, A. P., Murray, J. C., Nelis, M., Ohlsson, C., Oldehinkel, A. J., Ong, K. K., Madden, P. A. F., Pasterkamp, G., Peden, J. F., Peters, A., Postma, D. S., Pramstaller, P. P., Price, J. F., Qi, L., Raitakari, O. T., Rankinen, T., Rao, D. C., Rice, T. K., Ridker, P. M., Rioux, J. D., Ritchie, M. D., Rudan,

I., Salomaa, V., Samani, N. J., Saramies, J., Sarzynski, M. A., Schunkert, H., Schwarz, P. E. H., Sever, P., Shuldiner, A. R., Sinisalo, J., Stolk, R. P., Strauch, K., Tönjes, A., Trégouët, D.-A., Tremblay, A., Tremoli, E., Virtamo, J., Vohl, M.-C., Völker, U., Waeber, G., Willemsen, G., Witteman, J. C., Carola Zillikens, M., Adair, L. S., Amouyel, P., Asselbergs, F. W., Assimes, T. L., Bochud, M., Boehm, B. O., Boerwinkle, E., Bornstein, S. R., Bottinger, E. P., Bouchard, C., Cauchi, S., Chambers, J. C., Chanock, S. J., Cooper, R. S., de Bakker, P. I. W., Dedoussis, G., Ferrucci, L., Franks, P. W., Froguel, P., Groop, L. C., Haiman, C. A., Hamsten, A., Hui, J., Hunter, D. J., Hveem, K., Kaplan, R. C., Kivimaki, M., Kuh, D., Laakso, M., Liu, Y., Martin, N. G., März, W., Melbye, M., Metspalu, A., Moebus, S., Munroe, P. B., Njølstad, I., Oostra, B. A., Palmer, C. N. A., Pedersen, N. L., Perola, M., Pérusse, L., Peters, U., Power, C., Quertermous, T., Rauramaa, R., Rivadeneira, F., Saaristo, T. E., Saleheen, D., Sattar, N., Schadt, E. E., Schlessinger, D., Eline Slagboom, P., Snieder, H., Spector, T. D., Thorsteinsdottir, U., Stumvoll, M., Tuomilehto, J., Uitterlinden, A., Uusitupa, M., van der Harst, P., Walker, M., Wallaschofski, H., Wareham, N. J., Watkins, H., Weir, D. R., Wichmann, H.-E., Wilson, J. F., Zanen, P., Borecki, I. B., Deloukas, P., Fox, C. S., Heid, I. M., O'Connell, J. R., Strachan, D. P., Stefansson, K., van Duijn, C. M., Abecasis, G. R., Franke, L., Frayling, T. M., McCarthy, M. I., Visscher, P. M., Scherag, A., Willer, C. J., Boehnke, M., Mohlke, K. L., Lindgren, C. M., Beckmann, J. S., Barroso, I., North, K. E., Ingelsson, E., Hirschhorn, J. N., Loos, R. J. F., and Speliotes, E. K. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518:197 EP –.

[MacArthur et al., 2017] MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., Pendlington, Z. M., Welter, D., Burdett, T., Hindorff, L., Flicek, P., Cunningham, F., and Parkinson, H. (2017). The new nhgri-ebi catalog of published genome-wide association studies (gwas catalog). *Nucleic Acids Res*, 45(D1):D896–D901.

[Mahajan et al., 2018] Mahajan, A., Taliun, D., Thurner, M., Robertson, N. R., Torres, J. M., Rayner, N. W., Payne, A. J., Steinthorsdottir, V., Scott, R. A., Grarup, N., Cook, J. P., Schmidt, E. M., Wuttke, M., Sarnowski, C., Mägi, R., Nano, J., Gieger, C., Trompet, S., Lecoeur, C., Preuss, M. H., Prins, B. P., Guo, X., Bielak, L. F., Below, J. E., Bowden, D. W., Chambers, J. C., Kim, Y. J., Ng, M. C. Y., Petty, L. E., Sim, X., Zhang, W., Bennett, A. J., Bork-Jensen, J., Brummett, C. M., Canouil, M., Ec kardt, K.-U., Fischer, K., Kardia, S. L. R., Kronenberg, F., Läll, K., Liu, C.-T., Locke, A. E., Luan, J., Ntalla, I., Nylander, V., Schönherr, S., Schurmann, C., Yengo, L., Bottinger, E. P., Brandslund, I., Christensen, C., Dedoussis, G., Florez, J. C., Ford, I., Franco, O. H., Frayling, T. M., Giedraitis, V., Hackinger, S., Hattersley, A. T., Herder, C., Ikram, M. A., Ingelsson, M., Jørgensen, M. E., Jørgensen, T., Kriebel, J., Kuusisto, J., Ligthart, S., Lindgren, C. M., Linneberg, A., Lyssenko, V., Mamakou, V., Meitinger, T., Mohlke, K. L., Morris, A. D.,

Nadkarni, G., Pankow, J. S., Peters, A., Sattar, N., Stančáková, A., Strauch, K., Taylor, K. D., Thorand, B., Thorleifsson, G., Thorsteinsdottir, U., Tuomilehto, J., Witte, D. R., Dupuis, J., Peyser, P. A., Zeggini, E., Loos, R. J. F., Froguel, P., Ingelsson, E., Lind, L., Groop, L., Laakso, M., Collins, F. S., Jukema, J. W., Palmer, C. N. A., Grallert, H., Metspalu, A., Dehghan, A., Köttgen, A., Abecasis, G. R., Meigs, J. B., Rotter, J. I., Marchini, J., Pedersen, O., Hansen, T., Langenberg, C., Wareham, N. J., Stefansson, K., Gloyn, A. L., Morris, A. P., Boehnke, M., and McCarthy, M. I. (2018). Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nature Genetics*, 50(11):1505–1513.

[Medvedeva et al., 2015] Medvedeva, Y. A., Lennartsson, A., Ehsani, R., Kulakovskiy, I. V., Vorontsov, I. E., Panahandeh, P., Khimulya, G., Kasukawa, T., Drabløs, F., Consortium, F., et al. (2015). Epifactors: a comprehensive database of human epigenetic factors and complexes. *Database*, 2015.

[Mishra and Macgregor, 2015] Mishra, A. and Macgregor, S. (2015). Vegas2: Software for more flexible gene-based testing. *Twin Research and Human Genetics*, 18(1):86–91.

[Neira-Fresneda and Potocki, 2015] Neira-Fresneda, J. and Potocki, L. (2015). Neurodevelopmental disorders associated with abnormal gene dosage: Smith–magenis and potocki–lupski syndromes. *Journal of Pediatric Genetics*, 4(03):159–167.

[Ni et al., 2021] Ni, G., Zeng, J., Revez, J. A., Wang, Y., Zheng, Z., Ge, T., Restuadi, R., Kiewa, J., Nyholt, D. R., Coleman, J. R., et al. (2021). A comparison of ten polygenic score methods for psychiatric disorders applied across multiple cohorts. *Biological psychiatry*, 90(9):611–620.

[Nicolae et al., 2010] Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E., and Cox, N. J. (2010). Trait-associated snps are more likely to be eqtls: Annotation to enhance discovery from gwas. *PLOS Genetics*, 6(4):1–10.

[O'Connor et al., 2019] O'Connor, L. J., Schoech, A. P., Hormozdiari, F., Gazal, S., Patterson, N., and Price, A. L. (2019). Extreme polygenicity of complex traits is explained by negative selection. *The American Journal of Human Genetics*, 105(3):456 – 476.

[Pardiñas et al., 2018] Pardiñas, A. F., Holmans, P., Pocklington, A. J., Escott-Price, V., Ripke, S., Carrera, N., Legge, S. E., Bishop, S., Cameron, D., Hamshere, M. L., et al. (2018). Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nature genetics*, 50(3):381–389.

[R Core Team, 2020] R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[Ripke et al., 2014] Ripke, S., Neale, B. M., Corvin, A., Walters, J. T., Farh, K.-H., Holmans, P. A., Lee, P., Bulik-Sullivan, B., Collier, D. A., Huang, H., et al. (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, 511(7510):421.

[Ruderfer et al., 2014] Ruderfer, D. M., Fanous, A. H., Ripke, S., McQuillin, A., Amdur, R. L., Gejman, P. V., O'Donovan, M. C., Andreassen, O. A., Djurovic, S., Hultman, C. M., et al. (2014). Polygenic dissection of diagnosis and clinical dimensions of bipolar disorder and schizophrenia. *Molecular psychiatry*, 19(9):1017–1024.

[Ruderfer et al., 2018a] Ruderfer, D. M., Ripke, S., McQuillin, A., Boocock, J., Stahl, E. A., Pavlides, J. M. W., Mullins, N., Charney, A. W., Ori, A. P., Loohuis, L. M. O., et al. (2018a). Genomic dissection of bipolar disorder and schizophrenia, including 28 subphenotypes. *Cell*, 173(7):1705–1715.

[Ruderfer et al., 2018b] Ruderfer, D. M., Ripke, S., McQuillin, A., Boocock, J., Stahl, E. A., Pavlides, J. M. W., Mullins, N., Charney, A. W., Ori, A. P., Loohuis, L. M. O., et al. (2018b). Genomic dissection of bipolar disorder and schizophrenia, including 28 subphenotypes. *Cell*, 173(7):1705–1715.

[Samocha et al., 2014] Samocha, K. E., Robinson, E. B., Sanders, S. J., Stevens, C., Sabo, A., McGrath, L. M., Kosmicki, J. A., Rehnström, K., Mallick, S., Kirby, A., et al. (2014). A framework for the interpretation of de novo mutation in human disease. *Nature genetics*, 46(9):944–950.

[Satterstrom et al., 2020] Satterstrom, F. K., Kosmicki, J. A., Wang, J., Breen, M. S., De Rubeis, S., An, J.-Y., Peng, M., Collins, R., Grove, J., Klei, L., et al. (2020). Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. *Cell*, 180(3):568–584.

[Schaid et al., 2018] Schaid, D. J., Chen, W., and Larson, N. B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics*, 19(8):491–504.

[Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014] Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, 511(7510):421–427.

[Scott et al., 2015] Scott, J. G., Kelly, R. C., Smith, M. A., Zhou, P., and Kass, R. E. (2015). False discovery rate regression: An application to neural synchrony detection in primary visual cortex. *Journal of the American Statistical Association*, 110(510):459–471.

[Seranski et al., 1999] Seranski, P., Heiss, N. S., Dhorne-Pollet, S., Radelof, U., Korn, B., Hennig, S., Backes, E., Schmidt, S., Wiemann, S., Schwarz, C. E., et al. (1999).

Transcription mapping in a medulloblastoma breakpoint interval and smith–magenis syndrome candidate region: identification of 53 transcriptional units and new candidate genes. *Genomics*, 56(1):1–11.

[Sey et al., 2020] Sey, N. Y. A., Hu, B., Mah, W., Fauni, H., McAfee, J. C., Rajarajan, P., Brennand, K. J., Akbarian, S., and Won, H. (2020). A computational tool (h-magma) for improved prediction of brain-disorder risk genes by incorporating brain chromatin interaction profiles. *Nature Neuroscience*, 23(4):583–593.

[Sievert, 2020] Sievert, C. (2020). *Interactive Web-Based Data Visualization with R, plotly, and shiny*. Chapman and Hall/CRC.

[Silva and Haggarty, 2020] Silva, M. and Haggarty, S. J. (2020). Tauopathies: Deciphering disease mechanisms to develop effective therapies. *International Journal of Molecular Sciences*, 21(23):8948.

[Stahl et al., 2019] Stahl, E. A., Breen, G., Forstner, A. J., McQuillin, A., Ripke, S., Trubetskoy, V., Mattheisen, M., Wang, Y., Coleman, J. R., Gaspar, H. A., et al. (2019). Genome-wide association study identifies 30 loci associated with bipolar disorder. *Nature genetics*, 51(5):793–803.

[Stefansson et al., 2005] Stefansson, H., Helgason, A., Thorleifsson, G., Steinthorsdottir, V., Masson, G., Barnard, J., Baker, A., Jonasdottir, A., Ingason, A., Gudnadottir, V. G., et al. (2005). A common inversion under selection in europeans. *Nature genetics*, 37(2):129–137.

[Steinberg et al., 2012] Steinberg, K. M., Antonacci, F., Sudmant, P. H., Kidd, J. M., Campbell, C. D., Vives, L., Malig, M., Scheinfeldt, L., Beggs, W., Ibrahim, M., et al. (2012). Structural diversity and african origin of the 17q21. 31 inversion polymorphism. *Nature genetics*, 44(8):872–880.

[Subramanian et al., 2005] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550.

[Supek et al., 2011] Supek, F., Bošnjak, M., Škunca, N., and Šmuc, T. (2011). Revigo summarizes and visualizes long lists of gene ontology terms. *PloS one*, 6(7):e21800.

[The Gene Ontology Consortium, 2018] The Gene Ontology Consortium (2018). The gene ontology resource: 20 years and still going strong. *Nucleic Acids Research*, 47(D1):D330–D338.

[Tian et al., 2021] Tian, J., Chen, X., Katsevich, E., Goeman, J., and Ramdas, A. (2021). Large-scale simultaneous inference under dependence. *arXiv preprint arXiv:2102.11253*.

[Watanabe et al., 2019] Watanabe, K., Stringer, S., Frei, O., UmićevićMirkov, M., de Leeuw, C., Polderman, T. J. C., van der Sluis, S., Andreassen, O. A., Neale, B. M., and Posthuma, D. (2019). A global overview of pleiotropy and genetic architecture in complex traits. *Nature Genetics*, 51(9):1339–1348.

[Watanabe et al., 2017] Watanabe, K., Taskesen, E., Van Bochoven, A., and Posthuma, D. (2017). Functional mapping and annotation of genetic associations with fuma. *Nature communications*, 8(1):1–11.

[Weiner et al., 2017] Weiner, D. J., Wigdor, E. M., Ripke, S., Walters, R. K., Kosmicki, J. A., Grove, J., Samocha, K. E., Goldstein, J. I., Okbay, A., Bybjerg-Grauholm, J., et al. (2017). Polygenic transmission disequilibrium confirms that common and rare variation act additively to create risk for autism spectrum disorders. *Nature genetics*, 49(7):978–985.

[Werling et al., 2018] Werling, D. M., Brand, H., An, J.-Y., Stone, M. R., Zhu, L., Glessner, J. T., Collins, R. L., Dong, S., Layer, R. M., Markenscoff-Papadimitriou, E., et al. (2018). An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. *Nature genetics*, 50(5):727–736.

[Werling et al., 2020a] Werling, D. M., Pochareddy, S., Choi, J., An, J.-Y., Sheppard, B., Peng, M., Li, Z., Dastmalchi, C., Santpere, G., Sousa, A. M., et al. (2020a). Whole-genome and rna sequencing reveal variation and transcriptomic coordination in the developing human prefrontal cortex. *Cell reports*, 31(1):107489.

[Werling et al., 2020b] Werling, D. M., Pochareddy, S., Choi, J., An, J.-Y., Sheppard, B., Peng, M., Li, Z., Dastmalchi, C., Santpere, G., Sousa, A. M., Tebbenkamp, A. T., Kaur, N., Gulden, F. O., Breen, M. S., Liang, L., Gilson, M. C., Zhao, X., Dong, S., Klei, L., Cicek, A. E., Buxbaum, J. D., Adle-Biassette, H., Thomas, J.-L., Aldinger, K. A., O'Day, D. R., Glass, I. A., Zaitlen, N. A., Talkowski, M. E., Roeder, K., State, M. W., Devlin, B., Sanders, S. J., and Sestan, N. (2020b). Whole-genome and rna sequencing reveal variation and transcriptomic coordination in the developing human prefrontal cortex. *Cell Reports*, 31(1):107489.

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

[Wilson, 2019] Wilson, D. J. (2019). The harmonic mean p-value for combining dependent tests. *Proceedings of the National Academy of Sciences*, 116(4):1195–1200.

[Yang, 2010] Yang, J. J. (2010). Distribution of fisher's combination statistic when the tests are dependent. *Journal of Statistical Computation and Simulation*, 80(1):1–12.

[Yang et al., 2016] Yang, J. J., Li, J., Williams, L. K., and Buu, A. (2016). An efficient genome-wide association test for multivariate phenotypes based on the fisher combination function. *BMC bioinformatics*, 17(1):19.

[Yurko, 2020] Yurko, R. (2020). *snpcombineR: Library for Simulating and Combining GWAS SNP Summary Statistics with Rcpp*. R package version 0.2.0.

[Yurko et al., 2020] Yurko, R., G'Sell, M., Roeder, K., and Devlin, B. (2020). A selective inference approach for false discovery rate control using multiomics covariates yields insights into disease risk. *Proceedings of the National Academy of Sciences*.

[Yurko et al., 2021a] Yurko, R., Roeder, K., Devlin, B., and G'Sell, M. (2021a). H-magma, inheriting a shaky statistical foundation, yields excess false positives. *Annals of Human Genetics*, 85(3-4):97–100.

[Yurko et al., 2021b] Yurko, R., Roeder, K., Devlin, B., and G'Sell, M. (2021b). An approach to gene-based testing accounting for dependence of tests among nearby genes. *Briefings in Bioinformatics*.

[Zhang and Horvath, 2005] Zhang, B. and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis a general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, 4(1).

[Zhang and Wu, 2020] Zhang, H. and Wu, Z. (2020). Accurate p-Value Calculation for Generalized Fisher's Combination Tests Under Dependence. *arXiv preprint arXiv:2003.01286*.

[Zhang et al., 2019] Zhang, M. J., Xia, F., and Zou, J. (2019). Fast and covariate-adaptive method amplifies detection power in large-scale multiple hypothesis testing. *Nature communications*, 10(1):1–11.

[Zhu and Stephens, 2018] Zhu, X. and Stephens, M. (2018). Large-scale genome-wide enrichment analyses identify new trait-associated genes and pathways across 31 human phenotypes. *Nature communications*, 9(1):1–14.