Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Aaron M. Holleman Date

Investigating genetic associations using power-optimizing analytic approaches

By

Aaron M. Holleman
Doctor of Philosophy
Epidemiology

_____
Jennifer G. Mulle, PhD, MHS
Advisor

_____
Viola Vaccarino, MD, PhD
Advisor

_____
David J. Cutler, PhD
Committee Member

_____
Michael P. Epstein, PhD, MS
Committee Member

_____
Timothy L. Lash, DSc, MPH
Committee Member

Accepted:

_____
Kimberly Jacob Arriola, PhD, MPH
Dean of the James T. Laney School of Graduate Studies

_____
Date

Investigating genetic associations using power-optimizing analytic approaches

By

Aaron M. Holleman
MPH, Johns Hopkins University, 2012
BA, Rice University, 2007

Advisors: Jennifer G. Mulle, PhD, MHS
Viola Vaccarino, MD, PhD

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Epidemiology
2021

Abstract

Investigating genetic associations using power-optimizing analytic approaches

By Aaron M. Holleman

Substantial progress has been made toward identifying genetic factors that contribute to many complex phenotypes, yet there remains an incomplete understanding of the genetics underlying such traits. This is partly due to insufficient study power. Increases in sample size will yield greater power, but may be challenging to accomplish given study constraints; and, at times, other approaches may be preferable for achieving power gains. In these situations, power-optimizing analytic techniques can be particularly useful. For this dissertation, we applied such techniques to more powerfully investigate genetic associations with phenotypes of interest.

In Aim 1, we employed polygenic risk score (PRS) methods to optimally examine the contribution of common genetic variation to atrioventricular septal defects (AVSD) in individuals with Down syndrome (DS). Using one of the largest available AVSD in DS case-control datasets, we constructed PRS based on large sets of common variants for each individual, using effect estimates from the largest available GWAS of congenital heart defects as weights. PRS were associated with AVSD with odds ratios ranging from 1.2 to 1.3 per standard deviation increase in PRS, suggesting at least a small contribution by common variants collectively to DS-associated AVSD.

In Aim 2, we evaluated the Gene Association with Multiple Traits (GAMuT) method as a potentially powerful approach to identify genes harboring common variants that influence psychiatric phenotypes. When applied to simulated data, GAMuT's multivariate modeling of Beck Depression Inventory (BDI) items demonstrated greater power for identifying common variant associations than univariate methods analyzing a summary BDI score. Application of GAMuT to Grady Trauma Project data identified common variant associations with the PTSD Symptom Scale and the BDI.

In Aim 3, we investigated associations of rare regulatory variants with gene expression, for genes with schizophrenia-associated expression levels. We employed a modified version of a burden method developed to increase power for investigating rare variant associations with gene expression, and consistently observed U-shaped patterns of estimated association whereby rare regulatory allele burden was increased at both low and high expression levels.

By applying certain power-optimizing analytic approaches, we have generated novel findings suggestive of genetic associations with phenotypes of interest.

Investigating genetic associations using power-optimizing analytic approaches

By

Aaron M. Holleman
MPH, Johns Hopkins University, 2012
BA, Rice University, 2007

Advisors: Jennifer G. Mulle, PhD, MHS
Viola Vaccarino, MD, PhD

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Epidemiology
2021

Acknowledgements

There are many individuals who have contributed to my success on this PhD journey. Foremost among them is my wife, Diana. She knows better than anyone how challenging these past several years were for me at various times, and she has been my most important source of support, providing love and encouragement to help me through the toughest times. This PhD program has been a journey for her as well. Her long hours at work have ensured comfortable living for our family, and enabled our daughter, Laila, to benefit from attending one of the best daycares in Atlanta. Diana has taken care of Laila on many weekends to allow me extra time for working on my dissertation, and has been patient and supportive as I've taken longer than expected to finish my PhD. I'm not sure that I would have persevered to the end of this program without her continued support and encouragement. Diana: Thank you. I couldn't ask for a better partner with whom to share life's journey.

I'm grateful to my advisors, Drs. Jennifer Mulle and Michael Epstein, for spending countless hours over the past several years meeting with me to provide constructive feedback on my work and facilitating my development as a scientist. I thank Jen for her leadership as my dissertation chair, her guidance as my advisor from day one, and her time spent helping me grow in my knowledge of genetics and genetic epidemiology concepts. I am also grateful to her for connecting me with numerous faculty and staff from Emory's Department of Human Genetics, including my co-advisor Mike Epstein.

I have benefited greatly from having Mike as an advisor. By working closely with him, I have gained a wealth of knowledge and skills in the areas of statistical genetics and genetic epidemiology. His guidance, teaching, and encouragement have been instrumental in my professional development. I am grateful for the time he has devoted to helping me grow, and feel fortunate to have had him as a close mentor along with Jen.

In addition to Jen and Mike, I thank my other committee members for their contributions. Dr. David Cutler has taught me so much about computational and quantitative genetics. I have peppered him with questions on many occasions, and he has always been happy to take time to help me learn, responding to my inquiries with impressive speed. Drs. Timothy Lash and Viola Vaccarino have provided valuable feedback on my projects, particularly with respect to the 'epidemiology' in my genetic epidemiology research. I thank Viola for serving as my dissertation co-chair, providing important leadership and guidance in this capacity, and thank Tim for his feedback, guidance, and encouragement.

I am grateful to Drs. Rich Johnston, Alex Kotlar, Alaine Broadaway, and Robert Arthur for their help with various components of this dissertation. More generally, I am thankful for the faculty, staff, and students from Emory's Departments of Epidemiology and Human Genetics, my interactions with whom have made me more knowledgeable and capable as a scientist.

I would not be at this point without the ever-present guidance, support and encouragement of my family. In particular, I thank my parents for all they have taught me and continue to teach me, for believing in me, and for providing encouragement and wisdom during the most challenging times. I thank my brother for his life-long support and ability to help me take things a bit less seriously when needed. In addition, I thank Diana's parents, who many times have stayed with us for extended durations to help take care of Laila, providing me with much needed additional time to work on my dissertation. And of course, I am grateful for my daughter Laila, whose big, beautiful smile can brighten the toughest of days. I am truly blessed to be part of such a wonderful family.

Finally, I am grateful to the many study participants who provided the data that I analyzed for this dissertation research. My hope is that that this work will be of benefit to them in at least in some small way.

Table of Contents

List of Tables

Chapter 2

Chapter 3

Chapter 4

List of Figures

Chapter 2

Chapter 3

Chapter 4

Chapter 1:

Introduction and Background

INTRODUCTION

Although genetic epidemiology investigations over the past decade have made substantial progress in identifying genetic factors that contribute to the heritability (the proportion of phenotypic variation due to genetic variation) of many complex diseases and phenotypes,[1] methodological and other practical limitations have prevented researchers from attaining a more complete understanding of the genetics underlying nearly all such traits. An important limitation, though by no means the only one, is inadequate power for identifying genetic associations of interest.

Power for detecting genetic signals is frequently considered in relation to sample size, whereby the sample size must be sufficiently large to have a given amount of power (often 80%) for detecting the expected association. The importance of sample size for identifying genetic associations has been demonstrated over time as genetic epidemiology studies with larger sample numbers have yielded greater quantities of robust genetic signals, with larger sample numbers particularly important for identifying genetic variants with smaller effect sizes and/or rarer minor allele frequencies (MAF).[1] Thus, an obvious means of improving the ability to detect genetic associations is to vastly increase sample numbers (Note: This assumes that systematic biases including selection bias, information bias, and confounding are sufficiently negligible to enable increasingly accurate effect size estimation with increases in sample size). This is precisely the approach currently being taken by multiple international consortia. For instance, the Schizophrenia Working Group of the Psychiatric Genomics Consortium has accumulated ~67,000 schizophrenia/schizoaffective disorder cases and ~94,000 controls,[2] with plans to assemble at least 150,000 cases.[3]

However, for a given study, large increases in sample size may not be feasible given resource limitations (financial or otherwise); or may be feasible but only over a several-year or longer time frame. For such a study, alternative analytical approaches may be employed to increase power for investigating an association of interest without the need to increase sample

size. Furthermore, in certain situations as will be described below, the gain in power for detecting genetic associations that results from using a more appropriate statistical modeling approach may exceed the additional power gained from increasing sample size.

Overarching goal and specific aims

This dissertation applied recently developed and cutting-edge genetic epidemiologic analysis methods to data from three study samples to more powerfully examine genetic associations of interest. In **Aim 1**, we employed polygenic risk score (PRS) methods to optimally examine the collective contribution of common (minor allele frequency [MAF] > 1%) genetic variation to atrioventricular septal defect (AVSD) in individuals with Down syndrome (DS). In **Aim 2**, we first used simulated datasets to examine the power of a cutting-edge multivariate analysis approach called the Gene Association with Multiple Traits (GAMuT) test for identifying common variant associations with multivariate questionnaire data, and then used this multivariate approach to examine genetic associations with two multivariate psychiatric phenotypes using real data. In **Aim 3**, we investigated associations of rare (MAF < 1%) regulatory variants with gene expression for a set of genes with schizophrenia-associated expression, gaining statistical power by using a modified version of a recently developed rare variant burden approach.

BACKGROUND

Aim 1

The primary objective of Aim 1 was to examine the collective contribution of common genetic variants to AVSD among individuals with DS. AVSD is a type of congenital heart defect with a substantially increased prevalence among those with DS as compared with the general population. Although the presence of a third copy of chromosome 21 appears to be a key factor

driving the increased risk for AVSD among those with DS, it is likely that additional variation throughout the genome, both common and rare, also contributes to elevated risk. Common genetic variant contributions to DS-associated AVSD were previously investigated in a genome-wide association study (GWAS) including 210 AVSD cases and 242 controls with structurally normal hearts, all of whom had DS.[4] This prior GWAS analyzed 606,195 autosomal single nucleotide polymorphisms (SNPs) genotyped using an Affymetrix SNP array, and did not identify any robust signals of association between common variants and DS-associated AVSD, despite adequate power to detect odds ratios (ORs) > 2; suggesting that large-effect common variants may not play an important role in DS-associated AVSD. However, a contribution by moderate- to low-effect common variants has yet to be sufficiently examined, primarily due to sample size limitations.

Due to challenges inherent in recruiting participants with a condition as rare as DS-associated AVSD, currently available case-control datasets are limited to hundreds of samples. However, sample sizes required to identify common variants with moderate to small effect sizes are on the order of thousands to tens of thousands of participants (assuming a 20% prevalence of AVSD in the DS population, and Bonferroni correction for 606,195 SNP tests).[5] Thus, the standard GWAS approach for identifying common variants associated with disease is severely underpowered for detecting anything but large effect common variants in current DS-associated AVSD study samples.

For the purposes of Aim 1, we were able to use a final analytic sample of 487 participants with DS, including 245 AVSD cases and 242 controls (including the majority of the 210 cases and 242 controls analyzed in the prior GWAS as just described). To examine the extent to which common genetic variants might play a role in increased risk for AVSD among those with DS, we decided to employ polygenic risk score (PRS) methods. Although standard GWAS tests each genetic variant individually, yielding hundreds of thousands to millions of separate tests across the entire genome, the PRS approach aggregates the contributions of

SNPs across the genome into a single score, named the polygenic risk score (also called a polygenic score or genetic risk score), which is then tested for association with the phenotype of interest. A PRS approach can offer increased power over a GWAS of common variants in at least two ways: 1) thousands of SNPs with small individual effects may be statistically undetectable by GWAS in relatively smaller sample sizes, whereas the cumulative effect of these thousands of SNPs, investigated using PRS, is more likely to be sufficiently large to enable detection; and 2) the huge quantity of SNP tests performed by GWAS requires a far stricter multiple testing correction as compared with PRS analyses that involve far fewer tests. A limitation of using PRS methods to examine the role of common genetic variation in DS-associated AVSD is that, since variants are considered collectively rather than individually, a PRS analysis will not be informative regarding the specific individual genetic variants that contribute to AVSD. However, an observed association between PRS and DS-associated AVSD might suggest that common variants in general do contribute to AVSD among those with DS, helping to advance our understanding of the genetic architecture of DS-associated AVSD; and it may inform the design of future studies to better understand the individual common variants that are most important.

Aim 2

The objectives of Aim 2 were twofold: 1) Use simulated data to evaluate the power of a previously developed multivariate genetic association method for identifying associations between common genetic variants and multivariate psychiatric phenotypes; and 2) apply this method to real data to examine genetic associations with two psychiatric phenotypes. Psychiatric disorders, such as depression or anxiety disorders, are characterized by the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) as behavior or psychological 'syndromes',[6] with a syndrome being defined as a set of correlated signs or symptoms, the cooccurrence of which may be said to constitute a particular disease or disorder (Note: The

DSM-5 actually uses the term mental disorder rather than psychiatric disorder).[7] As syndromes, clinicians diagnose psychiatric disorders based on specific criteria (such as those set forth by the DSM-5 or the International Classification of Diseases [ICD-10]), which often involves examining the number of syndrome-related symptoms an individual is presenting with, in combination with additional criteria such as whether the symptoms cause significant distress or impairment and are not better accounted for by a different condition.[6] Psychiatric disorders and mental health status are also frequently assessed using questionnaires (often self-reported). For instance, the PTSD Symptom Scale (PSS) is a 17-item questionnaire designed to assess and diagnose post-traumatic stress disorder (PTSD). Each item of the PSS corresponds to a PTSD symptom and is rated on an ordinal scale from 0 to 3, with higher scores indicating greater symptom frequency/intensity.[8,9] PTSD severity is then determined by totaling the scores for the 17 items (scores can range from 0 to 51), and a PTSD diagnosis can be made based on the reported presence of at least a minimum number of symptoms in three PSS subcategories. Similarly, the Beck Depression Inventory (BDI-II) is a 21-item questionnaire for assessing major depressive disorder (MDD), with each item scored on an ordinal scale from 0 to 3 (higher scores indicating more severe symptoms), and with the ability to sum across items and generate a cumulative score reflecting depression severity.[10]

These psychiatric disorder diagnoses (presence or absence of the disorder) or cumulative scores from questionnaires like the PSS and BDI are frequently analyzed as univariate outcomes in psychiatric research. However, as syndromes consisting of multiple correlated yet discrete symptoms, psychiatric disorders would perhaps more accurately be analyzed as multivariate phenotypes. For genetic epidemiology research seeking to identify genetic factors that may underlie various psychiatric phenotypes, analyzing a univariate measure that is a summary across multivariate symptom data has the potential to decrease power for detecting genetic associations. Specifically with respect to genetic analyses of multivariate ordinal data, it has been shown that a univariate summary measure will fail to

function as an adequate summary of the multivariate data under circumstances that include the genetic factor not having identical effects on all of the multivariate items.[11] A scenario such as this is entirely plausible for most psychiatric disorders, given the variety of symptoms that collectively constitute each syndrome. As one example, a genetic factor may have different effects on BDI items that are more somatic-related (e.g., items assessing sleep or appetite) as compared with items that are more mood-related (e.g., those assessing feelings of sadness or guilt). In turn, if the univariate measure does not adequately summarize the multivariate data, then the choice of using the univariate measure for analysis has the potential to decrease power for identifying genetic associations with the psychiatric phenotype.[11] In such a circumstance, use of analytic techniques that allow for proper modeling of the multivariate phenotype data may provide increased power for detecting genetic signals as compared with univariate analysis approaches.

Multivariate analysis methods that enable modeling of ordinal data which is commonly generated using questionnaires are currently suboptimal.[12] We therefore evaluated the effectiveness of a novel multivariate analysis approach in identifying genetic associations with multivariate psychiatric questionnaire data. The method we examined, named the Gene Association with Multiple Traits (GAMuT) test, was developed previously as a means of testing for rare variant pleiotropy.[13] We repurposed GAMuT and evaluated it as a potentially powerful method for identifying common variant associations with multivariate psychiatric phenotypes, specifically those assessed with ordinal questionnaire items, with special attention paid to scenarios in which the genetic effect differs across the various phenotypes assessed by the questionnaire items. In the first part of Aim 2, we simulated genetic and BDI data under a variety of scenarios (e.g., varying the causal SNP; varying the proportion of BDI items affected by the causal SNP), and used these simulated datasets to evaluate GAMuT with respect to Type I error control and power for identifying SNP effects. We also applied two univariate analysis approaches, kernel machine regression (KMR) and standard linear regression, to the simulated

datasets to compare univariate analysis of the cumulative BDI score (summed across all items) with GAMuT. In the second part of Aim 2, we applied GAMuT and the two univariate analysis methods to real genetic and phenotypic data accumulated through the Grady Trauma Project (GTP). For these applied analyses, we specifically examined common variant associations with psychiatric phenotypes interrogated by the PSS and BDI questionnaires.

Aim 3

In Aim 3, we had the objective of examining associations of rare regulatory genetic variants with gene expression levels, for a set of genes enriched for having expression levels previously identified to be associated with schizophrenia (SZ). In recent years, numerous genes have been identified to have expression levels that are associated with SZ.[14-16] The role of various factors in regulating the expression of these genes is presently not well understood, and merits investigation to advance understanding of upstream elements that may affect SZ risk through modification of gene expression. Considering SZ's high heritability (estimates are as high as ~80% for SZ heritability)[17,18] and the established role for both common and rare genetic variants in influencing SZ risk,[19-23] it is particularly important to explore the potential impact of genetic variants on the expression of these SZ-linked genes.

A role for common variants in modifying the expression of genes across the genome is well-supported through a collection of many studies.[24] However, the impact of rare variants on expression, which has potential to be greater than that of common variants, has been much less well studied, both for genes with SZ-associated expression and for genes in general. Rare variants located within regulatory sequences, which may be especially involved in the regulation of gene expression, are particularly understudied. The lag in studying rare variants, as compared with common variants, is largely due to technological and power (related to sample size) limitations: identification of rare variants, particularly small rare variants like single

nucleotide variants (SNVs), requires DNA sequencing of large sample sets, which has historically been quite costly.

A limited number of studies have investigated associations of rare (defined by some of the studies as MAF < 0.05) non-coding variants with gene expression, yielding findings consistent with a role for rare regulatory variants in modifying gene expression.[25-30] Investigations specifically focused on the contribution of rare regulatory variants to gene expression variation for genes with SZ-associated expression have been even more limited, as well as underpowered.[14] Aim 3 sought to help fill this knowledge gap, by using a power-optimizing approach to investigate the contribution of rare regulatory variants to gene expression for genes with SZ-associated expression levels.

For this aim, we analyzed a set of 725 samples that had undergone both targeted DNA sequencing and genome-wide RNA sequencing. The targeted DNA sequencing had been performed for 1) exonic and regulatory sequence (including 2,000 bases upstream of the first exon) for 64 genes previously identified as having SZ-associated expression levels, and 2) exonic sequence for 172 genes or gene regions with prior evidence for involvement in SZ due to being located within a SZ-associated large CNV interval. As approaches traditionally employed for studying common variant associations (e.g., expression quantitative trait loci analyses, genome-wide association studies) tend to suffer from low power when analyzing rare variants, we employed a modified version of an alternative analytic method that was recently developed specifically for examining associations of rare genetic variation with gene expression.[27] This approach first assigns rare alleles within a gene region to ordered expression bins (ranging from low to high) for the gene, and then gains substantial power by aggregating rare alleles for each expression bin across all genes being considered and examining association of rare allele burden with gene expression level for this aggregated dataset. We employed this basic approach to more powerfully examine associations of rare regulatory variants with gene expression, specifically considering rare variants located within gene promoter regions, as well

as those in the 5' untranslated region (5'UTR) or the 3' untranslated region (3'UTR). These analyses have the potential to be informative with respect to the role of rare regulatory variation in altering expression levels specifically for genes with SZ-associated expression, as well as for genes more broadly.

The remainder of this dissertation is organized as follows. Chapter 2 details work related to Aim 1, which involved examining the common variant contribution to DS-associated AVSD, through application of PRS methods. Chapter 3 presents the Aim 2 work, which explored the GAMuT method as a potentially powerful alternative approach for identifying common variant associations with multivariate psychiatric phenotypes. Chapter 4 describes the research conducted for Aim 3, which examined associations of rare regulatory variants with expression for genes with SZ-associated expression levels. Chapter 5 serves as the concluding chapter to this dissertation, and provides a summary of the main dissertation findings, as well as an interpretation of these findings in relation to existing knowledge and a discussion of potential future studies which can build on these results.

Chapter 2:

Employing polygenic risk score methods to examine the contribution of common genetic variants to atrioventricular septal defects in infants with Down syndrome

Research described in **Chapter 2** has been published:

Trevino CE*, Holleman AM*, Corbitt H, Maslen CL, Rosser TC, Cutler DJ, Johnston HR, Rambo-Martin BL, Oberoi J, Dooley KJ, Capone GT, Reeves RH, Cordell HJ, Keavney BD, Agopian AJ, Goldmuntz E, Gruber PJ, O'Brien JE Jr, Bittel DC, Wadhwa L, Cua CL, Moskowitz IP, Mulle JG, Epstein MP, Sherman SL, Zwick ME. **Identifying genetic factors that contribute to the increased risk of congenital heart defects in infants with Down syndrome.** Sci Rep. 2020 Oct 22;10(1):18051. doi: 10.1038/s41598-020-74650-4. Erratum in: Sci Rep. 2021 Jul 20;11(1):15164. PMID: 33093519; PMCID: PMC7582922. [Published by Springer Nature]

*Joint first authors

Note: With a few necessary exceptions, work described in the above publication that I did not perform is not included in Chapter 2 of this dissertation. The exceptions include the original subject recruitment and data collection, and WGS mapping and variant calling, which are directly related to my subsequent work. Text in the above publication for which I was not the original author is not included in Chapter 2 of this dissertation.

Abstract

Background: Individuals with Down syndrome (DS), which is also referred to as trisomy 21, have a substantially increased risk for congenital heart defects as compared with the general population. This increased risk is particularly pronounced for a subtype of congenital heart defect called atrioventricular septal defect (AVSD): among those with DS, AVSD is over 300 times more prevalent compared with the general population and over 2,000 times more prevalent compared with non-syndromic individuals. While it is evident that the extra copy of chromosome 21 plays an important role in this increased AVSD risk, it is also likely that additional genetic factors throughout the genome are involved. Prior studies of the role of common genetic variation in DS-associated AVSD have not identified common variants with large effect sizes (e.g., odds ratios > 2.0), and have been underpowered to investigate a potential contribution to risk by small- to moderate-effect common variants, with the result that the extent to which common variants may play a role in DS-associated AVSD remains unknown.

Methods: We examined the contribution by common variants to DS-associated AVSD using a case-control dataset including 245 AVSD cases and 242 controls, all with DS. Samples had undergone either whole genome sequencing or array-based genotyping followed by genome-wide imputation. Rather than using a standard genome-wide association study approach to examine associations between individual variants and AVSD, which would suffer from inadequate power given our small sample size, we used polygenic risk score (PRS) methods to examine the collective contribution of common genetic variants to DS-associated AVSD. We performed primary analyses that examined the genome-wide common variant polygenic contribution to DS-associated AVSD, and secondary analyses specifically examining the additional polygenic contribution made by variants on chromosome 21.

Results: Constructing PRS using weights based on the largest genome-wide association study of congenital heart defects available (2,594 cases and 5,159 controls; all without DS), we found genome-wide PRS to be associated with AVSD with odds ratios ranging from 1.2 to 1.3 per standard deviation increase in PRS, with PRS explaining approximately 1% of variance in outcome on the liability scale. Results from the secondary analyses suggested that common variants on chromosome 21 contribute negligibly to polygenic risk.

Conclusions: Results from the genome-wide PRS analyses suggest at least a small common variant polygenic contribution to DS-associated AVSD. Supplemental power analyses indicated that a PRS explaining 1% of variance on the liability scale is near the maximum contribution detectable given our small training GWAS size, and that if a larger polygenic contribution exists, it should be detectable with larger training GWAS sizes. Thus, future studies using larger training datasets are needed to more accurately quantify the collective contribution of common genetic variants to DS-associated AVSD.

INTRODUCTION

Individuals with Down syndrome (DS), which is also referred to as trisomy 21 given the presence of a third copy of chromosome 21, have a substantially elevated risk for congenital heart defects (CHD) as compared with the general population. Specifically, it has been estimated that in 44% of DS live births the infant has a CHD,[31] as compared with an approximately 1% CHD prevalence for births in the general population.[32,33]

Of the different types of CHDs affecting individuals with DS, atrioventricular septal defects (AVSD) are the most common.[31] AVSD is a severe heart defect that involves the presence of holes between the heart's chambers and valves, which may be malformed, resulting in problems with blood flow and oxygenation.[34] AVSD usually requires surgical repair early in life, with those undergoing surgery still subject to potential lifelong complications.[34] Among infants with DS and a CHD, approximately 45% have an AVSD; meaning that in about 20% of DS live births the infant has AVSD.[31] In comparison, AVSD occurs with a prevalence of about 1 in 1,859 births in the general population;[35] it occurs in 0.83/10,000 live births when excluding individuals with chromosomal abnormalities (including DS) or single-gene disorders.[36] Thus, the prevalence of AVSD among those with DS is > 300 times that of the general population and > 2,000 times that of the non-syndromic population.

The dramatically increased prevalence of AVSD among those with DS strongly indicates involvement of the extra copy of chromosome 21 in DS-associated AVSD. Considering that 80% of infants with DS do not have AVSD, it is also likely that additional genetic factors throughout the genome contribute to AVSD among those with DS, potentially including both common (minor allele frequency [MAF] > 1%) and rare (MAF < 1%) genetic variants. The role played by such other genetic factors, however, is largely unknown. The effort to more completely understand the genetic etiology of DS-associated AVSD, including the contribution by factors other than the trisomy 21, is important as it may inform the development of strategies to reduce the burden of AVSD among those with DS. Furthermore, it has potential to provide insights into

the biological underpinnings of CHD more generally, with the possibility of yielding benefits that extend beyond the community of individuals with DS and into the wider population.

For this study, we focused on attempting to advance understanding regarding the potential role of common variants in DS-associated AVSD. Multiple studies have been conducted to examine the contribution by common variants to DS-associated AVSD, the largest such study being a genome-wide association study (GWAS) involving 210 complete AVSD cases and 242 controls with structurally normal hearts, all with DS. These multiple studies have not identified any robust common variant associations (based on exceeding the genome-wide significance threshold), despite being sufficiently powered to detect common variants with large effect sizes.[4,37,38] These results suggest that large-effect common variants (e.g., those with odds ratios > 2.0) do not play a considerable role in elevating AVSD risk among those with DS. However, a potential role for small- to moderate-effect common variants has yet to be adequately examined.

The limited success in identifying individual common variants associated with AVSD among those with DS may be due to small sample sizes that have been underpowered for discovery. If this is the case, then greatly increasing sample sizes should yield power increases sufficient to allow robust detection of common variant associations, including for common variants with smaller effects on AVSD. For a condition with the rarity of DS-associated AVSD, however, attaining a sufficiently powered sample size would take substantial resources and time.

An alternative approach for investigating the contribution of common variants to DS-associated AVSD, which involves little additional resources and time, is to use the methods of polygenic risk score (PRS) analysis to examine whether and to what extent common variants may be collectively contributing to AVSD in DS. PRS methods are designed to quantify and analyze potential polygenicity for disorders and traits, with polygenicity referring to the contribution by many genes or genetic variants to a phenotype. PRS methods typically focus on

common genetic variants (usually SNPs) to examine the extent to which large numbers of common variants, with exceedingly small individual effects, collectively contribute to the disorder or trait. Such methods have already indicated that complex disorders including schizophrenia and cardiovascular disease possess sizable polygenic components consisting of hundreds to thousands of common variants.[19,39]

Although enabling examination of polygenicity and thus informing on the genetic architecture of a given disorder or trait, PRS methods also act as an alternative approach for assessing the role of common genetic variation in a phenotype that can overcome certain limitations of the standard genome-wide association study (GWAS) approach for examining SNPs. Specifically, a PRS approach can offer increased power over a GWAS of common variants, which is accomplished in at least two ways: 1) thousands of SNPs with small individual effects may be statistically undetectable by GWAS in relatively smaller sample sizes, whereas the cumulative effect of these thousands of SNPs, investigated using PRS, is more likely to be sufficiently large to enable detection; and 2) the huge quantity of SNP tests performed by GWAS requires a far stricter multiple testing correction as compared with PRS analyses, which involve far fewer tests.

PRS methods therefore offer a means of gaining information about the potential role of common genetic variation in DS-associated AVSD, which standard individual-variant-level approaches like GWAS are not able to provide due to current sample size limitations. By applying PRS methods to our DS dataset, which includes 245 AVSD cases and 242 normal heart controls (sample size after quality control), we can more powerfully examine whether common variants collectively are associated with AVSD. We can also assess whether, contrary to original hypotheses, the genetic architecture of DS-associated AVSD might exhibit complexity similar to that observed for numerous other disorders that research suggests involve individually small contributions by up to thousands of common variants. Such a polygenic component for

AVSD, if it exists, might be particularly relevant when combined with the dysregulation of many genes due to trisomy 21.

A limitation of the PRS approach is that it does not pinpoint individual common variants as associated with the phenotype. However, an observed association between PRS and DS-associated AVSD may suggest that common variants in general do contribute to AVSD among those with DS, helping to advance our understanding of the genetic architecture of DS-associated AVSD and informing the design of future studies to better understand which particular common variants, proximal to which genes, may be making the greatest contributions. With this in mind, we applied the PRS approach to our DS dataset to examine whether and to what extent common variants may be contributing to DS-associated AVSD. We performed two main sets of PRS analyses: 1) PRS analyses to investigate the genome-wide common variant contribution to DS-associated AVSD; and 2) PRS analyses to examine the additional common variant contribution specifically due to the trisomic chromosome 21.

METHODS

Overview of the PRS method

PRS analysis requires a “target” dataset that includes the individuals for which PRS will be constructed and analyzed, and a “discovery” or “training” dataset that typically takes the form of summary results from a GWAS, with the GWAS variant effect estimates used as weights for the PRS construction in the target dataset. In our study, the target dataset was our case-control sample of individuals with DS and with or without AVSD, and we employed two discovery datasets generated from independent GWAS of CHD. Following PRS construction, analyses are performed to examine the association between PRS and target phenotype, with the results informing whether common variants associated with the discovery GWAS phenotype may be contributing to the target phenotype.

Target dataset sources

We generated our target dataset by merging samples from two DS AVSD case-control datasets, one for which participants had undergone whole genome sequencing (WGS), while participants in the other dataset had undergone SNP array genotyping (with subsequent genome-wide imputation by us).

Whole genome sequencing dataset

The WGS dataset originally included 169 AVSD cases and 39 normal-heart controls, all with Down syndrome. Participants were recruited through two projects: the Down Syndrome Heart Project (DSHP) and the Pediatric Cardiac Genomic Consortium (PCGC). Details regarding participant ascertainment and assessment for the DSHP[4,37] and the PCGC[40,41] have been described previously.

Briefly, the DSHP recruited subjects through sites in multiple locations across the United States. Participants were required to have Down syndrome, based on diagnosis of full trisomy 21. Cases (n = 122) were defined as individuals with a complete, balanced AVSD, as determined through review of echocardiogram or surgical reports. Controls (n = 39) were defined as individuals with a structurally normal heart, patent foramen ovale, or patent ductus arteriosus, with determination primarily based on echocardiogram and in some cases medical records. We note that all controls in our WGS dataset were ascertained through the DSHP.

The PCGC recruited subjects with various heart defects from multiple locations in the United States and the United Kingdom. This study determined AVSD in the same manner as the DSHP — based on echocardiogram or surgical reports documenting a complete, balanced AVSD. From this study, we only used data from the subset of individuals with AVSD and trisomy 21 (n = 47).

Genome-wide imputation dataset

Our dataset of samples with genome-wide imputed genotypes originated from a set of 459 individuals with DS, including 211 AVSD cases and 248 controls, all of whom had been assayed using the Affymetrix Genome-Wide Human SNP 6.0 genotype array. These 459 samples included the 210 cases and 242 controls analyzed in the prior genome-wide association study (GWAS) of DS-associated AVSD.[4] These samples were also ascertained through the DSHP, and thus recruitment and determination of AVSD case or control status were identical to that described above for the DSHP WGS samples. As described in detail in a later section, we ultimately performed genome-wide genotype imputation for these samples.

For ease of reporting, we refer to cases as "DS+AVSD cases" and controls as "DS+NH controls" throughout this manuscript. We also assert that all recruitment of participants and collection of data and biological samples accomplished through the DSHP and PCGC were performed in accordance with experimental protocols approved by the Institutional Review Boards of the participating sites. This includes ascertainment of informed consent from the parent or guardian of each minor participant prior to completing assessments and obtaining biological samples.

Target dataset preparation

Primary analyses

Whole genome sequencing dataset

Paired-ended WGS was performed on 169 DS+AVSD cases and 39 DS+NH controls by Hudson Alpha to a target depth of 30x. We mapped raw sequence reads to the most recent human genome build (hg38) using PEMapper.[42] Average read depth was 30.2 (with standard deviation [SD] of 4.1), indicating good coverage. We then called variant sites using PECaller,[42] and used the online tool Bystro to annotate the variants.[43] In all, 12,302,231 variants were identified across the 169 cases and 39 controls. Mean ratio of transitions to transversions

considering all samples was 2.05 (SD = 0.007), squarely within the range for mean transition/transversion that is expected for a set of high-quality WGS calls for a human dataset. We note that, due to trisomy, sequence data for chromosome 21 were variant called and quality controlled (QC'd) separately from the rest of the genome.

Our sample QC involved removing samples identified as outliers for certain variant-based metrics reported by Bystro, including samples with theta < 3 SD below the mean, transition/transversion ratio < 3 sd below the mean, and heterozygosity/homozygosity ratio > 4 sd above the mean. We also excluded samples with excess missing genotypes (missing > 1% of genotypes), those with discordant reported and inferred sex as determined using PLINK1.9's 'check-sex' flag, and one individual from each pair of samples identified as related based on having a proportion of alleles shared identical by descent (IBD) > 0.1875.[44,45] This combination of sample QC steps resulted in the identification and exclusion of 16 low quality WGS samples. We ultimately removed one other sample in the process of additionally preparing this WGS dataset for certain analyses distinct from the PRS analyses described in this chapter. Variant QC of the WGS dataset involved removing variants missing for > 10% of samples, and those significantly deviating from Hardy-Weinberg equilibrium (HWE) expectation based on a HWE exact test p-value < 10^{-12} when considering cases and controls combined.

We applied principal component analysis (PCA) to identify and remove sample outliers with respect to ancestry. Using PLINK1.9, we first limited our data to a set of sufficiently common (MAF > 0.05) and independent variants (pruned to have linkage disequilibrium [LD] r^2 < 0.2 with nearby common variants), strictly for the purpose of PCA. Then, through three rounds of PCA, we identified 16 additional samples meriting exclusion due to being PC outliers.

Following these sample and variant QC steps, our WGS dataset contained 175 samples (148 cases and 27 controls) and 12,279,101 variants. At this point, we applied additional variant filters to more closely match the variant QC procedures that were applied to the imputed dataset (described directly below). We removed variants with MAF < 0.01, those missing for > 2% of

samples, and indels (insertions and deletions), leaving a WGS dataset with 175 samples and 4,173,676 autosomal SNPs (excluding chromosome 21).

Genome-wide imputation dataset

Affymetrix SNP array genotype data were available for 211 DS+AVSD cases and 248 DS+NH controls. Using PLINK1.9 and R (version 3.4.1),[46] we applied standard GWAS QC procedures, excluding subjects for sex discordance, outlier heterozygosity rates (+/- 3 SD from the mean), missing > 3% of genotypes, and one subject from each pair with proportion of alleles shared IBD > 0.1875. Variants were excluded if they were missing for > 5% of samples, had minor allele frequency (MAF) < 0.01, yielded a HWE mid-p-value < 0.00001 (among controls), or showed significantly different rates of missingness in cases versus controls ($p < 0.00001$). We then used PCA to identify and remove any population outliers, which involved identifying and removing non-European samples using the HapMap3[47] dataset as a population reference (we identified ancestral outliers based on the Anderson et al. 2010 protocol[48]). All together, these QC steps yielded a dataset with 207 DS+AVSD cases and 234 DS+NH controls, and 612,125 autosomal single nucleotide polymorphisms (SNP), excluding chromosome 21.

For these samples, we then performed genotype imputation using the Michigan Imputation Server.[49] Prior to imputation, all alleles were aligned to the (+) strand, and we used a program[50] written by the McCarthy Group to check our dataset against the Haplotype Reference Consortium (HRC) panel and ensure that our data were properly configured for imputation using the HRC panel. We then submitted the DS dataset to the Michigan Imputation Server, for imputation based on the HRC panel (version r1-1 2016),[51] which includes 32,470 samples predominantly of European ancestry.

The post-imputation files included 38,596,402 autosomal variants (all SNPs). Mean correlation between true and imputed genotypes for the ~600,000 genotyped SNPs was 0.990, suggesting high quality imputation. Considering all post-imputation variants, those with MAF ≥

0.05 (5,349,403 variants) had mean imputation $r^2 = 0.971$, those with $0.01 \leq MAF < 0.05$ (2,300,344 variants) had mean $r^2 = 0.882$, and those with $MAF < 0.01$ (30,946,655 variants) had mean $r^2 = 0.180$. This indicates good imputation quality for variants with common MAF. We decided to drop variants with $MAF < 0.01$, those missing for more than 2% of samples, and those with imputation $r^2 < 0.80$; and we set to missing genotypes with a maximum imputed genotype probability < 0.80.

We then applied standard GWAS QC to the imputed dataset. We dropped one sample with an outlying heterozygosity rate (> 3 SDs below the mean). No samples were dropped for excess missing genotypes (all had < 1% missingness). Following removal of the single sample, we again excluded variants missing for > 2% of individuals and those with MAF < 0.01, and also dropped variants with HWE mid-p-value < 0.00001 and those with significant differences in missing genotype rate between cases and controls ($p < 0.05$). We also removed variants with A/T, T/A, C/G, and G/C alleles which can be difficult to match between datasets due to strand ambiguity. This was done in preparation for merging this imputed dataset with unique WGS samples, to create a larger sample for the PRS analyses. This left a dataset with 440 samples (206 DS+AVSD cases, 234 DS+NH controls) and 5,079,537 autosomal SNPs.

Merging WGS and imputed samples

Coordinates for the WGS dataset were based on human genome build 38 (hg38), while those for the imputed dataset were based on human genome build 19 (hg19). Before merging the datasets, we used the University of California Santa Cruz (UCSC) Genome Browser LiftOver[52] tool to convert the WGS data coordinates from hg38 to hg19, and also modified dbSNP Reference numbers (rsIDs) for each variant as needed using an external file based on HRC panel variants containing hg19 rsIDs and coordinates. We chose to convert the WGS data to hg19 rather than converting the imputed data to hg38 as a matter of convenience, given the PRS training files we used had hg19 coordinates.

As one additional step before merging the WGS and imputed datasets, we compared allele frequencies for SNPs in each dataset to identify any instances in which allele frequency for a SNP in one dataset differed considerably from its allele frequency in the other dataset, which could indicate genotyping error for the variant. We identified and removed 77 SNPs with allele frequencies that differed by at least 0.20 between the WGS and imputed datasets.

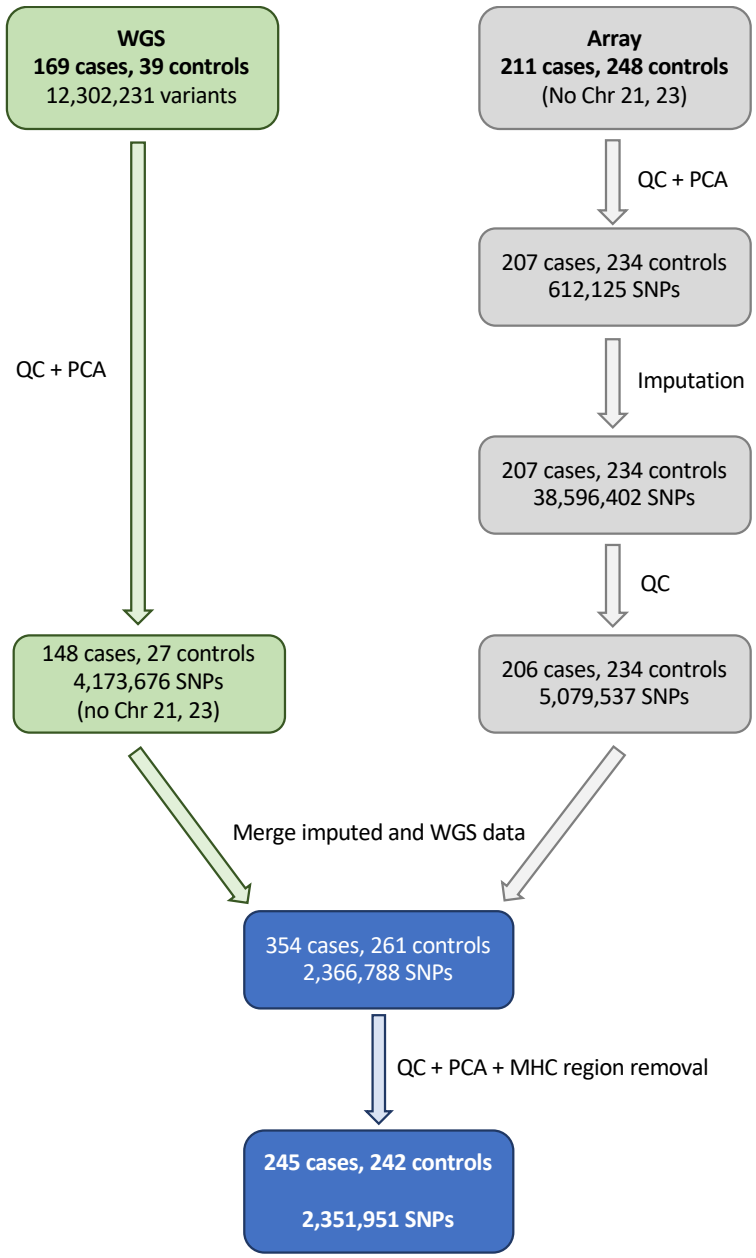
We then merged the WGS and imputed datasets on rsID, position, and alleles (using PLINK1.9), yielding a single dataset with 615 samples and 2,366,788 SNPs. For all 615 samples missingness was < 1%. An IBD check identified 90 sample duplicates and 1 sample pair with a sibling or child/parent relation. Each of these related pairs involved a WGS sample and an imputed sample (i.e., the duplicates were the result of each sample being represented in both the imputed and WGS datasets; this amount of sample overlap was expected). For these samples, we kept the data from the WGS dataset as it appeared to be of slightly better quality overall, and we dropped the imputed duplicates (Note: We ultimately performed sensitivity analyses using a dataset that kept the imputed duplicates rather than the WGS samples, and obtained highly similar PRS results). No additional variant QC filters were needed — all SNPs had missingness ≤ 2% among all samples and ≤ 3% among both cases and controls, all had MAF approximately ≥ 1% (we applied stricter MAF filters during PRS construction), and no SNPs required dropping for HWE violation. Thus, this intermediate data set included 524 samples (263 cases, 261 controls) and 2,366,788 autosomal SNPs.

We next performed PCA, first anchoring our dataset in the HapMap3 dataset and constructing PCs to identify and remove DS samples with PC values outside of the HapMap3 CEPH/Utah (CEU) cluster (in order to match the European ancestry of the discovery datasets). We then removed the HapMap samples and performed further outlier removal based only on the DS samples. We constructed PCs for just the DS samples, and removed samples with values > 3 SD from the mean for PC1 or PC2 (which explained most of the genetic variation in the sample). We then reconstructed PCs for the remaining samples and again identified 3 SD

outliers for removal, repeating this PCA process until all substantial outliers had been identified and removed. This PCA approach identified 37 sample outliers for removal.

As a final step in preparing the DS target dataset for PRS analysis, we removed the extended major histocompatibility complex region (chromosome 6, ~25000000-34000000, human genome build 19), which is a region of extended high linkage disequilibrium that can overly influence PRS results. Our final data set included 487 samples (245 DS+AVSD cases, 242 DS+NH controls) and 2,351,951 autosomal SNPs (excluding chromosome 21). The multiple steps involved in generating this final data set for the primary PRS analyses are presented as a flowchart in **Figure 2.1**.

Figure 2.1. Flowchart showing the multiple steps involved in generating the final data set for the primary PRS analyses.



Secondary analyses

Our secondary PRS analyses examined the contribution by alleles on the trisomic chromosome 21 to a polygenic component for DS-associated AVSD. To do this, we compared PRS results based on polygenic scores generated using all autosomes (including chromosome 21) to PRS results based on scores using all autosomes except for chromosome 21.

We analyzed the same set of target samples as for the primary analyses (245 DS+AVSD cases, 242 DS+NH controls). We were able to do this because, though imputed data were not available for chromosome 21 (due to the complexities of imputing trisomic genotypes), all 329 imputed samples did have Affymetrix Genome-Wide Human SNP 6.0 array genotypes for chromosome 21. Furthermore, all 158 WGS samples had sequencing data for chromosome 21. For all target samples analyzed in the primary analyses, we therefore obtained SNP-array-level data for the trisomic chromosome 21

Given that trisomic data cannot be represented by the PLINK1.9 binary format, we handled these chromosome 21 data separately from the other chromosomes. Before merging chromosome 21 data for these WGS and array samples, we applied certain QC filters. None of the 158 WGS samples nor the 329 array samples had an excess of missing genotypes for chromosome 21 (all had approximately 5% or less missingness). For variant QC, we excluded SNPs missing for > 5% of samples, as well as SNPs with A/T, T/A, C/G, and G/C alleles that can be difficult to match between datasets due to strand ambiguity. We also removed SNPs with substantially different allele frequencies between the WGS and array datasets (we determined that a frequency difference of ≥ 0.125 was an appropriate threshold for these chromosome 21 datasets). Post-merger, we removed SNPs with excess missingness specifically among cases or controls (missing for > 3% of cases or > 3% of controls), and we also excluded SNPs that were monoallelic in the full sample. These steps yielded a merged chromosome 21 dataset with 487 samples and 3,984 SNPs.

We then took the dataset used for the primary analyses (487 samples and 2,351,951 autosomal SNPs, excluding chromosome 21), and limited it to SNPs on the Affymetrix Genome-Wide Human SNP 6.0 array, leaving 389,544 SNPs. This was done since the chromosome 21 data were also necessarily limited to the array SNPs. We used these SNP-array-level genotype data, both with and without the chromosome 21 data, in order to perform the secondary PRS analyses.

Analytic approach

We have grouped our PRS analyses into primary and secondary analyses. The primary analyses had the goal of examining the genome-wide polygenic contribution to DS-associated AVSD, while the secondary analyses had the goal of estimating the additional polygenic contribution specifically due to the trisomic chromosome 21. These primary and secondary PRS analyses used slightly different target datasets (as described above) and slightly different processes for generating and analyzing the PRS (as described below), but employed the same discovery datasets for weighting alleles in the PRS.

Discovery data used to define weights for the PRS

For discovery datasets, there were no GWAS of AVSD or other congenital heart defects (CHD) among individuals with DS that were independent of our target dataset nor were there any GWAS specifically for non-syndromic AVSD. Thus, we used results from two of the largest available independent GWAS of mixed CHD, diagnosed among those without DS who were ancestrally matched to our target samples.

The first discovery dataset was a GWAS of 2,594 cases with a mixture of CHD diagnoses (see **Table 2.1**) and 5,159 population-based controls, all of European ancestry. Genotyping was performed using the Illumina Human660W-Quad array for cases and the Illumina 1.2M chip for controls. The GWAS results included summary statistics for 501,899

autosomal SNPs. GWAS of particular diagnostic CHD subsets of this dataset have been published previously.[53,54]

Table 2.1. First discovery dataset: diagnoses for 2,594 mixed CHD cases. For a more complete list of included CHD diagnosis, see [53,54]

CHD diagnosis	Number (%) of samples
Tetralogy of Fallot	835 (32.2)
Left-sided malformations	387 (14.9)
Ostium secundum atrial septal defect	340 (13.1)
Transposition of the great arteries	207 (8.0)
Ventricular septal defect	191 (7.4)
Conotruncal malformations	151 (5.8)
Double outlet right ventricle	96 (3.7)
AVSD (partial and complete)	73 (2.8)
Other CHD*	314 (12.1)

The second discovery dataset was a GWAS of 406 mixed CHD cases (**Table 2.2**) and 2,976 pediatric controls, all recruited from the same hospital and self-reporting as non-Hispanic Caucasian.[55] Samples were genotyped with Illumina arrays (550 v1/v3, 610, or 2.5M chip), and genome-wide imputation was then carried out using the 1000 Genomes Project data as a reference. The GWAS results included summary statistics for 4,612,359 autosomal SNPs, all of which had imputation $r^2 > 0.80$.

Table 2.2. Second discovery dataset: diagnoses for 406 mixed CHD cases[55]

CHD diagnosis	Number (%) of samples
Tetralogy of Fallot	134 (33.0)
Ventricular septal defect	109 (26.8)
D-transposition of the great arteries	80 (19.7)
Double outlet right ventricle	25 (6.2)
Isolated aortic arch anomalies	22 (5.4)
Truncus arteriosus	19 (4.7)
Other CHD	17 (4.2)

We used each of these discovery datasets separately as training data for the PRS analyses. We also meta-analyzed the summary results from these two GWAS using Genome-Wide Association Meta-Analysis (GWAMA) software,[56] and used the resulting estimates as training data.

Generating PRS for the primary analyses

For the primary PRS analyses, PRSice-2 (version 2.1.6)[57] was used to generate PRS for each sample in the target dataset. Before PRS construction, PRSice performs clumping on the discovery dataset to obtain a set of independent SNPs for scoring (clumping parameters: 500-kilobase window, r^2 threshold 0.10). The clumped SNPs are then used to generate PRS, which are calculated as

$$PRS_j = \sum_i \frac{\beta_i \times EA_{ij}}{N_j}$$

where the subscript *i* denotes a specific SNP contributing to the PRS, the subscript *j* denotes a particular individual in the target dataset, *β* is the estimated effect from the discovery GWAS (e.g., the natural logarithm of the odds ratio), *EA* is the number of effective alleles possessed by

the target individual (0,1 or 2 for a disomic chromosome; the effective allele is the same as the allele for which an increase corresponds to β from the discovery GWAS), and N is the total number of alleles considered for scoring. To facilitate interpretation of results, we applied an option in PRSice to standardize the PRS. We constructed multiple PRS for each target individual using different subsets of the set of clumped SNPs, with subsets determined by applying different p-value thresholds to the discovery GWAS results (e.g., PRS may be constructed using SNPs with discovery p-value < $1x10^{-6}$, < 0.05, < 1).

With respect to PRS analyses, the sample sizes for our two discovery GWAS were rather small, both individually and combined. GWAS effect estimates based on smaller sample sizes are more subject to random error, and will frequently be more likely to miss the mark in terms of capturing true effect sizes as compared with estimates produced using similar, rigorous genetic epidemiologic methods that are based on larger sample sizes. Specifically, assuming negligible bias (i.e., systematic error), increases in sample size will increase the probability of obtaining an effect estimate close to the truth. SNP effect estimates that are less precise and accurate due to being derived from smaller GWAS will, in turn, result in PRS that perform more poorly in capturing the true polygenicity of a phenotype. This has been demonstrated by PRS analyses of schizophrenia performed by the Psychiatric Genomics Consortium (PGC): Using the same schizophrenia case-control target dataset, larger discovery datasets yielded greater maximum Nagelkerke's r^2 values (variance in schizophrenia case/control status explained by the PRS), with a discovery GWAS of 2,615 schizophrenia cases and 3,338 controls producing an r^2 of 3% and a discovery GWAS of 32,838 schizophrenia cases and 44,357 controls yielding an r^2 of 18.4%.[19] Considering this limitation of smaller discovery datasets along with the relatively small sample sizes used to generate our discovery GWAS datasets, we decided to perform PRS analyses by first applying minor allele frequency (MAF) filters ranging from 0.10 to 0.40 to the discovery datasets. Our rationale for this approach was that SNPs with higher MAFs are likely to have more accurately estimated GWAS effect sizes, with poor estimation of lower

MAF SNPs particularly likely for GWAS with smaller sample sizes. Thus, for our smaller discovery GWAS datasets, only using SNPs with MAF > 0.25, for instance, may yield more discriminating PRS than using all discovery SNPs with MAF > 0.10. Assuming a polygenic component is present, applying a filter such as MAF > 0.25 will result in PRS that almost certainly will not capture the full extent of the polygenic contribution, since it is expected that common variants with lower MAF would also contribute; however, if effects for the less common variants are poorly estimated, then such a filter may reduce the contribution of noise to the PRS and result in a better assessment of polygenicity than would otherwise be achieved. Therefore, for each discovery dataset, we applied MAF filters in addition to the aforementioned filtering based on discovery GWAS p-value thresholds, resulting in separate PRS construction and analyses for each unique combination of these two filtering parameters.

Generating PRS for the secondary analyses

For the secondary PRS analyses, which involved analyses both with and without the trisomic chromosome 21 data, we constructed PRS using PLINK1.9. The PLINK1.9 binary, which is the file format that we used in conjunction with PRSice for the primary PRS analyses, is not able to represent trisomic genotype data. However, we were able to modify the chromosome 21 genotype data to fit the PLINK1.9 dosage file format, which can be used in conjunction with PLINK's allelic scoring flag to generate PRS. This involved dividing each allele count by 3 and thereby converting allele counts of 0, 1, 2 and 3 to values of 0, 1/3, 2/3 and 1 (interpreted by PLINK as dosages ranging from 0 to 1). We then used this chromosome 21 dosage format file in combination with the clumped training data (clumped using PRSice) to generate PRS, which were generated by PLINK as a simple sum score (a sum of the products of SNP weight times transformed allele count for each scoring SNP). Finally, we multiplied each outputted PRS by 3, yielding PRS that accurately reflected allele counts of 0, 1, 2 and 3 for the trisomic chromosome 21.

Separately, we used PLINK1.9 to construct PRS for the remaining autosomes. Given that these remaining autosomes were diploid, we were able to use the standard PLINK1.9 binary in combination with the allelic scoring flag to generate PRS. To be consistent with the chromosome 21 PRS, we used an option to generate these PRS as sum scores. For the analyses including chromosome 21, we then summed the chromosome 21 PRS and the PRS for the remaining autosomes for each target individual, yielding a PRS based on alleles from all autosomes combined. The analyses excluding chromosome 21 only utilized the PRS based on all autosomes minus chromosome 21. As for the primary PRS analyses, we standardized the final PRS, and generated multiple PRS for each target individual based on different discovery GWAS p-value and MAF thresholds.

Testing association of PRS with DS+AVSD

We used logistic regression to test associations of PRS with the outcome; this was performed by PRSice for the primary analyses and within R for the secondary analyses. We included sex, platform (WGS vs. imputed), and the top five principal components of ancestry as covariates in the analyses. Tests were two-tailed. Given the multiple testing involved in these PRS analyses (394 tests for different combinations of MAF filter, p-value threshold, and discovery and target datasets, considering the primary and secondary PRS analyses together), we employed the P-value Adjusted for Correlated Tests (P_{ACT})[58] method to generate p-values corrected for multiple correlated tests.

RESULTS

Primary analyses

Over a range of MAF filters and discovery GWAS p-value thresholds for constructing PRS, the analyses using the GWAS of 2,594 mixed CHD cases and 5,159 controls as the

discovery dataset (501,899 autosomal SNPs) tended to yield maximum odds ratios (ORs) of 1.2 to 1.3 for association of PRS with AVSD among those with DS, meaning that a 1 standard deviation increase in PRS was associated with a 20-30% greater odds of having AVSD in the DS target sample (**Figure 2.2**). Corresponding Nagelkerke's r^2 values ranged from 0.75-1.25% (calculated as Nagelkerke's r^2 for the model with PRS and covariates minus Nagelkerke's r^2 for the model with only covariates), with p-values that were non-significant following adjustment for multiple correlated tests (adjusted p-values > 0.15; unadjusted p-values approximately 0.01-0.09). These maximum results were most evident at higher MAF filters (i.e., MAF ≥ 0.30, ≥ 0.35, ≥ 0.40) and discovery GWAS p-value thresholds between 0.001 and 0.3. **Figure 2.3** and **Table 2.3** present results when PRS were constructed using SNPs with MAF ≥ 0.35, which are representative of the maximum PRS results achieved when using this particular discovery dataset.

PRS results when using the GWAS of 406 CHD cases and 2,976 pediatric controls as the discovery dataset (4,612,359 autosomal SNPs) exhibited a different pattern than when using the GWAS of 2,594 mixed CHD cases and 5,159 controls as training data. Across various MAF filters and p-value thresholds, ORs tended to hover near the null and on both sides of the null, indicating that these PRS were minimally associated with AVSD (**Figure 2.4**). A few results were stronger, with ORs in the 1.2 to 1.3 range (adjusted p-values > 0.15); these results occurred when using MAF filters of ≥ 0.10 and ≥ 0.15 in combination with the smallest discovery GWAS p-value thresholds for selecting scoring SNPs.

In addition, we performed a meta-analysis of the two GWAS datasets, yielding a single discovery dataset with association estimates for 4,684,854 autosomal SNPs, of which 429,336 SNPs had estimates based on both studies (meta-analysis sample size of 3,000 CHD cases and 8,135 controls), while the remainder had estimates based on just one of the two studies. In constructing PRS based on this meta-analysis discovery dataset, we applied an inverse variance weighting approach such that SNP association estimates based on a larger sample

size (e.g., two studies) were weighted more heavily. Using the meta-analysis dataset in this manner produced results which, as might be expected, were a mixture of the PRS results obtained when using each discovery GWAS dataset separately (**Figure 2.5**). In general, maximum ORs for association of AVSD in DS with PRS and corresponding Nagelkerke's r^2 values were slightly attenuated compared with results when using the GWAS of 2,594 mixed CHD cases and 5,159 controls as the discovery dataset.

We also performed sensitivity analyses using the GWAS of 2,594 mixed CHD cases and 5,159 controls as the discovery dataset and using a target dataset that had excluded the AVSD cases obtained from the PCGC study, leaving a set of 217 cases and 242 controls who had all been recruited through the DSHP study. We did this in order to examine the potential influence of including PCGC participants among the cases but not among the controls (recall that our PCGC subjects were all AVSD cases). These sensitivity analyses yielded results that were effectively identical to those displayed in **Figure 2.2**, which were obtained from analyses of our full set of 487 samples.

Figure 2.2. PRS results using discovery GWAS of 2,594 mixed CHD cases and 5,159 controls and various MAF thresholds. MAF thresholds were applied to the discovery GWAS; SNPs with MAF below the threshold were excluded from PRS construction. Top row: Each plot displays odds ratio per standard deviation in PRS and the corresponding 95% confidence interval (y-axis) for PRS constructed based on particular discovery GWAS p-value thresholds (x-axis). P_{adj} are adjusted p-values (corrected for multiple correlated tests). 95% CIs correspond to unadjusted p-values. Bottom row: Each plot displays Nagelkerke's r^2 (y-axis) for PRS constructed based on particular discovery GWAS p-value thresholds (x-axis). Numbers above each r^2 bar are the number of SNPs used to construct PRS at that particular p-value threshold and MAF filter combination.

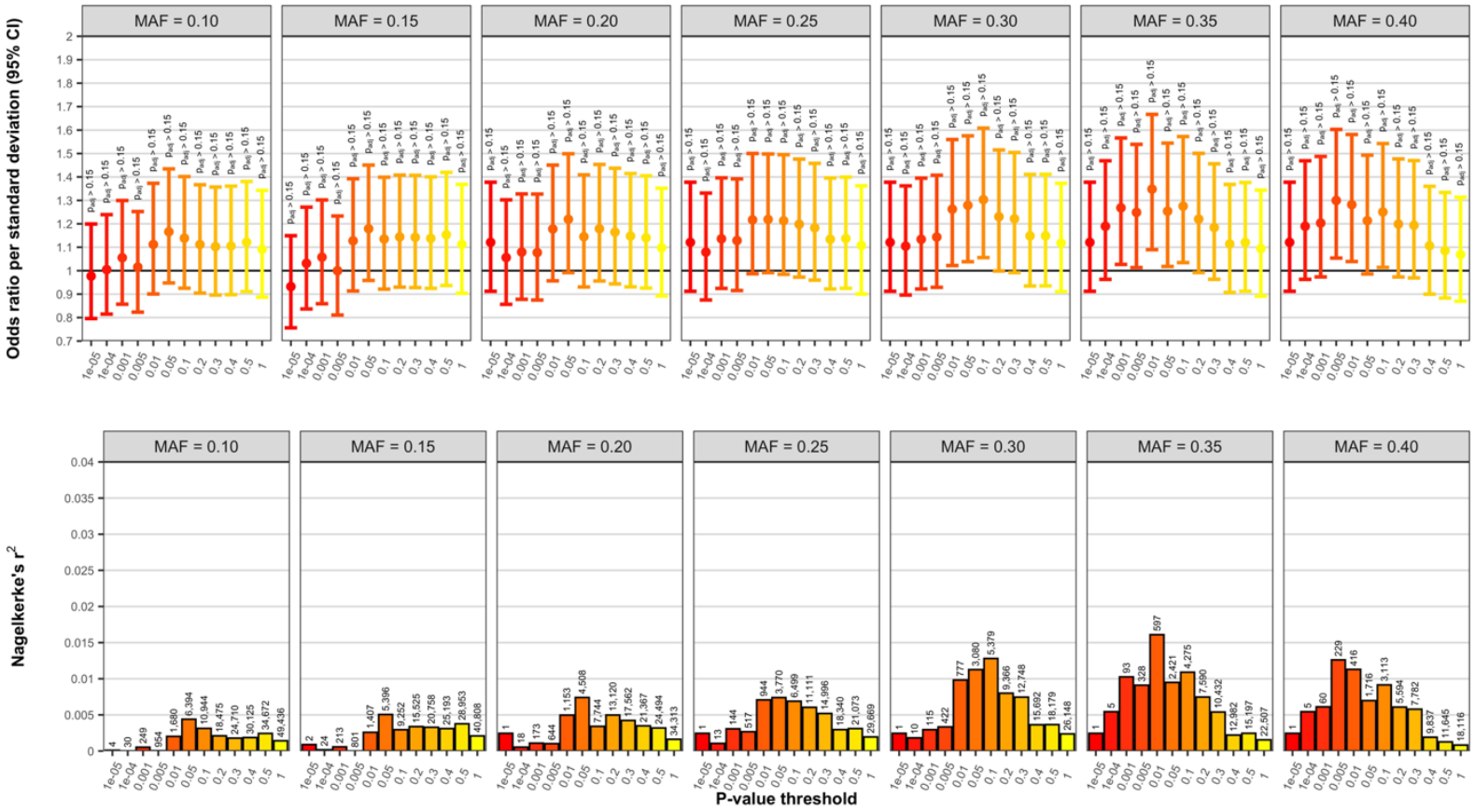


Figure 2.3. PRS results using discovery GWAS of 2,594 mixed CHD cases and 5,159 controls and SNPs with MAF ≥ 0.35. Plot shows odds ratio per standard deviation increase in PRS, with corresponding 95% confidence interval (CI). 'P-value threshold' indicates that SNPs with discovery GWAS p-values below the threshold were used for PRS construction. P_{adj} is the p-value after correction for multiple correlated tests.

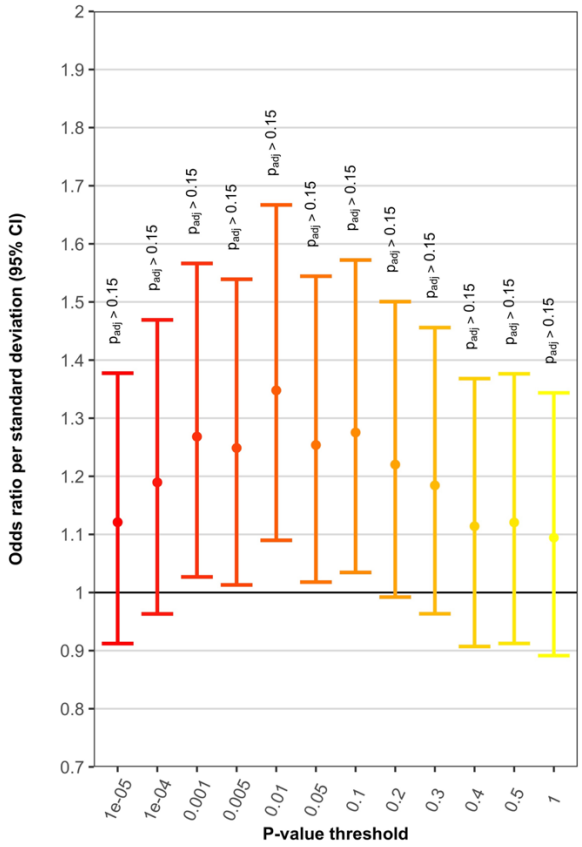


Table 2.3. PRS results using discovery GWAS of 2,594 mixed CHD cases and 5,159 controls and SNPs with MAF ≥ 0.35. 'Threshold' indicates that SNPs with discovery GWAS p-values below the threshold were used for PRS construction, and 'No. SNP' is the corresponding number of SNPs used for scoring. OR: Odds ratio per standard deviation increase in PRS, CI: Confidence interval (corresponding to uncorrected p-value), Nag. r^2: Nagelkerke's r^2, P_{unadj}: Uncorrected p-value, P_{adj}: P-value corrected for multiple correlated tests.

Threshold	No. SNP	OR	95% CI	Nag. r^2	P_{unadj}	P_{adj}
1e-05	1	1.12	0.91-1.38	0.24%	0.278	> 0.15
1e-04	5	1.19	0.96-1.47	0.54%	0.107	> 0.15
0.001	93	1.27	1.03-1.57	1.03%	0.027	> 0.15
0.005	328	1.25	1.01-1.54	0.91%	0.037	> 0.15
0.01	597	1.35	1.09-1.67	1.61%	0.006	> 0.15
0.05	2,421	1.25	1.02-1.54	0.95%	0.033	> 0.15
0.1	4,275	1.28	1.03-1.57	1.09%	0.023	> 0.15
0.2	7,590	1.22	0.99-1.50	0.75%	0.059	> 0.15
0.3	10,432	1.18	0.96-1.46	0.54%	0.108	> 0.15
0.4	12,982	1.11	0.91-1.37	0.22%	0.303	> 0.15
0.5	15,197	1.12	0.91-1.38	0.25%	0.278	> 0.15
1	22,507	1.09	0.89-1.34	0.15%	0.389	> 0.15

Figure 2.4. PRS results using discovery GWAS of 406 mixed CHD cases and 2,976 controls and various MAF thresholds. MAF thresholds were applied to the discovery GWAS; SNPs with MAF below the threshold were excluded from PRS construction. Top row: Each plot displays odds ratio per standard deviation in PRS and the corresponding 95% confidence interval (y-axis) for PRS constructed based on particular discovery GWAS p-value thresholds (x-axis). P_{adj} are adjusted p-values (corrected for multiple correlated tests). 95% CIs correspond to unadjusted p-values. Bottom row: Each plot displays Nagelkerke's r^2 (y-axis) for PRS constructed based on particular discovery GWAS p-value thresholds (x-axis). Numbers above each r^2 bar are the number of SNPs used to construct PRS at that particular p-value threshold and MAF filter combination.

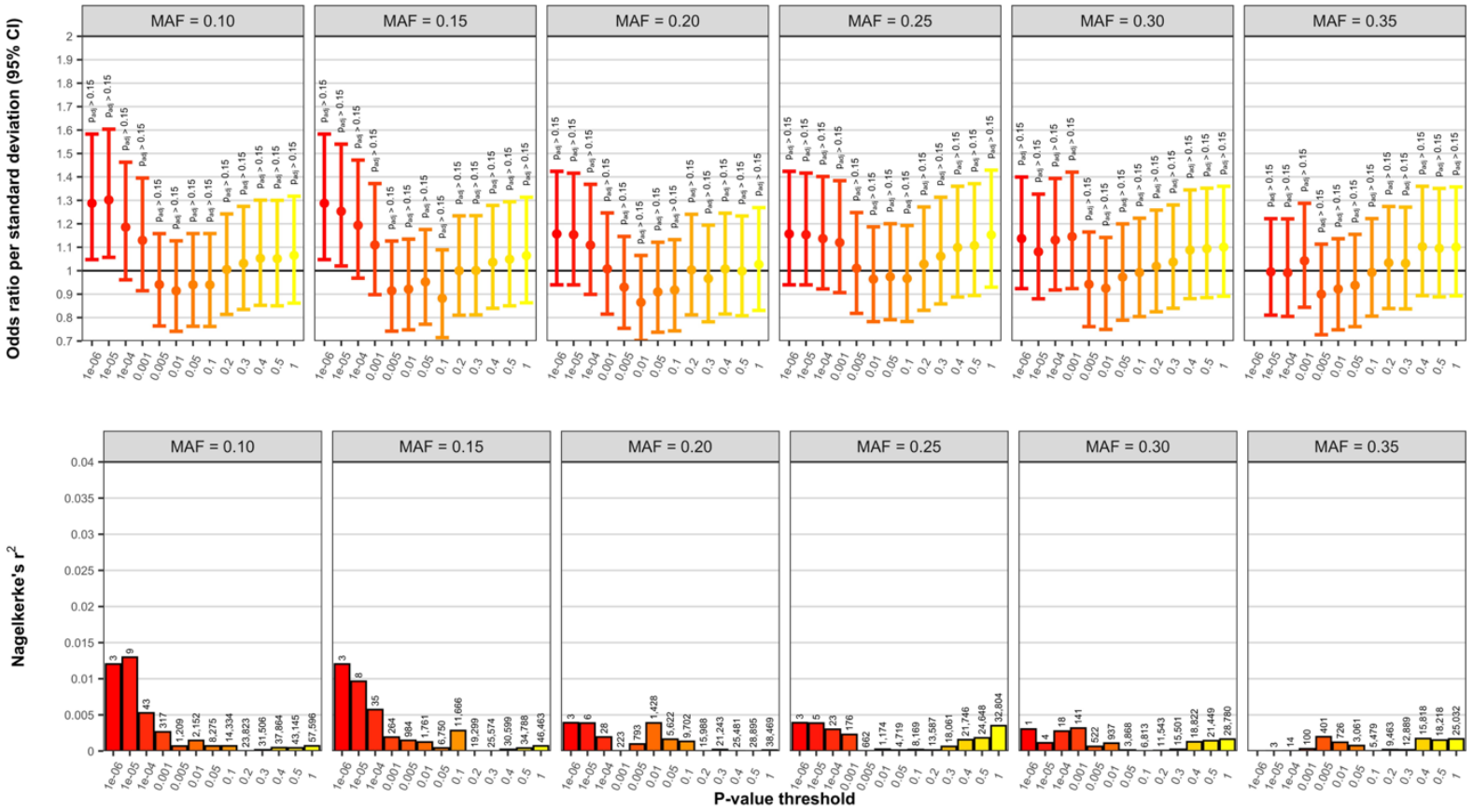
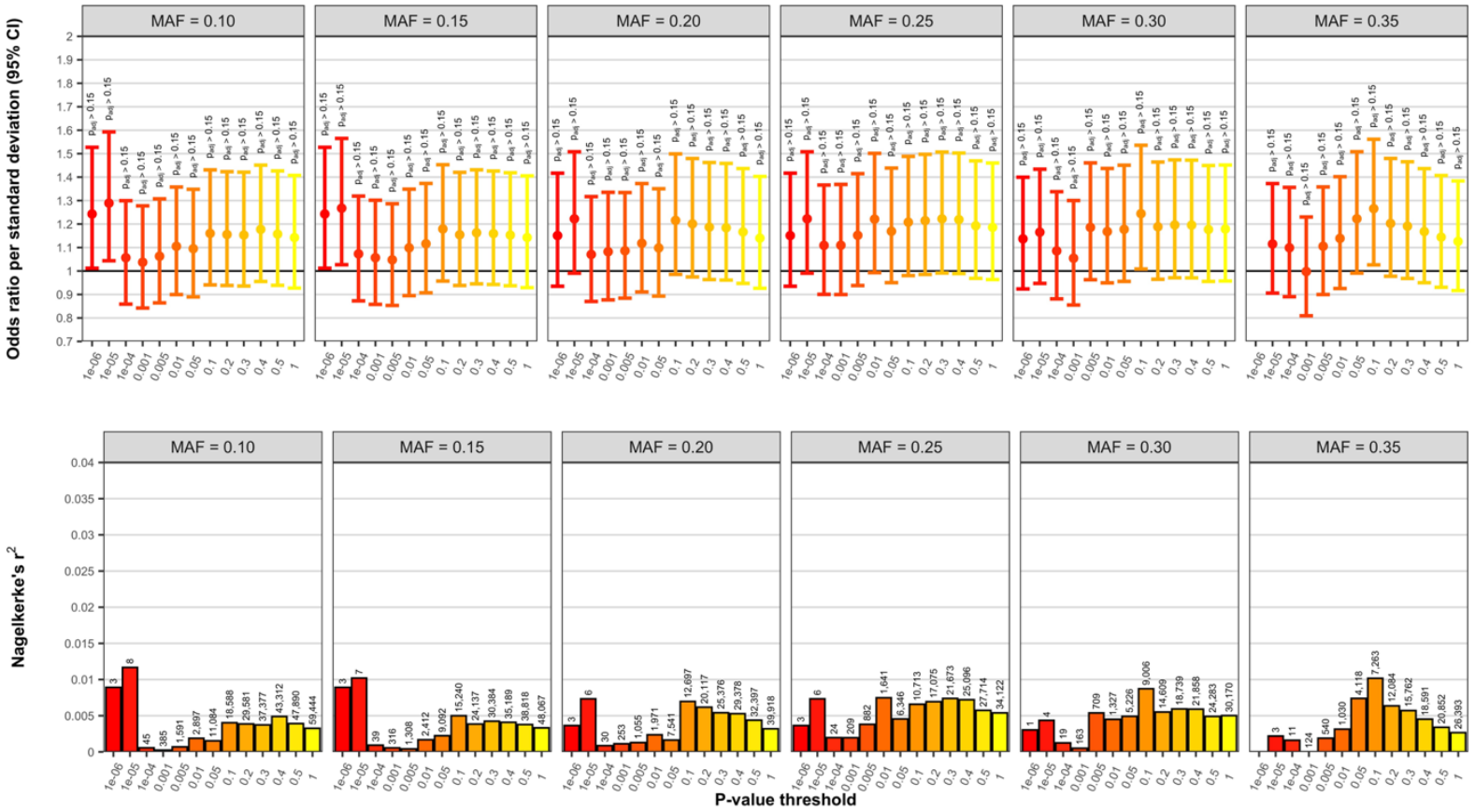


Figure 2.5. PRS results using meta-analysis of two GWAS as discovery dataset and employing inverse-variance-weighted SNP effects for scoring, for various MAF thresholds. MAF thresholds were applied to the discovery GWAS; SNPs with MAF below the threshold were excluded from PRS construction. Top row: Each plot displays odds ratio per standard deviation in PRS and the corresponding 95% confidence interval (y-axis) for PRS constructed based on particular discovery GWAS p-value thresholds (x-axis). P_{adj} are adjusted p-values (corrected for multiple correlated tests). 95% CIs correspond to unadjusted p-values. Bottom row: Each plot displays Nagelkerke's r^2 (y-axis) for PRS constructed based on particular discovery GWAS p-value thresholds (x-axis). Numbers above each r^2 bar are the number of SNPs used to construct PRS at that particular p-value threshold and MAF filter combination.



Secondary analyses

We performed the secondary analyses using only the training dataset derived from 2,594 mixed CHD cases and 5,159 controls, since using these training data produced the best results for the primary PRS analyses. The results from PRS analyses including and excluding chromosome 21 were essentially the same, with only slight fluctuations in ORs and corresponding Nagelkerke's r^2 values (**Figures 2.6** and **2.7**). These results generally followed a similar pattern to those observed for the primary PRS analysis using the same discovery dataset (**Figure 2.2**), wherein use of greater MAF filters yielded larger associations. However, the results from these secondary analyses fluctuated more across discovery GWAS p-value thresholds and included more outlier OR estimates, which was likely a result of the smaller number of SNPs used for scoring in the secondary analyses (which were limited to SNPs included on the Affymetrix array).

Figure 2.6. PRS results for all autosomes *excluding* chromosome 21. These analyses used the discovery GWAS of 2,594 mixed CHD cases and 5,159 controls. Various MAF thresholds were applied to the discovery GWAS; SNPs with MAF below the threshold were excluded from PRS construction. Top row: Each plot displays odds ratio per standard deviation in PRS and the corresponding 95% confidence interval (y-axis) for PRS constructed based on particular discovery GWAS p-value thresholds (x-axis). P_{adj} are adjusted p-values (corrected for multiple correlated tests). 95% CIs correspond to unadjusted p-values. Bottom row: Each plot displays Nagelkerke's r^2 (y-axis) for PRS constructed based on particular discovery GWAS p-value thresholds (x-axis). Numbers above each r^2 bar are the number of SNPs used to construct PRS at that particular p-value threshold and MAF filter combination.

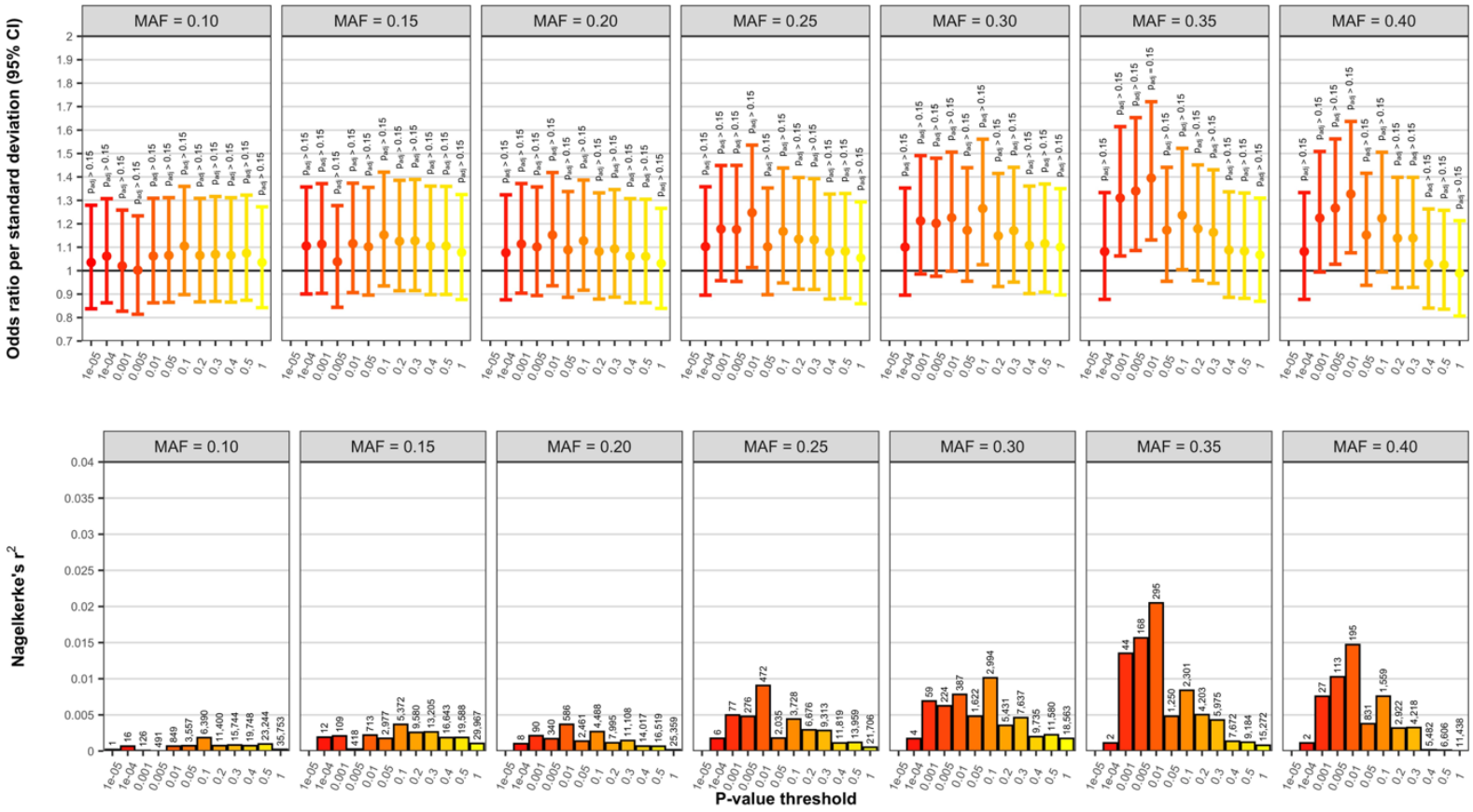
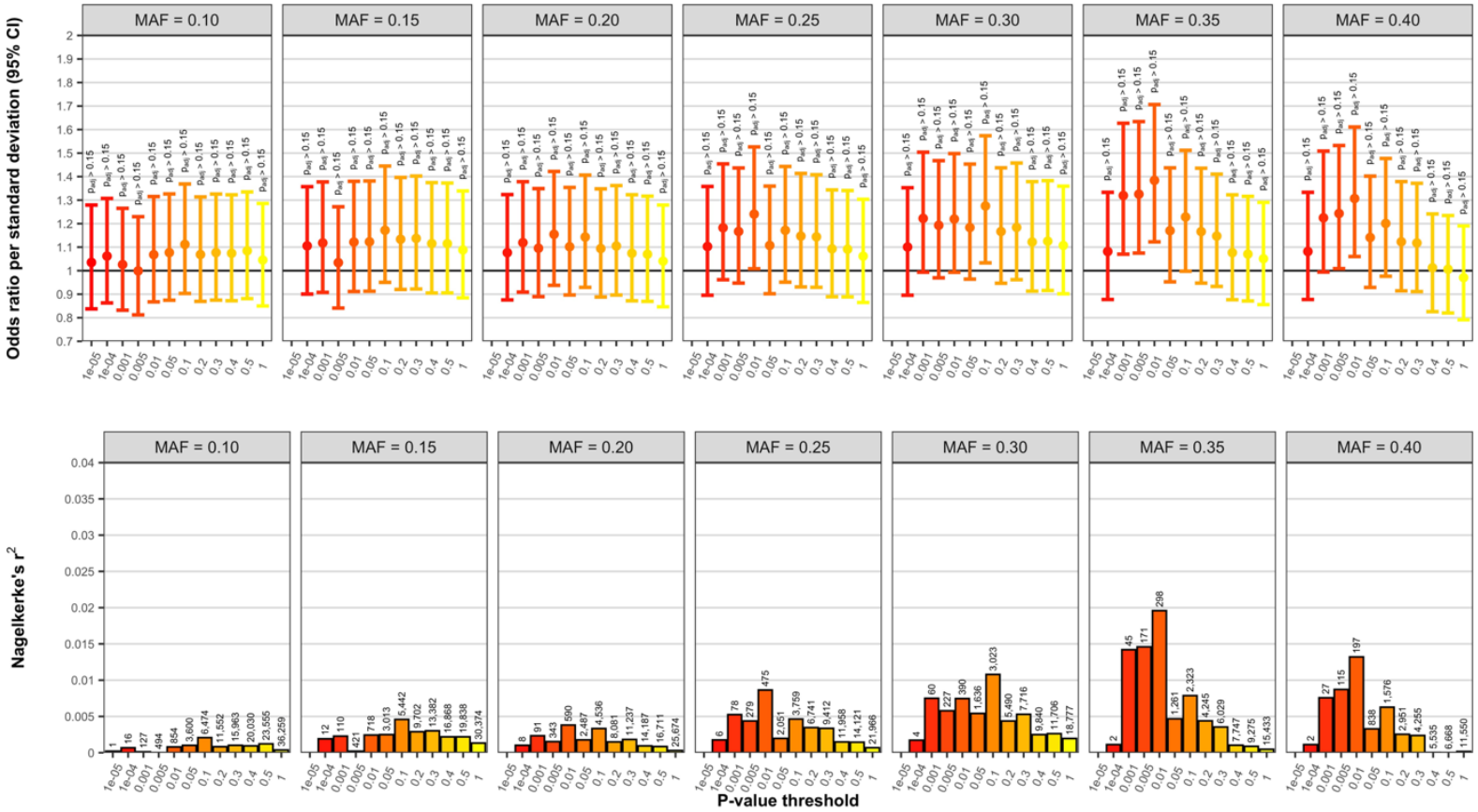


Figure 2.7. PRS results for all autosomes *including* chromosome 21. These analyses used the discovery GWAS of 2,594 mixed CHD cases and 5,159 controls. Various MAF thresholds were applied to the discovery GWAS; SNPs with MAF below the threshold were excluded from PRS construction. Top row: Each plot displays odds ratio per standard deviation in PRS and the corresponding 95% confidence interval (y-axis) for PRS constructed based on particular discovery GWAS p-value thresholds (x-axis). P_{adj} are adjusted p-values (corrected for multiple correlated tests). 95% CIs correspond to unadjusted p-values. Bottom row: Each plot displays Nagelkerke's r^2 (y-axis) for PRS constructed based on particular discovery GWAS p-value thresholds (x-axis). Numbers above each r^2 bar are the number of SNPs used to construct PRS at that particular p-value threshold and MAF filter combination.



DISCUSSION

Prior research attempts to illuminate the role of common genetic variants in DS-associated AVSD have yielded no robust common variant associations.[4,37,38] While these studies were adequately powered to detect common variants with large effects on AVSD, they did not have sufficient power for investigating common variants with small- to moderate-sized effects. At present, small sample sizes continue to be a limiting factor for investigating the contribution by common variants to AVSD in DS, particularly for single-variant analysis approaches like GWAS.

In the current study, we used PRS methods as an alternative means of assessing the role of common variants in DS-associated AVSD. The PRS approach involves examining the collective contribution of many common variants across the genome to AVSD. By focusing on the aggregate effect of numerous SNPs, this approach provides increased power for investigating the overall role played by common variants in DS-associated AVSD, particularly given our small sample size.

These PRS analyses are the first such analyses of AVSD in DS, and to the best of our knowledge they are also the first use of PRS methods to examine polygenicity of CHD generally. Our analyses of PRS calculated from GWAS studies of non-syndromic CHD suggest at minimum a small polygenic contribution by common variants to AVSD among individuals with DS. When using dense SNP data (WGS or imputed data) for the 487 individuals in the target sample and excluding chromosome 21, a single standard deviation increase in PRS was associated with a 20-30% increased odds for having AVSD, with Nagelkerke's r^2 values for PRS of around 1% (**Figures 2.2** and **2.3**); this occurred when using the larger of the two independent discovery datasets. Assuming a population prevalence of 20% for AVSD among those with DS, these Nagelkerke's r^2 values are quite similar to the corresponding liability scale r^2 values (correcting for case-control ascertainment). For instance, the PRS analyses depicted in **Figure 2.3** yielded a Nagelkerke's r^2 of 1.03% when applying MAF ≥ 0.35 and discovery GWAS p-value

≤ 0.001 thresholds; the corresponding liability r^2 estimate is 1.11%.[42] As demonstrated by the PRS results presented in **Figures 2.6** and **2.7**, which involved the use of array SNPs only, inclusion of dense genotype data for chromosome 21 is unlikely to substantially alter these estimates for the association of PRS with DS-associated AVSD; SNPs on chromosome 21 are perhaps not a key factor driving AVSD in DS.

Given the small sample sizes for the discovery GWAS datasets and prior research demonstrating that variance explained by PRS tends to increase as discovery GWAS sample size increases,[19] which is attributable to increased accuracy of the SNP effect estimates used as weights for the PRS, it seems likely that the use of a larger discovery GWAS of CHD will uncover a greater polygenic contribution of common variants to AVSD in DS. Furthermore, use of a large discovery GWAS that only includes the particular CHD subtypes that are most closely genetically related to AVSD (perhaps a GWAS including only AVSD and septal defect cases) may reveal a polygenic contribution to DS-associated AVSD that exceeds what we have identified. We demonstrate this in **Figure 2.8**, showing that under reasonable assumptions, using a discovery GWAS of phenotypes that are highly genetically correlated with the target phenotype (AVSD) will result in PRS r^2 values that increase as discovery GWAS sample size increases; the discovery samples similar in size to those used for the current PRS analyses are only able to capture a portion of the true polygenic component (plots generated using the 'avengeme' R package).[59]

The finding of an association of AVSD in DS with PRS constructed based on SNPs identified as having some measure of association with CHD in mixed CHD samples suggests the possibility of genetic overlap between AVSD and various other subtypes of CHD. This is consistent with the potential for investigations of DS-associated AVSD to shed light on fundamental biology relevant to CHD more generally. To further examine this potential genetic overlap, including which CHD subtypes may have the greatest shared genetic architecture with

AVSD, it will be important to utilize large GWAS datasets of specific CHD subtypes rather than a mixture of CHD types.

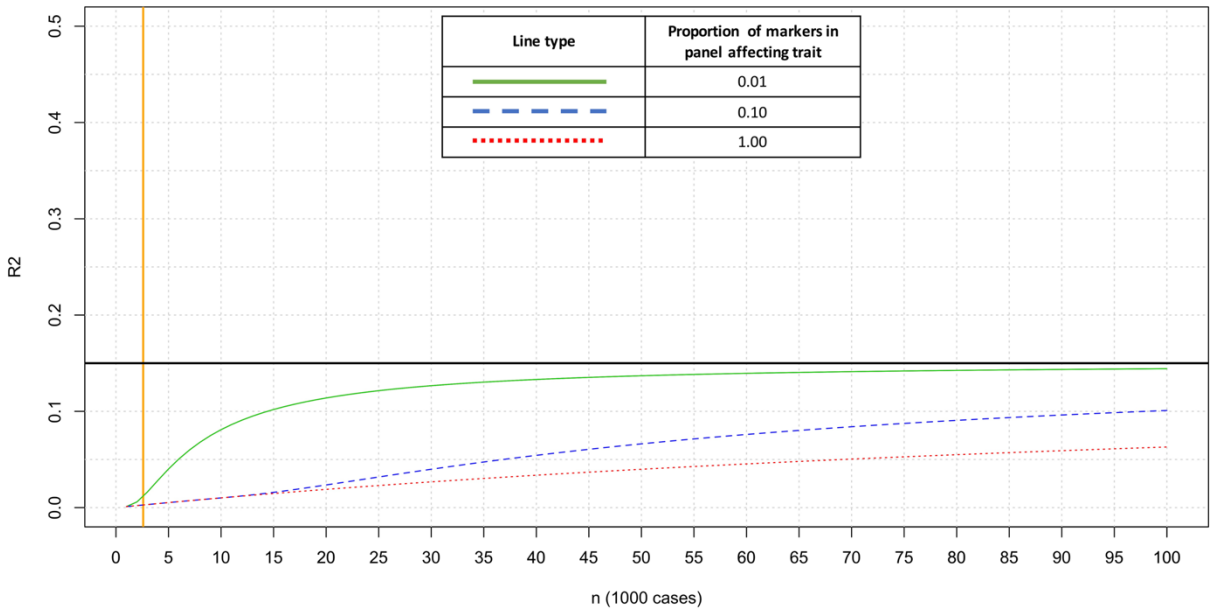
We observed that PRS constructed based on the discovery GWAS of 2,594 mixed CHD cases and 5,159 controls consistently yielded ORs > 1 (indicating, as expected, that increased PRS was associated with increased AVSD risk). In contrast, PRS constructed using the discovery GWAS of 406 CHD cases and 2,976 pediatric controls yielded OR estimates generally quite close to the null, and on both sides of the null. One possible reason for this difference is that the smaller-sized discovery GWAS had more imprecisely estimated SNP associations, leading to less informative PRS. Another possibility is that particular CHD diagnoses included within the larger discovery GWAS may be more genetically related to AVSD in DS than the CHD diagnoses in the smaller discovery GWAS. Indeed, the larger GWAS included 73 cases with AVSD, while in the smaller GWAS there were only seven instances of AVSD (six of the cases with double outlet right ventricle also had AVSD, and a single case had tetralogy of Fallot with atrioventricular canal septal defect).

In conclusion, our PRS analyses yielded association estimates that are consistent with common variants, acting collectively through a polygenic component, playing a role in increasing AVSD risk among those with DS. Future conduct of larger CHD-focused GWAS will enable the use of more accurate weights in PRS construction, which in turn will allow more accurate quantification of the collective common variant contribution to DS-associated AVSD. As PRS become more accurate, we will also be better positioned to investigate whether polygenic risk due to common variants is especially pronounced in the presence of particular environmental influences. In addition, should methods such as the PRS approach continue to support a role for common variants in DS-associated AVSD, GWAS and SNP-set analyses of larger DS-associated AVSD sample sets may help identify which particular common variants, acting through which particular genes, are making the greatest contribution to AVSD risk, which would further our understanding of the biology underlying AVSD. Lastly, continued accumulation

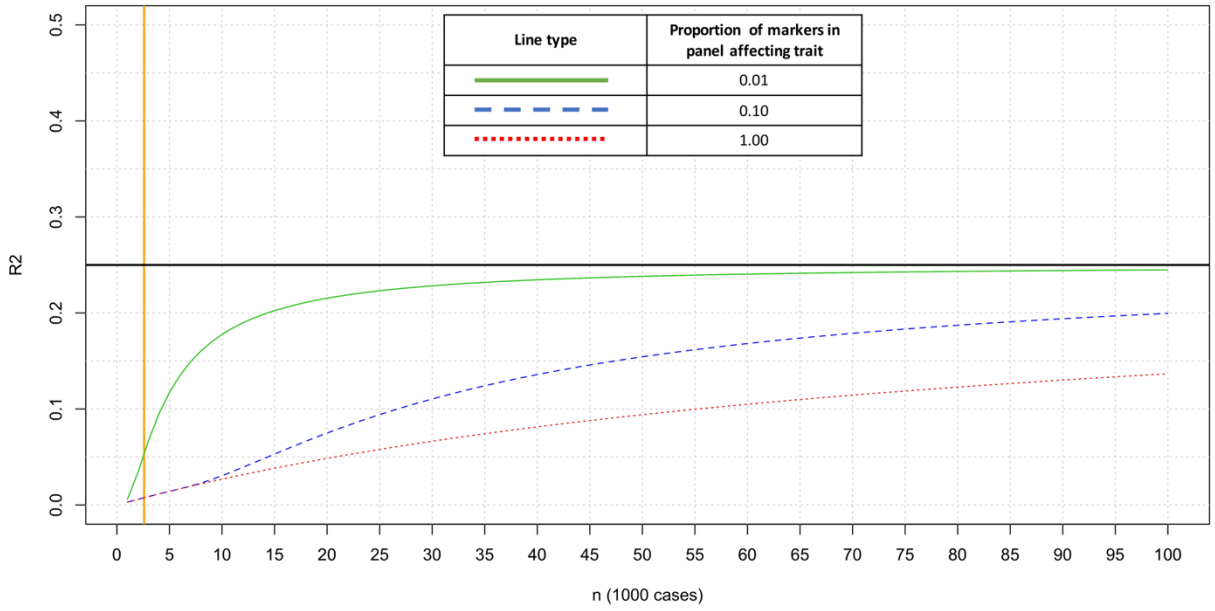
of findings regarding the genetic architecture of DS-associated AVSD, including the contribution by common variants both collectively and individually, has the potential to inform investigation and understanding of CHD in the general population, with the possibility and hope of benefiting individuals both with and without DS.

Figure 2.8. Maximum variance in target phenotype that can be explained by PRS (y-axis: liability scale r^2) given a range of training sample sizes (x-axis: number of cases in thousands). Assumptions: training sample with case:control ratio of 1:2 (same as ratio for larger of the two independent CHD discovery datasets); target sample with case:control ratio of 1:1 (same as ratio for DS target dataset); prevalence of CHD in training population is 1%; prevalence of AVSD in DS target population is 20%; 100,000 independent variants in the training SNP panel; genetic effects for training and target samples are identical (correlation = 1); proportion of SNPs in the training set panel that affect the training phenotype is 1%, 10% or 100%. For plot **A**, amount of variance in the training phenotype explained by the training set SNP panel (Vg_{train}) is 15%; for plot **B** Vg_{train} is 25%; for plot **C** Vg_{train} is 35%. Solid black horizontal line marks the maximum r^2 that can be explained by PRS using an infinitely large training sample size (given the assumed parameters). Vertical orange line marks the number of CHD cases in the larger of the two independent discovery datasets (2,594 cases).

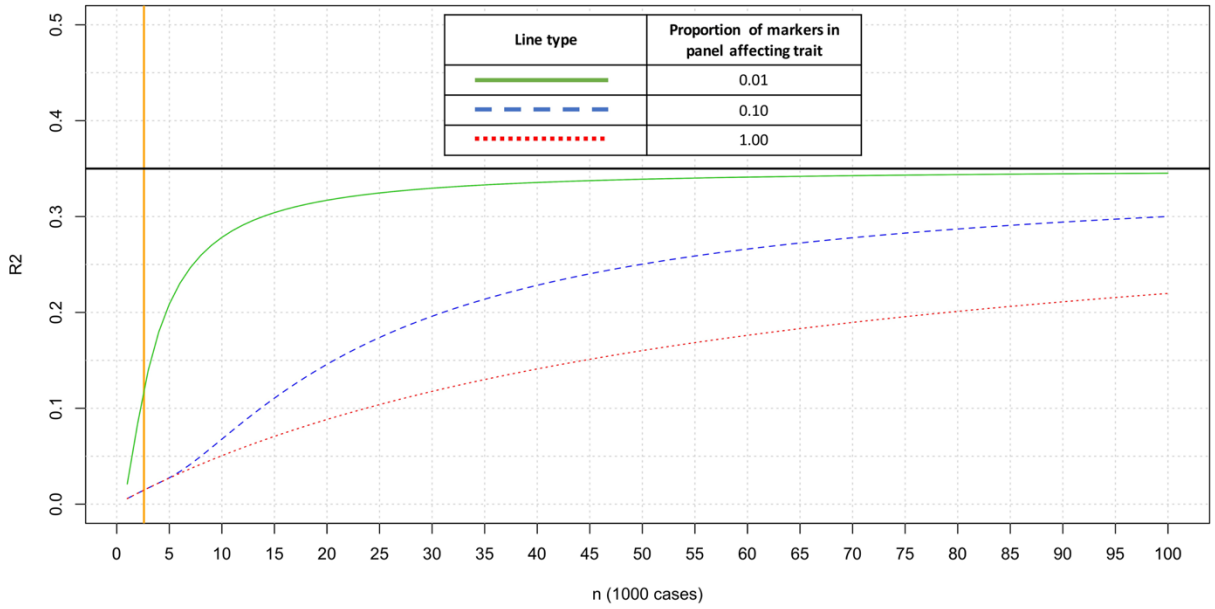
A.



B.


Line type
Proportion of markers in panel affecting trait
0.01
0.10
1.00
R2
0.0
0.1
0.2
0.3
0.4
0.5
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100
n (1000 cases)

C.


Line type
Proportion of markers in panel affecting trait
0.01
0.10
1.00
R2
0.0
0.1
0.2
0.3
0.4
0.5
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100
n (1000 cases)

Chapter 3:

A powerful multivariate method for examining genetic associations with psychiatric phenotypes

Research described in **Chapter 3** has been published:

Holleman AM*, Broadaway KA*, Duncan R, Todor A, Almli LM, Bradley B, Ressler KJ, Ghosh D, Mulle JG, Epstein MP. **Powerful and Efficient Strategies for Genetic Association Testing of Symptom and Questionnaire Data in Psychiatric Genetic Studies.** Sci Rep. 2019 May 17;9(1):7523. doi: 10.1038/s41598-019-44046-0. PMID: 31101869; PMCID: PMC6525248. [Published by Springer Nature]

*Joint first authors

Note: With a few necessary exceptions, work described in this publication that I did not perform is not included in Chapter 3 of this dissertation. The exceptions include the original GTP subject recruitment and data collection, and the initial GWAS quality control of the GTP dataset (after which 4,607 subjects remained), which are directly related to my subsequent work. Text in the above publication for which I was not the original author is not included in Chapter 3 of this dissertation.

Abstract

Background: Psychiatric disorders are syndromes involving the co-occurrence of multiple correlated yet discrete symptoms. They are assessed based on multivariate diagnostic criteria (e.g., DSM or ICD criteria). In research settings, this assessment is often accomplished by administering psychiatric questionnaires, such as the 17-item PTSD Symptom Scale (PSS) and the 21-item Beck Depression Inventory (BDI). Responses to such questionnaires are commonly aggregated (e.g., through summation) to generate a single overall measure of the psychiatric condition of interest which is then analyzed using univariate methods. However, with respect to genetic epidemiology research, if a genetic factor only affects a subset of the symptoms assessed by the psychiatric questionnaire items, the genetic effect may be very challenging to detect with univariate approaches that analyze a single, aggregated score.

Methods: We evaluated GAMuT, a multivariate method previously developed for tests of rare variant pleiotropy, as a potentially powerful approach for identifying common (MAF > 1%) variant associations with multivariate psychiatric phenotypes. We applied GAMuT to simulated datasets of SNP genotypes and BDI responses to examine Type I error control and power for GAMuT, and we applied univariate kernel machine regression (KMR) and standard linear regression (both of which analyze a single overall score) to compare with GAMuT. We then utilized GAMuT to examine associations of common variants within gene regions with both the PSS and the BDI for a sample of 3,826 African-American participants in the Grady Trauma Project.

Results: Our simulated data analyses demonstrated that GAMuT properly controls Type I error, and that it is substantially more powerful than standard univariate approaches for identifying

common variant associations with the BDI, particularly in scenarios where a SNP only affects half of the BDI items or fewer. When applied to the GTP datasets, GAMuT identified common variants in or near *SIRPA* and *ZHX2* to be significantly associated with the PSS and BDI, respectively. In comparison, univariate KMR and linear regression detected no study-wise significant associations.

Conclusions: Through analyses of simulated and real data, we have demonstrated GAMuT to be a powerful method for detecting common variant associations with multivariate psychiatric phenotypes. Application of GAMuT in future psychiatric genetics studies has good potential to facilitate the identification of robust common variant associations which have often evaded detection by traditional analytic approaches.

INTRODUCTION

Psychiatric disorders are etiologically complex and result from a combination of genetic and environmental risk factors.[60] Recently performed meta-analyses of twin studies estimate the heritability (broadly defined as the overall genetic contribution to phenotypic variance) for psychiatric disorders collectively to be 46%.[61,62] Heritability estimates for specific psychiatric disorders range from 34% for major depression and 30-46% for posttraumatic stress disorder (PTSD), to > 60% for bipolar disorder and schizophrenia.[61-67] These heritability estimates reflect contributions by all genetic variants, including common (minor allele frequency [MAF] > 1%) and rare (MAF < 1%) variants. Heritability estimates that only consider the additive contributions by common variants, termed the single nucleotide polymorphism (SNP; a type of common variant) heritability, are 12% and 15% for major depression and PTSD, respectively; and 21% and 24% for bipolar disorder and schizophrenia, respectively, indicating that common variants play an important role in psychiatric disorders.[68,69] However, the proportion of phenotypic variance explained by common variants that have been identified as robustly associated with various psychiatric disorders (based on exceeding genome-wide significance thresholds) falls well short of the estimated SNP-heritabilities. For instance, the 128 SNPs identified as significantly associated with schizophrenia in a recent genome-wide association study (GWAS) consisting of 36,989 cases and 113,075 controls collectively explained 3.4% of variation in schizophrenia status,[19] which is just a small proportion of the 24% SNP-heritability estimated for schizophrenia. As GWAS sample sizes grow ever larger, it is likely that more robust common variant associations will be identified. However, another possible explanation for the observed challenges in identifying the common variants contributing to psychiatric disorders relates to the ways in which psychiatric phenotypes are typically modeled in genetic epidemiology analyses.

Psychiatric disorders, such as depression or anxiety disorders, are characterized by the Diagnostic and Statistical Manual of Mental Disorders (DSM-5)[6] as behavior or psychological ‘syndromes’, with a syndrome defined as a set of correlated signs or symptoms, the

cooccurrence of which may be said to constitute a particular disease or disorder (Note: The DSM-5 uses the term mental disorder in place of psychiatric disorder).[7] As syndromes, clinicians diagnose psychiatric disorders based on specific criteria (such as those set forth by the DSM-5 or the International Classification of Diseases [ICD-10]), which often involves examining the number of syndrome-related symptoms an individual is presenting with, in combination with additional criteria such as whether the symptoms cause significant distress or impairment and are not better accounted for by a different condition.[6] Psychiatric disorders and mental health status are also frequently assessed using questionnaires, which may be administered or self-reported. For instance, the PTSD Symptom Scale (PSS) is a 17-item questionnaire designed to assess and diagnose PTSD based on DSM-IV criteria. Each item of the PSS corresponds to a PTSD symptom and is rated on an ordinal scale from 0 to 3, with higher scores indicating greater symptom frequency/intensity.[8,9] PTSD severity is then determined by totaling the scores for the 17 items (scores can range from 0 to 51), and a PTSD diagnosis can be made based on the reported presence of a certain number of symptoms in the three PSS subcategories. Similarly, the Beck Depression Inventory (BDI-II) is a 21-item questionnaire for assessing major depressive disorder (MDD) according to DSM-IV criteria, with each item scored on an ordinal scale from 0 to 3 (higher scores indicative of more severe symptoms), and with the ability to sum across items and generate a cumulative score reflecting depression severity.[10]

These psychiatric disorder diagnoses (presence or absence of the disorder) or cumulative scores from questionnaires like the PSS and BDI are commonly analyzed as univariate outcomes in psychiatric research, including genetic studies of psychiatric disorders. However, as syndromes consisting of multiple correlated yet discrete symptoms, psychiatric disorders perhaps would be more accurately analyzed as multivariate phenotypes. This perspective is consistent with the National Institute of Mental Health's (NIMH) recent focus on Research Domain Criteria (RDoC), which were developed in response to recognition of the limitations of research based on traditional diagnostic categories, such as the observation that

two individuals can share the same psychiatric diagnosis yet have few symptoms in common, with these differing symptom manifestations potentially influenced by different underlying mechanisms.[70,71] To address these limitations, RDoC emphasize research focused on basic functional dimensions or mechanisms involved in psychopathology (e.g., fear, reward-seeking, attention, perception, arousal) rather than DSM or ICD diagnostic categories.

With regard to genetic epidemiology research seeking to identify genetic factors that may underlie various psychiatric phenotypes, analyzing a univariate measure that is a summary across multivariate symptom data has the potential to decrease power for detecting genetic associations. Specifically with respect to genetic analyses of multivariate ordinal data, it has been shown that a univariate summary measure will fail to function as an adequate summary of the multivariate data under circumstances that include the genetic factor not having identical effects on all of the multivariate items.[11] A scenario such as this is entirely plausible for most psychiatric disorders, given the variety of symptoms that collectively constitute each syndrome. As one example, a genetic factor may have different effects on BDI items that are more somatic-related (e.g., items assessing sleep or appetite) as compared with items that are more mood-related (e.g., those assessing feelings of sadness or guilt). In turn, if the univariate measure does not adequately summarize the multivariate data, then the choice of using the univariate measure for analysis has the potential to decrease power for identifying genetic associations with the psychiatric phenotype.[11] In such a circumstance, use of analytic techniques that allow for proper modeling of the multivariate phenotype data are likely to provide increased power for detecting genetic signals as compared with univariate analysis approaches.

Multivariate analysis methods that enable modeling of ordinal data that is commonly generated using questionnaires are currently suboptimal in various respects. For instance, principal component analysis (PCA) is frequently employed for multivariate analysis of correlated measures such as items in a psychiatric questionnaire, but the standard practice of analyzing the top principal components (PCs) (those which capture the majority of phenotypic

variance) was recently shown to often yield low power.[12] With these current limitations in mind, we decided to examine the effectiveness of a novel multivariate analysis approach for identifying genetic associations with multivariate psychiatric questionnaire data. The method we examined, named the Gene Association with Multiple Traits (GAMuT) test, was developed recently as a means of testing for rare variant pleiotropy.[13] GAMuT enables high-dimensional modeling of multiple traits and multiple rare variants (e.g., within a gene), and tests for association between these high-dimensional phenotype and genotype data through use of a kernel distance-covariance (KDC) framework. We repurposed GAMuT and evaluated it as a potentially powerful method for identifying common variant associations with multivariate psychiatric phenotypes, specifically those assessed with ordinal questionnaire items, with special attention paid to scenarios in which the genetic effect differs across the various phenotypes assessed by the questionnaire items.

In the first part of **Aim 2**, we simulated genetic and BDI data under a variety of scenarios (e.g., varying the causal SNP; varying the proportion of BDI items affected by the causal SNP), and used these simulated datasets to evaluate GAMuT with respect to Type I error control and power for identifying SNP effects. We also applied two univariate analysis approaches, standard linear regression and kernel machine regression (KMR), to the simulated datasets to compare univariate analysis of the cumulative BDI score (summed across all items) with GAMuT. Univariate linear regression is a standard, SNP-level analysis approach used in GWAS for identifying associations between individual common variants and continuous phenotypes. KMR, on the other hand, is like GAMuT in that it models multiple genetic variants together and thereby enables gene-level analyses.

In the second part of **Aim 2**, we applied GAMuT and the two univariate analysis methods to real genetic and phenotypic data accumulated through the Grady Trauma Project (GTP). For these applied analyses, we first examined common variant associations with the PSS among 3,826 individuals with African-American ancestry. The PSS consists of 17 questionnaire items

that can be considered all together to provide an overall assessment of PTSD, or can be partitioned into three non-overlapping subscales, including PSS subscales assessing the re-experiencing of trauma (PSS Intrusive; 5 items), avoidance-numbing (PSS AvoidNumb; 7 items), and hyperarousal (PSS Hyperarousal; 5 items). We performed separate analyses for the PSS overall and also for each of the three PSS subscales. We then also performed analyses to examine common variant associations with the BDI, using the subset of 3,520 individuals with complete BDI data.

METHODS

Overview of GAMuT

A detailed technical description of the GAMuT method has been provided in the original GAMuT publication.[13] Here, we provide a summary of GAMuT. For a sample of *N* unrelated subjects, GAMuT examines the association between a set of *Q* questions (which may be continuous and/or ordinal categorical variables with an arbitrary number of levels) and a set of *V* genetic variants such as the set of variants defined by a gene. GAMuT is motivated by the idea that, for a pair of individuals, increased genetic similarity at phenotype-influencing variant sites across a gene should lead to increased similarity in the phenotype data. Consequently, GAMuT employs a KDC framework to construct two separate similarity matrices, one for the multivariate phenotype outcomes and the other for the genetic variants within a gene. Each similarity matrix has *N* rows and *N* columns, and each element of the matrix is a measure representing the similarity in multivariate data (phenotypic or genetic) between two individuals. The similarity matrices can be modeled in different ways depending on user preference. For instance, similarity in phenotypes can be modeled using a projection matrix[72] or various kernel functions;[73,74] while similarity in genetic data can be modeled using the same kernel functions used for the phenotypes or using genetic-specific kernel functions.[75] The genetic similarity

matrix can also be modeled using weighted kernel functions, allowing different genetic variants to carry different weight based on prior information, with the potential result of increased power. Using the KDC framework, GAMuT then tests for association between the individual elements in the genetic similarity matrix and the individual elements in the phenotype similarity matrix. The resulting test statistic follows a known asymptotic distribution, enabling accurate and rapid calculation of p-values (making GAMuT analyses faster than alternative multivariate approaches that employ permutations to generate p-values). The GAMuT framework is also amenable to adjustment for covariates, which can be accomplished by regressing the multivariate phenotypes onto the covariates of interest, and then using the resulting regression residuals as the phenotypes for GAMuT analysis.

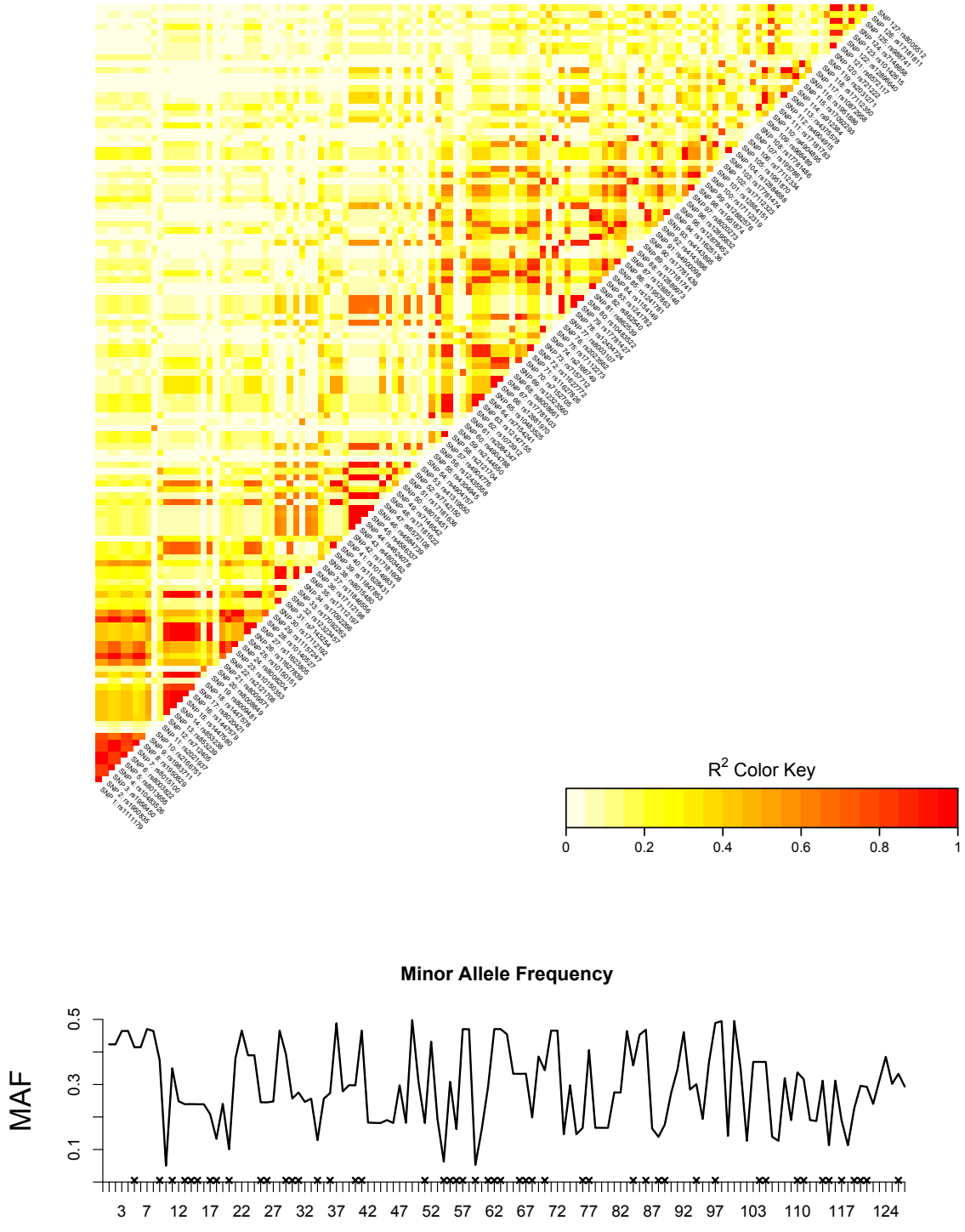
Simulated data analyses

We applied GAMuT to simulated datasets of 1,000 or 2,500 unrelated subjects to examine Type I error control and power when using GAMuT to examine associations between common variants within a gene and multivariate psychiatric questionnaire data. For comparison, we also applied KMR and standard linear regression to the simulated datasets; these methods differ from GAMuT in that they model the phenotype as univariate, analyzing the cumulative score which results from summing across all questionnaire items.

We simulated common variant data for the gene *LRFN5* (leucine rich repeat and fibronectin type III domain containing 5) on chromosome 14. This gene was recently identified as potentially involved in MDD.[76] *LRFN5* is a relatively large gene: it is 297 kilobases (kb) in length and includes > 120 common variants with MAF > 5% (we applied a 5% MAF threshold for our analyses in this study).[77] In the top of **Figure 3.1**, we present a heatmap showing the linkage disequilibrium (LD) structure for 127 common variants within or in close proximity to *LRFN5* (close proximity defined as within 2 kb of either gene end). The bottom of **Figure 3.1** is a plot of the MAF for each of these 127 common variants (all variants have MAF > 5%). The

SNPs as shown in the heatmap are ordered according to genomic position, and the ordering of SNPs in the MAF plot (labeled numerically) matches that of the heatmap.

Figure 3.1: Pairwise LD (R^2) heatmap (top) and MAF (bottom) for 127 common variants in the *LRFN5* gene. SNPs are ordered by genomic position. Numeric label on x-axis of MAF plot corresponds to SNP number in heatmap. The 50 SNPs present on the Illumina array are denoted with an 'x' in the MAF plot.



As a larger gene, *LRFN5* is more likely to include certain combinations of SNPs that are in lower LD with one another (namely, the SNPs that are located farther away from one another); whereas for a shorter gene, LD between SNPs would tend to be higher given the SNPs would generally be located in closer proximity to each other. We intentionally selected *LRFN5*, as opposed to a smaller gene, based on its relatively weaker LD structure, whereby a substantial proportion of its SNPs are in relatively low LD with one another. Our reason for this selection criteria was that such a gene would present a greater challenge for GAMuT to successfully detect a causal SNP association, and would therefore enable a more convincing demonstration of GAMuT's utility for identifying common variant associations with multivariate psychiatric phenotypes. The challenge would be greater for a gene like *LRFN5* because the particular SNP assigned as causal for the phenotype (which we varied across simulations) would be in relatively low LD with many other SNPs in the gene, making the causal genetic signal more difficult to detect once these "low-LD SNPs" were grouped together with the causal SNP and the other remaining common variants for the gene-based analyses. This setting is contrasted with a smaller gene, for which the SNPs would be in closer proximity to each other and would thus be expected to generally be in higher LD with one another, resulting in a greater boost in the genetic signal and making detection by GAMuT easier.

We used the HAPGEN2 package to simulate realistic SNP data for *LRFN5*.[78] HAPGEN2 enables the generation of large quantities of simulated haplotypes that mirror the LD structure of participants in the International HapMap Project.[47] We generated 20,000 haplotypes for the region corresponding to *LRFN5* for use as our simulation haplotype pool, which we sampled to generate genotypes. For the purposes of these simulated data analyses, we included common variants within the genomic window ranging from 2kb upstream to 2kb downstream of *LRFN5*. This window, which we simply refer to as the gene *LRFN5*, included 127 common variants.

For phenotypes, we simulated BDI questionnaire responses. BDI responses were simulated under a null genetic model for the Type I error analyses, and under a variety of causal genetic models for power analyses (described in detail below). For all analyses, we simulated realistic BDI datasets that approximated the BDI correlation structure (pairwise correlations between different BDI items ranged from 0.22 to 0.60) and distribution of ordinal responses observed among the GTP participants.

Type I error

For the Type I error analyses, we simulated datasets of 1,000 and 2,500 unrelated subjects under the assumption that none of the 127 SNPs within *LRFN5* had a causal effect on any of the 21 BDI items. For each simulated subject, we randomly sampled genotypes from our haplotype pool.

For phenotypes, we simulated 21 ordinal BDI responses for each subject by first randomly sampling values from a multivariate normal distribution with dimension 21, mean vector 0, and sigma equal to the 21x21 pairwise correlation matrix generated by calculating Spearman rank correlations between BDI items using data from all GTP participants. Once this process was completed for all simulated subjects, for each of the 21 phenotype items, we transformed the continuous values to ordinal BDI responses (with values of 0, 1, 2 or 3) by a process that yielded response distributions matching the observed response distribution for each BDI item among the GTP participants. For example, considering BDI item 4 (which assesses loss of pleasure), the proportions of GTP participants selecting 0 (no loss of pleasure), 1, 2 and 3 (high loss of pleasure) were 49%, 34%, 12% and 4%, respectively. For this item, we took the continuous values simulated for all subjects and assigned the lowest 49% of values a response of 0, the next lowest 34% of values a response of 1, the next lowest 12% of values a response of 2, and the highest 4% of values a response of 3. In this manner, we generated realistic BDI questionnaire responses for all simulated subjects.

Using this process, we simulated 10,000 null datasets of 1,000 subjects and 10,000 null datasets of 2,500 subjects. For each simulated dataset, we applied GAMuT to test for association between common variants within *LRFN5* and BDI phenotypes. We applied GAMuT with a weighted linear genotype kernel, with weights based on estimated MAF for each common variant (rarer variants carrying greater weight); and we modeled phenotypic similarity with a projection matrix as well as a linear kernel. We also applied gene-based KMR and SNP-based linear regression to compare GAMuT with standard univariate approaches. We applied KMR with a weighted linear genotype kernel with MAF-based weights, in order to match GAMuT.

While we simulated genotypes for all 127 common variants within or near *LRFN5*, our association tests only considered the subset of 50 SNPs present on the Illumina HumanOmni1-Quad genotyping array. This was done to mirror the realistic scenario wherein a limited set of SNPs would be available for consideration in a study employing array-based genotyping without subsequent imputation. This decision to only test common variants on the SNP array perhaps has greater relevance for the power analyses, as described in the next section.

We computed Type I error for α values of 0.05, 0.01 and 0.001. For a method that properly controls Type I error, we would expect that approximately 5% of the 10,000 p-values (outputted by applying the method to 10,000 simulated datasets) have $p < 0.05$, approximately 1% have $p < 0.01$, and approximately 0.1% have $p < 0.001$. GAMuT and KMR are both gene-based (or variant set) methods that yield a single p-value per simulated dataset, while linear regression is applied at the SNP-level and therefore involves 50 tests and 50 resulting p-values for each simulated dataset, one for each of *LRFN5*'s 50 SNPs present on the Illumina array. For each simulated dataset, we selected the linear regression p-value (out of 50 total) that was the smallest, used an approach called P_{ACT}[58] to adjust this p-value for the multiple correlated tests performed across the gene, and stored this adjusted p-value as the single value representing the linear regression results, thus yielding the 10,000 linear regression p-values used for estimating Type I error.

Power

For the power analyses, we generated simulated datasets of 2,500 subjects. We simulated genotypes for the 127 SNPs in *LRFN5* by sampling from our HAPGEN2-generated haplotype pool. We then simulated BDI responses for each subject by setting one of the 127 SNPs as causal for a specified proportion of the BDI items. We considered scenarios where 18/21, 12/21, and 6/21 BDI items were affected by the causal SNP. The effect size of the causal SNP on each associated BDI item was drawn from a normal distribution with mean 0.10 and variance 0.03, resulting in an overall modest effect of SNP on the cumulative BDI score. For instance, these SNP effect parameters yield $r^2 = 0.009$ when a SNP with MAF = 0.30 is causally associated with all BDI items.

As with the simulated phenotypes for the Type I error analyses, we applied a process to ensure simulated BDI responses displayed a correlation structure (pairwise BDI item correlations) which mirrored that of the GTP samples. However, different from the Type I error simulations, we included consideration of trait-specific heritability (the relative variance in responses for a given BDI item explained by the causal SNP) in the process of controlling residual correlation among simulated BDI items. We calculated trait-specific heritability for a given BDI item as

$$h_q = \beta_{v,q}^2 * 2 * MAF_v * (1 - MAF_v)$$

where $\beta_{v,q}$ is the effect size of the causal variant *v* on the BDI questionnaire item *q*, MAF_v is the minor allele frequency of the causal variant *v*, and h_q is the heritability for BDI questionnaire item *q*. The correlation between two BDI items *q* and *q'*, adjusted for trait specific heritability, is then

$$E_{q,q'} = \sqrt{1 - h_q} * \sqrt{1 - h_{q'}} * \Sigma_{q,q'}$$

where $\Sigma_{q,q'}$ is the pairwise correlation between BDI items *q* and *q'*, estimated using the GTP samples.

For each simulated subject, we then sampled from a multivariate normal distribution with dimension 21, and with mean equal to the vector of SNP effect sizes described at the beginning of this section, and sigma equal to $E_{q,q'}$. We subsequently transformed the resulting values to ordinal BDI scores using the approach described in the Type I error section above, ultimately yielding simulated datasets in which a subset of BDI questionnaire items were causally associated with a given SNP, with BDI responses mirroring the correlation structure and distribution of BDI responses observed among the GTP samples.

We simulated datasets for the power analyses with each of the 127 SNPs within *LRFN5* set as causal, one at a time. For each combination of causal SNP (127 total) and proportion of BDI items affected by the causal SNP (18/21, 12/21, 6/21), we simulated 500 datasets of 2,500 samples. We applied GAMuT to each of these simulated datasets, using a weighted linear genotype kernel (with MAF-based weights), and both a projection matrix and linear kernel for modeling phenotypic similarity. For comparison with univariate approaches, we also applied KMR and standard linear regression, using a weighted linear genotype kernel (with MAF-based weights) for KMR.

Though we simulated genotypes for all 127 *LRFN5* SNPs and we set each of these SNPs as causal, one at a time, when simulating BDI phenotypes, we only considered the 50 SNPs present on the Illumina array for association testing (as described in the Type I error section above). Therefore, if the causal SNP was not among the 50 SNPs included for testing, detection of the causal association relied entirely on LD of the causal SNP with the testing SNPs. This mirrors the realistic scenario wherein the array SNPs may tag the causal SNP but are not causal themselves, with the observed association between tag SNP and phenotype likely attenuated (depending on extent of LD) compared with the association that would be observed for the causal SNP.

For each unique causal scenario (each unique combination of causal SNP and proportion of associated BDI items) we calculated power as the proportion of p-values from the

500 simulated data analyses that were < 0.001. As with the Type I error analyses, for linear regression, we selected the smallest p-value resulting from 50 SNP-level tests, used P_{ACT} to adjust this p-value for multiple correlated tests, and used the resulting set of 500 linear regression p-values for power calculations.

Applied analyses

As described previously, major depression has an estimated heritability of 34%,[61,62] with approximately one-third of this heritability thought due to common genetic variants;[68] and PTSD following trauma shows heritability estimates ranging from 30-46%,[63-67] with SNP-heritability estimated to be approximately 15%.[69] Despite strong evidence for an important role by common variants, however, presently identified robust common variant associations with depression and PTSD fall well short of accounting for the expected full contribution by common variants.

To more powerfully investigate common variant associations with these two psychiatric conditions, we applied GAMuT to genetic and phenotypic data collected as part of the Grady Trauma Project. A key objective of the GTP is to advance understanding of the role played by genetics in PTSD and related psychiatric disorders.[79] To this end, GTP staff recruit study participants from Grady Memorial Hospital in Atlanta, GA, approaching potential subjects in the waiting rooms of primary care clinics, obstetrics and gynecology clinics, as well as other clinics, and consenting participants. GTP participants are majority African-American, city-dwelling, and are of relatively low socioeconomic status. Participants provide an Oragene salivary sample for DNA extraction and subsequent genotyping on the Illumina HumanOmni1-Quad array. They are also assessed with regard to demographics, history of stressful experiences, and psychiatric symptoms. The examination of psychiatric symptoms includes completion of the PSS and the BDI. Recruitment and assessment of subjects for the GTP has been carried out according to protocols approved by the IRBs of Emory University School of Medicine and Grady Memorial Hospital.

PSS analyses

Applying standard GWAS quality control filters left 4,607 African-American subjects with good quality genotype data. Further removal of subjects not reporting having experienced a traumatic event, missing PSS data, or with incomplete covariate data (age, gender, and the top ten principal components to account for ancestry) yielded a final sample size of 3,826 subjects for the PSS analyses.

For this sample of 3,826 individuals, we identified 769,270 common variants (MAF > 5%) corresponding to 22,067 autosomal genes (we assigned SNPs to genes using the Illumina annotation files). We then dropped 'small' genes, which we defined as having fewer than 5 common variants, resulting in a set of 19,609 'big' genes and a total of 765,580 corresponding SNPs.

As described previously, the PSS consists of 17 questionnaire items designed to assess PTSD symptoms, which can be analyzed all together or split into three non-overlapping subscales, including subscales assessing the re-experiencing of trauma (PSS Intrusive; 5 items), avoidance-numbing (PSS AvoidNumb; 7 items), and hyperarousal (PSS Hyperarousal; 5 items).[8,9] We performed separate analyses for the PSS overall and also for each of the three PSS subscales.

We applied GAMuT with a linear kernel to model genotype similarity, and performed both unweighted analyses and weighted analyses. For the weighted analyses, we employed weights based on variants' MAF, with rarer variants carrying greater weight, as in the simulated data analyses. In addition, we employed weights based on each SNP's reported association (log odds ratio) with particular psychiatric disorders as estimated from external and independent large-scale GWAS of MDD, bipolar disorder, and schizophrenia; these estimates are available from the Psychiatric Genomics Consortium.[19,80,81] Results from comparably large-scale GWAS of PTSD were not available at the time of these analyses. Using external weights derived from

association studies of psychiatric disorders that are different from the conditions we were examining likely has utility, considering studies that have demonstrated sizable genetic overlap between multiple psychiatric disorders.[82] To model phenotypic similarity in the PSS responses, we performed GAMuT analyses both with a projection matrix and with a linear kernel. For comparison with GAMuT, we applied SNP-based linear regression and gene-based KMR to the cumulative, univariate PSS, PSS Intrusive, PSS AvoidNumb, and PSS Hyperarousal scores. For KMR, we utilized the same linear kernels to model genotypic similarity and applied the same genotype weighting schemes as used in the GAMuT analyses.

Prior to GAMuT and KMR analyses, we controlled for age, gender and the top 10 genetic PCs (which capture ancestry) by regressing the relevant phenotype (each questionnaire item in the case of GAMuT; cumulative scores in the case of KMR) onto these variables, then extracting the regression residuals and using these as the phenotypes to be analyzed by GAMuT and KMR. For the linear regression analyses, we controlled for these same variables by including them as covariates within the regression model.

BDI analyses

We also utilized GAMuT to examine associations of common genetic variants in gene regions with phenotypes assessed using the 21-item BDI questionnaire. For these analyses, we took the group of 3,826 subjects from the PSS analyses and limited it to subjects with complete BDI data, leaving 3,520 subjects, and we considered the same 19,609 genes as mentioned above containing 765,580 SNPs. We then applied GAMuT in the same fashion as described for the PSS analyses, using both a projection matrix and linear kernel for the phenotype and employing the same genotype weighting schemes under a linear genotype kernel. For comparison with GAMuT, we performed SNP-based linear regression and gene-based KMR on the univariate, cumulative BDI score.

Multiple testing differences and correction

As noted above, the PSS and BDI analyses using GAMuT considered 19,609 genes and 765,580 common variants for analysis. Analyses employing no genotype weights and analyses employing MAF-based weights involved association testing for all of these 19,609 genes. However, when employing the external PGC GWAS-derived weights, we observed that not all of the 765,580 common variants available for analyses were present within the PGC GWAS results, thus necessitating that analyses utilizing these external weights include fewer SNPs and corresponding genes than the analyses using MAF-based weights or no weights. Specifically, GAMuT analyses using PGC MDD weights involved 16,716 genes containing 469,582 SNPs; analyses using PGC bipolar disorder weights involved 16,761 genes containing 586,505 SNPs; and analyses using PGC schizophrenia weights involved 18,067 genes containing 661,879 SNPs.

As stated above, the univariate KMR analyses employed the same genotyping weighting schemes as used for GAMuT, and therefore tested the exact same genes as tested in the GAMuT analyses. For standard linear regression, we individually tested 775,255 common variants for association with each cumulative phenotype.

Since GAMuT and KMR analyze genes whereas linear regression analyzes SNPs, the multiple-testing adjusted significance thresholds differed for the former and latter approaches. For each GAMuT and KMR analysis, we used a Bonferroni correction procedure to establish a study-wise significance threshold, calculating this threshold as 0.05 divided by the number of genes analyzed. Thus, the study-wise significance threshold differed depending on the particular genotype weights used, ranging from a threshold of $0.05/16{,}716 = 2.99 \times 10^{-6}$ for PGC MDD weights to $0.05/19{,}609 = 2.55 \times 10^{-6}$ for MAF-based weights and no weights. For all GAMuT and KMR analyses we selected $p < 1 \times 10^{-4}$ as a suggestive significance threshold. For SNP-based linear regression, which involved testing 775,255 SNPs across the genome, we used a

study-wise significance threshold of 0.05/775,255 = 6.45×10^{-8} and a suggestive significance threshold of $p < 1 \times 10^{-6}$. We note that for these linear regression analyses we could have used the standard GWAS significance threshold of 5×10^{-8}, but we decided against this given that this standard threshold is more conservative than a Bonferroni correction based on the number of SNPs tested.

RESULTS

Simulated data analyses

Type I error

Application of GAMuT to 10,000 null simulated datasets, for which none of the 127 SNPs within *LRFN5* had a causal effect on any of the 21 BDI items, revealed that GAMuT properly controls Type I error. This was observed for simulated datasets of both 1,000 and 2,500 subjects. Proper control of Type I error was also observed for univariate KMR and linear regression.

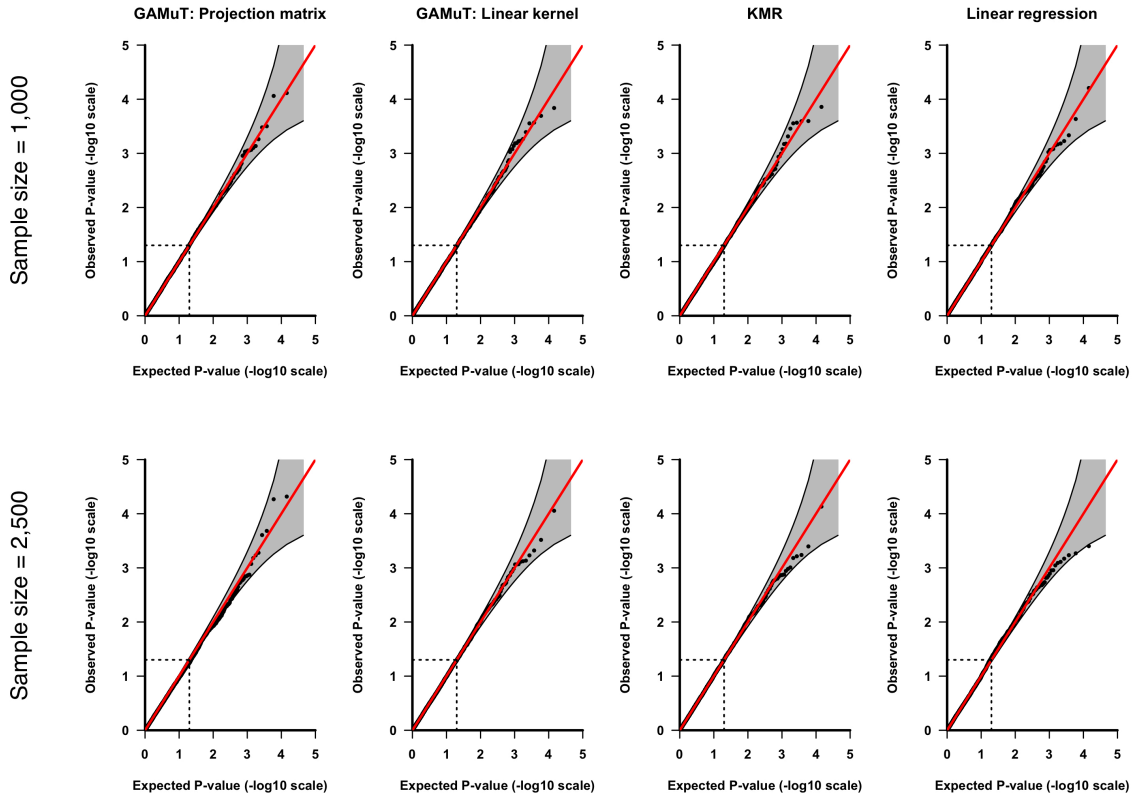
Table 3.1 presents empirical Type I error rates for GAMuT, as well as for KMR and linear regression. As an example of GAMuT's success in controlling Type I error, we can consider alpha (α) level = 0.001. For this α, a statistical test that properly controls Type I error should generate significant p-values ($p < 0.001$) approximately 0.1% of the time when applied to null datasets. This indeed is what we observed for GAMuT when setting $\alpha = 0.001$: GAMuT using a projection matrix for modeling phenotypic similarity yields 12/10,000 = 0.12% significant results when analyzing 1,000 simulated samples and 8/10,000 = 0.08% significant results when analyzing 2,500 samples; while GAMuT using a linear kernel for modeling phenotypic similarity produces 14/10,000 = 0.14% significant findings for simulated datasets of 1,000 samples and 10/10,000 = 0.1% significant results for datasets of 2,500 samples.

The quantile-quantile (QQ) plots shown in **Figure 3.2**, which plot the -log10 transformation of the 10,000 observed p-values against 10,000 p-values expected from a null distribution, also demonstrate proper Type I error control by GAMuT, KMR and linear regression. Inadequate Type I error control would result in an obvious departure of points from the diagonal red line (which has slope = 1 and thus is the line defined by observed p-values equaling expected p-values), particularly beginning at larger p-values (those points relatively closer to the bottom left corner of the plots). Such worrisome deviation is not observed for GAMuT, nor for KMR and linear regression.

Table 3.1. Empirical Type I error rates are presented for GAMuT (with projection matrix or linear kernel for modeling phenotypic similarity), univariate KMR and linear regression, for different combinations of sample size and significance (α) level. Error rates are calculated as the proportion of p-values less than the specified significance threshold given 10,000 null simulations. All GAMuT and KMR analyses used a weighted linear genotype kernel, with weights based on sample MAF. GAMuT, KMR and linear regression properly control Type I error across all scenarios tested.

	Sample Size = 1,000			**Sample Size = 2,500**		
	α = 0.05	**α = 0.01**	**α = 0.001**	**α = 0.05**	**α = 0.01**	**α = 0.001**
GAMuT: Projection Matrix	0.0506	0.0095	0.0012	0.0445	0.0080	0.0008
GAMuT: Linear Kernel	0.0480	0.0096	0.0014	0.0492	0.0105	0.0010
KMR	0.0491	0.0102	0.0010	0.0500	0.0107	0.0006
Linear Regression	0.0508	0.0111	0.0011	0.0533	0.0104	0.0007

Figure 3.2. Quantile-quantile (QQ) plots of p-values resulting from applying GAMuT (with projection matrix or linear kernel for modeling phenotypic similarity), univariate KMR, and standard linear regression to 10,000 simulated null data sets of either 1,000 (top) or 2,500 (bottom) samples. GAMuT and KMR analyses used a weighted linear genotype kernel, with weights based on sample MAF.



Power

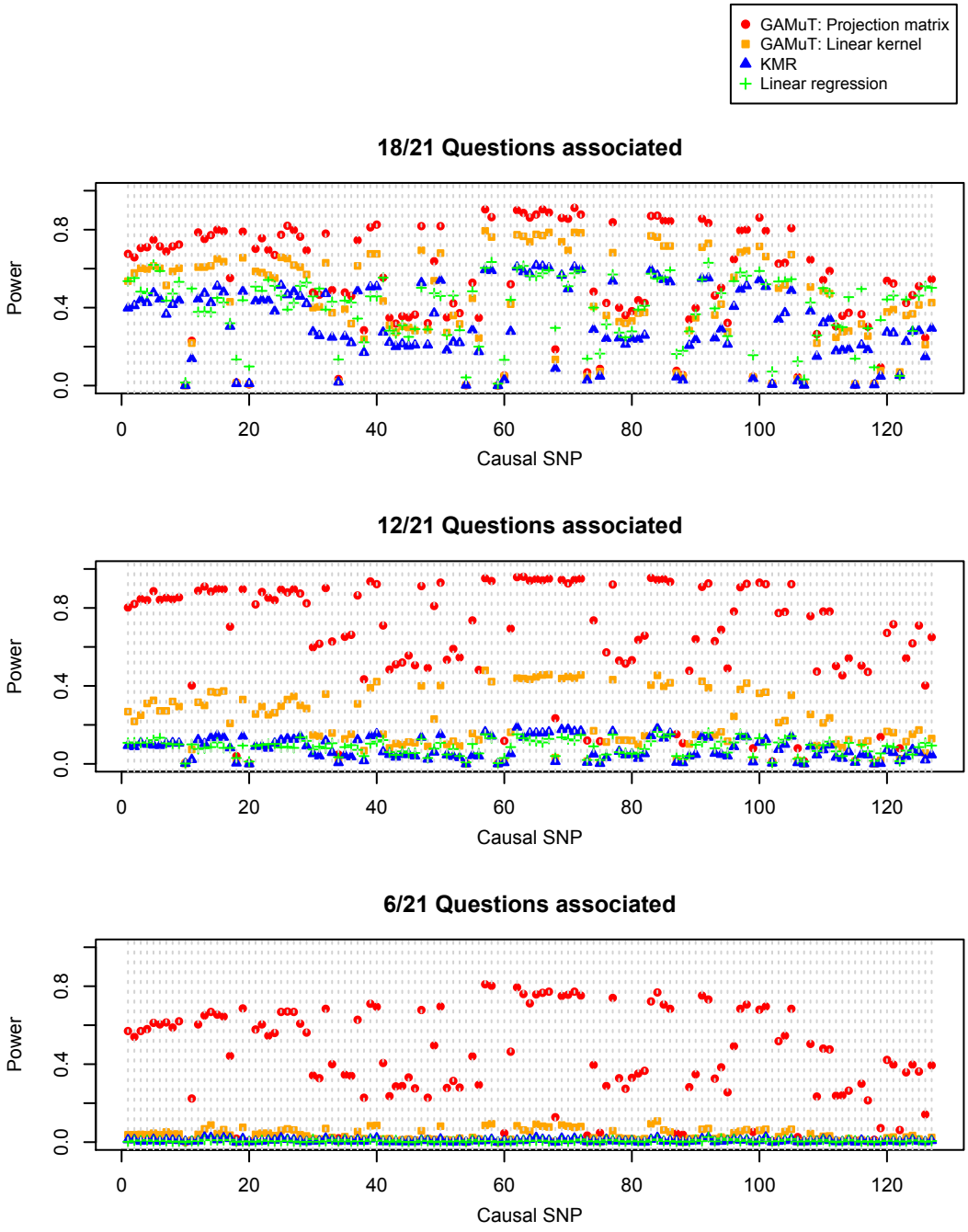
To evaluate the power of GAMuT to detect causal associations between common variants in a gene and multivariate psychiatric phenotypes, relative to the power of standard univariate approaches, we applied GAMuT, KMR and linear regression to datasets simulated under the variety of causal scenarios corresponding to each unique combination of causal SNP (127 SNPs within *LRFN5*, set as causal one at a time) and proportion of BDI questionnaire items affected by the causal SNP (18/21, 12/21 or 6/21). Each dataset included simulated genetic and BDI data for 2,500 subjects. **Figure 3.3** plots power for each of these unique causal

scenarios, with power defined as the proportion of p-values < 0.001 based on 500 simulated data analyses for the unique causal scenario. In this figure, the particular SNP set as causal (out of 127 SNPs) is noted on the x-axis, with SNPs ordered by genomic position and numerically labeled to correspond to the ordering and labeling of SNPs as presented in **Figure 3.1**. We remind the reader that while each of the 127 SNPs in *LRFN5* was set as causal in turn, only the 50 SNPs directly represented on the Illumina SNP array were included in association analyses.

Figure 3.3 shows that, across the various causal scenarios considered, GAMuT has a clear tendency to outperform the univariate approaches of KMR and linear regression with respect to power. The power differential favoring GAMuT is particularly pronounced for scenarios in which the causal SNP is associated with approximately half (12/21) or fewer of the 21 BDI items, and is also more evident when applying GAMuT with a projection matrix for modeling phenotypic similarity. When the causal SNP affects 12 of the 21 BDI items, univariate KMR and linear regression both show less than 20% power for detecting all 127 causal SNPs; whereas GAMuT employing a projection matrix to model phenotypic similarity observes greater than 80% power to detect 55 of the causal SNPs and greater than 50% power to detect 96 causal SNPs. When only 6 of the 21 BDI items are affected by the causal SNP, the univariate approaches have almost zero power for detecting all 127 causal SNPs, while GAMuT with a projection matrix maintains greater than 50% power for detecting 57 of the causal SNPs.

These results demonstrate the benefits of using the multivariate GAMuT framework to examine associations of common genetic variants with psychiatric phenotypes such as those assessed by the BDI. Such phenotypes are correlated but may be heterogeneous with respect to risk factors like genetic variants, and therefore aggregating these multivariate phenotypes into a single score for analysis using univariate methods can greatly reduce power for detecting associations.

Figure 3.3 (next page). Power for GAMuT (with projection matrix or linear kernel for modeling phenotypic similarity), univariate KMR, and standard linear regression, across various causal scenarios defined by unique combinations of causal SNP (127 SNPs within *LRFN5*) and proportion of BDI questionnaire items affected by the causal SNP (18/21, 12/21 or 6/21 questions associated with the causal SNP). GAMuT and KMR used a weighted linear genotype kernel, with MAF-based weights. Simulated datasets had sample size of 2,500. Power was calculated as the proportion of p-values < 0.001, based on 500 simulated data analyses for the unique causal scenario. In the plots, the particular SNP set as causal is noted on the x-axis, with SNPs ordered by genomic position and numerically labeled to correspond to the ordering and labeling of SNPs as presented in **Figure 3.1**.


GAMuT: Projection matrix
GAMuT: Linear kernel
KMR
Linear regression
18/21 Questions associated
Power
0.0
0.4
0.8
0
20
40
60
80
100
120
Causal SNP
12/21 Questions associated
Power
0.0
0.4
0.8
0
20
40
60
80
100
120
Causal SNP
6/21 Questions associated
Power
0.0
0.4
0.8
0
20
40
60
80
100
120
Causal SNP

Applied analyses

PSS

We used GAMuT to examine associations of common genetic variants in gene regions with PSS questionnaire items for a sample of 3,826 African-American individuals who were participants in the GTP. To model phenotypic similarity for GAMuT, we employed a projection matrix as well as a linear kernel. We also performed analyses using univariate KMR and linear regression. We modeled genotypic similarity in the same way for both GAMuT and KMR (linear genotype kernel) and employed the same genotype weighting schemes for both approaches. In addition to analyzing the PSS overall (17 items), we performed analyses for the three non-overlapping subscales: PSS Intrusive (5 items), PSS AvoidNumb (7 items), and PSS Hyperarousal (5 items).

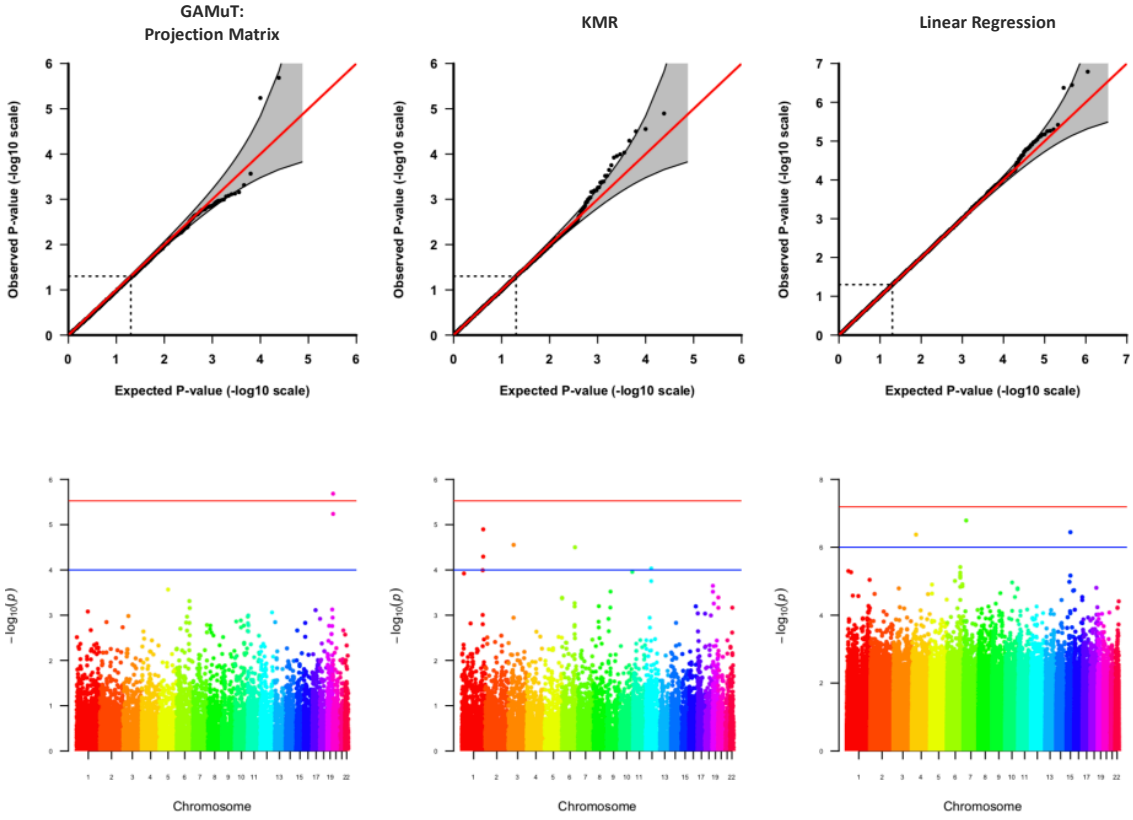
We provide QQ and Manhattan plots for all GAMuT, KMR, and linear regression analyses of overall PSS, PSS Intrusive, PSS AvoidNumb, and PSS Hyperarousal in **Supplementary Figures 3.1-3.4**, respectively. We also present genes identified by GAMuT, KMR or linear regression to be associated with PSS or its subscales at study-wise or suggestive significance levels within **Supplementary Tables 3.1-3.4**. The GAMuT analyses of PSS and its subscales identified one gene that exceeded the study-wise significance threshold. In comparison, univariate KMR and linear regression found no genes or SNPs to be associated with PSS or its subscales at a study-wise significant level. GAMuT identified *SIRPA*, a gene on chromosome 20, as significantly associated with the PSS AvoidNumb subscale ($p = 2.07x10^{-6}$), when using a projection matrix to measure phenotypic similarity and genotype weights based on estimated log odds ratios from the PGC GWAS for bipolar disorder. The first column of **Figure 3.4** displays QQ and Manhattan plots for this particular GAMuT analysis. *SIRPA* also showed suggestive association in the analysis of PSS AvoidNumb using a linear kernel for phenotypic similarity and weights based on the PGC GWAS of bipolar disorder (see **Supplementary Table**

3.3), and in the analysis of overall PSS when using a projection matrix and PGC GWAS bipolar disorder weights (see **Supplementary Table 3.1**). To the best of our knowledge, there are no previous reports of *SIRPA* being associated with neuropsychiatric phenotypes. However, research shows that *SIRPA* has highly elevated expression levels in the brain.[83] In comparison, analysis of the cumulative PSS AvoidNumb score using KMR with PGC GWAS bipolar disorder weights did not identify *SIRPA* as even suggestively associated (**Figure 3.4**, middle column; **Supplementary Figure 3.3c**), and univariate linear regression identified no SNPs suggesting association within *SIRPA* or any other gene on chromosome 20 (**Figure 3.4, last column; Supplementary Figure 3.3d**).

To more fully examine the detection of *SIRPA* by GAMuT and not by standard univariate methods, we used KMR to analyze the associations between *SIRPA* and each of the 7 items comprising the PSS AvoidNumb subscale (i.e., we evaluated these 7 items one-by-one rather than collapsing them into a cumulative PSS AvoidNumb score). As in the above analyses, we weighted the genotype matrices by estimated log odds ratios from the PGC GWAS for bipolar disorder. These analyses identified *SIRPA* as associated with one PSS AvoidNumb item with $p = 6.37 \times 10^{-9}$ (the item "Have you persistently been making efforts to avoid thoughts or feelings associated with the event(s) we've talked about"), associated with 4 items with p-values ranging from approximately 0.001 to 0.10, and unassociated with 2 items based on p-values greater than 0.50. These findings suggest that *SIRPA* is associated with only a subset of the PSS AvoidNumb items. As we previously showed using simulated data, in such a situation, standard univariate methods using collapsed, cumulative phenotypes provide inadequate power for detecting genetic associations, while GAMuT maintains good power. The finding that SIRPA is associated with only a subset of PSS AvoidNumb items therefore provides an explanation for the failure of KMR using the cumulative PSS AvoidNumb score to identify *SIRPA*, while GAMuT succeeded in detecting this gene association.

As shown in **Supplementary Table 3.3** and **Figure 3.4**, the same GAMuT analysis that identified *SIRPA* as study-wise significant also found *PDYN*, located nearby *SIRPA* on chromosome 20, as having strongly suggestive evidence of association ($p = 5.78 \times 10^{-6}$). *PDYN* was also identified as suggestively significant in the analysis of PSS AvoidNumb with a projection matrix and no genotype weights (**Supplementary Table 3.3**). *PDYN* has been found to be associated with a variety of neuropsychiatric phenotypes, including mood disorders,[84] schizophrenia,[85,86] memory disorders,[87] epilepsy,[88] and substance use disorders,[84,89-93] and it shows expression that is restricted toward the brain.[83] Analysis of cumulative PSS AvoidNumb scores using KMR with PGC GWAS bipolar disorder weights did not identify *PDYN* as even suggestively associated with PSS AvoidNumb (**Figure 3.4,** middle column; **Supplementary Figure 3.3c**), and univariate linear regression identified no SNPs suggesting association within *PDYN* (**Figure 3.4,** last column; **Supplementary Figure 3.3d**).

Figure 3.4. QQ and Manhattan plots for GAMBITS, KMR, and linear regression analyses of PSS AvoidNumb. The GAMBITS analysis used a projection matrix to model phenotypic similarity and genotype weights derived from results of the PGC GWAS for bipolar disorder. The KMR analysis also used weights based on the PGC GWAS for bipolar disorder. In the Manhattan plots, the red line represents the study-wise significance threshold and the blue line represents the suggestive significance threshold. The study-wise significance thresholds for the GAMBITS and KMR analyses are based on a Bonferroni correction for 16,761 genes tested, while the study-wise significance threshold for the linear regression analysis is based on a Bonferroni correction for 775,255 SNPs tested. In the Manhattan plot for the GAMBITS results, the point exceeding the study-wise significance threshold is the $-\log_{10}$(p-value) for *SIRPA*, a gene on chromosome 20. These analyses used a sample of n = 3,826.



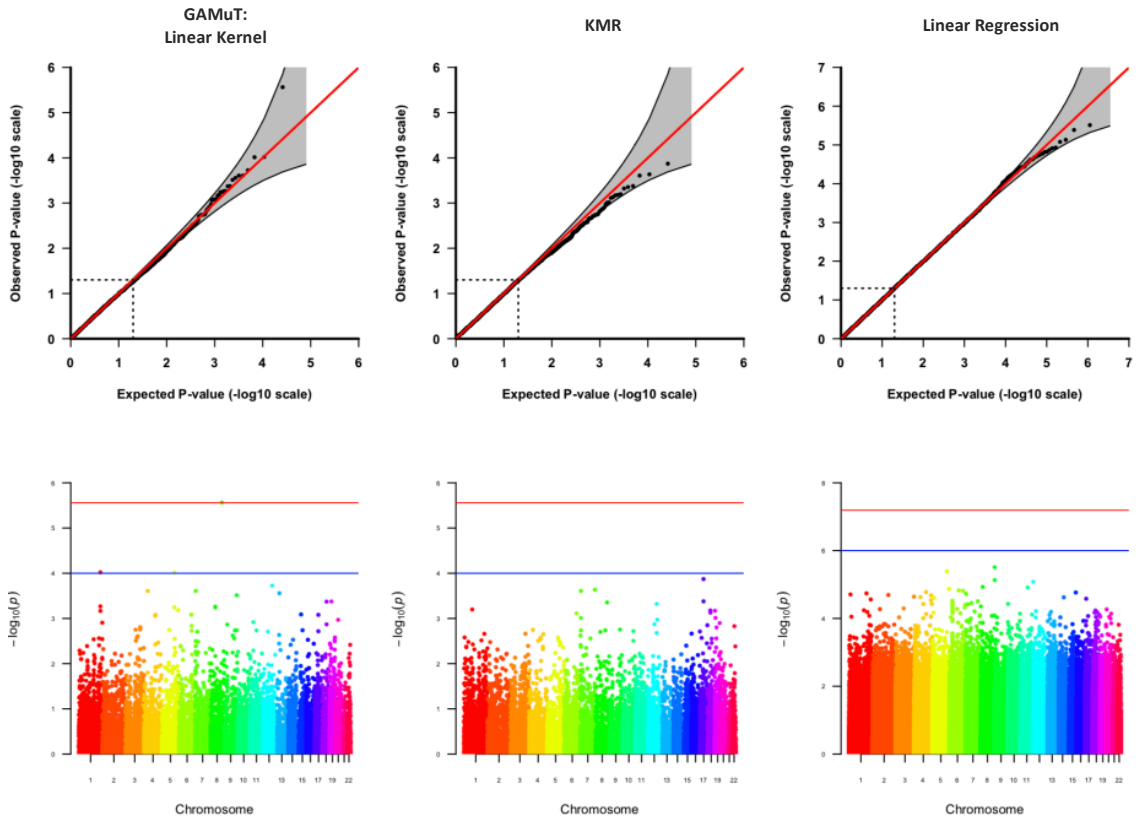
BDI

We also applied GAMuT to investigate associations of common genetic variants in gene regions with BDI questionnaire items. This was done using the subset of 3,520 African-American GTP participants with complete analysis BDI information. As with the PSS analyses,

GAMuT employed both a projection matrix and linear kernel to model the phenotypic similarity matrix, and a weighted or non-weighted linear kernel to model genotypic similarity (employing the same weighting schemes as used for the PSS analyses). We also performed univariate KMR and standard linear regression to compare with GAMuT.

We present QQ and Manhattan plots for all GAMuT, KMR, and linear regression analyses of BDI in **Supplementary Figure 3.5**. We also present the full listing of genes found by GAMuT, KMR or linear regression to be associated with BDI at study-wise or suggestive significance levels in **Supplementary Table 3.5**. The GAMuT analyses of BDI identified one gene exceeding study-wise significance, whereas univariate KMR and linear regression of BDI did not detect any study-wise significant genes or SNPs. GAMuT found *ZHX2*, on chromosome 8, to be significantly associated with BDI ($p = 2.73x10^{-6}$, study-wise significant after Bonferroni correction for 18,067 genes tested), when using a linear kernel to measure phenotypic similarity and genotype weights based on estimated log odds ratios from the PGC GWAS for schizophrenia. We present QQ and Manhattan plots for this particular analysis in the first column of **Figure 3.5**. *ZHX2* was also found to be highly suggestively associated with BDI when using a linear kernel for the phenotype and employing genotype weights based on the PGC GWAS of MDD. Previous research suggests a possible link between *ZHX2* and autism spectrum disorder.[94] In comparison with the GAMuT analyses, univariate KMR analyzing the cumulative BDI score did not identify *ZHX2* as having even suggestive association (**Figure 3.5,** middle column; **Supplementary Figure 3.5c**), and univariate linear regression revealed no SNPs suggestively associated with BDI within *ZHX2* or anywhere else across the genome (**Figure 3.5,** last column; **Supplementary Figure 3.5d**).

Figure 3.5. QQ and Manhattan plots for GAMuT, KMR, and linear regression analyses of BDI. The GAMuT analysis used a linear kernel to model phenotypic similarity and genotype weights derived from results of the PGC GWAS for schizophrenia. The KMR analysis also used weights based on the PGC GWAS for schizophrenia. In the Manhattan plots, the red line represents the study-wise significance threshold and the blue line represents the suggestive significance threshold. The study-wise significance thresholds for the GAMuT and KMR analyses are based on a Bonferroni correction for 18,067 genes tested, while the study-wise significance threshold for the linear regression analysis is based on a Bonferroni correction for 775,255 SNPs tested. In the Manhattan plot for the GAMuT results, the point exceeding the study-wise significance threshold is the $-\log_{10}$(p-value) for *ZHX2*, a gene on chromosome 8. These analyses used a sample of n = 3,520.



DISCUSSION

We have examined GAMuT, a multivariate method previously developed for tests of rare variant pleiotropy, as a potentially powerful approach for identifying common variant associations with multivariate psychiatric phenotypes such as those assessed using commonly administered psychiatric questionnaires like the PSS and BDI. Responses to items in these types of questionnaires are commonly summed or otherwise aggregated into a single, overall score, which is then analyzed using univariate techniques. However, for the realistic circumstance in which a genetic factor affects only a subset of the phenotypes assessed by the psychiatric questionnaire items, or affects the various phenotypes differently, the genetic effect may be very challenging to detect with univariate approaches that analyze a single, aggregated score.

By analyzing simulated SNP genotypes within the gene *LRFN5* and simulated responses for 21 BDI items, we have shown that GAMuT properly controls Type I error and that GAMuT's joint modeling of the multivariate BDI items offers substantially greater power for detecting genetic associations than standard univariate methods like KMR and linear regression which analyze a single, cumulative score. GAMuT's gain in power over univariate approaches was especially pronounced for scenarios in which only half of the questionnaire items or fewer were affected by the causal SNP.

We then applied GAMuT, univariate KMR and standard linear regression to data accumulated by the GTP to examine associations of common variants in gene regions with both the PSS and BDI. We employed various genotype weighting schemes, demonstrating that the GAMuT framework is able to incorporate prior biological information that may facilitate identifying genetic associations. In these applied analyses, GAMuT identified a strong association between the PSS subscale for avoidance-numbing (PSS AvoidNumb) and *SIRPA* (p

= $2.07x10^{-6}$), which is a gene on chromosome 20 that shows high expression in the brain.[83] GAMuT also found a strong association between the BDI and a gene on chromosome 8 called *ZHX2* ($p = 2.73x10^{-6}$), which previous research suggests might be associated with autism spectrum disorder.[94] In comparison, univariate KMR and linear regression did not identify the *SIRPA* gene or SNPs within *SIRPA* to be associated with PSS AvoidNumb, nor did they identify *ZHX2* or SNPs within it to be associated with BDI, at even suggestive levels. These two genes require further, independent investigation before making any conclusions about their role in PTSD and depression symptomology. These applied analyses demonstrate through use of real-world data the capacity for GAMuT to detect genotype-phenotype associations that would be missed using standard cumulative univariate approaches.

As a powerful, computationally efficient method for detecting genetic associations with multivariate phenotype data, GAMuT has potential to facilitate the identification of robust common variant associations with psychiatric phenotypes, which have often evaded detection by traditional analytic approaches. It is worth noting that the GAMuT framework is also amenable to analyses of rare genetic variants and other omics data types like methylation data, and we expect the findings from this study to generalize to these other variant classes and data types.

GAMuT is well-positioned to facilitate the types of investigations and analyses promoted by NIMH's RDoC. As genetic studies of psychiatric phenotypes increasingly shift to the study of high-dimensional symptom data, in greater alignment with RDoC, multivariate methods like GAMuT are expected to grow in importance and use within the domain of psychiatric genetics research. The findings from this study support GAMuT as a method which can help meet this growing need for powerful multivariate analysis techniques.

WEB RESOURCES

R script for implementing the GAMuT method is available at https://github.com/epstein-software

SUPPLEMENT

Supplementary Table 3.1. Full GAMuT results for PSS overall (17 items). Genes with $p < 1x10^{-4}$ identified in the GAMuT and KMR analyses are shown (p-values in bold). GAMuT and KMR utilized a linear genotype kernel (possibly weighted) for all analyses. For this PSS overall phenotype, standard linear regression identified no SNPs exceeding genome-wide significance, and only two SNPs with suggestive significance ($p < 1x10^{-6}$): one SNP assigned to *PRR15* ($p = 1.63x10^{-7}$) and one SNP assigned to *MBTPS1* ($p = 6.68x10^{-7}$). PGC MDD, PGC BPD, PGC SZ denote weights based on log odds ratios from the Psychiatric Genomics Consortium GWAS of major depressive disorder, bipolar disorder, and schizophrenia, respectively; MAF-based = weights based on minor allele frequencies of variants calculated using the Grady Trauma Project genotype data.

Gene	Chr	Number of variants	Genotype weights	GAMuT Phenotypic Similarity Matrix		KMR	Linear Regression (minimum p-value of SNP in gene)
				Projection Matrix	Linear Kernel		
TNFAIP3	6	296	PGC BPD	$1.54x10^{-4}$	$1.32x10^{-4}$	**$1.23x10^{-5}$**	$3.60x10^{-6}$
PSEN2	1	36	PGC BPD	$2.32x10^{-2}$	$1.42x10^{-4}$	**$1.30x10^{-5}$**	$5.18x10^{-5}$
ADAD1	4	25	PGC SZ	**$1.34x10^{-5}$**	$3.76x10^{-2}$	$9.22x10^{-2}$	$1.16x10^{-2}$
		17	PGC BPD	**$4.71x10^{-5}$**	$6.75x10^{-2}$	$9.78x10^{-2}$	$1.16x10^{-2}$
		57	MAF-based	**$8.73x10^{-5}$**	$1.09x10^{-1}$	$4.86x10^{-1}$	$1.16x10^{-2}$
ZNF410	14	6	PGC MDD	$1.10x10^{-1}$	$4.24x10^{-4}$	**$1.47x10^{-5}$**	$1.05x10^{-5}$
		7	PGC SZ	$2.15x10^{-1}$	$1.79x10^{-3}$	**$7.36x10^{-5}$**	$1.05x10^{-5}$
NFIC	19	33	PGC BPD	$3.43x10^{-4}$	**$5.61x10^{-5}$**	**$1.90x10^{-5}$**	$1.38x10^{-5}$
BRUNOL5	19	38	PGC BPD	$3.04x10^{-3}$	**$7.45x10^{-5}$**	**$2.13x10^{-5}$**	$1.38x10^{-5}$
CABC1	1	24	PGC BPD	$7.03x10^{-3}$	$1.33x10^{-4}$	**$2.63x10^{-5}$**	$5.18x10^{-5}$
SMOX	20	76	PGC MDD	$1.09x10^{-2}$	**$3.11x10^{-5}$**	$8.22x10^{-4}$	$7.85x10^{-4}$
RNF24	20	65	PGC MDD	$1.10x10^{-2}$	**$3.12x10^{-5}$**	$8.19x10^{-4}$	$7.85x10^{-4}$
SHB	9	120	PGC MDD	$2.36x10^{-3}$	**$3.69x10^{-5}$**	$3.91x10^{-4}$	$1.67x10^{-3}$

OR2S2	9	56	No weights	2.72×10^{-1}	**4.37×10^{-5}**	**8.74×10^{-5}**	5.54×10^{-5}
TSPAN19	12	10	PGC MDD	1.34×10^{-1}	1.40×10^{-3}	**4.46×10^{-5}**	4.55×10^{-5}
		13	PGC BPD	8.87×10^{-2}	9.50×10^{-4}	**5.05×10^{-5}**	4.55×10^{-5}
LOC100289485	12	9	PGC MDD	1.35×10^{-1}	1.40×10^{-3}	**4.51×10^{-5}**	4.55×10^{-5}
		12	PGC BPD	9.00×10^{-2}	9.51×10^{-4}	**5.03×10^{-5}**	4.55×10^{-5}
EPHB2	1	87	PGC BPD	6.82×10^{-2}	5.73×10^{-4}	**4.65×10^{-5}**	6.05×10^{-5}
LOC256483	1	10	PGC SZ	1.18×10^{-2}	1.47×10^{-3}	**5.27×10^{-5}**	4.43×10^{-5}
LTBP4	19	15	PGC MDD	**5.31×10^{-5}**	1.37×10^{-1}	4.29×10^{-1}	5.46×10^{-2}
UROS	10	10	PGC BPD	2.25×10^{-2}	1.24×10^{-4}	**5.47×10^{-5}**	8.77×10^{-6}
		12	PGC SZ	1.63×10^{-2}	1.36×10^{-4}	**5.84×10^{-5}**	8.77×10^{-6}
CCDC66	3	22	PGC BPD	9.74×10^{-3}	**5.52×10^{-5}**	2.59×10^{-4}	1.69×10^{-4}
SIPA1L3	19	54	PGC SZ	**5.70×10^{-5}**	9.39×10^{-2}	7.45×10^{-1}	4.46×10^{-2}
LRIG3	12	301	PGC BPD	3.21×10^{-1}	1.14×10^{-4}	**6.63×10^{-5}**	2.82×10^{-4}
GPR151	5	18	PGC MDD	**6.83×10^{-5}**	3.81×10^{-2}	4.59×10^{-2}	6.88×10^{-3}
LOC286177	8	64	PGC BPD	2.11×10^{-1}	1.48×10^{-3}	**6.84×10^{-5}**	5.26×10^{-6}
B4GALT7	5	5	PGC BPD	**7.32×10^{-5}**	1.65×10^{-1}	9.82×10^{-1}	6.98×10^{-1}
PRG1	19	8	PGC MDD	**7.56×10^{-5}**	7.08×10^{-2}	4.77×10^{-2}	3.97×10^{-2}
		9	PGC BPD	**8.48×10^{-5}**	7.19×10^{-2}	5.00×10^{-2}	3.97×10^{-2}
PSG9	19	8	PGC MDD	**7.56×10^{-5}**	7.08×10^{-2}	4.77×10^{-2}	3.97×10^{-2}
		9	PGC BPD	**8.48×10^{-5}**	7.19×10^{-2}	5.00×10^{-2}	3.97×10^{-2}
SIRPA	20	73	PGC BPD	**7.73×10^{-5}**	3.20×10^{-4}	3.42×10^{-3}	9.50×10^{-4}
PKIB	6	49	PGC MDD	**8.62×10^{-5}**	3.67×10^{-2}	3.09×10^{-1}	1.33×10^{-2}
FABP7	6	21	PGC MDD	**8.84×10^{-5}**	3.68×10^{-2}	3.04×10^{-1}	1.33×10^{-2}
IL2	4	161	MAF-based	**9.57×10^{-5}**	6.23×10^{-2}	3.94×10^{-1}	3.29×10^{-2}
PSCA	8	16	PGC BPD	**9.96×10^{-5}**	1.62×10^{-1}	7.17×10^{-1}	4.00×10^{-1}

Supplementary Table 3.2. Full GAMuT results for PSS Intrusive (5 items). Genes with p < $1x10^{-4}$ identified in the GAMuT and KMR analyses are shown (p-values in bold). GAMuT and KMR utilized a linear genotype kernel (possibly weighted) for all analyses. For this PSS Intrusive phenotype, standard linear regression identified no SNPs of suggestive significance (p < $1x10^{-6}$). PGC MDD, PGC BPD, PGC SZ denote weights based on log odds ratios from the Psychiatric Genomics Consortium GWAS of major depressive disorder, bipolar disorder, and schizophrenia, respectively; MAF-based = weights based on minor allele frequencies of variants calculated using the Grady Trauma Project genotype data.

Gene	Chr	Number of variants	Genotype weights	GAMuT Phenotypic Similarity Matrix		KMR	Linear Regression (minimum p-value of SNP in gene)
				Projection Matrix	Linear Kernel		
SF3B3	16	18	PGC MDD	$1.32x10^{-3}$	**$2.72x10^{-5}$**	$1.87x10^{-4}$	$1.58x10^{-4}$
TNFAIP3	6	296	PGC BPD	$4.81x10^{-4}$	$1.02x10^{-4}$	**$3.54x10^{-5}$**	$4.16x10^{-5}$
CCDC66	3	22	PGC BPD	$3.68x10^{-3}$	**$5.65x10^{-5}$**	**$9.76x10^{-5}$**	$3.31x10^{-4}$
SNX1	15	5	PGC BPD	$8.68x10^{-4}$	**$6.14x10^{-5}$**	$1.69x10^{-4}$	$1.64x10^{-4}$
MAST2	1	32	No weights	$9.01x10^{-3}$	$3.92x10^{-3}$	**$6.62x10^{-5}$**	$7.25x10^{-5}$
SMOX	20	76	PGC MDD	$2.33x10^{-3}$	**$7.39x10^{-5}$**	$8.62x10^{-4}$	$8.12x10^{-4}$
RNF24	20	65	PGC MDD	$2.34x10^{-3}$	**$7.40x10^{-5}$**	$8.62x10^{-4}$	$8.12x10^{-4}$
LRIG3	12	301	PGC BPD	$4.55x10^{-2}$	$1.22x10^{-3}$	**$7.93x10^{-5}$**	$5.18x10^{-4}$
SLC22A5	5	49	No weights	$2.45x10^{-4}$	**$9.48x10^{-5}$**	$1.05x10^{-4}$	$2.02x10^{-5}$
FAM13A	4	59	PGC MDD	$7.37x10^{-4}$	**$9.87x10^{-5}$**	$4.29x10^{-3}$	$1.73x10^{-3}$

Supplementary Table 3.3. Full GAMuT results for PSS AvoidNumb (7 items). Genes with p < $1x10^{-4}$ identified in the GAMuT and KMR analyses are shown (p-values in bold). GAMuT and KMR utilized a linear genotype kernel (possibly weighted) for all analyses. For this PSS AvoidNumb phenotype, standard linear regression identified no SNPs exceeding genome-wide significance, and only three SNPs with suggestive significance (p < $1x10^{-6}$): one SNP assigned to *PRR15* (p = $1.63x10^{-7}$), one SNP assigned to *RORA* (p = $3.59x10^{-7}$), and one assigned to *APBB2* (p = $4.23x10^{-7}$). PGC MDD, PGC BPD, PGC SZ denote weights based on log odds ratios from the Psychiatric Genomics Consortium GWAS of major depressive disorder, bipolar disorder, and schizophrenia, respectively; MAF-based = weights based on minor allele frequencies of variants calculated using the Grady Trauma Project genotype data. ** denotes that the result exceeds the study-wise significance threshold.

Gene	Chr	Number of variants	Genotype weights	GAMuT Phenotypic Similarity Matrix		KMR	Linear Regression (minimum p-value of SNP in gene)
				Projection Matrix	Linear Kernel		
SIRPA	20	73	PGC BPD	**$2.07x10^{-6}$ ****	**$7.74x10^{-5}$**	$4.06x10^{-4}$	$2.11x10^{-4}$
PDYN	20	71	PGC BPD	**$5.78x10^{-6}$**	$1.31x10^{-4}$	$6.87x10^{-4}$	$2.11x10^{-4}$
		90	No weights	**$6.78x10^{-5}$**	$5.74x10^{-4}$	$5.61x10^{-3}$	$2.11x10^{-4}$
CABC1	1	24	PGC BPD	$5.35x10^{-3}$	$1.25x10^{-4}$	**$1.27x10^{-5}$**	$5.34x10^{-5}$
CCL4	17	7	MAF-based	**$1.98x10^{-5}$**	$6.97x10^{-2}$	$6.38x10^{-1}$	$2.18x10^{-1}$
		7	No weights	**$4.63x10^{-5}$**	$6.40x10^{-2}$	$6.44x10^{-1}$	$2.18x10^{-1}$
FOXC1	6	79	PGC MDD	$5.62x10^{-3}$	**$1.98x10^{-5}$**	$2.34x10^{-4}$	$1.68x10^{-4}$
		99	PGC BPD	$5.51x10^{-3}$	**$2.46x10^{-5}$**	$4.12x10^{-4}$	$1.68x10^{-4}$
FOXF2	6	80	PGC MDD	$5.72x10^{-3}$	**$2.20x10^{-5}$**	$2.43x10^{-4}$	$1.68x10^{-4}$
		102	PGC BPD	$5.63x10^{-3}$	**$2.53x10^{-5}$**	$4.25x10^{-4}$	$1.68x10^{-4}$
CCDC66	3	22	PGC BPD	$1.04x10^{-3}$	**$8.67x10^{-5}$**	**$2.80x10^{-5}$**	$1.62x10^{-5}$
		18	PGC MDD	$2.14x10^{-2}$	$4.34x10^{-4}$	**$7.73x10^{-5}$**	$1.62x10^{-5}$
ZNF410	14	6	PGC MDD	$1.25x10^{-2}$	$1.65x10^{-4}$	**$2.95x10^{-5}$**	$3.57x10^{-5}$
TNFAIP3	6	296	PGC BPD	$6.96x10^{-4}$	$8.95x10^{-4}$	**$3.16x10^{-5}$**	$5.60x10^{-6}$
SHB	9	120	PGC MDD	$4.42x10^{-3}$	**$3.34x10^{-5}$**	**$3.48x10^{-5}$**	$4.57x10^{-5}$
		159	MAF-based	$1.94x10^{-1}$	$1.13x10^{-3}$	**$5.96x10^{-5}$**	$4.57x10^{-5}$
		159	No weights	$1.45x10^{-1}$	$1.29x10^{-3}$	**$7.28x10^{-5}$**	$4.57x10^{-5}$
		146	PGC SZ	$4.72x10^{-2}$	$4.15x10^{-4}$	**$7.48x10^{-5}$**	$4.57x10^{-5}$

PSEN2	1	36	PGC BPD	$4.95x10^{-2}$	$2.86x10^{-4}$	**$5.07x10^{-5}$**	$5.34x10^{-5}$
TBX5	12	109	PGC MDD	**$7.38x10^{-5}$**	$4.60x10^{-3}$	$1.40x10^{-1}$	$1.10x10^{-2}$
PROX1	1	267	PGC SZ	$5.33x10^{-3}$	**$8.64x10^{-5}$**	$7.48x10^{-4}$	$1.37x10^{-4}$
RPL3	22	31	MAF-based	$4.06x10^{-3}$	**$8.96x10^{-5}$**	$1.00x10^{-3}$	$2.76x10^{-4}$
UROS	10	10	PGC BPD	$5.45x10^{-3}$	**$9.01x10^{-5}$**	$1.11x10^{-4}$	$1.61x10^{-5}$
RBM17	10	37	No weights	$2.58x10^{-3}$	**$9.22x10^{-5}$**	$7.79x10^{-4}$	$2.38x10^{-3}$
LRIG3	12	301	PGC BPD	$8.30x10^{-2}$	$6.91x10^{-4}$	**$9.27x10^{-5}$**	$5.55x10^{-5}$
NAP1L5	4	9	MAF-based	$4.29x10^{-3}$	**$9.67x10^{-5}$**	$1.12x10^{-3}$	$3.27x10^{-3}$

Supplementary Table 3.4. Full GAMuT results for PSS Hyperarousal (5 items). Genes with $p < 1x10^{-4}$ identified in the GAMuT and KMR analyses are shown (p-values in bold). GAMuT and KMR utilized a linear genotype kernel (possibly weighted) for all analyses. For this PSS Hyperarousal phenotype, standard linear regression identified no SNPs of suggestive significance ($p < 1x10^{-6}$). PGC MDD, PGC BPD, PGC SZ denote weights based on log odds ratios from the Psychiatric Genomics Consortium GWAS of major depressive disorder, bipolar disorder, and schizophrenia, respectively; MAF-based = weights based on minor allele frequencies of variants calculated using the Grady Trauma Project genotype data.

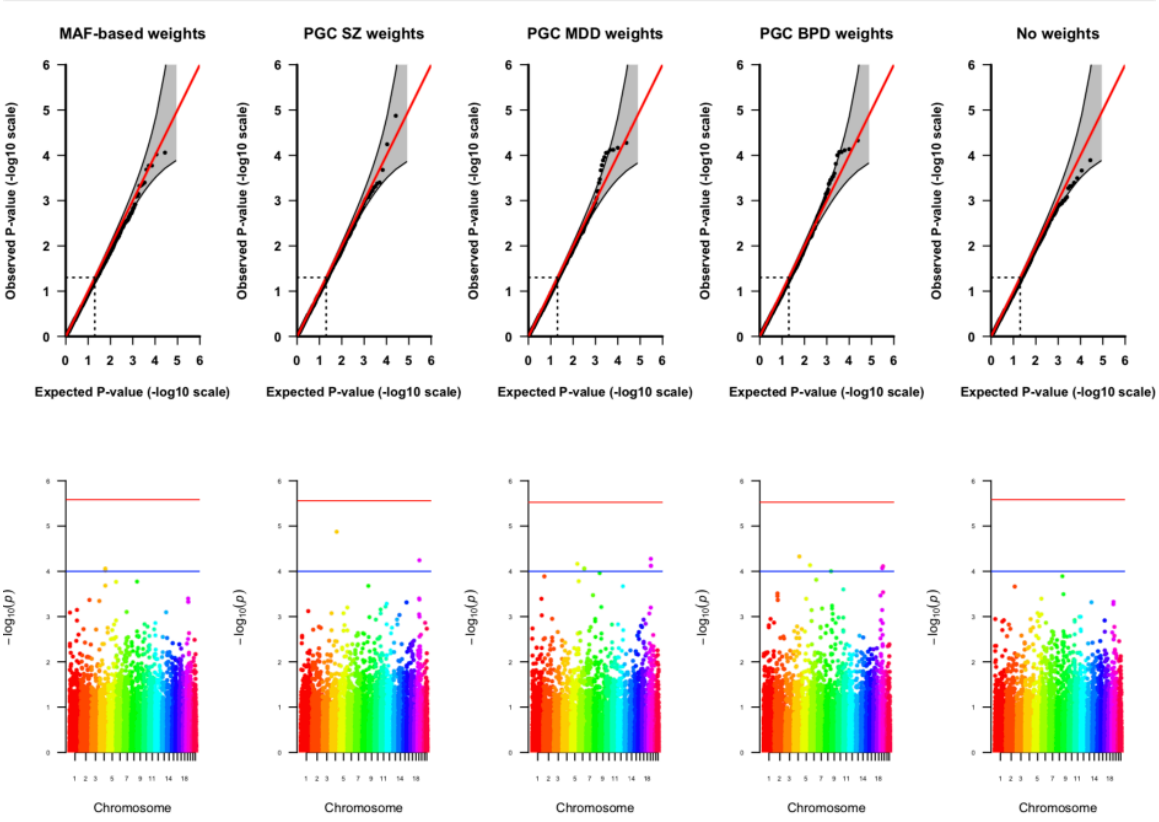
Gene	Chr	Number of variants	Genotype weights	GAMuT Phenotypic Similarity Matrix		KMR	Linear Regression (minimum p-value of SNP in gene)
				Projection Matrix	Linear Kernel		
GLTSCR1	19	35	PGC SZ	$8.59x10^{-3}$	$2.63x10^{-4}$	**$1.38x10^{-5}$**	$2.61x10^{-4}$
SMOX	20	76	PGC MDD	$6.88x10^{-4}$	**$1.79x10^{-5}$**	**$5.57x10^{-5}$**	$5.31x10^{-5}$
RNF24	20	65	PGC MDD	$6.94x10^{-4}$	**$1.80x10^{-5}$**	**$5.58x10^{-5}$**	$5.31x10^{-5}$
GPC1	2	103	No weights	$7.31x10^{-4}$	**$3.12x10^{-5}$**	$1.03x10^{-3}$	$1.79x10^{-3}$
NFIC	19	33	PGC BPD	$3.05x10^{-4}$	$1.09x10^{-4}$	**$3.21x10^{-5}$**	$2.53x10^{-5}$
BRUNOL5	19	38	PGC BPD	$1.12x10^{-3}$	$1.04x10^{-4}$	**$3.47x10^{-5}$**	$2.53x10^{-5}$
OTOS	2	104	No weights	$9.31x10^{-4}$	**$3.58x10^{-5}$**	$1.16x10^{-3}$	$1.79x10^{-3}$
LOC100287311	19	7	PGC MDD	$5.31x10^{-4}$	**$5.85x10^{-5}$**	$4.79x10^{-3}$	$4.33x10^{-3}$
PDE1C	7	287	PGC SZ	$1.05x10^{-2}$	$1.52x10^{-3}$	**$6.61x10^{-5}$**	$3.17x10^{-4}$
OR2S2	9	56	No weights	$4.05x10^{-2}$	**$6.72x10^{-5}$**	$1.19x10^{-4}$	$2.53x10^{-5}$
CAMTA1	1	231	PGC MDD	**$7.16x10^{-5}$**	$4.87x10^{-2}$	$3.31x10^{-1}$	$2.33x10^{-2}$
EPHB2	1	87	PGC BPD	$1.05x10^{-2}$	$2.34x10^{-4}$	**$8.54x10^{-5}$**	$4.38x10^{-5}$
LBH	2	97	PGC SZ	**$9.75x10^{-5}$**	$3.96x10^{-4}$	$2.76x10^{-3}$	$7.08x10^{-5}$

Supplementary Table 3.5. Full GAMuT results for BDI (21 items). Genes with $p < 1x10^{-4}$ identified in the GAMuT and KMR analyses are shown (p-values in bold). GAMuT and KMR utilized a linear genotype kernel (possibly weighted) for all analyses. For this BDI phenotype, standard linear regression identified no SNPs of suggestive significance ($p < 1x10^{-6}$). PGC MDD, PGC BPD, PGC SZ denote weights based on log odds ratios from the Psychiatric Genomics Consortium GWAS of major depressive disorder, bipolar disorder, and schizophrenia, respectively; MAF-based = weights based on minor allele frequencies of variants calculated using the Grady Trauma Project genotype data. ** denotes that the result exceeds the study-wise significance threshold.

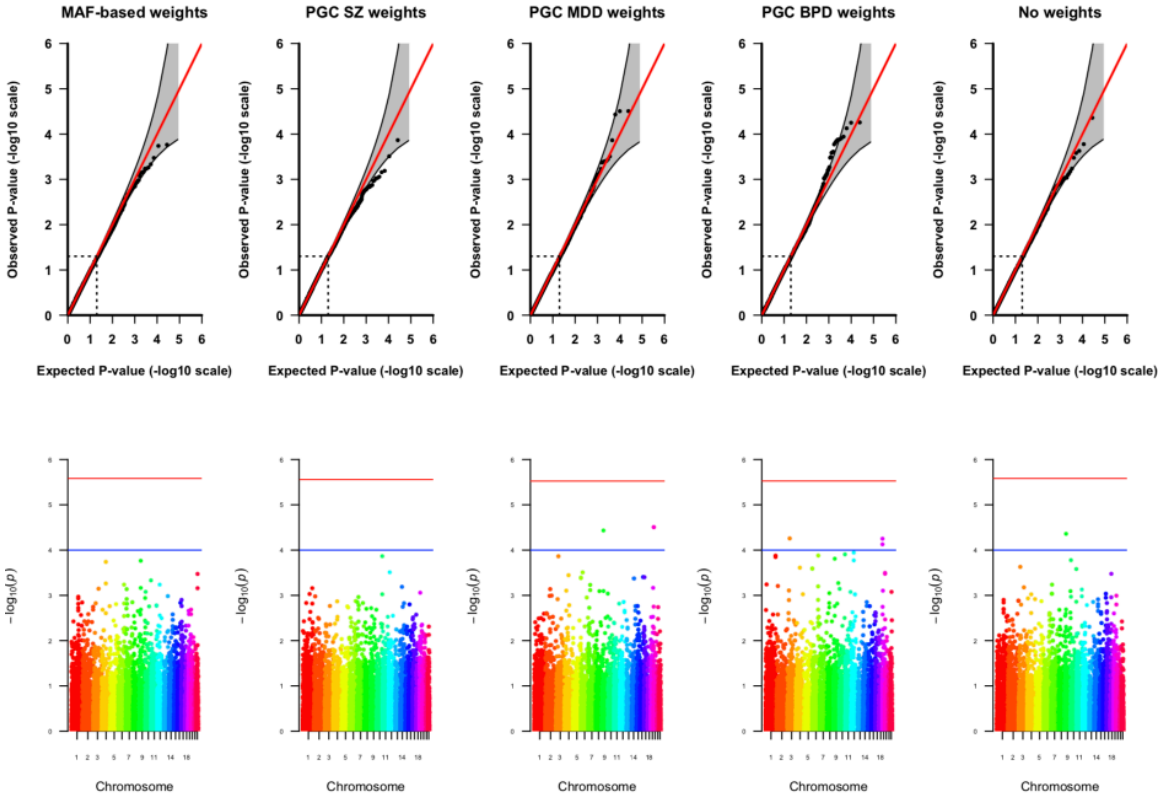
Gene	Chr	Number of variants	Genotype weights	GAMuT Phenotypic Similarity Matrix		KMR	Linear Regression (minimum p-value of SNP in gene)
				Projection Matrix	Linear Kernel		
ZHX2	8	97	PGC SZ	$1.64x10^{-2}$	**$2.73x10^{-6}$** **	$4.42x10^{-4}$	$1.00x10^{-3}$
		76	PGC MDD	$5.41x10^{-3}$	**$8.59x10^{-6}$**	$1.36x10^{-3}$	$1.00x10^{-3}$
SLC2A4	17	6	PGC MDD	**$1.88x10^{-5}$**	$2.22x10^{-2}$	$3.78x10^{-1}$	$2.86x10^{-1}$
BRWD2	10	147	PGC MDD	**$3.91x10^{-5}$**	$3.13x10^{-4}$	$2.16x10^{-3}$	$2.27x10^{-3}$
C10orf85	10	80	PGC MDD	**$5.04x10^{-5}$**	$3.47x10^{-4}$	$2.29x10^{-3}$	$2.27x10^{-3}$
PAXIP1	7	15	MAF-based	$2.76x10^{-2}$	$1.21x10^{-3}$	**$5.37x10^{-5}$**	$1.34x10^{-4}$
		15	No weights	$9.29x10^{-3}$	$1.39x10^{-3}$	**$8.27x10^{-5}$**	$1.34x10^{-4}$
PIK3CG	7	143	PGC SZ	**$5.59x10^{-5}$**	$1.07x10^{-1}$	$4.09x10^{-1}$	$6.27x10^{-3}$
FLJ36031	7	219	PGC SZ	**$5.95x10^{-5}$**	$1.71x10^{-1}$	$1.83x10^{-1}$	$1.65x10^{-3}$
LRP1B	2	620	PGC MDD	$2.49x10^{-2}$	$2.06x10^{-4}$	**$6.76x10^{-5}$**	$1.10x10^{-3}$
TXNIP	1	11	MAF-based	**$6.81x10^{-5}$**	$1.77x10^{-1}$	$2.01x10^{-1}$	$6.59x10^{-2}$
		11	No weights	**$7.60x10^{-5}$**	$1.38x10^{-1}$	$2.08x10^{-1}$	$6.59x10^{-2}$
FAM43A	3	115	PGC BPD	$1.69x10^{-1}$	**$7.35x10^{-5}$**	$4.15x10^{-4}$	$2.27x10^{-5}$
NUP214	9	19	MAF-based	$5.49x10^{-1}$	**$7.97x10^{-5}$**	$5.16x10^{-3}$	$5.57x10^{-3}$
E2F6	2	29	PGC MDD	$3.37x10^{-2}$	**$9.45x10^{-5}$**	$4.91x10^{-2}$	$3.32x10^{-2}$
GUK1	1	6	PGC SZ	$1.64x10^{-1}$	**$9.54x10^{-5}$**	$5.52x10^{-3}$	$2.34x10^{-3}$
SLC22A5	5	38	PGC SZ	$4.63x10^{-3}$	**$9.69x10^{-5}$**	$3.20x10^{-3}$	$3.98x10^{-4}$

Supplementary Figures 3.1a-3.1d. Application of GAMuT, univariate KMR, and standard linear regression to overall PSS (17 items). Supplementary Figure 3.1a includes plots for the GAMuT analyses that used a projection matrix to model phenotypic similarity, showing the results for each genotype weighting method. In the Manhattan plots, the red line represents the study-wise significance threshold (based on a Bonferroni correction for the number of genes tested), and the blue line represents the suggestive significance threshold. Supplementary Figure 3.1b provides analogous GAMuT results using a linear kernel for the phenotype. Supplementary Figures 3.1c and 3.1d show results from the corresponding univariate KMR (gene-level testing) and linear regression (SNP-level testing) analyses.

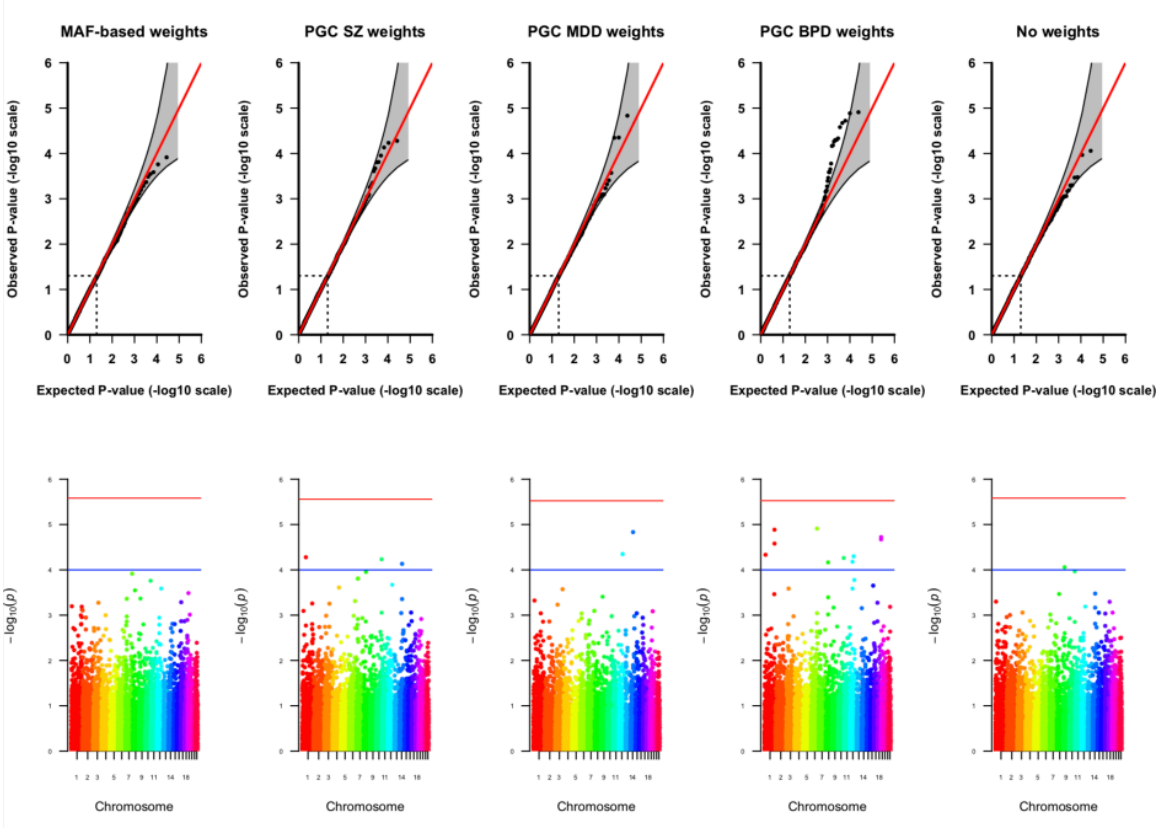
Supplementary Figure 3.1a. Overall PSS (17 items), GAMuT with Projection Matrix for modeling phenotypic similarity



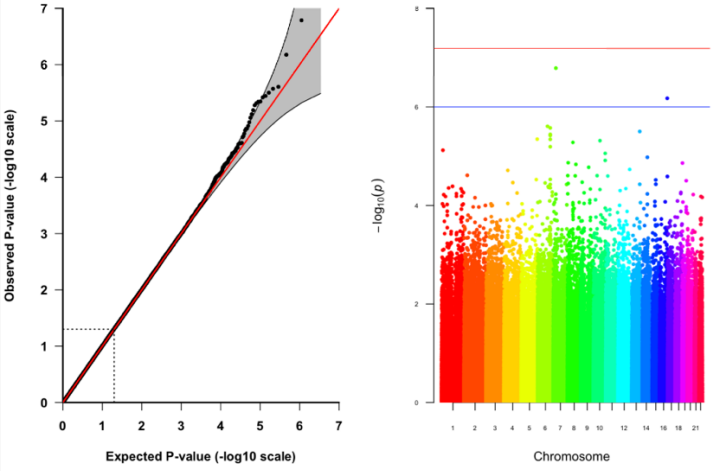
Supplementary Figure 3.1b. Overall PSS (17 items), GAMuT with Linear Kernel for modeling phenotypic similarity



Supplementary Figure 3.1c. Overall PSS (cumulative score), univariate KMR

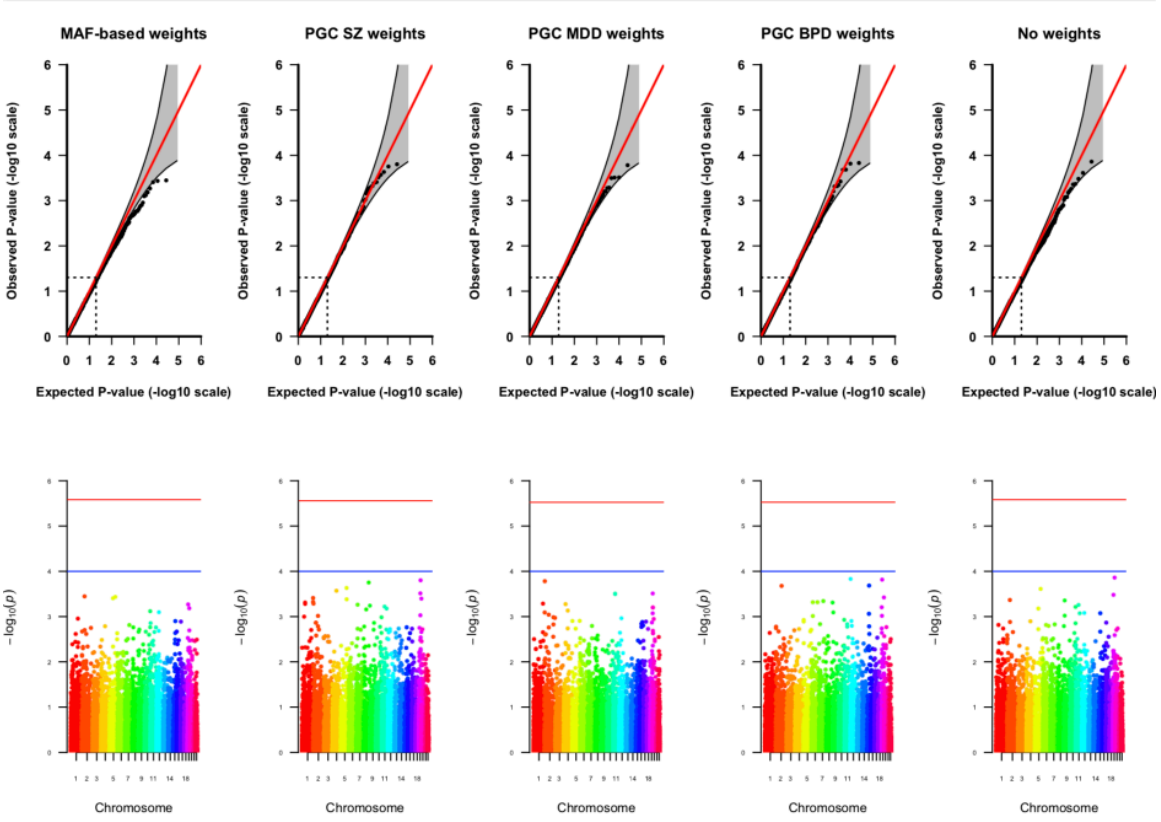


Supplementary Figure 3.1d. Overall PSS (cumulative score), standard linear regression

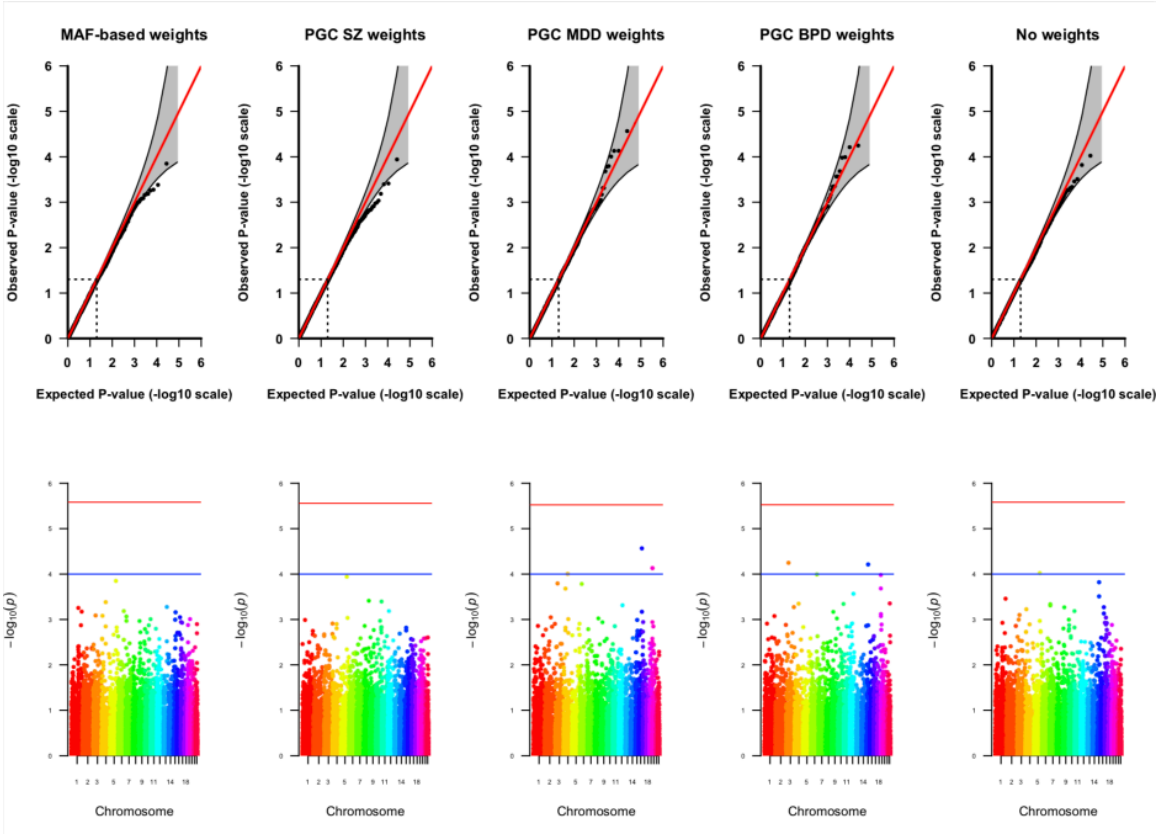


Supplementary Figures 3.2a-3.2d. Application of GAMuT, univariate KMR, and standard linear regression to PSS Intrusive (5 items). Supplementary Figure 3.2a includes plots for the GAMuT analyses that used a projection matrix to model phenotypic similarity, showing the results for each genotype weighting method. In the Manhattan plots, the red line represents the study-wise significance threshold (based on a Bonferroni correction for the number of genes tested), and the blue line represents the suggestive significance threshold. Supplementary Figure 3.2b provides analogous GAMuT results using a linear kernel for the phenotype. Supplementary Figures 3.2c and 3.2d show results from the corresponding univariate KMR (gene-level testing) and linear regression (SNP-level testing) analyses.

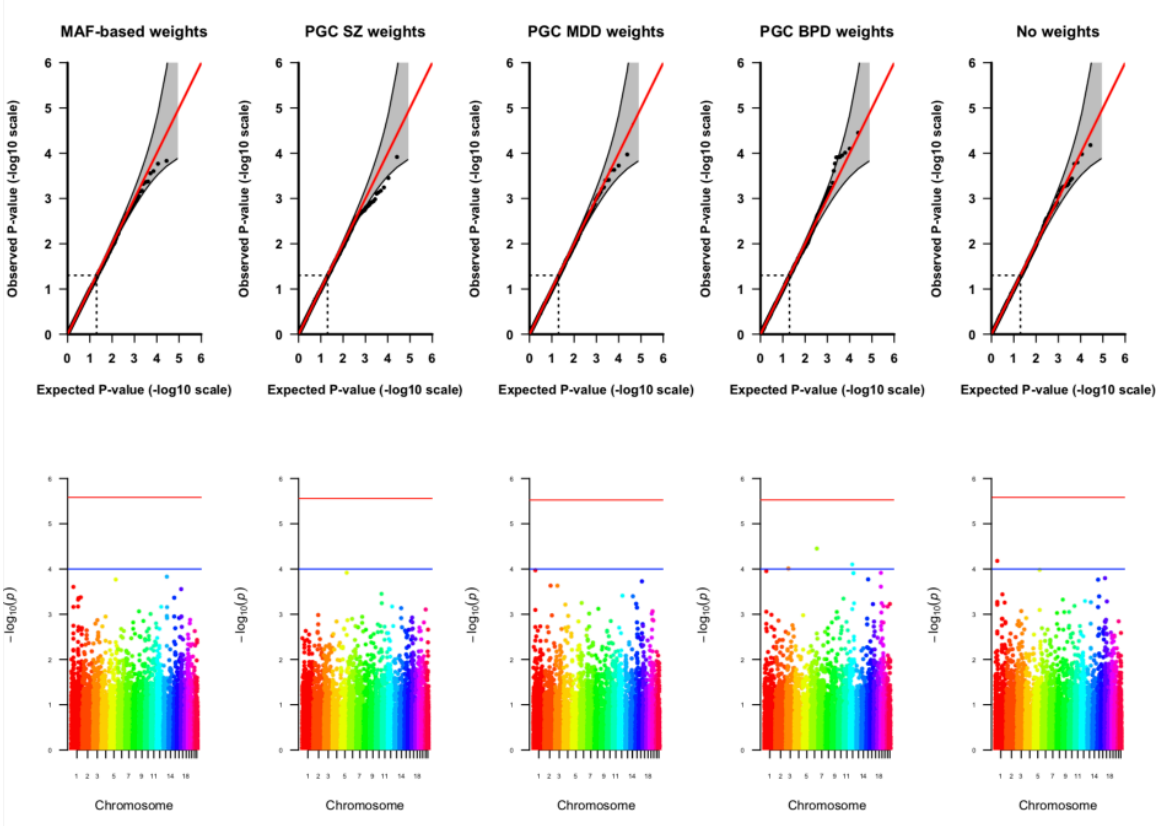
Supplementary Figure 3.2a. PSS Intrusive (5 items), GAMuT with Projection Matrix for modeling phenotypic similarity



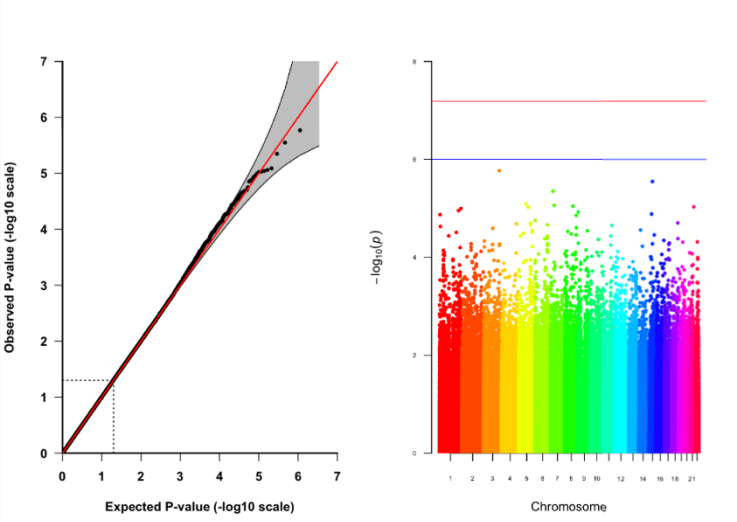
Supplementary Figure 3.2b. PSS Intrusive (5 items), GAMuT with Linear Kernel for modeling phenotypic similarity



Supplementary Figure 3.2c. PSS Intrusive (cumulative score), univariate KMR

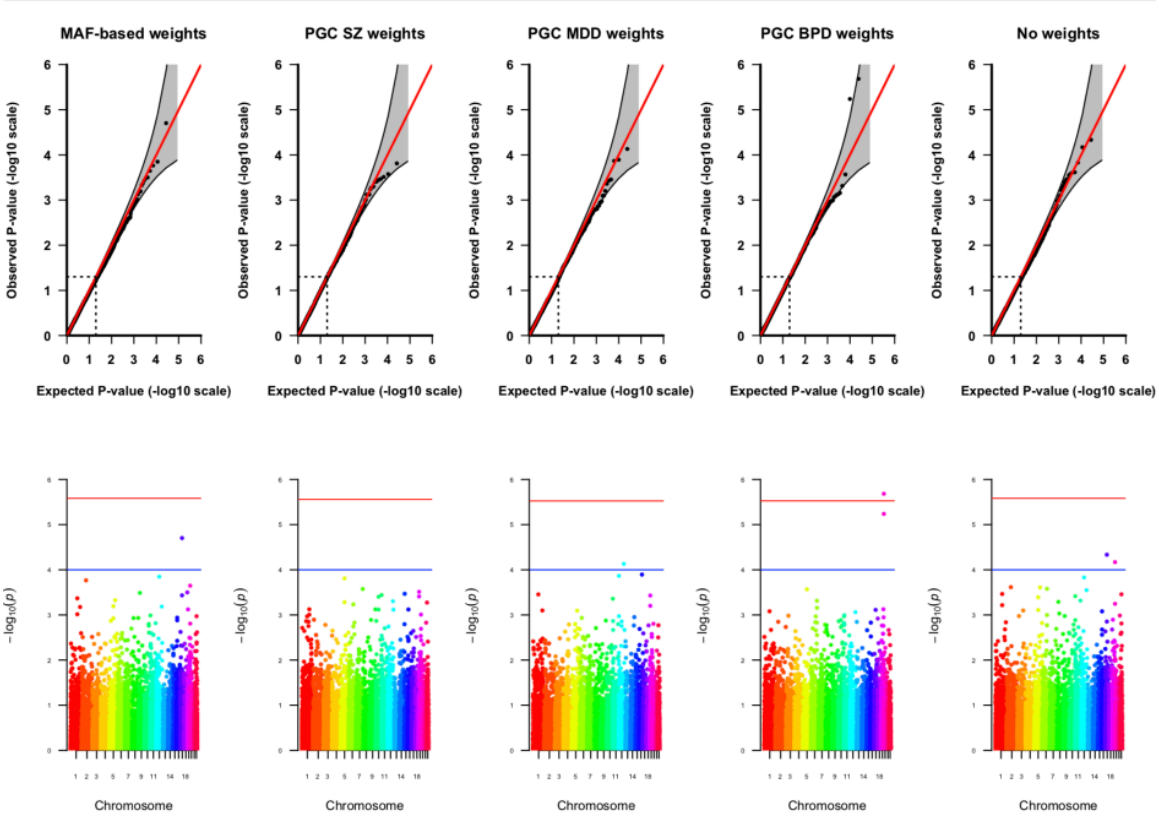


Supplementary Figure 3.2d. PSS Intrusive (cumulative score), standard linear regression

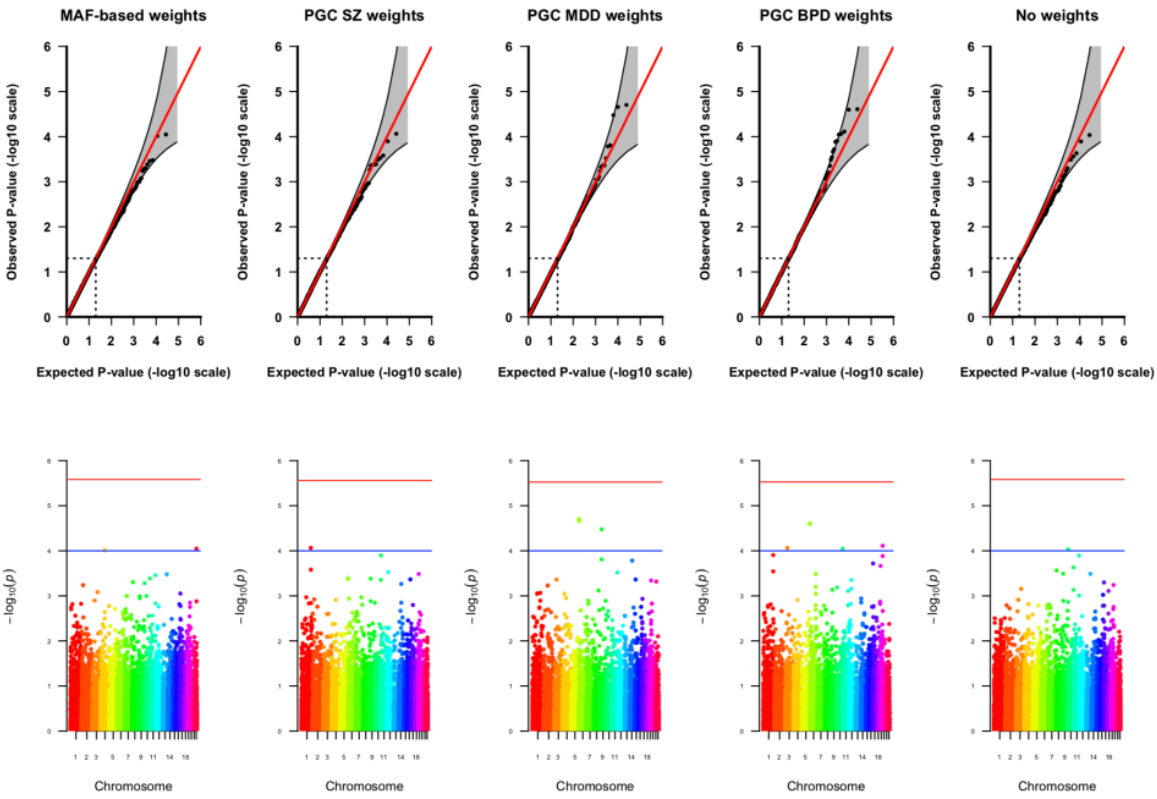


Supplementary Figures 3.3a-3.3d. Application of GAMuT, univariate KMR, and standard linear regression to PSS AvoidNumb (7 items). Supplementary Figure 3.3a includes plots for the GAMuT analyses that used a projection matrix to model phenotypic similarity, showing the results for each genotype weighting method. In the Manhattan plots, the red line represents the study-wise significance threshold (based on a Bonferroni correction for the number of genes tested), and the blue line represents the suggestive significance threshold. Supplementary Figure 3.3b provides analogous GAMuT results using a linear kernel for the phenotype. Supplementary Figures 3.3c and 3.3d show results from the corresponding univariate KMR (gene-level testing) and linear regression (SNP-level testing) analyses.

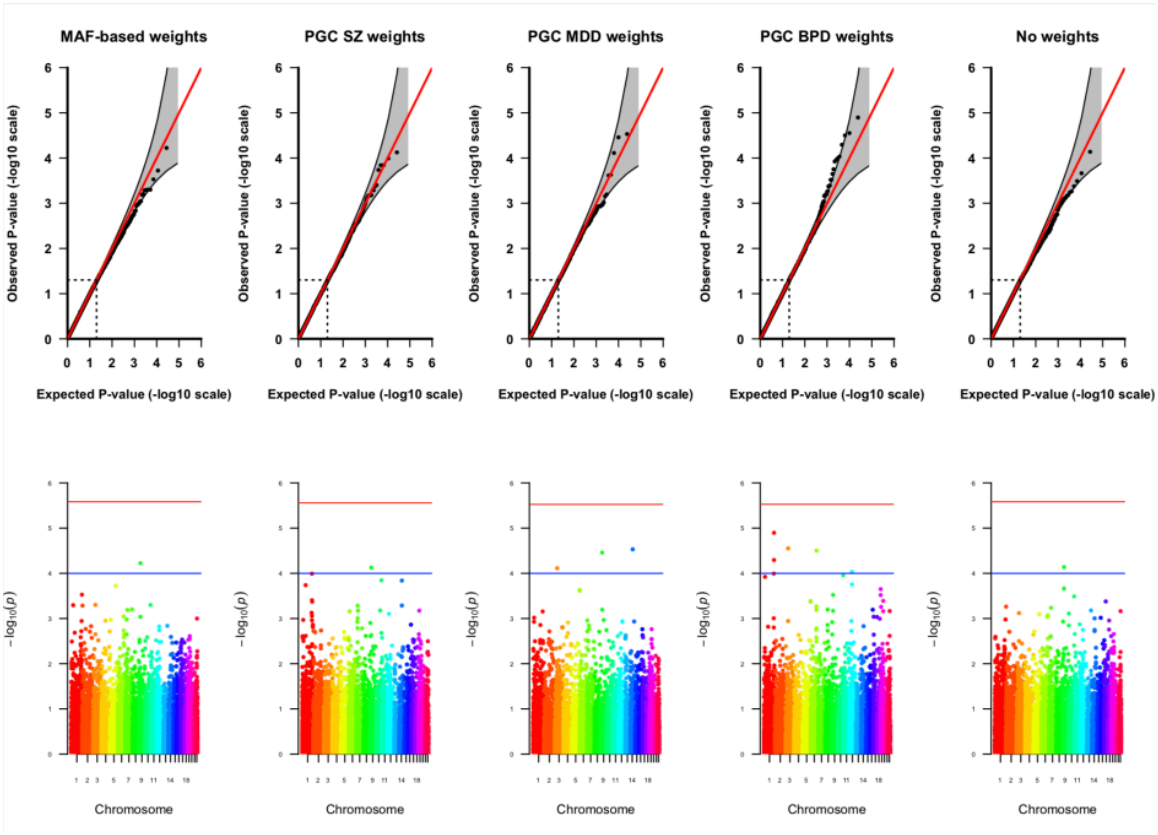
Supplementary Figure 3.3a. PSS AvoidNumb (7 items), GAMuT with Projection Matrix for modeling phenotypic similarity



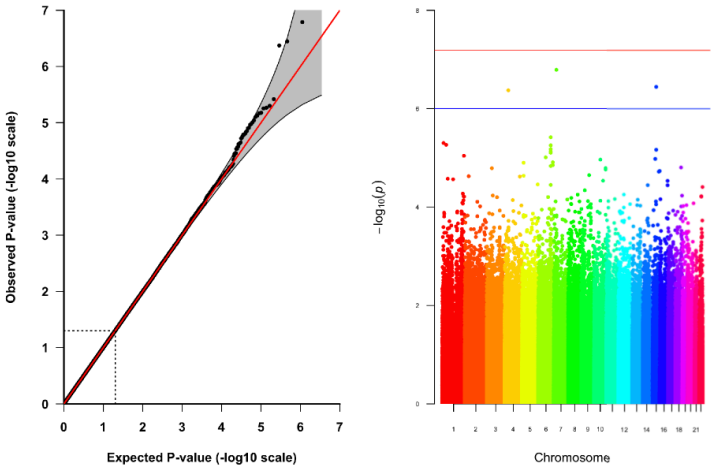
Supplementary Figure 3.3b. PSS AvoidNumb (7 items), GAMuT with Linear Kernel for modeling phenotypic similarity



Supplementary Figure 3.3c. PSS AvoidNumb (cumulative score), univariate KMR

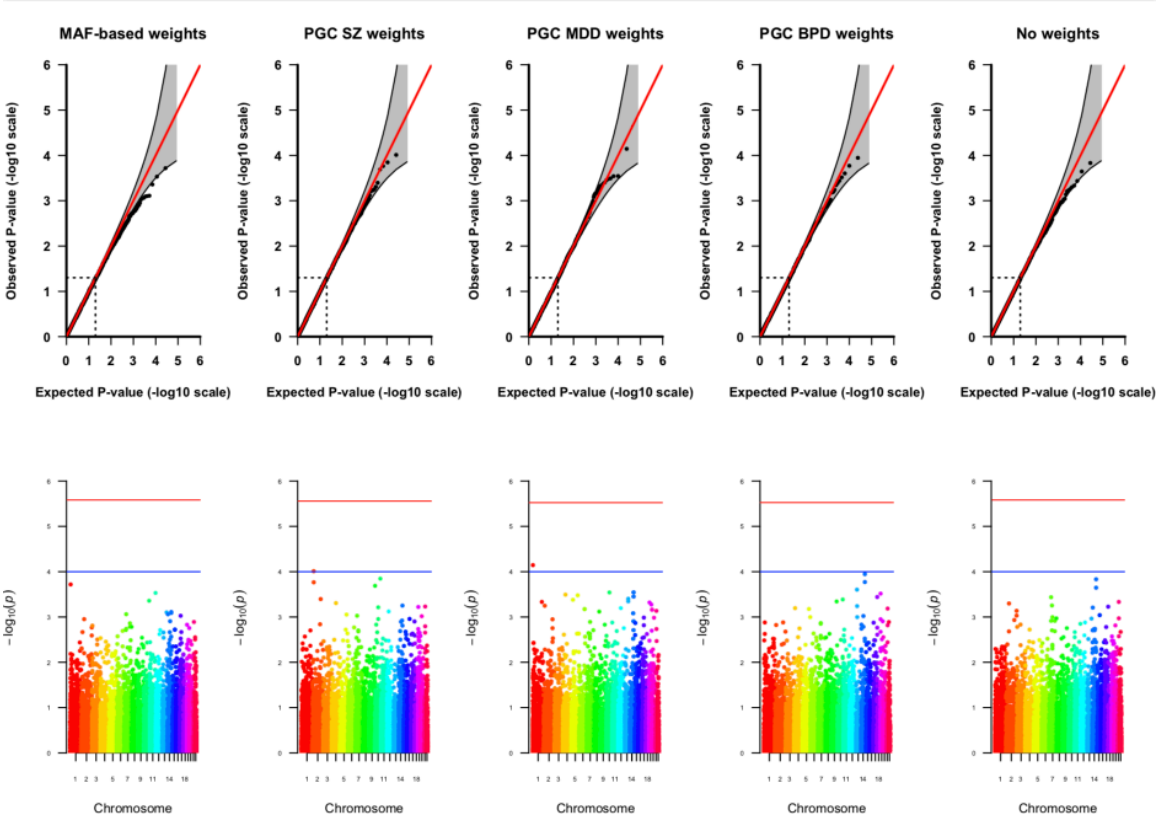


Supplementary Figure 3.3d. PSS AvoidNumb (cumulative score), standard linear regression

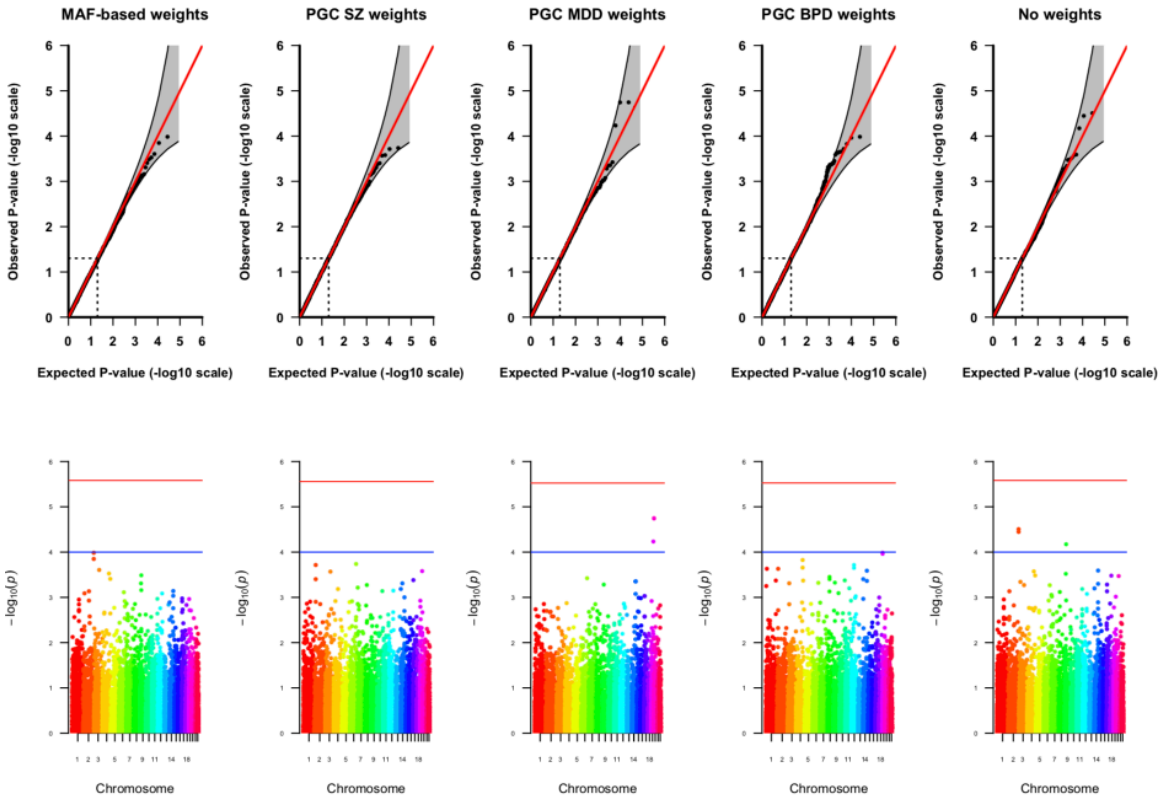


Supplementary Figures 3.4a-3.4d. Application of GAMuT, univariate KMR, and standard linear regression to PSS Hyperarousal (5 items). Supplementary Figure 3.4a includes plots for the GAMuT analyses that used a projection matrix to model phenotypic similarity, showing the results for each genotype weighting method. In the Manhattan plots, the red line represents the study-wise significance threshold (based on a Bonferroni correction for the number of genes tested), and the blue line represents the suggestive significance threshold. Supplementary Figure 3.4b provides analogous GAMuT results using a linear kernel for the phenotype. Supplementary Figures 3.4c and 3.4d show results from the corresponding univariate KMR (gene-level testing) and linear regression (SNP-level testing) analyses.

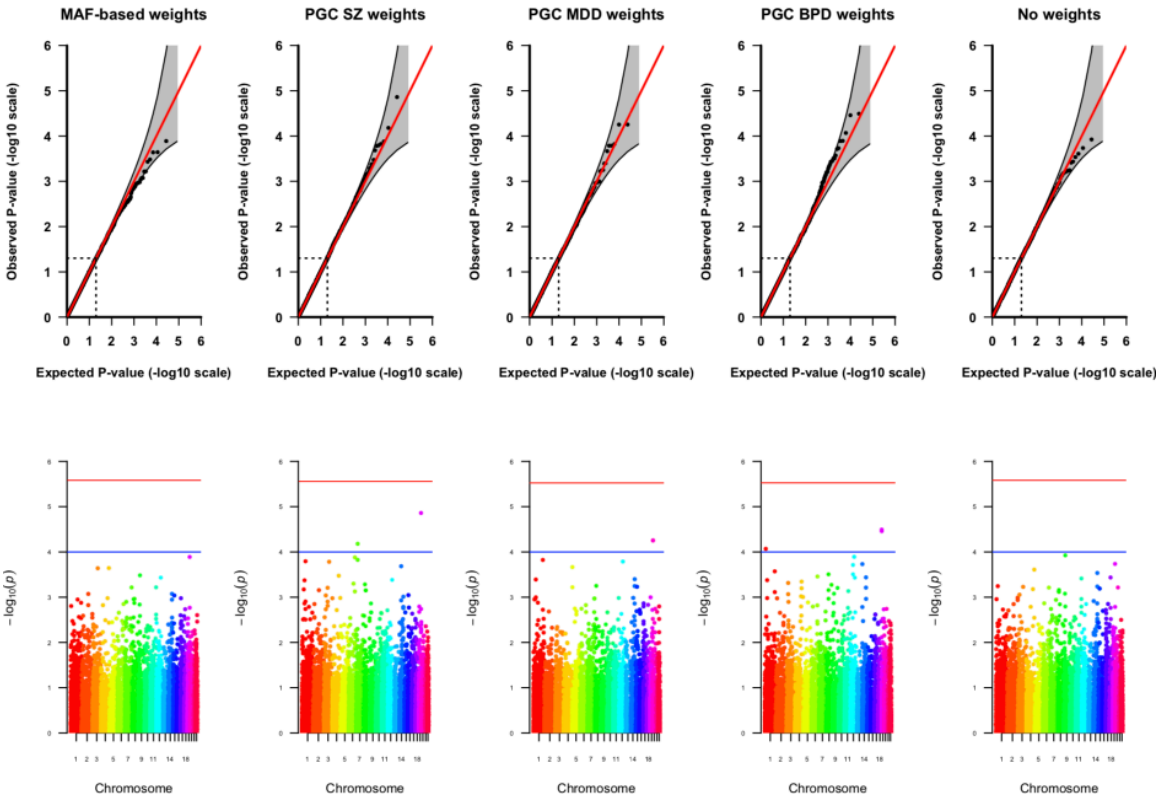
Supplementary Figure 3.4a. PSS Hyperarousal (5 items), GAMuT with Projection Matrix for modeling phenotypic similarity



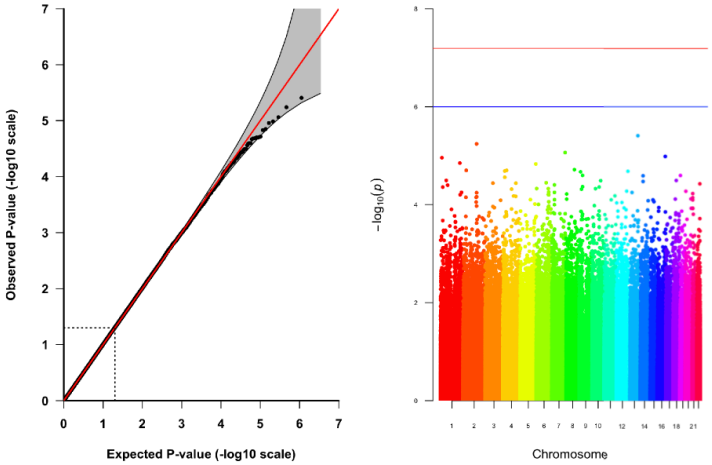
Supplementary Figure 3.4b. PSS Hyperarousal (5 items), GAMuT with Linear Kernel for modeling phenotypic similarity



Supplementary Figure 3.4c. PSS Hyperarousal (cumulative score), univariate KMR

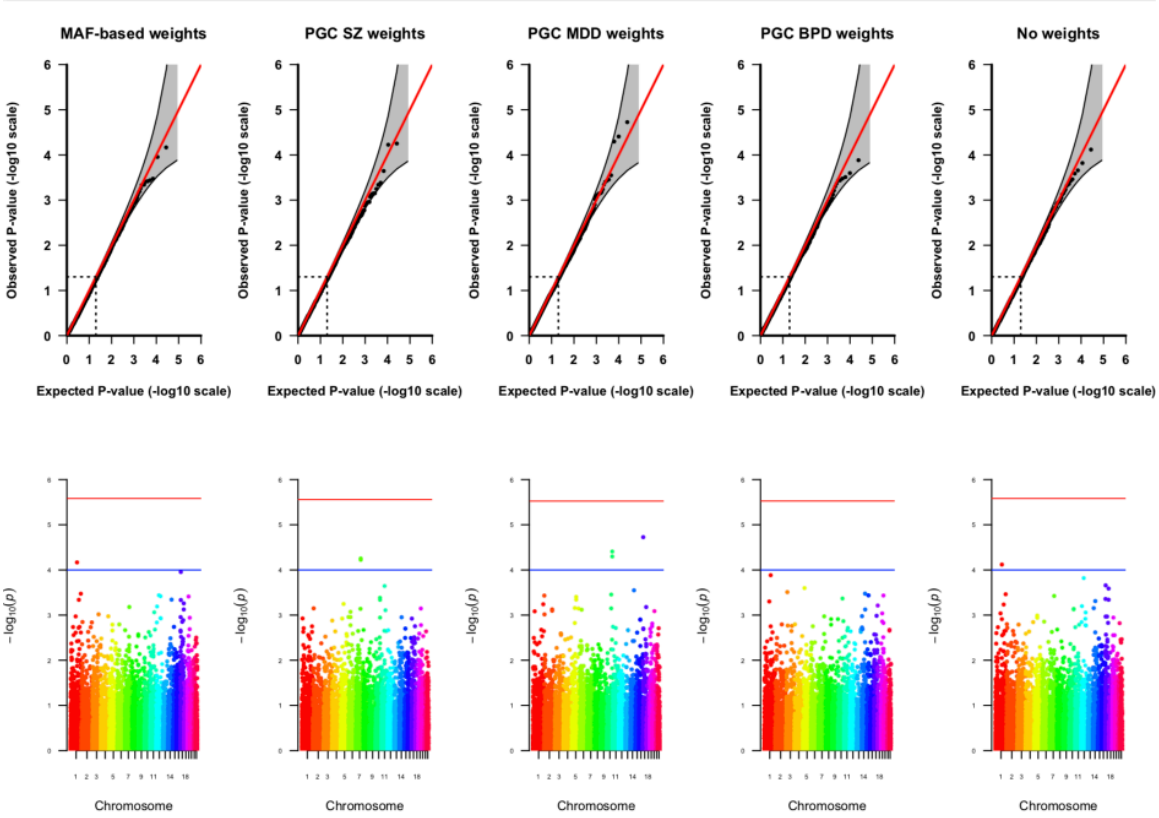


Supplementary Figure 3.4d. PSS Hyperarousal (cumulative score), standard linear regression

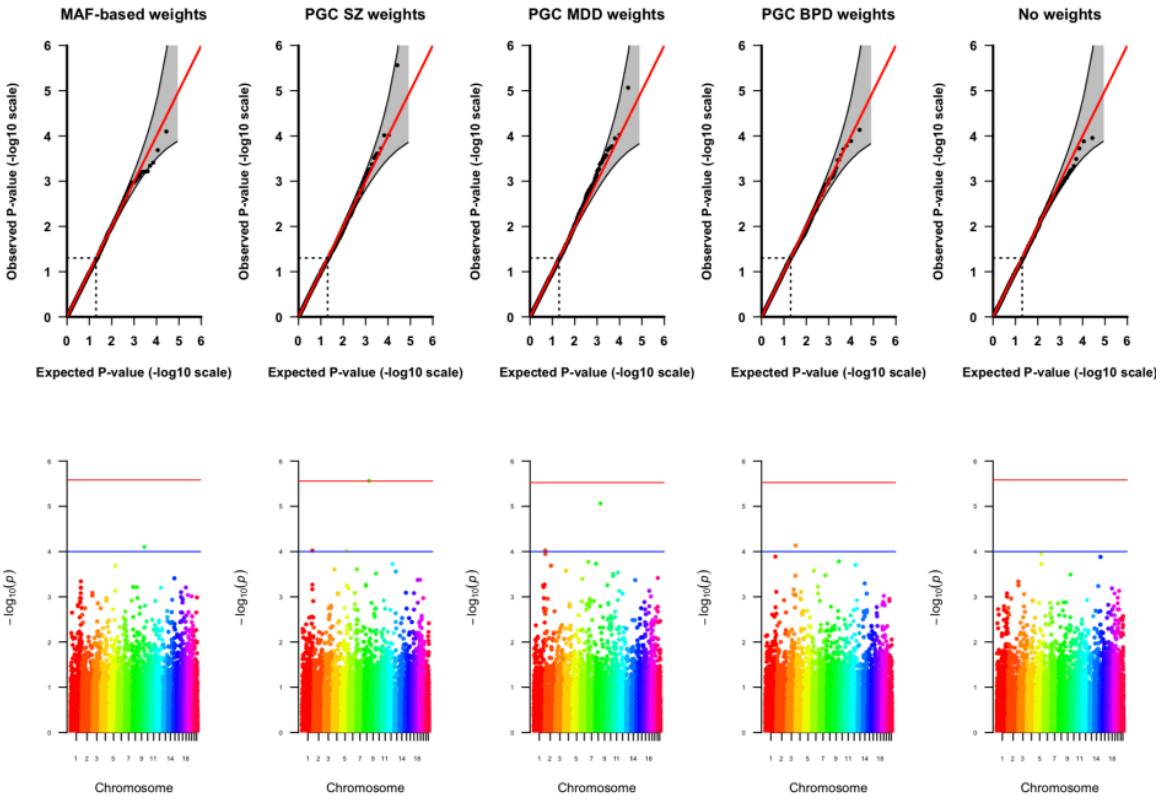


Supplementary Figures 3.5a-3.5d. Application of GAMuT, univariate KMR, and standard linear regression to BDI (21 items). Supplementary Figure 3.5a includes plots for the GAMuT analyses that used a projection matrix to model phenotypic similarity, showing the results for each genotype weighting method. In the Manhattan plots, the red line represents the study-wise significance threshold (based on a Bonferroni correction for the number of genes tested), and the blue line represents the suggestive significance threshold. Supplementary Figure 3.5b provides analogous GAMuT results using a linear kernel for the phenotype. Supplementary Figures 3.5c and 3.5d show results from the corresponding univariate KMR (gene-level testing) and linear regression (SNP-level testing) analyses.

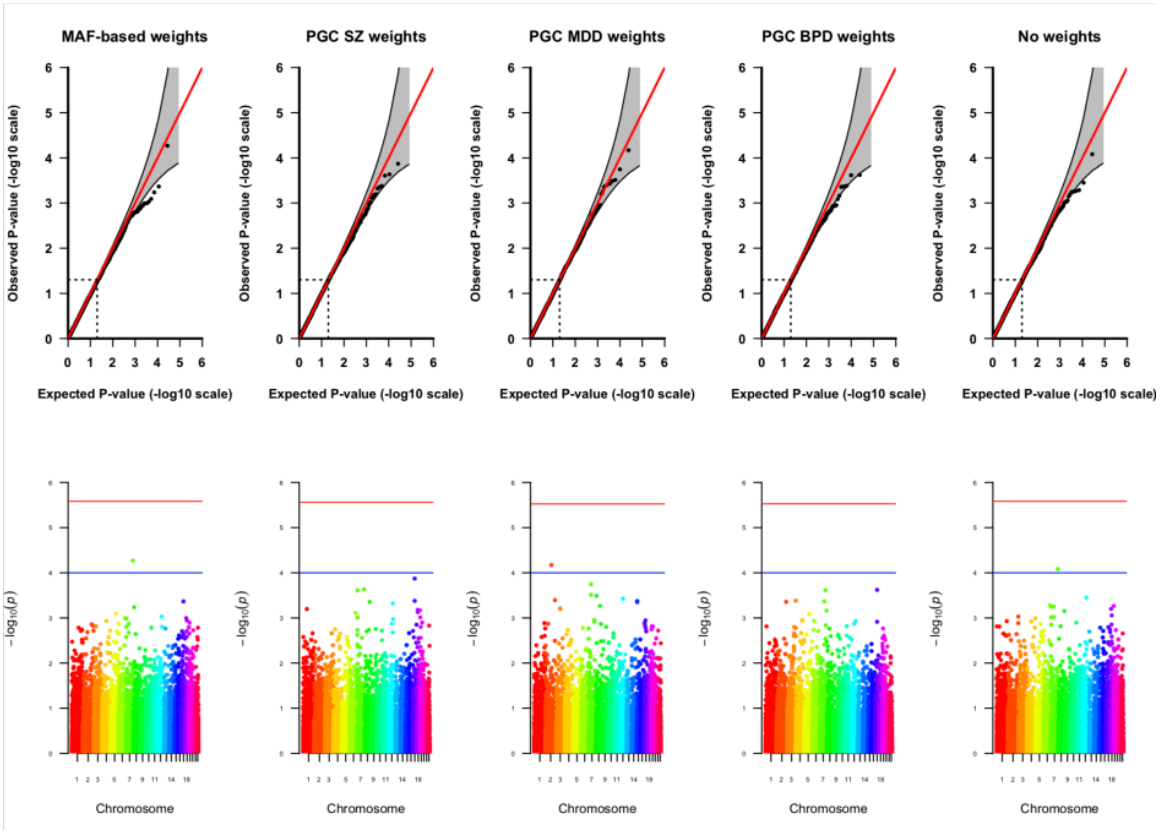
Supplementary Figure 3.5a. BDI (21 items), GAMuT with Projection Matrix for modeling phenotypic similarity



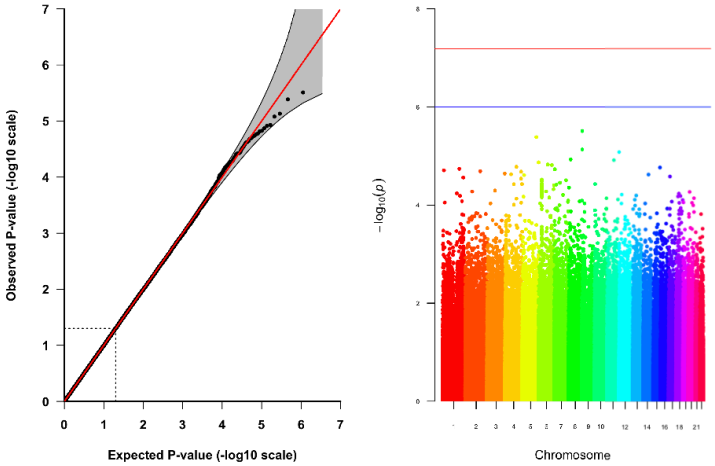
Supplementary Figure 3.5b. BDI (21 items), GAMuT with Linear Kernel for modeling phenotypic similarity



Supplementary Figure 3.5c. BDI (cumulative score), univariate KMR



Supplementary Figure 3.5d. BDI (cumulative score), standard linear regression



Chapter 4:

Investigating the association of rare regulatory variation and gene expression among genes with schizophrenia-associated expression levels

Abstract

Background: Schizophrenia is known to involve a substantial genetic component, with contributions to risk made by both common (MAF > 0.01) and rare (MAF < 0.01) genetic variants. In recent years, numerous genes have been identified to have expression levels that are associated with schizophrenia, potentially playing a role in the causal pathway leading to this disorder. It is likely that genetic factors are involved in regulating the expression of these genes, with rare regulatory variants perhaps having especially large effects on gene expression levels. However, investigations of rare regulatory variation in relation to gene expression have been limited, particularly for genes with SZ-associated expression. We sought to help fill this research gap, employing next generation DNA and RNA sequencing and a modified version of a recently developed burden approach to more powerfully examine associations of rare regulatory variants with gene expression levels for genes with SZ-associated expression.

Methods: Our analytic sample consisted of 725 individuals, including 355 schizophrenia cases and 370 controls. These individuals had undergone targeted DNA sequencing for 64 genes previously identified to have schizophrenia-associated expression and 172 gene regions located within or near schizophrenia-associated large CNV intervals. They had also undergone genome-wide RNA sequencing. We analyzed these data using a modified version of a burden method that was recently developed for the specific purpose of increasing power for investigating rare regulatory variant associations with gene expression. We used this approach to examine associations of rare promoter, 5'UTR and 3'UTR variants with gene expression levels for our set of schizophrenia-associated genes.

Results: We consistently observed U-shaped patterns of estimated association between rare regulatory allele burden and gene expression, whereby rare regulatory alleles were most likely to be observed at the extremes (low and high) of gene expression. We also observed a consistent tendency for the U-shaped estimated associations to be more pronounced when limiting analyses to the rarest variants (MAF < 0.001), variants more likely to be deleterious (CADD ≥ 5), and when only considering variants in the 5'UTR. As one example, when analyzing genes with SZ-associated expression, and considering only 5'UTR variants with MAF < 0.001 and CADD ≥ 5, we observed a U-shaped estimated association between rare regulatory allele burden and gene expression, with OR = 0.56 (95% CI: 0.33, 0.97) for the odds of observing a rare allele at a medium expression level versus the lowest expression level. Additional analyses revealed that estimated associations between rare regulatory variants and gene expression were weaker for genes intolerant to LoF or missense variation as compared with genes tolerant to these variant types, possibly reflecting selection against variants with strong influences on expression for highly constrained genes.

Conclusions: Our findings are consistent with a potential effect of rare regulatory variants on the expression levels of genes with SZ-associated expression, whereby rare regulatory alleles may cause decreased or increased expression. Our results also suggest that such effects might be particularly strong for variants that are rarer, more likely deleterious, and located in the 5'UTR. Although many of our estimates were rather imprecise, their plausibility is supported by prior studies. Future research that considers a larger number of genes and/or employs larger sample sizes will enable more accurate and precise association estimates.

INTRODUCTION

Schizophrenia (SZ) is a chronic and severe psychiatric disorder characterized by delusions, hallucinations, disorganized speech or behavior, as well as other symptom manifestations.[6] It causes substantial functional impairment, and is associated with a greatly elevated risk for suicide.[6,95] SZ has a lifetime risk of ~1% in the general population, thus impacting substantial numbers of individuals.[6] With an estimated heritability of ~80%, genetic factors play a particularly important role in SZ.[17,18] Investigation of the genetics underlying SZ, including the biological pathways through which genetic factors increase SZ risk, is therefore important to gain a better understanding of its etiology and to facilitate identification of novel biological targets for treatment and prevention.

Both common (minor allele frequency [MAF] > 1%) and rare (MAF < 1%) genetic variants are known to contribute to SZ, though the role of common variation has been more extensively and robustly investigated and is better characterized as compared with the role of rare variants. At present, it is estimated that common variants collectively explain one-quarter to one-third of SZ risk variance, acting through a polygenic component whereby hundreds to thousands of common variants with exceedingly small individual effects contribute to SZ en masse.[19] Common variants thus make an important contribution to SZ risk, yet appear to account for only a minority of SZ's ~80% heritability. It is thought that the remaining heritability (sometimes termed the "missing heritability") may be largely explained by rare variants. Known associations between rare variants and SZ include at least nine rare, large (> 100kb) copy number variants (CNVs, which are genomic deletions or duplications at least 1 kilobase in size) that have been identified as strongly associated with SZ, with odds ratios ranging from 2 to > 40.[20-23] Smaller rare variants, including single nucleotide variants (SNVs), are also expected to contribute to SZ risk. However, progress in characterizing the role played by such genetic variants has been slow due to the need for large sample sets with DNA sequencing data to have sufficient power and genomic interrogation for identifying rare variant associations. Large-

scale studies employing DNA sequencing approaches for rare variant investigation have only recently become more feasible with the introduction and decreasing costs of next generation sequencing (NGS).[96]

Genetic variants, both common and rare, likely affect SZ risk in part by regulating gene expression. Over the past several years, numerous genes have been identified as having expression (i.e., transcription) levels that are associated with SZ.[14-16] For instance, a meta-analysis of results from differential expression analyses of two independent SZ case-control studies identified 647 genes differentially expressed by affection status.[15] Another study examined enrichment of SZ cases (versus controls) within the extreme tails of a gene's expression distribution (> 2 standard deviations from the expression mean), and identified numerous genes for which extremes of expression were associated with SZ.[14] These studies examined gene expression using tissue from lymphoblastoid cell lines (LCLs), which the investigators note are expected to be fairly removed from environmental and state aspects of the individual, in theory increasing the likelihood that any causal pathway is directed from gene expression to SZ and not vice versa.

The role of various factors in modifying the expression levels of these genes with SZ-associated expression is presently not well understood, and is important to investigate to advance knowledge of the upstream elements that may affect SZ through modification of expression levels. While numerous factors, both biological and environmental, are known to affect gene expression, genetic factors are particularly important to consider for the set of genes with SZ-associated expression levels, given the prominent role of genetics in SZ risk. Common variants are known to affect gene expression for many genes across the genome;[24] and many of the more than 100 individual common variants that have been identified as robustly associated with SZ are located in non-coding sequences and have been found to be associated with the expression of nearby genes.[19,97] Rare variants have the potential to exert larger influences on expression, but have been much less well studied in relation to gene expression,

both for genes with SZ-associated expression and for genes overall. This is in large part due to the technological and power (related to sample size) limitations mentioned above. Furthermore, rare regulatory variants in particular have been understudied, due a greater emphasis historically on identifying rare coding variants versus regulatory variants, as well as the poorer characterization of regulatory regions as compared with coding regions.

A small number of studies have investigated associations of rare non-coding variants with gene expression (sometimes defining rare as MAF < 0.05 rather than the current standard definition of rare as MAF < 0.01).[25-29] These studies have yielded findings consistent with a role for rare regulatory variants in modifying gene expression. Investigations specifically focused on the contribution of rare regulatory variants to gene expression variation for genes with SZ-associated expression have been more limited. One such investigation (to our knowledge, the only study examining associations of rare regulatory variants with SZ-associated expression levels) focused on 17 genes with SZ-associated expression based on having expression outliers (individuals with expression levels > 2 standard deviations from the mean expression level) enriched for SZ cases as compared with controls.[14] For these 17 genes, the researchers examined whether the subjects with outlier expression levels (expression levels > 2 SD from the mean) were enriched for rare variants within coding and regulatory sequences (DNase I hypersensitive sites within 50 kb upstream of the gene) as compared with non-outlier subjects. They found that, compared with coding variants, rare putative regulatory variants showed the strongest association with being an expression outlier, with 7/17 genes showing nominally significant associations. This study involved a small sample size, including only 157 SZ cases and 118 controls. To more thoroughly examine associations of rare regulatory variants with expression for genes with SZ-associated expression levels, larger sample sizes are needed. In addition, the use of analytical approaches that are specifically designed to increase power for identifying rare variant associations with gene expression is warranted.

For this study, we had a main objective of investigating the contribution of rare regulatory variants to gene expression for genes with SZ-associated expression levels. With this objective in mind, we set out to analyze two independent datasets derived from SZ case-control studies, one with 725 samples (a combination of 355 SZ cases and 370 controls) and the other with 400 samples (a combination of 265 SZ cases and 135 controls). All samples had undergone targeted DNA sequencing of 1) exonic and regulatory sequence (including 2,000 bases upstream of the first exon) for 64 genes previously identified as having SZ-associated expression levels, and 2) exonic sequence for 172 genes and gene fusions (regions spanning two or more genes) with prior evidence for involvement in SZ due to being located within or in close proximity to a SZ-associated large CNV interval. They had also all undergone genome-wide expression profiling, with RNA sequencing employed for the sample set of 725 individuals and microarray profiling used for the sample set of 400. We sought to analyze these datasets separately and then perform meta-analyses.

Efforts to identify rare variant associations using analytical methods that have traditionally been employed for studying common variant associations (e.g., genome-wide association studies, expression quantitative trait loci [eQTL] studies) tend to suffer from low power, both due to the very low frequency of the rare variants and the massive multiple testing burden corresponding to individually testing rare variants that theoretically may be present at any base in the genome. To address these challenges, investigators often employ methods such as burden tests or the sequence kernel association test (SKAT),[74] which are approaches that group rare variants together by region (e.g., by gene) and test these variant sets for association with a phenotype. This has the dual benefit of testing the aggregate effects of many rare variants within a region, which should facilitate identification of associations, and also reducing the multiple testing burden in comparison with testing each rare variant individually. Against this background of variant set approaches for investigating rare variant associations, a unique burden approach was recently developed for examining associations of rare genetic

variation specifically with gene expression. This approach, originally described in Zhao et al. (2016),[27] first assigns rare alleles within a gene region to ordered expression bins (ranging from low to high) for the gene, and then gains substantial power by aggregating rare alleles for each expression bin across all genes being considered and examining association of rare allele burden with gene expression level for this aggregated dataset. We have employed this approach, making certain modifications as detailed in the **Methods** section below, to more powerfully examine associations of rare regulatory variants with gene expression for our set of genes with SZ-associated expression. We have specifically considered rare variants located within gene promoter regions, as well as those in the 5' untranslated region (5'UTR) or the 3' untranslated region (3'UTR).

In addition to analyzing the set of genes with SZ-associated expression levels, we decided to perform analyses that incorporated the larger set of genes that were targeted for sequencing due to being located within or near a large SZ-associated CNV interval. These CNVs are genomic deletions or duplications, implying that reduced or increased transcript dosages for certain genes within these CNV intervals are important contributors to SZ risk. Consistent with this, prior research has found that for genes located within SZ-associated CNV intervals, low outlier expression levels (> 2 SD below mean expression) are more frequently observed among SZ cases as compared with controls.[14] Genes within SZ-associated CNV regions therefore seem likely to be enriched for genes with transcript dosage that is associated with SZ. Considering this, grouping the 64 genes with SZ-associated expression together with the 172 genes located within or near SZ CNV intervals is expected to yield an overall gene set enriched with genes for which expression level is associated with SZ. Analysis of this larger, overall gene set, which is better powered for examining rare regulatory variant associations with gene expression levels, should therefore yield informative results with regard to the association of rare regulatory variants with SZ-linked expression levels, which supplement the analyses that strictly focus on the 64 genes previously identified to have SZ-associated expression. Analysis

of the overall gene set should also be informative with respect to the association between rare regulatory variants and gene expression more broadly, as such investigations have been limited to date.

METHODS

Data sources

Analyses were carried out using data from the Molecular Genetics of Schizophrenia (MGS) case-control sample, which was assembled during the mid 2000's.[98] The MGS study includes genetic (GWAS) and phenotypic data for 8,257 samples. Of these, 3,503 samples with European ancestry have undergone targeted DNA sequencing (by Emory University and NorthShore University HealthSystem), and a partially overlapping set of 2,171 samples with European ancestry have undergone RNA sequencing or microarray expression profiling. We have analyzed samples with targeted DNA sequencing data, expression data, and genome-wide SNP data to examine associations between rare regulatory variation and gene expression.

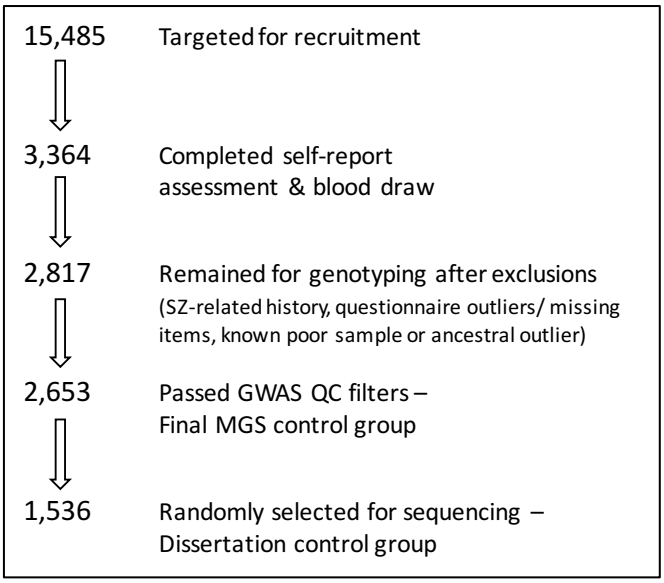
MGS Cases: Cases were recruited at 9 sites in the U.S. and 1 site in Australia, and were identified through clinics, hospitals, physician referrals, advocacy organizations, and media advertisements. They were 18 or older with a primary diagnosis of 1) DSM-IV SZ or 2) DSM-IV schizoaffective disorder with at least six months of meeting DSM Criterion A for SZ (e.g., delusions, hallucinations). The inclusion of individuals with schizoaffective disorder and a prolonged period of SZ Criterion A symptoms is standard in SZ genetics research, as it enables recruitment of the full range of SZ presentations, including those with and without concurrent mood symptoms (schizoaffective and SZ cases, respectively).[98] We refer to all cases as SZ cases (following standard practice). For each participant, two senior clinicians independently assigned diagnoses based on clinical information gathered through interview using the Diagnostic Interview for Genetic Studies, Family History Interview for Genetic Studies, and

medical records. Inclusion as a SZ case required consensus diagnoses by the two clinicians. Subjects with moderate or severe mental disability were ineligible, as were subjects who did not consent to provide a blood sample. In all, 2,873 SZ cases of European ancestry were recruited, of which 2,681 remained following GWAS quality control filters to form the final European ancestry case group for the MGS sample.

MGS Controls: Controls were drawn from a nationally representative online participant panel assembled by a survey and marketing research company. Participants in the panel were recruited from across the U.S., including from case recruitment areas, through random-digit dialing of residential phone numbers. In total, a member from 30% of targeted households joined the participant panel. Communication with panel members was performed via internet, but initial recruitment did not require internet access; web-based equipment was provided to participants lacking internet access. Additionally, weighting techniques were applied to reduce bias due to nonresponse and non-sampling of residences without telephones. The final online participant panel was representative of the U.S. population in terms of important demographic dimensions including age, sex, race/ethnicity, education, and urban/rural residence. Out of approximately 60,000 individuals of European ancestry in the panel, 15,485 were randomly selected and sent information about participation in the MGS study, of which 3,364 (21.7%) completed the required self-report clinical assessment and blood draw. Weighting adjustments were applied to this group of 3,364 individuals to reduce non-sampling error. Individuals who endorsed or did not answer items assessing SZ, schizoaffective disorder, bipolar disorder, hallucinations, or delusions were then excluded, as were individuals who were outliers in the number of missing items (4 or more missing items out of 69 items) or in the number of “yes” responses (50 or more “yes” responses out of 69 items), which reduced the sample of 3,364 by approximately 8%. A small portion of the remaining individuals had been studied previously by the MGS investigators, and were known to have biological samples that performed poorly or were known to be ancestral outliers based on ancestry-informative genotype data. These

individuals were also excluded, leaving 2,817 for GWAS genotyping. Subsequent GWAS quality control procedures resulted in 2,653 European ancestry controls remaining for the final MGS sample. A flowchart of the control selection process is shown in **Figure 4.1**.

Figure 4.1. Flowchart demonstrating selection of European ancestry controls.



Targeted DNA sequencing: From the set of 2,681 European-ancestry cases, 1,962 were randomly selected for DNA sequencing; and from the set of 2,653 European-ancestry controls, 1,536 were randomly selected for DNA sequencing. These randomly selected samples underwent targeted DNA sequencing for 172 genes and gene fusions (regions spanning two or more genes) with prior evidence for involvement in SZ due to being located within or near a SZ-associated large CNV interval (see **Table 4.1**); sequencing was performed for all exons within these genes, including 5'UTR and 3'UTR sequence. Targeted DNA sequencing was also performed for 64 genes with heightened probability for involvement in SZ due to exhibiting gene expression levels associated with SZ based on prior studies.[14,16] For these 64 genes, all exons were sequenced, and sequencing was also performed for the genomic interval spanning the transcription start site (TSS) to 2 kilobases (kb) upstream of the TSS (promoter sequence).

Targeted DNA sequencing was carried out with an Illumina Hi-Seq 2000 instrument in ten separate batches, including a pilot batch and 9 batches with approximately 384 samples each.

Table 4.1. SZ-associated CNV intervals containing 172 genes sequenced.

CNV Interval	Variant Type	Size	# Genes	SZ Odds Ratio (95% CI)
1q21.1	Deletion	1.35 Mb	10	8.35 (4.65-14.99)[21]
3q29	Deletion	1.6 Mb	23	41.1 (5.6-1,953.6)[22]
7q11.23	Duplication	1.5 Mb	29	10.78 (1.46-79.62)[23]
15q13.3	Deletion	1.58 Mb	8	7.52 (3.98-14.19)[21]
16p11.2	Duplication	600 kb	31	11.52 (6.86-19.34)[21]
22q11.2	Deletion	1.6 or 3 Mb	70	INF (28.27-INF)[21]
CNTNAP2	Deletion	220 kb	1	n.d.
			172	

RNA profiling: A subset of the 2,681 European-ancestry cases and the 2,653 European-ancestry controls (i.e., the final MGS GWAS dataset) was previously selected for microarray expression profiling.[16] Specifically, Illumina HT-12v4 microarrays were used to obtain expression profiles based on LCLs. Previous examination of these microarray expression data indicated good data quality, with mean correlations of 0.99 and 0.98 for expression level among technical and biological replicates, respectively. Microarray expression profiles for the previously analyzed set of 859 samples (413 SZ cases and 446 controls) were accessed through dbGaP following required approvals.

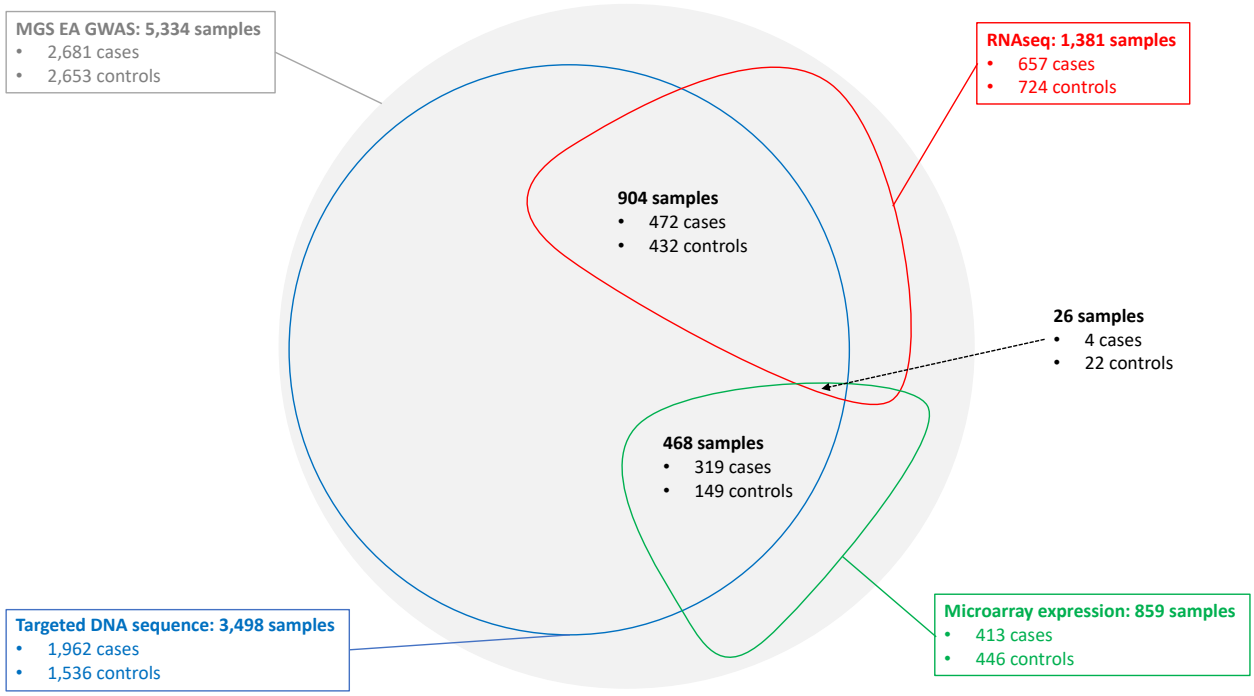
A subset of the European-ancestry cases and controls in the final MGS GWAS dataset was also previously selected for expression profiling using RNA sequencing technology.[14,15] This RNA sequencing subset largely did not overlap with the microarray expression samples (intentionally). For this RNA sequencing dataset, the investigators note that both the case group and control group were selected to have approximately equal proportions of females and males, and that cases and controls were roughly matched based on 5-year age brackets, with the goal of reducing confounding for their analyses. Sequencing was carried out on LCLs using an

Illumina Hi-Seq 2000 to a depth of 10 million reads per sample, generating 50-bp single-ended reads. Previous examination of these data found mean correlations of ~0.99 for expression level among technical and biological replicates, indicating high data quality.[14] FASTQ files containing raw RNA sequence reads for the previously analyzed set of 1,381 samples (657 SZ cases and 724 controls) are available through the National Center for Biotechnology Information's (NCBI) Database of Genotypes and Phenotypes (dbGaP), and were downloaded following approval of our data accession request.

Although we had the impression that the microarray expression samples and the RNA sequencing samples likely were selected from the final MGS GWAS dataset through random selection processes (random other than the gender and age matching employed for the RNA sequencing dataset), we were unable to confirm this. Concerns related to the impact of non-random selection on our results are addressed in a later section.

Figure 4.2 below depicts the overlap between the final MGS European-ancestry GWAS dataset, the targeted DNA sequenced samples, the RNA sequenced samples, and the samples with microarray expression profiles.

Figure 4.2. Overlap in samples between four datasets: the final MGS European-ancestry GWAS dataset, the targeted DNA sequenced samples, the RNA sequenced samples, and the samples with microarray expression profiles.



Processing and QC of targeted DNA sequencing and expression datasets

Targeted DNA sequencing dataset

Sample processing and QC

We mapped the raw targeted DNA sequencing reads for 3,503 samples to the human genome (hg38) using PEMapper.[42] Note that this is 5 more samples than the 3,498 total targeted DNA sequenced samples described above. We were unsuccessful in linking 5 of the 3,503 DNA sequenced samples back to their corresponding phenotype and covariate data, and therefore above we describe the case-control breakdown for 3,498 targeted sequenced samples.

We then called variant sites using PECaller.[42] We first excluded 85 samples with mean or median coverage < 10x. Due to observed variation in average coverage level across different

sequencing batches, we then grouped samples by coverage decile and batch-called samples separately for each of these groupings, including approximately equal numbers of samples in each coverage decile grouping. We subsequently merged the called datasets, yielding a single dataset with 3,418 samples and 24,463 variant sites. We set any genotypes with confidence score < 95% to missing.

Among these 3,418 samples, we removed 92 with genotype completion rate less than 3 standard deviations (SD) below the mean. We also removed 22 samples that were apparent duplicates with another targeted sequencing sample based on sharing 2 alleles identical-by-state (IBS2) for ≥ 90% of sites.

As a further quality check, we examined concordance between samples in the targeted DNA sequence dataset and those in the MGS GWAS dataset (recall that the former samples are a subset of the latter). We merged these datasets, limiting sites to up to ~180 high-quality overlapping variants with MAF > 1%. Using both custom scripts and PLINK1.9[44] functions, we computed IBS2 for all sample pairs, specifying IBS2 proportion ≥ 90% as indicating a genetic match (i.e., same sample). This process identified 56 sample pairs involving a targeted sequencing dataset sample and a GWAS dataset sample that were genetic matches but that were not expected to be matches based on sample ID. We also identified 10 instances in which the targeted sequencing and GWAS samples were expected to match based on sample ID, yet were not genetic matches. We excluded the targeted sequenced sample involved in each of the 66 discordant instances.

We performed additional sample QC based on various sample-level metrics computed by the annotation program Bystro.[43] We excluded 40 samples with heterozygosity/homozygosity ratio beyond +/- 3 SD from the mean; 1 sample with silent/replacement ratio < 3 SD below the mean; 2 samples with theta < 3 SD below the mean; 1 sample with transition/transversion ratio < 3 SD below the mean; and 3 samples with heterozygosity/homozygosity ratio < 1.6 who were

also a 3 SD outlier on another Bystro metric (one sample with high transition/transversion; one with low exonic theta; and one with high silent/replacement).

We used the GWAS dataset corresponding to our targeted sequenced samples to identify and remove 6 samples with discordant reported and inferred sex. We used PLINK's --check-sex to infer natal sex from inbreeding coefficients from both X and Y chromosome data. We also used the GWAS data to confirm that, after removing the sample duplicates mentioned above, no additional targeted sequenced samples required removal due to excessive relatedness with another sample (i.e., all sample pairs had IBD proportion < 0.1875). In the process of performing this sample QC, we also identified 5 targeted sequencing samples that we were unable to successfully map to phenotype and covariate data (as mentioned previously). We therefore excluded these 5 samples.

These various QC steps left 3,180 high-quality targeted sequencing samples for analysis. We confirmed European ancestry for these post-QC samples by anchoring this targeted sequencing dataset to the HapMap3 dataset, constructing principal components (PCs) for all samples, and plotting PC1 versus PC2 for all samples. All of the targeted sequencing samples clearly clustered on or very near the HapMap3 CEPH/Utah (CEU) cluster, confirming European ancestry for this targeted sequencing dataset (see **Figure 4.3** below). **Table 4.2** offers a summary of the steps leading to our final set of 3,180 samples.

Figure 4.3: PCA plot for post-QC targeted sequencing samples. Black crosshairs represent targeted sequencing samples. Red cluster (mostly covered by black crosshairs) is the CEU population (European ancestry). Purple cluster (top) is the CHB+JPT population (Asian ancestry). Green cluster is the YRI population (African ancestry).

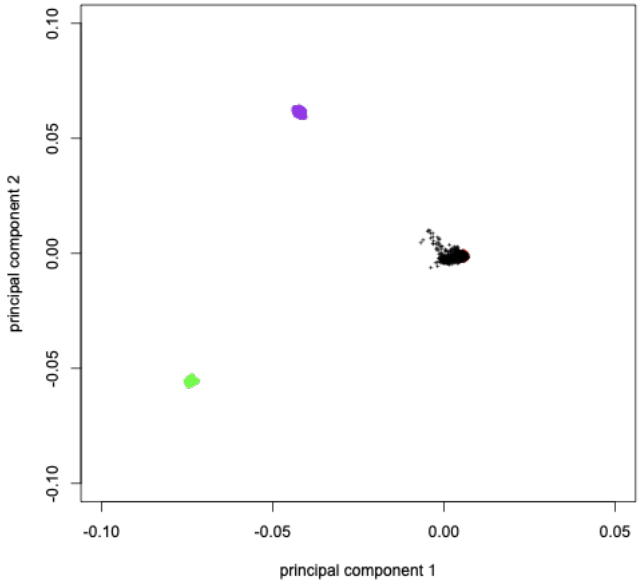


Table 4.2: Sample QC for the targeted DNA sequencing dataset.

		No. Samples
Total samples that underwent targeted DNA sequencing		**3,503**
Sample QC	Mean/median coverage < 10x	85
	Genotype completion rate < 3 SD below mean	92
	Duplicate (IBS2 ≥ 90%)	22
	Discordant with MGS GWAS dataset	66
	Bystro metric outlier	47
	Discordant sex	6
	Unable to match with phenotype/covariate data	5
Final post-QC sample		**3,180**

Variant QC

The targeted DNA sequencing dataset initially included 24,463 variant sites. We removed sites that were monomorphic among the cleaned set of 3,180 samples, as well as sites missing for more than 5% of samples, leaving 19,949 variant sites.

As an extra step to ensure maximal quality of variant calls to be used for analysis, we compared sample alternate allele frequencies (based on all 3,180 samples) with frequencies based on version 2 of the Genome Aggregation Database[99] (gnomAD v2; frequencies are based on samples that include a sizable minority of individuals with various neuropsychiatric disorders) and excluded variants with sample-based frequencies that differed significantly from gnomAD frequencies. The 19,949 variant sites included a total of 20,470 alternate alleles (due the presence of some sites with more than one alternate allele). We used Bystro to obtain gnomAD allele frequencies for the non-Finnish European (NFE) population (which matched the ancestry of our 3,180 targeted DNA sequencing samples): 11,577 alternate alleles had gnomAD NFE allele frequencies reported (based either on the gnomAD exome sample or genome sample), while the remaining 8,893 alternate alleles were missing from gnomAD. For each of the 11,577 alleles with gnomad NFE frequencies, we used a binomial test (R's binom.test() function) to calculate the probability of getting the observed number of alternate alleles (considering all 3,180 samples) or an observation more extreme, given the total number of observed alleles and assuming an alternate allele frequency equal to the gnomAD NFE frequency (two-sided test). We decided that alternate allele tests yielding Bonferroni-corrected p-values < 0.05/11,577 indicated sample-based allele frequencies that were significantly different from gnomAD frequencies, meriting removal to reduce the likelihood of including low-quality calls in our final analytic dataset. Furthermore, we decided to remove any alternate (or minor) allele observed in our sample set more than twice but with gnomad NFE frequency = 0%, regardless of the binomial test p-value. These steps resulted in 324 alternate alleles being

flagged for removal. For the 8,893 alternate alleles missing from the gnomAD NFE database (which may have been missing from gnomAD due to being extremely rare), we decided to remove such alleles if more than 2 were observed in our sample set; this resulted in 1,855 alternate alleles being flagged for removal. In total, these 2,179 alleles flagged for removal corresponded to 2,098 variant sites (among multiallelic sites, if one of the alternate alleles was flagged for removal, we decided to remove the entire variant site). Excluding these sites left 17,851 variants.

We then used PLINK2[44,100] to identify and remove 35 variant sites with Hardy-Weinberg equilibrium mid-p-value < 1e-6, leaving a total of 17,816 variants. In order to further maximize quality of the variant call set used for our analyses, we also decided to remove all sites involving insertions or deletions, calls for which tended to be lower confidence as compared with SNVs. This yielded a dataset with 3,180 samples and 16,573 SNVs (some multiallelic).

RNA sequencing dataset

We used FastQC (version 0.11.9)[101] and MultiQC (version 1.9)[102] to perform initial quality checks of the FASTQ files for the 1,381 samples with raw RNA sequence data. These checks indicated that the FASTQ files contained a small amount of Illumina adapter content (primarily adapter dimers). We removed this adapter content using Trimmomatic (version 0.39),[103] using a comprehensive list of adapter sequences that included those provided with the BBmap program.[104] We then ran FastQC and MultiQC once again, this time on the adapter-trimmed FASTQ files, confirming that Illumina adapter content had been successfully removed and determining that none of the 1,381 samples required removal for poor read quality before read mapping. Among the 1,381 samples, total read count ranged from 5.5 to 95.1 million, with median 11.9 million and mean 13.6 million (a small number of samples with known schizophrenia-associated copy number variants were intentionally sequenced to greater depth, yielding higher total read counts for these samples).

We mapped the raw RNA sequence reads to the human genome (build hg38) using STAR (version 2.7.0f),[105] requiring at least 16 matched bases between read and reference, and that the ratio of mismatched bases to total mapped length not exceed 5%. Among the 1,381 samples, percent of uniquely mapped reads (among all mapped and unmapped reads) ranged from 58.8% to 85.3%, with median 83.0% and mean 82.7%. Twenty samples had proportions of uniquely mapped reads less than 3 SD below the mean proportion of uniquely mapped reads for all 1,381 samples. We removed these 20 mapping outliers, leaving a minimum percent of uniquely mapping reads equal to 78.0%. We also examined other mapping quality metrics (e.g., mismatch rate per base among uniquely mapped reads; total unique reads crossing splice junctions; proportion of unmapped reads) and identified no additional outliers for removal. For the 1,361 remaining samples, we quantified read counts for each gene using the htseq-count script from the HTseq package (version 0.12.4).[106] Only uniquely mapping reads that overlapped exons for a single gene were counted.

Next, we performed additional sample QC using the R package DESeq2.[107] We applied a variance stabilizing transformation (VST) to the count data, as the DESeq2 user manual recommends for sample visualization and clustering. We then performed PCA on the transformed expression matrix (using R's prcomp()). Considering PCs 1 through 10, we identified 20 outlier samples with PC value beyond 3.5 SDs from the mean PC value. We also obtained Euclidian distance measures for each pair of samples (using the VST expression matrix), and identified outliers as samples with distance ≥ 3 SD above the mean (mean and SD for distance based on all sample pairs, considering 1,361 samples) with at least 200 other samples (this 200 sample threshold corresponds to 3 SD above the average number of samples with which an individual has distance ≥ 3 SD above the mean). This distance outlier analysis identified 3 additional samples for removal. Removing these 23 outliers left 1,338 samples with RNA sequencing data.

We applied edgeR's[108] trimmed mean of M-values (TMM) normalization to the count matrix, as has been done in previous association analyses of genetic variants with RNA sequencing expression data.[24] Before applying TMM normalization, we limited the dataset to only include genes for which at least 20% of samples had counts-per-million ≥ 0.1 and read count ≥ 6. This filtering out of lowly expressed genes reduced the dataset from 26,485 to 13,584 genes. We then TMM normalized the count data for the matrix of 13,584 genes and 1,338 samples.

Microarray expression dataset

The microarray dataset that we accessed through dbGaP included expression data for 859 samples and 27,118 probes. These 859 samples had been cleaned and outliers removed, and formed the sample set previously analyzed in Sanders et al. (2013).[16] The 27,118 probes had been filtered to exclude lowly expressed probes (after starting from 47,231 initial probes), and raw expression intensities had been transformed and normalized using background noise subtraction, log2 transformation, and quantile normalization (see Sanders et al. (2013) for more details on filtering and normalization of this microarray expression dataset).

Generating final analytic datasets

RNA sequencing samples

There were 819 samples that overlapped between the normalized RNA sequencing dataset (which included 1,338 samples) and the cleaned targeted DNA sequencing dataset (which included 3,180 samples). For these 819 samples, we performed PCA using available GWAS SNP data to identify and remove PC outliers and create a more homogenous sample set (outlier PC values often indicate that a subject differs from the rest of the sample with respect to potential confounders such as ancestry). PCA was performed with PLINK1.9's --pca flag, using a set of ~50,000 common (MAF > 5%), LD-pruned SNPs (pruning was accomplished using the

following flag: --indep-pairwise 50 5 0.11). Outliers were defined as individuals with PC1, PC2 or PC3 value beyond +/- 3 SD from the sample mean for the same PC. Removal of PC outliers was performed in an iterative fashion that involved computing PCs, identifying and removing outliers, recomputing PCs using the remaining samples, identifying and removing outliers based on these new PCs, and so on until no major outliers remained in the dataset. Three rounds of PCA collectively identified 94 outliers for removal, leaving a post-PCA dataset with 725 samples (including 355 SZ cases and 370 controls). A final set of PCs was generated using just these 725 samples, and the first 5 PCs were used as covariates in the rare variant analyses.

Microarray expression samples

There were 455 samples that overlapped between the normalized microarray expression dataset (which included 859 samples) and the cleaned targeted DNA sequencing dataset (which included 3,180 samples). Among these 455 samples, 23 were represented in the final set of 725 RNA sequenced samples; removing these left 432 unique samples with microarray expression data. We used available GWAS SNP data for these samples to perform PCA outlier analysis in the same manner as described above for the RNA sequencing samples. Through 2 rounds of PCA we identified 32 PC outliers for removal, leaving a set of 400 samples with microarray expression data for analysis (including 265 SZ cases and 135 controls). PCs 1 through 5 computed using these 400 samples were included as covariates in the rare variant analyses.

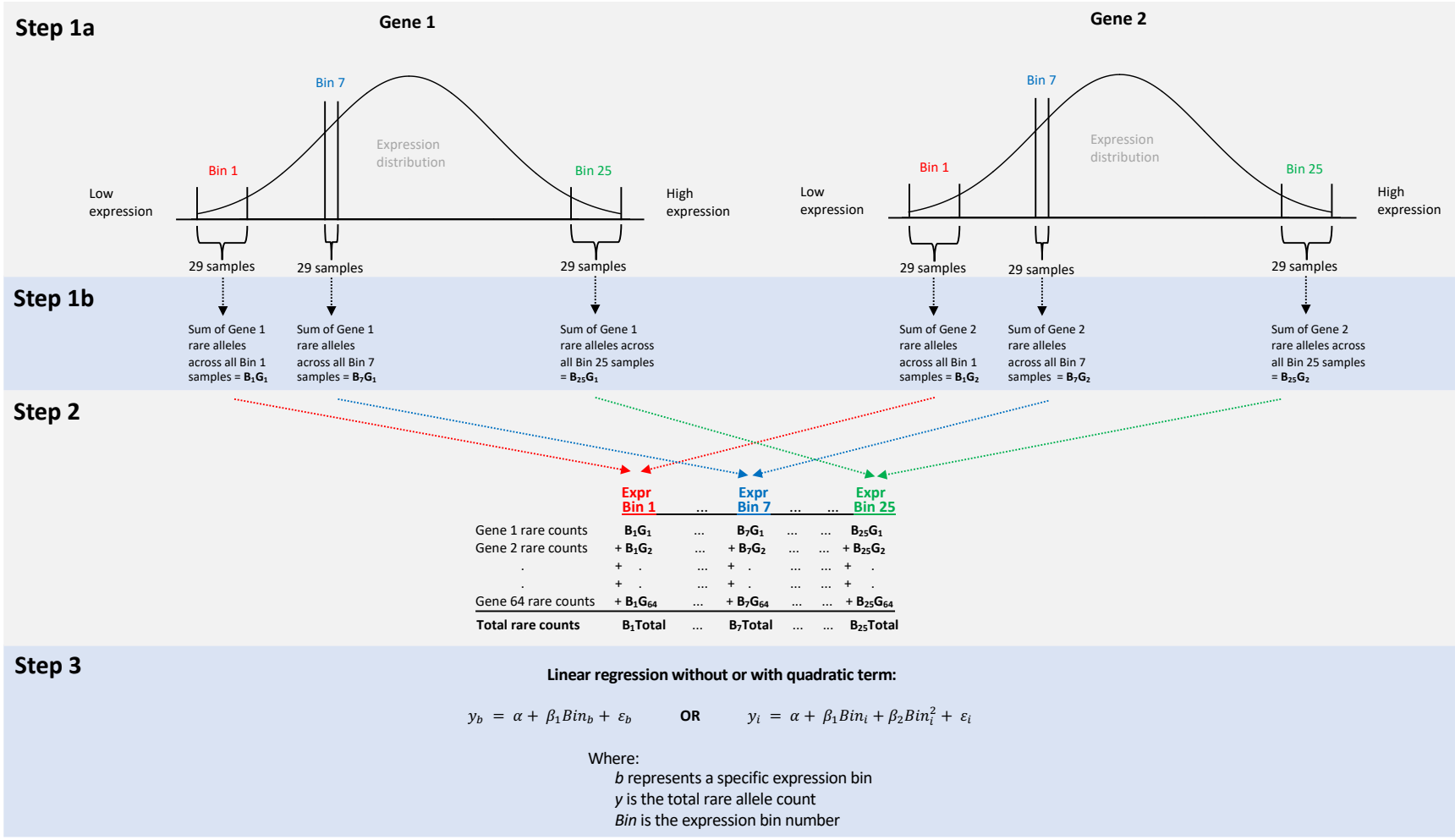
Analytic approach

Overview of burden method

To examine associations of rare regulatory variation with gene expression level, we employed an approach based on a burden method originally developed and applied by Zhao et al. (2016).[27] The primary steps of this method are depicted in **Figure 4.4**. For a given gene,

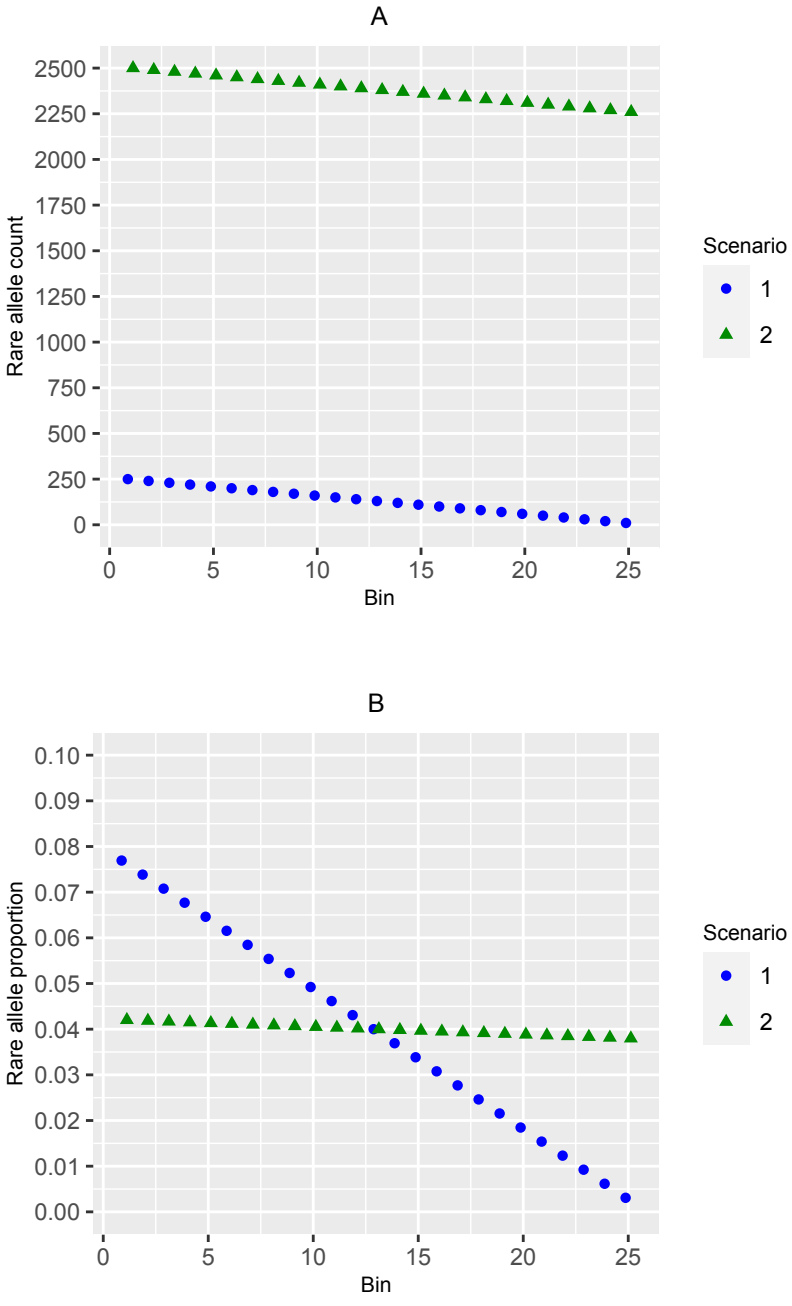
samples are ordered based on expression level, from lowest to highest expression; samples are then assigned to equally sized bins, whereby the lowest bin contains the samples with the lowest expression values for this gene, and the highest bin contains the samples with the highest expression for the gene (Step 1a). For instance, for the set of 725 samples with RNA sequencing data, we performed analyses that employed 25 bins that each contained 29 samples. For the samples within each bin, rare alleles within or near the gene being considered are summed, yielding a total rare allele count for each bin for this gene (Step 1b). This process is repeated for all genes to be analyzed. Then, to increase power for detecting rare variant associations with gene expression, the rare allele totals for each bin are summed across all genes (Step 2), yielding an overall rare allele total for each bin (i.e., sum rare allele totals for Bin 1 across all genes, yielding an overall rare allele total for Bin 1; do the same for all bins). Finally, association between rare allele total and bin number (with low bin number corresponding with low expression and high bin number corresponding with high expression) is examined using a linear model that regresses rare allele total onto bin number, with and without a quadratic term for bin, to evaluate different possible relationships between rare allele burden and gene expression (Step 3). Although the rare alleles are expected to affect gene expression, and not vice versa, regression of rare allele total onto expression bin number is a valid means of examining the nature of the association between rare allele burden and gene expression bin, and is preferable to regressing bin number on rare allele total as it facilitates interpretation of results (e.g., the change in rare allele burden corresponding to a 1-unit increase in expression bin is more easily interpreted than the fractional change in expression bin corresponding to an additional rare allele). Confounders and other covariates are adjusted before Step 1 by regressing normalized expression values for each gene onto covariates, and then taking the residualized expression values forward to Step 1.

Figure 4.4. Demonstration of steps involved in the Zhao et al. (2016)[27] rare allele burden method. Figure adapted from Figure 1 of Zhao et al. (2016).



The approach of analyzing cumulative rare allele counts has certain undesirable properties. For one, the parameters estimated from a linear regression of rare allele count onto bin number are limited in their informativeness and may lead to incorrect conclusions regarding rare allele associations with gene expression. For instance, in **Figure 4.5** below, the beta for a linear regression of rare allele count (y-axis) on expression bin (x-axis) and is -10 for scenarios 1 and 2, meaning that in both instances an increase in bin number by 1 corresponds to a decrease in rare allele count by 10. In addition, the standard errors and p-values for these two analyses are identical. However, these identical estimates mask the important fact that rare alleles are distributed quite differently for scenarios 1 and 2, with the proportion of rare alleles per bin very different across the bins in scenario 1 (e.g., bin 1 contains 7.7% of all rare alleles and bin 25 contains 0.3% of rare alleles), and not so different across the bins in scenario 2; compared with scenario 2, the distribution of rare alleles in scenario 1 is actually consistent with a stronger association of rare alleles and gene expression.

Figure 4.5. Illustration of one undesirable property of analyzing cumulative rare allele counts for the burden analysis. Figure 4.5 A (top) shows the distribution of rare allele counts (y-axis) across expression bins (x-axis) for scenarios 1 and 2. Linear regression of rare allele count on expression bin yields identical betas (β = -10), standard errors and p-values for these two scenarios, masking differences that are revealed when considering rare allele proportions as shown in Figure 4.5 B (bottom). The distribution of rare allele proportions for scenario 1 is consistent with a stronger association of rare alleles with gene expression.

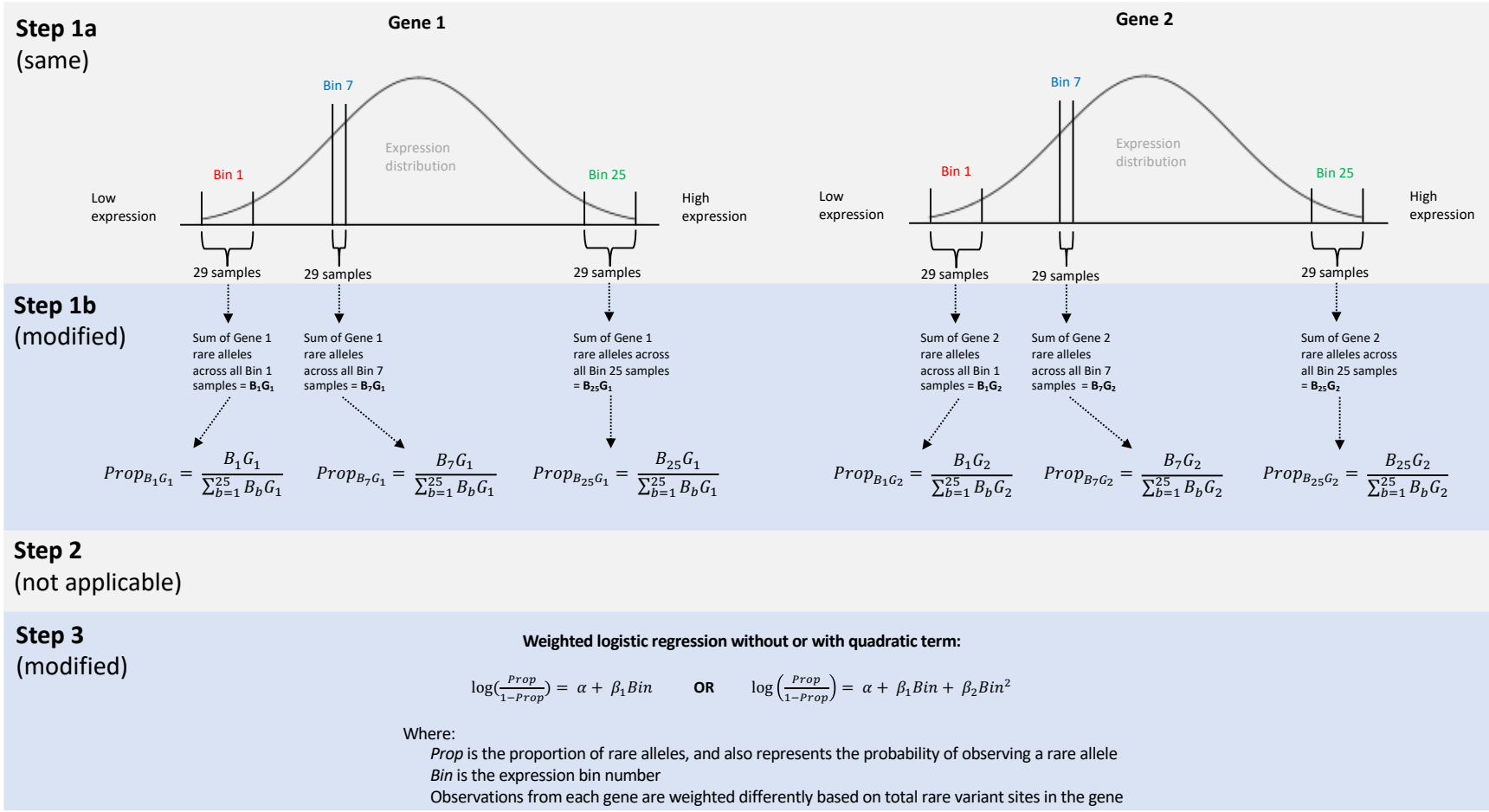


A second undesirable property of analyzing cumulative rare allele counts is the counterintuitive and potentially misleading observation that, when performing multiple separate analyses, as the total number of rare alleles considered in an analysis increases (for instance, due to increasing the number of genes considered, or relaxing the MAF threshold or other variant filters), the variance in rare allele counts across all bins also tends to increase, along with increases in the standard error (SE) of the estimated beta. This is counterintuitive because we would expect that a larger quantity of rare alleles considered for the analysis would typically correspond with increased precision of the estimate, in the same way that increases in sample size yield increased precision. As with the first undesirable property described in the preceding paragraph, this tendency of cumulative allele count analyses to exhibit oddly behaving variance has the potential to lead to incorrect conclusions regarding the associations of interest, especially when comparing results from separate analyses based on differing numbers of rare alleles.

These two undesirable properties related to performing the burden analyses using cumulative rare allele counts can be overcome by modeling rare allele proportions rather than counts. We therefore modified the burden approach by converting rare allele counts for each gene into rare allele proportions, effectively normalizing the counts for all genes, and we then directly analyzed these proportions. Specifically, for each gene, we divided the rare allele count for each bin (as obtained in Step 1b of **Figure 4.4**) by the total rare allele count for the gene, yielding rare allele proportions for each bin that collectively summed to 1 for the gene. We then used R's rbind() function to stack each gene's set of rare allele proportions atop one another (keeping the data for each gene separate, as opposed to the Zhao et al. (2016) approach of summing rare allele counts across all genes as shown in Step 2 of **Figure 4.4**), and we used logistic regression to model rare allele proportion as a function of expression bin number. We accomplished this through R's glm() function with family=quasibinomial(link=logit) and with observations for each gene weighted according to the number of rare variant sites for the gene.

Modeling the response variable using a quasibinomial distribution is often recommended when the outcome data are proportions. Weighting data according to the total number of rare sites in the corresponding gene tends to result in genes with a larger number of rare allele observations carrying greater weight, which is generally desirable. We decided not to simply weight based on the total observed rare alleles for each gene since doing so could result in genes harboring a greater amount of higher frequency variants carrying the greatest weight, and such genes may exhibit a different association between rare alleles and gene expression than other genes. The steps of our modified burden approach are depicted in **Figure 4.6**.

Figure 4.6. Demonstration of steps involved in our modified rare allele burden approach.



Genes for analyses

To be eligible for analysis, a gene needed to be present in the normalized, post-QC expression datasets and to have SNVs present in the post-QC targeted DNA sequencing dataset. The post-QC targeted DNA sequencing dataset had 16,573 total SNVs. We used Bystro to obtain annotation information for these variants. The 16,573 SNVs were annotated to a total of 268 unique genes. Of these 268 genes, 211 were included among the set of genes that had undergone DNA sequencing due to having SZ-associated expression (64 of the 211 genes) or being located within or near a SZ CNV interval (147 of the 211 genes, non-overlapping with the 64 SZ-associated expression genes). Taking the intersection of these 211 targeted DNA sequencing dataset genes and the 13,584 genes present in the final RNA sequencing dataset yielded 160 genes with both SNV and RNA sequencing expression data.

To determine the overlap between the 211 targeted DNA sequencing dataset genes and genes represented in the microarray expression dataset, we first mapped the 27,118 filtered microarray expression probes to their corresponding genes using the manifest provided by Illumina for their HT-12v4 chip. Then, limiting to probes mapping to genes among the 211 targeted sequencing genes yielded 226 probes that collectively mapped to 157 unique genes. For genes represented by multiple probes, we averaged expression values across the probes to obtain a single expression value for the gene, which is consistent with the approach used by Zhao et al. (2016).[27] Thus, we had 157 genes with both SNV and microarray expression data.

Original meta-analytic plan and subsequent modification:

Of the 160 genes with SNV and RNA sequencing expression data, 139 genes were also included among the 157 genes with SNV and microarray expression data. Our original analytic plan was to only analyze these 139 overlapping genes, first performing separate rare allele burden analyses for these 139 genes using the RNA sequencing dataset and the microarray dataset, and then combining the results via meta-analysis. However, as described in depth in

the **Results** section below, we came to the realization that for a large proportion of genes the microarray dataset appeared limited in its ability to accurately order individuals with respect to expression level, which is essential for our analytic approach. As a result, we modified our analysis plan. After considering various options, we decided that the best approach for our purposes would be to focus only on the RNA sequencing data, which we had reason to believe were overall of higher quality and better able to accurately place samples with respect to expression ordering.[109,110] Thus, we ultimately set aside the microarray expression dataset, and moved forward with analyzing the 160 genes overlapping between the targeted DNA sequencing dataset and the RNA sequencing dataset.

Covariate adjustment

We considered numerous covariates for adjustment, including potential confounders of the estimated association of rare allele burden with gene expression level, as well as covariates that are unlikely to be confounders but which contribute to increased variance in gene expression level and as such may make associations between rare alleles and gene expression more difficult to detect and estimate. The potential confounders that we ultimately controlled included fine-scale ancestry (controlled by adjusting for genomic PCs 1 through 5) and targeted DNA sequencing batch (10 batches). Covariates controlled for the purpose of reducing unwanted variance in gene expression included expression batch (analyses of the RNA sequencing data adjusted for 5 RNA sequencing batches; analyses of the microarray expression data adjusted for 2 microarray batches), LCL growth rate, LCL energy level (ATP level), LCL viral load (Epstein-Barr virus [EBV] load), LCL EBV transformation site, sex and age (at the time of sample collection). These variables were included as covariates in previous analyses of these RNA sequencing and microarray expression data,[14-16] and/or have been controlled for in other previous investigations of associations between genetic variants and gene expression levels.[24,27]

As was done by the Zhao et al. (2016) group, we also adjusted for common SNVs (MAF > 1%) found to be associated with gene expression levels. These SNVs, known as eQTLs, are an important contributor to gene expression variability, and also may at times be associated with nearby rare variants, warranting adjustment for them to gain a better estimate of the association between rare alleles and gene expression. We identified eQTLs by performing eQTL analyses using our expression datasets, as described in the subsection directly below.

In addition, based on scientific literature related to the proper analysis of secondary phenotypes using a combined sample of cases and controls, we adjusted for case-control status. We hypothesized that SZ case-control status might be a collider variable, with gene expression level (our outcome) expected to affect SZ risk, and rare regulatory variation (our predictor) possibly impacting SZ in part through pathways independent of gene expression. If this were the case, the standard epidemiologic approach would be to not condition on case-control status, as conditioning on a collider has the potential to open a biasing path between variables of interest. However, numerous studies examining the conduct of secondary analyses in combined case-control datasets have demonstrated through both simulated and real data analyses that, if the disease defining case status is rare (prevalence < 2%), then *not* adjusting for case-control status when it is a collider is likely to yield biased estimates of the association between the predictor and the secondary phenotype of interest, whereas adjusting for case-control status is expected to yield results that are not biased by the case-control ascertainment and that are good approximations of the true association (assuming other sources of bias are small).[111-115] As SZ has a prevalence of ~1%, it meets this rare disease criteria, and as such adjusting for case-control status is the recommended approach. We note that other methods, including weighting observations by the inverse probability of selection, are also able to correct bias resulting from oversampling of cases; however, these alternative methods are not clearly applicable to our analyses given the unique burden approach which we employed. We further

discuss the issue of properly accounting for case-control ascertainment in our analyses in the **Sensitivity Analyses** section below.

A number of prior studies investigating associations between genetic variants and gene expression have used computational approaches to infer hidden factors or principal components from gene expression data, and then included these factors as covariates in their analyses in order to control for unmeasured sources of expression variability (e.g., unmeasured batch effects) which may confound or otherwise impede estimation of associations.[24,27,116] We tested several such approaches, including inferring hidden determinants of gene expression using probabilistic estimation of expression residuals (PEER)[117] as well as computing PCs from the expression data (using R's prcomp()), and then including the inferred factors as covariates in our eQTL analyses. Such approaches actually served to increase inflation in eQTL results, therefore we decided against controlling for these inferred expression factors in our rare allele burden analyses and in our final eQTL analyses.

Identifying eQTLs for inclusion as covariates

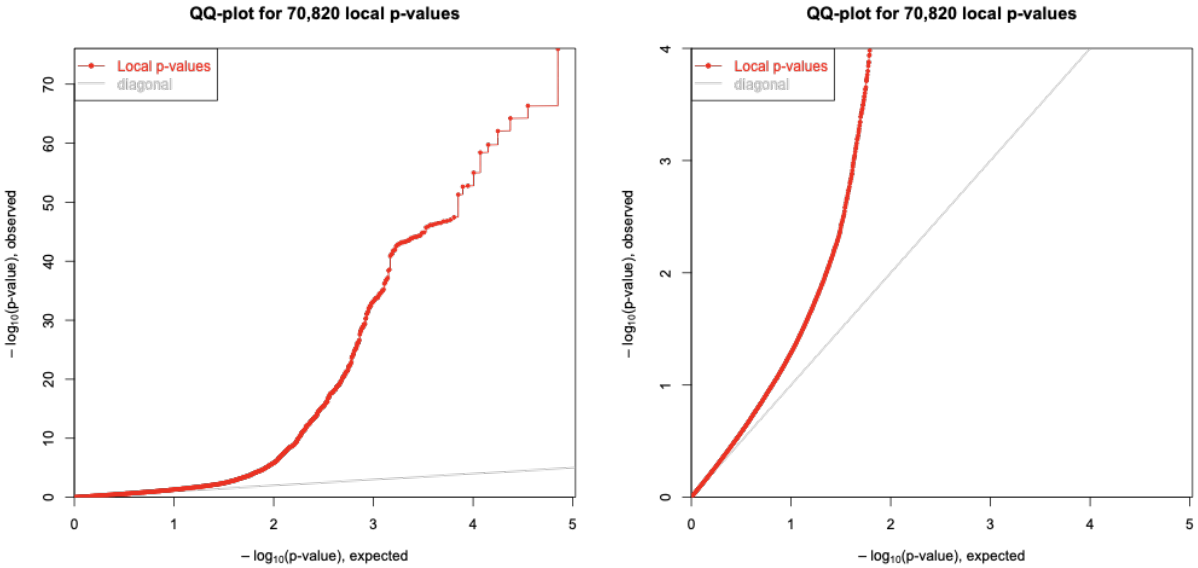
We originally performed eQTL analyses separately for the RNA sequencing dataset (725 samples) and the microarray expression dataset (400 samples), and then meta-analyzed the results to identify eQTLs that would be used as covariates. However, after realizing the questionable quality of the microarray expression data for our study purposes and deciding to move forward only with the RNA sequencing dataset, we also moved forward with only identifying eQTLs based on the RNA sequencing dataset. We describe these RNA sequencing dataset eQTL analyses in the following paragraphs.

We carried out eQTL analyses using the software MatrixEQTL.[118] Expression data input were the fully cleaned and normalized RNA sequencing data (160 genes). Given the use of linear regression for the eQTL analyses, we transformed the expression values for each gene to be normally distributed. Genotype data were the cleaned GWAS SNP data limited to variants

with MAF > 1%. We only performed cis-eQTL analyses, and considered as in-cis any SNP within the window of 1 Mb upstream of the transcription start site for the gene to 1 Mb downstream of the transcription stop site. Transcription start and stop sites for each gene were determined using a combination of the RefSeq Select+MANE table and the GENCODE v32 knownCanonical from UCSC's Table Browser (hg38 positions).[119] We also used UCSC's LiftOver tool to update SNP coordinates to hg38, in order to match the gene coordinates.[52] Covariates adjusted were the same as those included for the rare allele burden analyses (other than adjusting for eQTLs), but instead of targeted DNA sequencing batch we controlled for genotyping batch.

The overall eQTL results were inflated, based on quantile-quantile plots (see **Figure 4.7** below for a depiction of inflation observed for the RNA sequencing sample) and genomic inflation factor lambda (λ_{RNAseq} = 1.20). We note that this inflation in eQTL results had also been observed for analyses of the microarray expression sample.

Figure 4.7. Quantile-quantile (QQ) plots for eQTL results using the RNA sequencing sample. Right plot is zoomed in at the bottom left corner to better show inflation.



As has been done in previous studies,[24,27,116] we estimated hidden factors from the RNA sequencing expression data (using PEER software and R's prcomp()) and controlled for different subsets of these hidden factors in the eQTL analyses, but this only served to increase the inflation. We also tried other approaches, such as only including certain subsets of the covariates in the analyses, but inflation in the eQTL results remained. We ultimately decided to correct for the inflation using the method of genomic control, which involves dividing all χ^2 test statistics for the eQTL results by the genomic inflation factor lambda. We then recomputed false discovery rate (FDR) q-values using these adjusted test statistics. We considered q-values < 0.05 as indicative of a possible eQTL; out of 70,820 SNP-gene pairs 1,312 had inflation-adjusted q < 0.05.

Furthermore, as was done by Zhao et al. (2016), we only adjusted for 'conditional eQTLs', which are SNPs that are associated with gene expression independent of nearby

SNPs. We identified conditional eQTLs for each gene through an iterative process that involved the following steps:

1. Regress gene expression values onto the most significant eQTL for the gene (the eQTL with the smallest q-value, among those eQTLs with inflation-adjusted $q < 0.05$) as well as all covariates included in the original eQTL analyses;
2. Extract the expression residuals from this regression, and regress these residuals onto each remaining potential eQTL for the gene one at a time;
3. For each additional potential eQTL tested in (2), compute an inflation-adjusted q-value (using $\lambda_{RNAseq} = 1.20$, the lambda value calculated using the original, complete RNA sequencing dataset results);
4. Take the eQTL from (3) with the smallest q-value (among those with inflation-adjusted $q < 0.05$), and regress the expression residuals generated in (1) onto this eQTL;
5. Extract the expression residuals from the regression in (4), and repeat steps (1) through (4) until there are no longer any eQTLs with inflation-adjusted $q < 0.05$.

The identified conditional eQTLs were then controlled for in the rare allele burden analyses.

Analysis subsets

We performed analyses for the set of 64 genes sequenced due to having expression levels previously identified to be associated with SZ. We also stratified these 64 genes into those with evidence of low expression associated with SZ, and those with evidence of high expression associated with SZ, and we performed separate analyses for these two gene sets to examine whether rare allele associations with gene expression may differ for genes with low versus high expression linked with SZ. We then combined these 64 SZ-associated expression genes with the genes sequenced due to being located within or near a large SZ CNV interval (which we expect to be enriched for genes with SZ-associated expression), yielding an overall

combined set of 160 genes. It turned out that 11 of these 160 genes had no remaining rare regulatory variants after limiting to the variants of interest for our combined set analyses; thus, 149 genes were ultimately analyzed for the overall combined gene set. We analyzed this combined gene set, and then split this set of 149 genes into those with suggestive evidence of low dosage associated with SZ (genes with low expression associated with SZ, or existing within or near a SZ-associated large deletion) and those with suggestive evidence of high dosage associated with SZ (genes with high expression associated with SZ, or existing within or near a SZ-associated large duplication), and performed separate burden analyses for these two gene sets. If a gene was located in or near a SZ duplication interval and was found to have low expression associated with SZ (this was the case for 2 genes), we only included this gene in the set of SZ-associated low dosage genes; there were no instances of a gene being located in or near a SZ deletion interval and having high expression associated with SZ.

In addition, we took the combined set of 149 genes and subsetted it based on evidence of intolerance to loss-of-function (LoF) variation, as well as intolerance to missense variation. Constraint metrics for each gene were identified using the gnomAD gene constraint table from UCSC's Table Browser. We created and analyzed gene sets for genes with probability of intolerance to LoF variation (pLi) < 0.10 (suggesting tolerance to LoF variants) and pLi ≥ 0.90 (suggesting intolerance to LoF variants). We also created and analyzed gene sets that included genes that were highly tolerant to missense variation, and those that were intolerant to missense variation. We defined high tolerance to missense variants as having a ratio of observed to expected missense variants (with MAF < 0.001) that was greater than 1, with a 90% confidence interval (CI) that did not overlap 1; while our set of missense-intolerant genes all had missense variant observed to expected ratios of less than 1, with the 90% CI not overlapping 1.

In addition to analyzing these various gene sets, we performed analyses that applied a variety of rare variant filters. We varied the MAF threshold for SNVs, applying filters of MAF <

0.01 and < 0.001. This was done to examine the extent to which very rare alleles (MAF < 0.001) may be driving associations between rare SNVs and gene expression. MAFs used for filtering were based on gnomAD frequencies if they were available (i.e., if they were not missing in the gnomAD v2 database); otherwise they were based on the post-QC targeted DNA sequencing dataset with 3,180 samples. For multiallelic SNVs, we required that the sum of the non-major alleles not exceed the MAF threshold.

We also filtered variants according to Combined Annotation Dependent Depletion (CADD) score, which is a prediction of variant deleteriousness; variants with higher scores are predicted to be more deleterious.[120] We performed analyses with no CADD filter, and analyses only including variants with CADD ≥ 5. We originally considered applying a stricter threshold of CADD ≥ 10, which is often used by researchers, but we realized that such a filter would generally result in quite small numbers of rare alleles available for analysis and consequently highly imprecise association estimates.

In addition, we performed analyses that varied site type for the variants. For the set of 64 genes with SZ-associated expression, we performed analyses that included any regulatory variants (those located within the promoter, 5'UTR or 3'UTR), only regulatory variants upstream of the coding sequence (promoter or 5'UTR), and only UTR variants (5'UTR or 3'UTR); as well as analyses that only included promoter, 5'UTR or 3'UTR variants. We defined a gene's promoter region (or promoter-proximal region) as the genomic window encompassing DNA sequence from 1 kb upstream to 1 kb downstream of the gene's TSS, which matches the promoter region window used by Zhao et al. (2016).[27] We identified TSS for each gene using UCSC's Table Browser, as described previously. 5'UTR and 3'UTR for each gene were based on UCSC's refGene,[119] with assignment of variants to these regions made by the Bystro annotation tool.[43] When analyzing gene sets that included the genes sequenced due to being located within or near a large SZ CNV interval (e.g., the combined set of 149 genes), we only performed analyses for UTR variants, including 5'UTR and 3'UTR variants together and

separately; we did not include promoter variants for these gene sets, since the SZ CNV genes had not been sequenced upstream of the TSS.

Considering the numerous analyses performed for various combinations of genes and variant filters, we employed a multiple testing correction procedure. Use of a traditional approach such as the Bonferroni correction would be overly strict, since the Bonferroni method assumes independence of analyses, and our analyses are highly correlated (e.g., many analyses involve genes and variants that are subsets of those used for other analyses). We therefore employed a permutation approach to multiple testing correction, performing our analyses for 10,000 randomly permuted datasets (randomly shuffling samples' sets of gene expression data) to determine a significance threshold that maintains the study-wise Type I error level (α) at 5%.

Sensitivity analyses

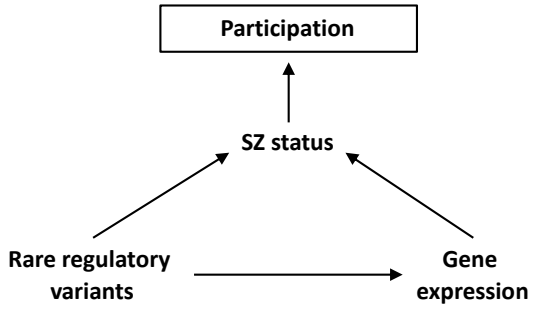
Overrepresentation of cases in the final analytic sample

Our goal in performing the analyses described in this manuscript was (to the extent possible) to accurately and precisely estimate the nature of the association between rare regulatory variants and gene expression for the population of individuals with European ancestry (the source population for our analytic sample). Our analytic sample of 725 participants was drawn from an existing case-control study of SZ, and included 355 SZ cases and 370 controls without SZ; thus, individuals with SZ were greatly overrepresented in our analytic sample as compared with the ~1% SZ prevalence observed for the source population.

This overrepresentation of cases in our analytic sample has the potential to bias estimates of the association of rare regulatory alleles with gene expression levels. Such bias might arise if rare regulatory alleles, gene expression, and SZ status are causally related as depicted in the directed acyclic graph (DAG) presented in **Figure 4.8**. This DAG shows our expectations that rare regulatory variants can cause deviations in gene expression level that

push expression levels toward the tails (lower and/or upper) of the expression distribution; and that, in turn, aberrant expression levels may increase risk for SZ. In addition, it is possible that rare regulatory variants increase SZ risk through pathways that do not involve gene expression, though the extent to which this may occur is unclear. Lastly, the DAG shows that whether or not an individual has SZ affects their probability of participating in our final analytic sample, since the prevalence of SZ in the source population is ~1%, while SZ cases make up 49% of our analytic sample. Our analyses necessarily condition on participation in the analytic sample (represented by the box around the 'Participation' variable in **Figure 4.8**), and as participation is a proxy for SZ status (in the DAG, 'SZ status' directly affects 'Participation'), we are also effectively conditioning on SZ status, reflected by the overrepresentation of cases in our sample. As SZ status is a collider in the DAG, this has the potential to open a biased pathway between rare regulatory variants and gene expression, which could result in bias that mixes with any causal effect.

Figure 4.8. Directed acyclic graph (DAG) depicting possible causal associations between variables in our study.



Another consideration related to the overrepresentation of SZ cases in our analytic sample is that, if the association between rare regulatory alleles and gene expression differs for individuals with and without SZ, then the overrepresentation of cases in our analytic sample could yield estimated associations that are different from the true association in the full source

population. However, we had little reason to think that the effect of a given rare regulatory variant on gene expression would differ meaningfully for individuals with and without SZ.

Accounting for the overrepresentation of cases in the final analytic sample

If SZ status were a collider variable with respect to our variables of interest, the typical epidemiologic guidance might be to apply a method such as inverse probability of selection weighting (IPW) to recreate the source population with respect to the distribution of those with and without SZ (and possibly other variables, as well), enabling valid estimates of the association of interest (assuming no bias from other sources).[111,121] However, IPW and similar techniques typically involve applying weights at the individual subject level, and there is no clear and validated approach for joining these techniques with our unique burden analysis method which examines associations at the level of expression bin.

Another approach for avoiding bias due to an overrepresentation of cases would be to simply limit our analyses to the 370 controls only. As SZ is rare condition (prevalence ~1%), the controls (who are free from SZ) would be expected to be a fairly good representation of the source population, assuming they were sampled at random. Indeed, numerous studies have demonstrated that, if the disease defining case status is rare, then performing secondary analyses using only the controls from a case-control study is expected to yield results that are very good approximations of the association in the full source population.[111,112,115,122] However, limiting analyses to controls only has the drawback of excluding a large amount of potentially good data (cases make up ~50% of our sample), and as a result would yield estimates that are less precise than those obtained from valid approaches that utilize data from cases and controls.

As described previously, an alternative approach to properly account for the overrepresentation of SZ cases in our final analytic sample is to directly condition on case-control status (e.g., include case-control status as a covariate). Though this may seem to be at

odds with the standard epidemiologic guidance to avoid conditioning on a potential collider variable, the success of this approach in yielding valid estimates for particular scenarios such as ours has been demonstrated by numerous studies employing both simulated and real data analyses to examine how to properly carry out secondary analyses of combined case and control data. Specifically, investigators have shown that, when the predictor and secondary phenotype independently affect the disease defining case status (i.e., case status is a collider variable, as shown in **Figure 4.8**), then if the disease is rare (prevalence < 2%), analyses that condition on case-control status (i.e., include it as a covariate) will yield estimates that closely approximate the true association of the predictor and the secondary phenotype in the source population (assuming minimal interaction between the predictor and the disease defining case status); if case-control status is *not* controlled in this scenario, the estimate is likely to be meaningfully biased.[111-115] These studies have also demonstrated that, in this rare disease scenario, analyses of controls only and cases only are both expected to yield estimates that closely approximate true associations in the overall source population (if there is meaningful interaction, the cases-only estimates will differ from the source population, but the controls-only estimates remain a good approximation due to the rarity of disease). However, the controls only or cases only approaches offer less statistical power than the combined approach, which adjusts for case-control status.

As SZ is a rare disease, and we think it unlikely that the effect of rare regulatory variation on gene expression would differ by SZ status, performing analyses using the combined case and control data and adjusting for case-control status is expected to yield estimated associations that closely approximate the association in the source population (assuming other sources of bias are small). Thus, our main analyses consider cases and controls together and adjust for case-control status. We decided to perform additional, sensitivity analyses using controls only, SZ cases only, and also using the combined set of cases and controls but not adjusting for case-control status. As an additional validated approach for secondary analyses of

case-control data when the disease is rare, controls-only estimates that are consistent with the estimated associations from our main analyses will further support our analytic approach. Estimates based on cases only can be compared with the controls-only estimates to examine whether our expectation of no meaningful interaction by SZ status is empirically supported by the data. Estimates from analyzing the combined set of cases and controls and not adjusting for case-control status, if consistent with associations estimated from our main analyses, may indicate that SZ status is in fact not a concern as a collider in our analyses.

Selection considerations for controls and cases in the final analytic sample

It is worth discussing the selection processes for the controls and SZ cases included in our final analytic sample, and the extent to which they may be representative of their respective source populations (individuals without and with SZ, respectively). These are important to consider since these were the samples that made up our full analytic dataset, and also since such consideration would inform interpretation of our sensitivity analyses of controls only and cases only. We previously mentioned that the individuals who underwent targeted DNA sequencing represent a random sample of the SZ cases and controls from the final MGS GWAS, which in turn is expected to be at least decently representative of individuals with and without SZ in the population of individuals with European ancestry. We also had the impression that the RNA sequencing dataset likely represented a randomly selected set of samples from the final MGS GWAS (aside from matching cases and controls on age and gender); in which case the overlapping samples between the targeted DNA sequencing sample set and the RNA sequencing sample set (i.e., the samples included in our final analytic dataset) would represent a conditionally (within strata of sex and age) random selection from the MGS GWAS participants. However, we were unable to verify whether the RNA sequenced individuals were indeed selected in this manner.

We decided to compare the distribution of certain demographic variables for the 355 SZ cases and 370 controls in our final analytic dataset with the distributions observed in the larger case and control datasets from which our final participants were selected, with the latter datasets thought to be fairly good representations of individuals with and without SZ among those with European ancestry (EA). We were unfortunately quite limited in the number of potentially relevant variables available for comparison, and so were unable to perform as thorough an examination as would have been liked.

Table 4.3 presents distributions of three variables for the set of EA SZ cases recruited by the MGS study (the full dataset before sample QC), and also for the set of EA SZ cases present in the final analytic dataset. The gender distribution is noticeably different for the MGS study and our set of SZ cases: the male to female ratio is 2.3:1 for the MGS study (reflecting known sex differences in the manifestation of SZ),[123] while it is approximately 1:1 for our cases. This difference is expected since, as noted previously, cases and controls were selected have 1:1 gender ratios for the RNA sequencing dataset. The other two variables, age and location of recruitment, are distributed very similarly for the MGS cases and the final set of cases. The observed distributions of these three variables provide at least minimal support for the idea of cases having been selected on gender but otherwise being representative of the full set of EA SZ cases recruited by the MGS study, possibly through a random selection process.

Table 4.3. Distributions of three variables for the set of EA SZ cases recruited by the MGS study (the full dataset prior to sample QC), and also for the set of EA SZ cases present in our final analytic dataset.

		2,873 MGS EA SZ cases (before GWAS QC)	355 EA SZ cases in final analytic dataset
Gender	Male	69.7%	47.3%
	Female	30.3%	52.7%
Age (years)	18-29	14.8%	13.8%
	30-44	37.8%	36.9%
	45-59	39.3%	39.4%
	60+	6.6%	9.9%
Recruitment location	Australia	23.9%	23.1%
	California	10.0%	9.9%
	Colorado	15.5%	18.6%
	Georgia	9.2%	8.2%
	Iowa	7.6%	9.3%
	Illinois	12.2%	9.3%
	Louisiana	3.3%	4.8%
	Massachusetts	1.1%	0.0%
	Montana	6.2%	7.9%
	New York	6.3%	4.8%
	Pennsylvania	4.5%	3.9%
	Texas	0.2%	0.3%

Table 4.4 presents distributions of five variables for four sets of samples: the initial group of 15,485 EA individuals who were randomly selected for potential participation as controls in the MGS study (selected from a large participant panel that was representative of the U.S population in terms of important demographic dimensions), the subset of 3,364 EA individuals (among the 15,485 targeted EA individuals) who completed the required self-report clinical assessment and blood draw, the set of EA controls after excluding ineligible participants (including those with a history of SZ) but before performing GWAS QC, and the set of EA controls present in our final analytic dataset. The distributions of gender and age appear to be rather different for our final control set compared with the three larger sample sets, with the three larger sample sets having quite similar distributions for these variables. The age distribution for our controls is actually expected, as it approximately matches the age distribution of our 355 cases, and we know that RNA sequenced cases and controls were matched on age categories. The distribution of gender among our final controls is rather unexpected, assuming

intended selection of females and males in a 1:1 ratio. The variable educational attainment is distributed somewhat differently across all sample sets, with the proportion of more highly educated individuals gradually increasing from left (the initial group targeted for recruitment) to right (the final analytic control set). Considering the final set of controls, the relatively higher proportion of individuals with greater educational attainment could partly be due to exclusion of individuals with SZ and SZ-related conditions, as SZ has been found to be associated with lower educational attainment;[124] while other unintended, non-random selective forces may have had a role as well. The distributions of region and urbanicity appear rather similar across these four sample sets. Considering these comparisons, particularly those for gender and educational attainment, it seems possible that there was some degree of unintended, non-random selection forces in operation during the processes which ultimately led to our final RNA sequenced control set.

Table 4.4. Distributions of five variables for four sets of samples: the initial group of 15,485 EA individuals who were randomly selected for potential participation as controls in the MGS study (selected from a large participant panel which was representative of the U.S population in terms of important demographic dimensions); the subset of 3,364 EA individuals (among the 15,485 targeted EA individuals) who completed the required self-report clinical assessment and blood draw; the set of EA controls after excluding ineligible participants (including those with a history of SZ) but before performing GWAS QC (this is slightly less that the previously mentioned 2,817 genotyped EA controls, because we were not able to ascertain information on all 2,817 participants); and the set of EA controls present in our final analytic dataset.

		15,485 EA randomly selected for control recruitment	3,364 EA completing assessment and blood draw	2,806 MGS EA controls (before GWAS QC)	370 EA controls in final analytic dataset
Gender	Male	48%	47%	48.0%	59.5%
	Female	52%	53%	52.0%	40.5%
Age (years)	18-29	13%	14%	13.0%	13.8%
	30-44	27%	27%	26.1%	41.9%
	45-59	30%	29%	29.1%	39.5%
	60+	30%	30%	31.8%	4.9%
Education	< HS	10%	8%	7.2%	5.7%
	HS	30%	27%	26.1%	23.5%
	Some college	31%	31%	30.0%	30.3%
	≥ Bachelor	29%	34%	36.2%	40.5%
Region of residence	Northeast	22%	18%	18.5%	15.1%
	Midwest	27%	27%	27.7%	30.0%
	South	35%	37%	36.0%	34.3%
	West	16%	18%	17.8%	20.5%
Urbanicity of residence	Non-metro	18%	15%	15.1%	19.5%
	Metro	82%	85%	84.9%	80.5%

Regarding the limited set of variables just considered, we do not expect gender or age to be collider variables in our analyses, and therefore do not expect non-representative gender and age distributions in our final analytic sample to result in biased estimates of the association of rare regulatory variants and gene expression. Educational attainment could potentially be a collider, in which case the overrepresentation of more highly educated individuals among our final controls might contribute some bias to associations estimated from the analyses of controls only. However, considering evidence of SZ's association with lower educational attainment, a comparison of the controls-only results with the all-samples (controls and SZ cases combined) results which shows little difference in patterns of estimated associations would seem to suggest that the overrepresentation of more highly educated participants among the final

controls has little biasing effects. Ultimately, we expect the control-only and case-only analyses to yield decent approximations of the associations that would be estimated in the source populations (EA individuals without and with SZ, respectively).

RESULTS

RNA sequencing and microarray expression discordance

In preliminary analyses of the set of genes overlapping between the RNA sequencing and microarray expression datasets, we observed unexpectedly discrepant results between the two datasets. We investigated a multitude of different potential explanations for these discrepancies, including differences in sample characteristics (distributions of sex, age, LCL transformation site, etc.) between the 725 RNA sequencing and 400 microarray samples, with no success in understanding why the discrepancies might exist.

We then identified a set of 68 samples who had both RNA sequencing data and microarray expression data; 23 of these samples were included among the 725 final RNA sequencing samples, while the remaining 45 were not included in the final RNA sequencing or microarray expression analysis datasets (most of these 45 samples were excluded due to not having targeted DNA sequencing data). We used these 68 samples to check concordance between RNA sequencing and microarray expression. For these concordance checks, we only considered the 139 genes described above which overlapped between RNA sequencing and microarray datasets, and we use the normalized expression data. Sample-wise correlations (generated by correlating the 139 RNA sequencing gene expression values for a particular individual with the 139 microarray gene expression values for the same individual) were moderate, ranging from approximately 0.50-0.60.

Gene-wise correlations (generated by correlating the RNA sequencing expression values of all 68 samples for a particular gene with the microarray expression values of all 68 samples for the same gene), however, were notably low. Median gene-wise correlation in expression values was 0.26. We also performed gene-wise correlation of the expression ranks, after ranking each of the 68 samples with respect to expression (separately for the RNA sequencing and the microarray expression datasets). Expression rank correlations are perhaps even more relevant for our purposes since accurate ordering of expression values is of primary importance to the rare allele burden approach we were employing. Gene-wise correlations for expression ranks exhibited a median correlation of 0.24.

In addition, we calculated gene-wise correlations by taking the normalized RNA sequencing and microarray expression values, regressing out all covariates controlled for in the final burden analyses aside from targeted DNA sequencing batch, and correlating these expression residuals. This is perhaps a better means of assessing expression concordance, as differences in normalized RNA sequencing and microarray expression rankings could be influenced by things like RNA sequencing or microarray batch, as well as LCL growth rate and LCL energy level which can differ between RNA sequencing and microarray datasets for the same sample. Median correlation between expression residuals was 0.18, and median correlation between ranks of expression residuals was also 0.18.

We attempted to find an explanation for why certain genes were more weakly correlated between RNA sequencing and microarray datasets than others. Things we considered included median microarray and RNA sequencing expression level for the gene; variance in microarray and RNA sequencing expression for the gene; number of microarray probes averaged to obtain the microarray gene expression level; whether any microarray probes have an original Illumina gene assignment that differs from a reannotation to the latest human genome build (hg38); transcript length; and whether applying higher expression threshold filters to the microarray and RNA sequencing data would impact correlations. None of the factors we considered explained

the large amount of low gene-wise correlations between the RNA sequencing and microarray expression datasets.

We ultimately concluded that the observed low correlations between RNA sequencing and microarray expression values likely reflected a reduced ability of microarray expression technology to accurately estimate relative expression levels in comparison with RNA sequencing technology. RNA sequencing is a newer technology for profiling gene expression that offers a number of advantages over the older microarray methods, including a greater dynamic range enabling more accurate profiling of lowly expressed genes.[109,110] The conclusion that the RNA sequencing data represented the better quality dataset was also supported by the observation that preliminary analyses of the RNA sequencing dataset yielded estimated associations between rare regulatory allele burden and gene expression that, in comparison with the microarray findings, tended to be more consistent with estimates from prior research;[27] as well as the observation that removing genes exhibiting low gene-wise correlations tended to change the RNA sequencing results minimally while bringing the microarray expression results more in line with the RNA sequencing results.

We considered several options for addressing the presumed lower quality microarray expression data, including multiple imputation of microarray expression values as well as subsetting the microarray genes to those exhibiting sufficiently high gene-wise correlations. In the end, though, we decided that simply excluding the microarray dataset and moving forward by only analyzing the RNA sequencing dataset was the approach that would enable the greatest confidence that our study results were based on the best quality of expression data available. The results sections that follow describe findings based on analyzing up to 149 genes using RNA sequencing data only.

RNA-sequencing-only analyses

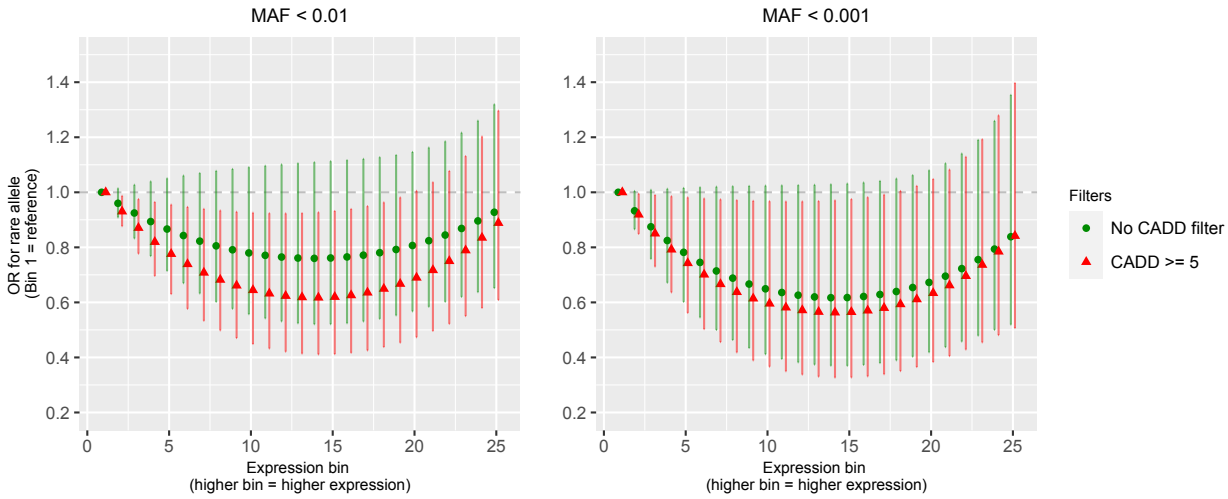
Genes with SZ-associated expression levels

We performed burden analyses for a set of 64 genes previously identified to have expression levels that were associated with SZ. For these genes, in addition to all exons being sequenced (which includes the 5'UTR and 3'UTR), the region 2kb upstream of the TSS had also been sequenced. This additional upstream sequencing allowed identification of rare variants within each gene's promoter region. Our analyses of rare regulatory variants for this set of 64 genes therefore included promoter variants, as well as variants within the 5'UTR or 3'UTR. We performed analyses that included any rare regulatory variants (promoter, 5'UTR or 3'UTR variants), only regulatory variants upstream of the coding sequence (promoter or 5'UTR), and only UTR variants (5'UTR or 3'UTR); as well as analyses that only included promoter, 5'UTR or 3'UTR variants. We also applied two MAF thresholds (MAF < 0.01 and MAF < 0.001) and two CADD filters (no CADD filter and CADD ≥ 5). In total, we performed rare allele burden analyses for each of the 24 unique filter combinations.

Across all analyses, results were consistent with a non-linear association between rare regulatory alleles and gene expression, whereby rare allele burden was greatest for the lowest and highest gene expression bins (see **Supplementary Table 4.1**); thus, a 'U-shaped' association was consistently observed. These U-shaped associations were more pronounced (greater concavity) when limiting analyses to the rarest variants (variant sites with MAF < 0.001), and also tended to be more pronounced when only analyzing variants with CADD ≥ 5. In addition, the U-shaped associations were most pronounced when only analyzing the 5'UTR rare variants, as compared with analyses that combined 5'UTR variants with other regulatory variants or that only considered promoter or 3'UTR variants. **Figure 4.9** shows an example of the tendency for rarer variants and those with CADD ≥ 5 to yield stronger estimated associations between rare allele burden and gene expression level. This figure was generated

based on results from the quadratic model analyses of 5'UTR variants only. In order to facilitate interpretation of association estimates outputted by these models, **Figure 4.9** plots rare allele ORs comparing each expression bin to bin 1, along with 95% CIs (ORs for the different comparisons were calculated using the beta estimates resulting from the corresponding analyses, which are presented in **Supplementary Table 4.1**). The aforementioned differences in estimated associations across MAF thresholds and CADD filters are apparent in this figure. For instance, considering the analyses that did not apply a CADD filter, using a MAF < 0.01 filter yielded a minimum OR = 0.760 (95% CI: 0.521, 1.109), which occurs for the comparison of expression bin 14 with bin 1 and is interpreted as middle-range expression bins being associated with a 24% reduced odds of having a rare allele as compared with the lowest expression bin; whereas using a MAF < 0.001 filter yielded a minimum OR = 0.617 (95% CI: 0.370, 1.029), also occurring for the comparison of bin 14 with bin 1. However, as can be seen by the 95% CIs plotted in **Figure 4.9**, these estimated associations are rather imprecise, the same being true for the other analyses of these 64 genes with SZ-associated expression levels. None of these analyses yielded a p-value exceeding the permutation-based multiple-testing-corrected statistical significance threshold of 2.61×10^{-4}. Though interesting, the random error of these estimates limits the ability to render more confident judgements regarding a potential association between rare alleles and gene expression for these genes, and whether such an association may differ across various MAF, CADD, and genomic region (e.g., 5'UTR versus promoter) filters.

Figure 4.9. Results from analyses of 64 genes with SZ-associated expression, when limiting to 5'UTR variants only. Points are ORs comparing each expression bin to bin 1, and vertical lines are corresponding 95% CIs. ORs and 95% CIs were calculated using estimates from the corresponding quadratic regression models.



Genes with low versus high expression associated with SZ

We then stratified these 64 genes based on whether SZ was associated with low expression (17 genes) or high expression (39 genes) (direction of association could not be determined for 8 of the 64 genes), and we performed separate rare allele burden analyses for these two gene sets. We observed that estimated associations tended to be rather different for the genes with low versus high expression associated with SZ (**Supplementary Tables 4.2** and **4.3**). Analyses of the genes with high expression associated with SZ yielded U-shaped estimated associations between rare allele burden and gene expression. In contrast, analyses of the genes with low expression associated with SZ tended to yield estimated associations that were more linear and with positive slopes, consistent with increases in expression bin being associated with greater odds of observing rare alleles. As an example of this observed difference, **Figure 4.10** plots ORs generated based on results from analyzing each gene set (genes with low versus high expression associated with SZ) when applying MAF < 0.001 and CADD ≥ 5 filters and including variants in any regulatory region (promoter, 5'UTR or 3'UTR).

Should these results reflect genuine differences in the association of rare allele burden with expression for genes with low versus high expression associated with SZ, a partial explanation might be that genes with low expression levels associated with SZ are depleted for rare regulatory variants that cause reduced expression, due to selection against the harmful consequences of such variants for these genes. However, the association estimates for these sets of analyses were quite imprecise, and observed differences might readily be explained by error in the estimates.

An additional interesting observation related to these analyses was that, while nearly all analyses yielded the basic result depicted in **Figure 4.10** of an apparent difference in association for genes with low versus high expression associated with SZ, the analyses including only 5'UTR variants produced similarly U-shaped estimates of association between rare allele burden and gene expression for both gene sets. This similarity is depicted in **Figure 4.11**, which presents ORs generated from analyses of 5'UTR variants with MAF < 0.01 and CADD ≥ 5.

Figure 4.10. Comparison of results from analyses of 17 genes with SZ-associated low expression and 39 genes with SZ-associated high expression, when including all available regulatory variants (promoter, 5'UTR and 3'UTR variants), and applying MAF < 0.001 and CADD ≥ 5 filters. Points are ORs comparing each expression bin to bin 1, and vertical lines are corresponding 95% CIs. ORs and 95% CIs were calculated using estimates from the corresponding quadratic regression models.

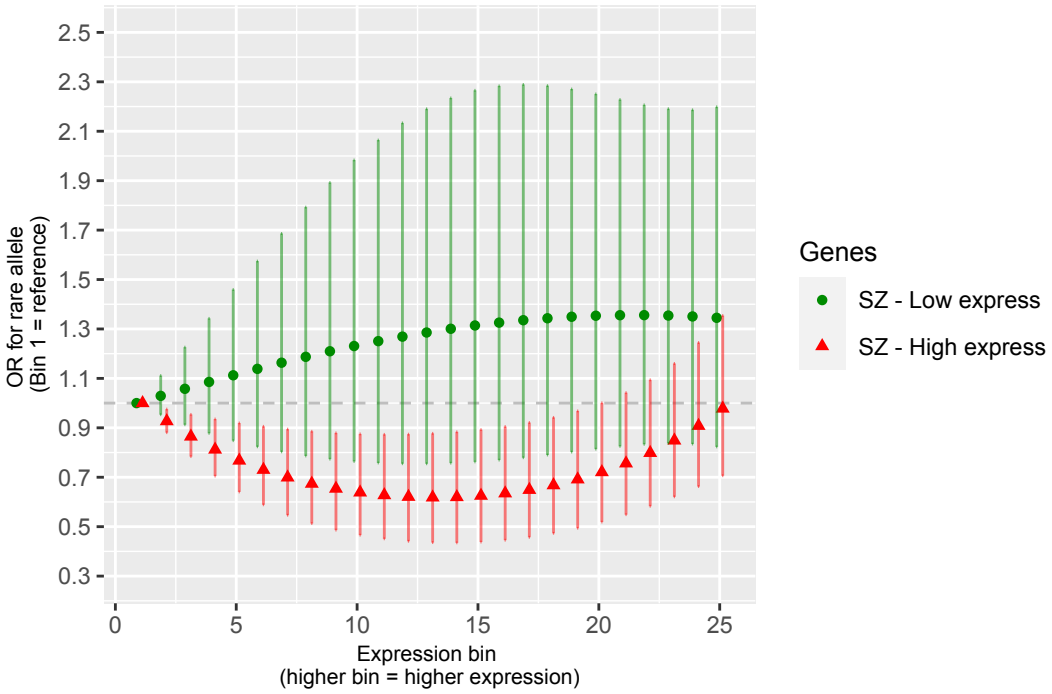
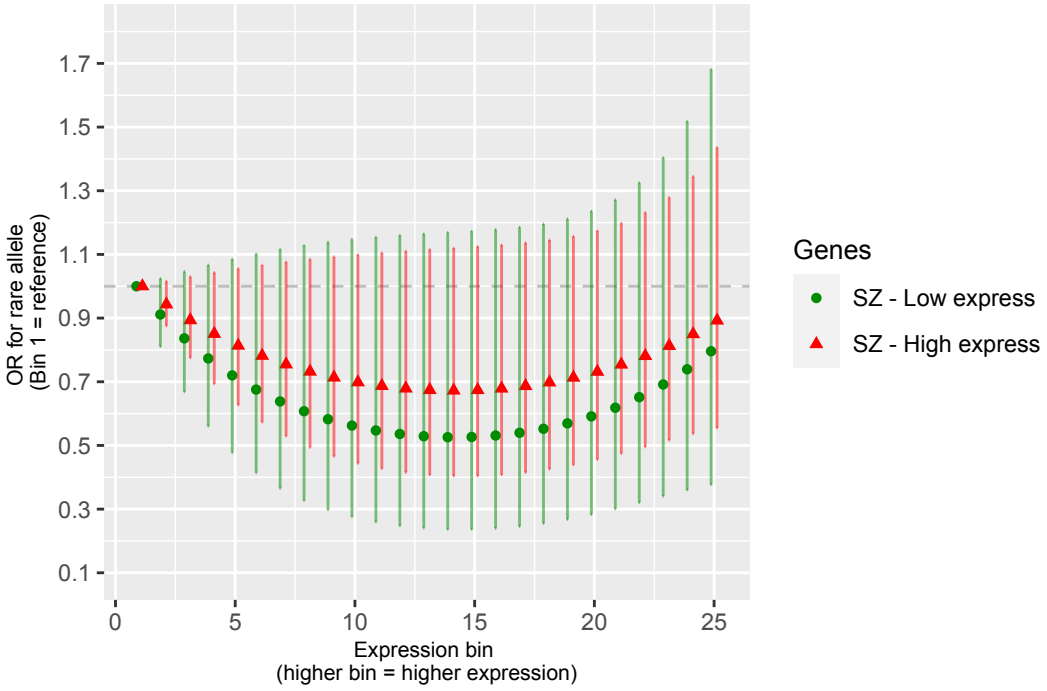


Figure 4.11. Comparison of results from analyses of 17 genes with SZ-associated low expression and 39 genes with SZ-associated high expression, when limited to 5'UTR variants only, and applying MAF < 0.01 and CADD ≥ 5 filters. Points are ORs comparing each expression bin to bin 1, and vertical lines are corresponding 95% CIs. ORs and 95% CIs were calculated using estimates from the corresponding quadratic regression models.



Genes with SZ-associated expression levels or within a SZ CNV

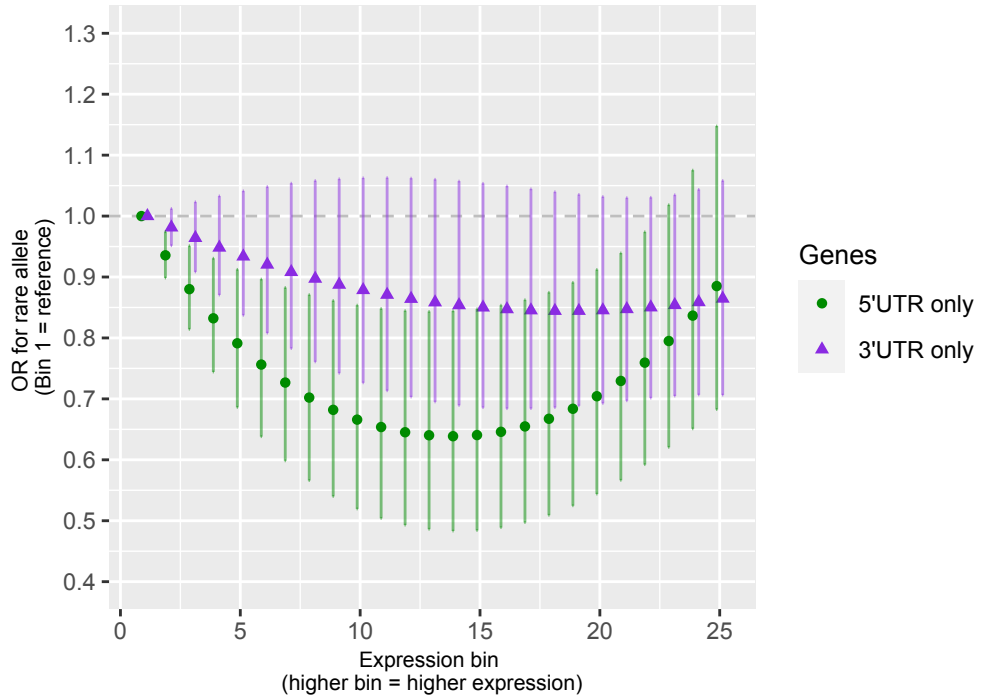
We then combined the set of 64 genes with SZ-associated expression with a larger set of genes that had been sequenced due to being located within or near a SZ-associated large CNV (deletion or duplication) interval. We theorized that just as transcript dosage is associated with SZ for the genes with SZ-associated expression, transcript dosage is also likely to be associated with SZ for numerous genes within the CNV intervals, and thus analyses of this "combined gene set" can serve as a useful supplement to the previously described analyses. Furthermore, these larger gene set analyses are expected to be of interest and value beyond genes with SZ-associated expression.

Unlike the 64 genes with SZ-associated expression levels, the genes that underwent targeted DNA sequencing due to existing within or near a SZ CNV interval only had exonic regions sequenced and were not sequenced to 2kb upstream of the TSS. Therefore, our analyses of this combined gene set did not include promoter region variants. The genomic regions analyzed were 5'UTR and 3'UTR, both combined and separately. We applied the same MAF and CADD filters (MAF < 0.01 and < 0.001; no CADD filter and CADD ≥ 5) as we did when only analyzing the genes with SZ-associated expression. We performed analyses for each of the 12 unique filter combinations. The maximum number of genes available for these analyses was 149, including 62 of the previously analyzed 64 genes with SZ-associated expression (2 of these previously analyzed genes no longer had variants available for analysis after excluding promoter variants), and 87 SZ CNV genes.

Results from these analyses were largely similar to those obtained from analyzing the 64 genes with SZ-associated expression, though these analyses yielded more precise estimates due to being based on a greater quantity of rare regulatory alleles, corresponding to the larger number of genes analyzed (see **Supplementary Table 4.4**). Estimated associations between rare regulatory allele burden and gene expression primarily exhibited U-shaped patterns, which tended to be somewhat more pronounced for the rarest variants (MAF < 0.001) and for variants more likely to be deleterious (CADD ≥ 5). In addition, U-shaped estimated associations were most pronounced when analyzing 5'UTR variants only. An example of this latter finding is depicted by **Figure 4.12**, which presents ORs generated from 5'UTR- and 3'UTR-only analyses which applied MAF < 0.01 and CADD ≥ 5 filters. For these filters, the analyses of 5'UTR variants yielded a minimum OR = 0.639 (for expression bin 14 versus bin 1; 95% CI: 0.483, 0.844; joint p-value for the quadratic model = 0.0064), while analyses of 3'UTR variants yielded a minimum OR = 0.844 (for expression bin 18 versus bin 1; 95% CI: 0.686, 1.040; joint p-value for the quadratic model = 0.2567). Imprecision of the estimated associations from these combined gene set analyses, though less than that of the estimates based analyzing the 64 genes with

SZ-associated expression, warrants caution in interpretation of results. No results from these analyses were statistically significant at the multiple-testing-corrected significance threshold.

Figure 4.12. Results from analyzing the combined set of 149 genes, including genes with SZ-associated expression and genes located within or near a large SZ-associated CNV interval. These results were obtained when limiting to variants with MAF < 0.01 and CADD ≥ 5. Points are ORs comparing each expression bin to bin 1, and vertical lines are corresponding 95% CIs. ORs and 95% CIs were calculated using estimates from the corresponding quadratic regression models.



Genes with low versus high dosage associated with SZ

We then stratified the genes within the combined set based on whether genes had suggestive evidence of low versus high transcript dosage associated with SZ. We analyzed 72 genes with evidence of low dosage linked with SZ, including 19 genes with low expression associated with SZ (16 of which were among the previously analyzed 64 genes with SZ-associated expression; 3 of which were not included amongst the previously analyzed 64 genes since promoter regions for these 3 genes had not be sequenced), and 54 genes within or near a

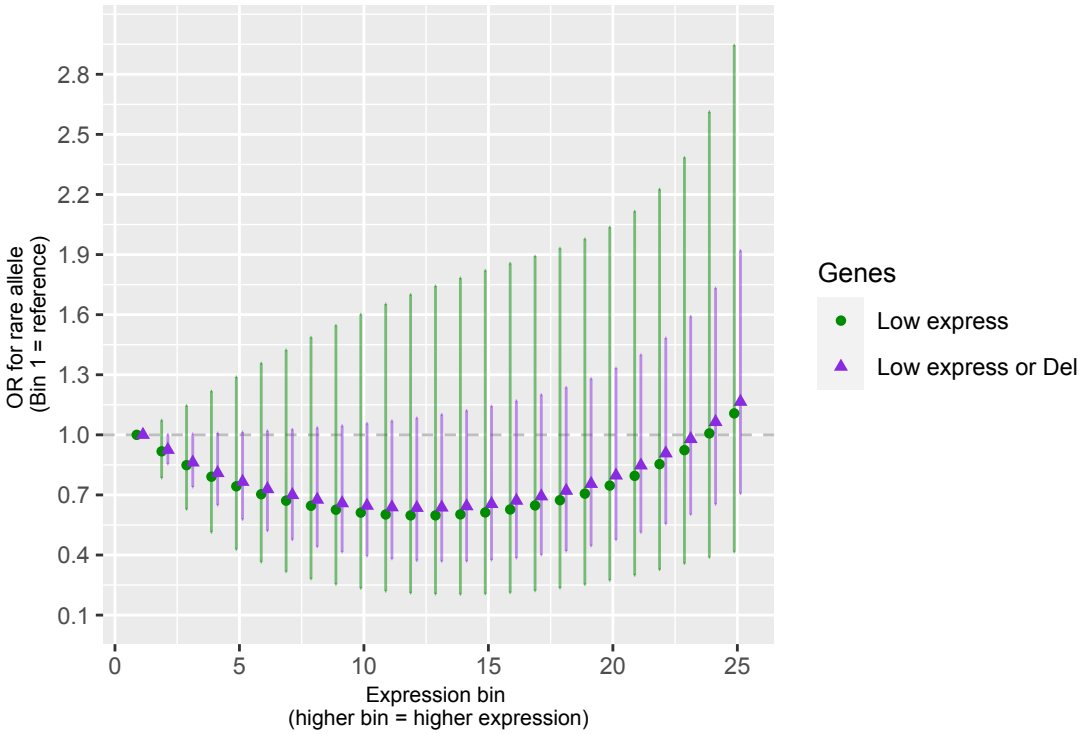
large SZ-associated deletion (1 of which also had low expression levels associated with SZ). We also analyzed a total of 69 genes with evidence of high dosage linked with SZ, including 43 genes with high expression associated with SZ (38 of which were included among the previously analyzed 64 genes with SZ-associated expression, with the remaining 5 not included due to lacking promoter sequence), and 31 genes within or near a large SZ-associated duplication (5 of which also had high expression associated with SZ). Two genes located within SZ duplications were previously identified to have low expression levels associated with SZ; we included these genes among the 72 genes with suggestive evidence of low transcript dosage linked with SZ and not among the genes with evidence of high dosage associated with SZ. For each gene set, we performed rare allele burden analyses corresponding to the 12 unique combinations of variant filters, as we did for the fully combined set of 149 genes.

Results from analyzing these two gene sets were rather different from the previously described results from only analyzing genes among the set of 64 genes with SZ-associated expression levels. The analyses of 72 genes with evidence of low dosage linked with SZ tended to yield U-shaped estimated associations between rare regulatory allele burden and gene expression (see **Supplementary Table 4.5**); this is in contrast to the more linear, positive sloped pattern of estimated associations observed when only analyzing the 17 genes with SZ-associated low expression levels. It is possible that the previously observed linear, positive sloped pattern does not reflect true underlying associations for genes with low expression associated with SZ; we note the imprecision of these previous estimates based on only 17 genes. It is also possible that the previously observed pattern does reflect actual underlying associations for genes with low expression associated with SZ, and that estimated associations based on all 72 genes are rather different due to including many genes that in fact may not have low expression levels associated with SZ. A result which was consistent between analyses of the 17 genes with SZ-associated low expression and the 72 genes with evidence of low dosage linked with SZ was an especially pronounced U-shaped estimated association between rare

allele burden and gene expression observed when only analyzing 5'UTR variants. An example of this consistent result is presented in **Figure 4.13**. As can be seen in this figure, the U-shaped estimated association for the larger set of genes with SZ-associated expression or within a SZ deletion is more precise, yielding a minimum OR = 0.635 (95% CI: 0.372, 1.085; joint p-value for quadratic model = 0.0835). This result is consistent with rare 5'UTR variants potentially playing a role in regulating gene expression for genes with increased likelihood of having SZ-associated low expression levels.

The analyses of up to 69 genes with evidence of high dosage linked with SZ tended to yield estimated associations between rare regulatory allele burden and gene expression which were rather linear and with negative slopes (indicating that decreases in gene expression were associated with greater rare allele burden; see **Supplementary Table 4.6**); this was somewhat different from the consistently more U-shaped estimated associations observed when only analyzing the 39 genes with high expression associated with SZ. However, these observed differences in estimates were generally rather minor and could readily be explained by statistical imprecision.

Figure 4.13. Comparison of results when analyzing the 17 genes with SZ-associated expression ('Low express') and the 72 genes with evidence of low dosage linked with SZ ('Low express or Del'). These results were obtained when including 5'UTR variants only, and limiting to variants with MAF < 0.001 and CADD ≥ 5. Points are ORs comparing each expression bin to bin 1, and vertical lines are corresponding 95% CIs. ORs and 95% CIs were calculated using estimates from the corresponding quadratic regression models.



Filtering the combined gene set based on gene constraint metrics

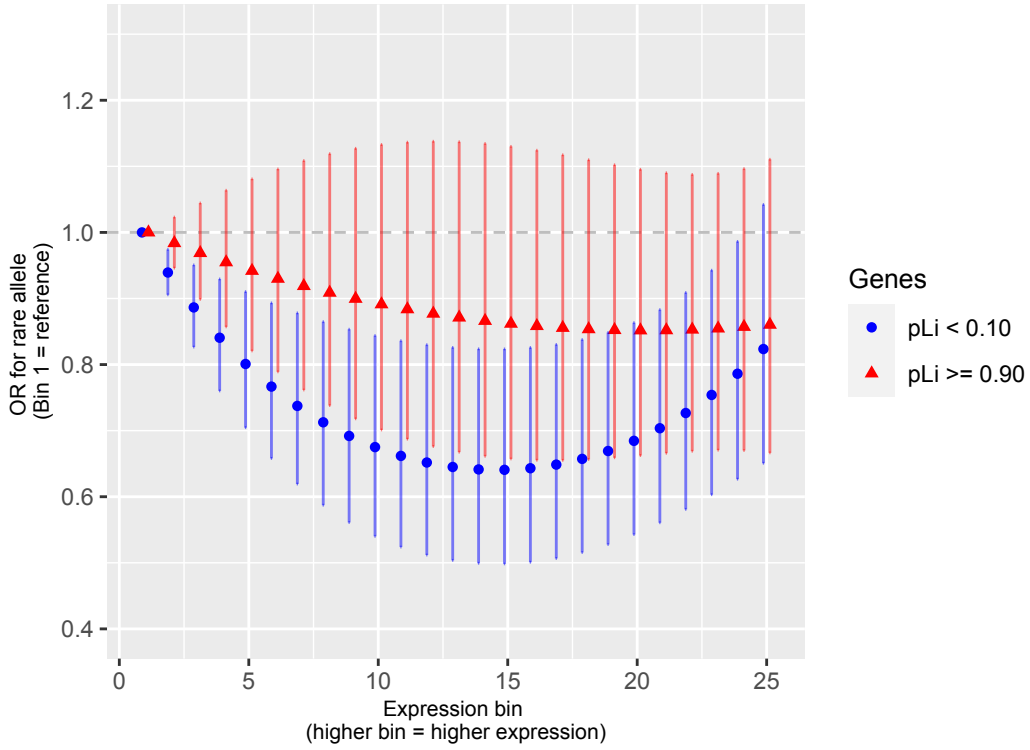
In order to examine whether associations between rare regulatory allele burden and gene expression may vary across different levels of gene constraint, we took the combined set of 149 genes and applied various gene constraint filters, and then performed rare allele burden analyses (for 12 unique filter combinations, as described above) for each resulting gene set.

pLi filters

From the combined set of 149 genes, we identified 87 genes with pLi < 0.10 (low intolerance to LoF mutations) and 33 genes with pLi ≥ 0.90 (high intolerance to LoF mutations). Results from analyzing these two gene sets were noticeably different (see **Supplementary**

Tables 4.7 and **4.8**). Genes with pLi < 0.10 tended to yield stronger estimated associations between rare regulatory alleles and gene expression (more pronounced U-shaped associations) as compared with the pLi ≥ 0.90 genes. As an example, **Figure 4.14** plots ORs generated from analyzing these two gene sets, with variants limited to those within 5'UTR or 3'UTR, and with MAF < 0.01 and CADD ≥ 5. For these specific analyses, the minimum OR when analyzing the pLi < 0.10 genes was 0.641 (when comparing expression bin 15 with bin 1; 95% CI: 0.498, 0.824; joint p-value for quadratic model = 0.0029); while the smallest OR obtained when analyzing the pLi > 0.90 genes was 0.852 (comparing bin 20 with bin 1; 95% CI: 0.662, 1.095; joint p-value for quadratic model = 0.4437). A tendency for genes with relatively low intolerance to LoF mutations to exhibit stronger U-shaped associations between rare regulatory allele burden and gene expression has plausibility, as extremes in expression level for such genes may be relatively more tolerable, and in turn there may be less negative selective pressure against rare variants with strong effects on gene expression. Consistent with findings from our prior analyses, the strongest estimated associations for the pLi < 0.10 genes were observed when analyzing 5'UTR variants only.

Figure 4.14. Comparison of results when analyzing genes with pLi < 0.10 and genes with pLi ≥ 0.90. These results were obtained when analyzing any UTR variants (5'UTR and 3'UTR variants), and limiting to variants with MAF < 0.01 and CADD ≥ 5. Points are ORs comparing each expression bin to bin 1, and vertical lines are corresponding 95% CIs. ORs and 95% CIs were calculated using estimates from the corresponding quadratic regression models.

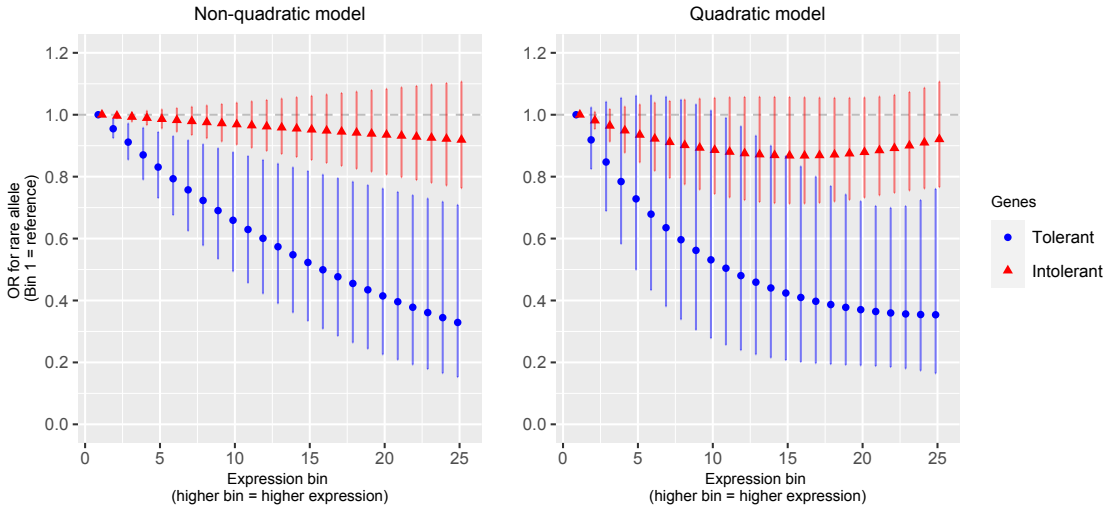


Missense variation constraint filters

Starting from the combined set of 149 genes, we then identified 89 genes intolerant to missense variation, and 7 genes extremely tolerant to missense variation. Of the 89 missense-intolerant genes, 32 genes had pLi ≥ 0.90 and 34 had pLi < 0.10 (we note that potential inconsistencies in the assignment of pLi and missense constraint are due to methodological differences for determining these metrics, and that incorporating both approaches into our analyses is of benefit). Of the 7 genes with evidence of extreme tolerance to missense variation, all had pLi < 0.10. We performed analyses for each of these two gene sets, for the same 12 unique filter combinations as described previously.

The analyses of genes intolerant to missense variation yielded U-shaped estimates of association between rare alleles and gene expression, while the analyses of genes extremely tolerant to missense variants tended to yield negatively sloped, linear association estimates, with the latter association estimates generally stronger than the former but becoming highly imprecise with stricter variant filters due to the small numbers of rare alleles available for analysis (see **Supplementary Tables 4.9** and **4.10**). An illustration of these differences are the results from analyzing these two gene sets limited to variants in any UTR (5'UTR or 3'UTR), MAF < 0.01 and CADD ≥ 5; plots based on these results are shown in **Figure 4.15**. The stronger, more linear pattern of estimated associations produced from analyzing the genes extremely tolerant to missense variation is evident in this figure, with the non-quadratic analysis of these genes yielding a minimum OR = 0.329 (comparing expression bin 25 with bin 1; 95% CI: 0.153, 0.709; p-value for the non-quadratic model = 0.0048); in contrast, results from analyzing the genes intolerant to missense variation are observed to exhibit a more U-shaped pattern, with the quadratic analysis of these genes generating a minimum OR = 0.868 (comparing expression bin 16 with bin 1; 95% CI: 0.713, 1.056; joint p-value for the quadratic model = 0.3697). Overall, the results from these analyses are consistent with a stronger association of rare regulatory allele burden with gene expression level for genes that are extremely tolerant to missense mutations as compared with those intolerant to such mutations. This finding is consistent with our LoF constraint findings, and seems similarly plausible in light of potentially stronger negative selection for highly intolerant genes.

Figure 4.15. Comparison of results when analyzing genes extremely tolerant versus intolerant to missense variants. These results were obtained when analyzing any UTR variants (5'UTR and 3'UTR variants), and limiting to variants with MAF < 0.01 and CADD ≥ 5. Points are ORs comparing each expression bin to bin 1, and vertical lines are corresponding 95% CIs. ORs and 95% CIs were calculated using estimates from the corresponding quadratic regression models.

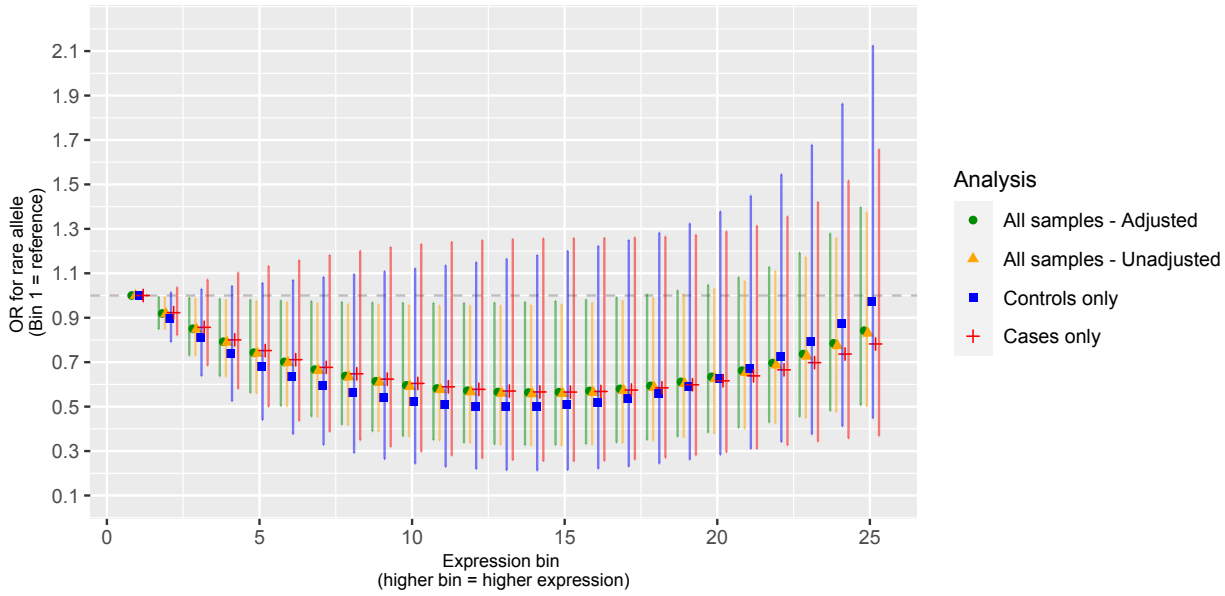


Sensitivity analyses

We reperformed the rare variant burden analyses of the 64 genes with SZ-associated expression, for the controls only, SZ cases only, and for all samples combined but without adjusting for case-control status as we did for the main analyses. In order to be consistent with our main analyses, which used 25 expression bins, we randomly excluded 20 of the 370 controls, and analyzed the remaining 350 controls (25 expression bins of 14 participants each); and we randomly excluded 5 of the 355 SZ cases, analyzing the remaining 350 cases. We observed patterns of estimated associations that were highly consistent across analyses based on all samples combined with adjustment for case-control status (our main analyses), all samples combined without adjustment for case-control status, controls only and cases only (see **Supplementary Tables 4.11, 4.12** and **4.13**). Analyzing all samples combined yielded the most precise estimates. As observed for our main analyses, the analyses of controls only, cases only, and all samples without adjustment for case-control status all produced U-shaped patterns of

association between rare regulatory allele burden and gene expression, which were frequently more pronounced when only including the rarest variants (MAF < 0.001), when limiting to variants with CADD ≥ 5, and when only including 5'UTR variants. **Figure 4.16** shows results from the quadratic model analyses of 5'UTR regulatory variants with MAF < 0.001 and CADD ≥ 5; the version of this analysis performed for generating our main results was plotted previously in **Figure 4.9**, with ORs from this main analysis shown again here. It is apparent from **Figure 4.16** that the patterns of estimated associations when analyzing controls only, cases only, and all samples with or without adjustment for case-control status do not exhibit evidence of meaningful differences. For the controls-only analysis, the OR at the bottom of the U-shaped curve was estimated as 0.499; dividing this by the minimum OR estimated for our main analysis yields a ratio of 0.499/0.563 = 0.887. For the cases-only analysis, the OR at the bottom of the U-shaped curve was estimated as 0.565; the ratio of this OR to the main analysis OR is 0.565/0.563 = 1.004. For the analysis of all samples without case-control adjustment, the minimum OR was 0.557, yielding a ratio of 0.557/0.563 = 0.989 when compared with the corresponding OR from the main analysis of all samples with case-control adjustment. These ratios indicate little difference in the smallest OR estimated across these different analyses.

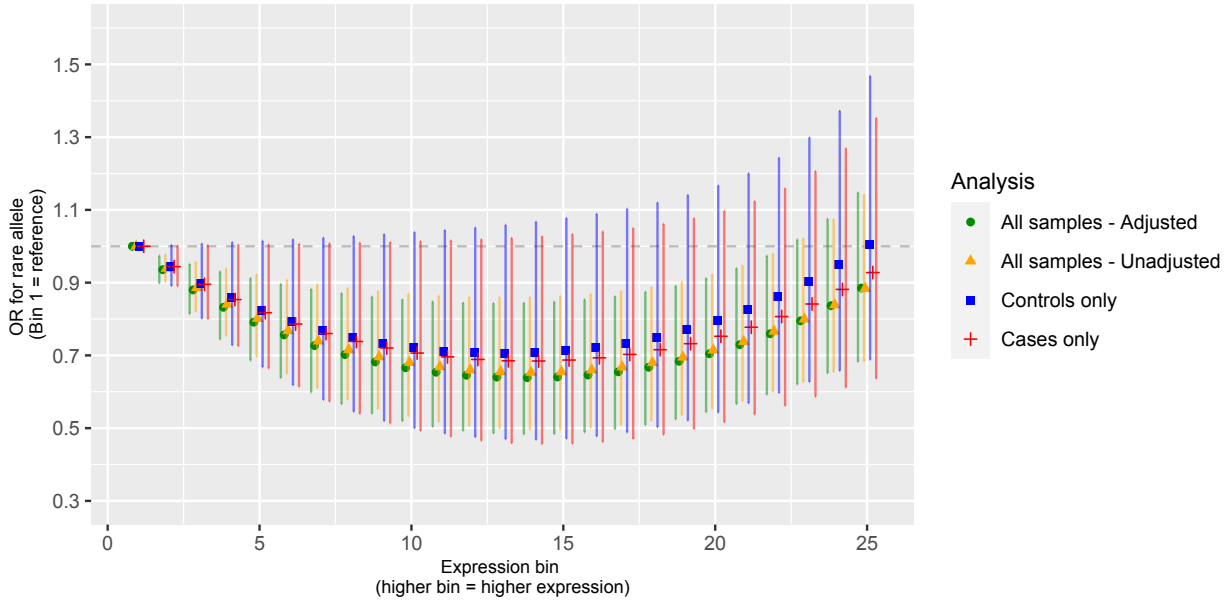
Figure 4.16. Comparison of results when performing burden analyses for all samples with adjustment for case-control status ('All samples - Adjusted'; main analyses), all samples without adjustment for case-control status ('All samples - Unadjusted'), controls only, and cases only. Analyses were performed for the 64 genes with SZ-associated expression, limiting to 5'UTR variants with MAF < 0.001 and CADD ≥ 5. Points are ORs comparing each expression bin to bin 1, and vertical lines are corresponding 95% CIs. ORs and 95% CIs were calculated using estimates from the corresponding quadratic regression models.



We also used the datasets of 350 controls, 350 SZ cases, and all samples unadjusted for case-control status to reperform the rare burden analyses of the full set of 149 genes either with SZ-associated expression and/or located within or near a large SZ CNV interval. For these largest gene set analyses, we again observed results that were highly consistent across analyses of the controls only, cases only, and all samples with or without adjustment for case-control status: non-linear estimated associations between rare regulatory allele burden and gene expression which tended to exhibit U-shaped patterns, which were most pronounced when analyzing 5'UTR variants only, and which were often somewhat more pronounced when applying MAF < 0.001 and CADD ≥ 5 filters (see **Supplementary Tables 4.14, 4.15** and **4.16**). We previously plotted results from the main analyses all 149 genes, including only 5'UTR variants with MAF < 0.01 and CADD ≥ 5 (see **Figure 4.12**). In **Figure 4.17**, we again plot these

results (which are based on all samples, with adjustment for case-control status), along with OR estimates from analyzing only the controls, only the cases, and all samples without adjusting for case-control status. The estimated associations from each of these datasets are very similar. The ratio comparing minimum OR from the controls-only analysis to minimum OR from the main analysis is 0.705/0.639 = 1.104; the ratio comparing minimum ORs from the cases-only analysis and the main analysis is 0.685/0.639 = 1.072; and the ratio of minimum OR from the analysis of all samples without case-control adjustment to minimum OR from our main analysis is 0.653/0.639 = 1.023.

Figure 4.17. Comparison of results when performing burden analyses for all samples with adjustment for case-control status ('All samples - Adjusted'; main analyses), all samples without adjustment for case-control status ('All samples - Unadjusted'), controls only, and cases only. Analyses were performed for the combined set of 149 genes, including genes with SZ-associated expression and those located within or near a large SZ-associated CNV. These results are based on analyzing 5'UTR variants with MAF < 0.01 and CADD ≥ 5. Points are ORs comparing each expression bin to bin 1, and vertical lines are corresponding 95% CIs. ORs and 95% CIs were calculated using estimates from the corresponding quadratic regression models.



In addition, we reperformed the analyses depicted in **Figures 4.10**, **4.14**, and **4.15** using controls only, cases only, and all samples without case-control adjustment. We observed that the patterns of estimated associations were quite similar across the analyses of all samples with or without case-control adjustment, controls only and cases only, with the differences previously described for the main analyses (differences in estimated associations between genes with low versus high expression associated with SZ, genes with pLi < 0.10 versus pLi ≥ 0.90, and genes extremely tolerant versus intolerant to missense variation) also observed for these supplemental analyses (see **Supplementary Figures 4.1, 4.2,** and **4.3**).

The approach of analyzing all cases and controls combined and adjusting for case-control status, employed for our main analyses, and the approach of analyzing controls only, have both been demonstrated to yield valid results for the secondary analysis of case-control data when the disease defining case status is rare.[111-115,122] The finding that these two approaches yielded highly similar estimated associations between rare regulatory allele burden and gene expression for our study is therefore unsurprising; nonetheless, this consistency lends additional credibility to our main results. The observed consistency between the results generated from analyzing controls only and SZ cases only is in line with our expectation that an effect of rare regulatory variation on gene expression would likely not be meaningfully different for individuals with and without SZ.

In addition, the fact that estimates were nearly identical when analyzing all samples with and without adjustment for case-control status may reflect that SZ is not a collider variable in our analyses, or that it is only a very weak collider; if it were a strong collider, prior studies indicate that it is likely the unadjusted analyses would yield results that are biased and therefore differ from the adjusted analyses. One possibility is that many or most of the rare regulatory variants included in our analyses do not affect SZ through pathways other than those involving gene expression, or do so only weakly, in which case the designation of SZ status as a collider in the DAG could be incorrect.

DISCUSSION

Recent studies have identified numerous genes to have expression levels that are associated with SZ.[14-16] Rare variants within regulatory regions may play an important role in modifying the expression of such genes. However, research studying the functional consequences of rare regulatory variants has been very limited, including for investigations of the contribution of rare regulatory variation to SZ-associated expression. Studies of the role of rare regulatory variants in modifying gene expression have lagged due to technological and power limitations. For the present study, we were able to overcome or reduce these limitations by making use of targeted DNA sequencing data, and by employing a modified version of a recently developed rare variant burden analysis technique specifically designed to offer more power for examining associations of rare alleles with gene expression. We analyzed a subset of 725 individuals from the MGS study, all of whom had both targeted DNA sequence and RNA sequence data. We had also originally planned to analyze an independent sample of 400 individuals with microarray expression data, but during preliminary work we realized that the microarray data appeared limited in its ability to accurately order individuals' expression levels for our analyses, and so we ultimately moved forward only analyzing the RNA sequencing dataset.

We first analyzed a set of 64 genes previously identified to have SZ-associated expression levels. All exons for these genes had been sequenced, along with a 2 kb region upstream of the TSS. Our regulatory variant analyses for these genes included rare variants within the promoter, 5'UTR or 3'UTR (or various combinations of these regions); and considered different MAF filters (MAF < 0.01 or < 0.001) and different CADD filters (no CADD filter or CADD ≥ 5). Across multiple analyses we consistently observed a U-shaped pattern of estimated association between rare regulatory allele burden and gene expression, whereby rare allele burden was greatest both at the lowest and highest expression levels. These U-shaped estimated associations tended to be more pronounced when analyses were limited to the rarest

variants (MAF < 0.001) and the variants more likely to be deleterious (CADD ≥ 5); and also when analyses were performed only for 5'UTR variants (as opposed to promoter or 3'UTR variants). These same basic findings were also observed when we analyzed a larger set of 149 genes, which combined the aforementioned SZ-associated expression genes with a set of genes likely to be enriched for SZ-associated expression due to being located within or near a SZ-associated large CNV (deletion or duplication). For this larger gene set, we only considered 5'UTR and 3'UTR variants since regions upstream of the TSS had not been sequenced for the SZ CNV genes. Among these particular analyses, the strongest statistical support for an association between rare regulatory alleles and gene expression was provided by the analysis of the larger gene set, when only including 5'UTR variants with MAF < 0.01 and CADD ≥ 5; this quadratic model analysis yielded OR = 0.639 for the odds of observing a rare allele in expression bin 14 versus bin 1 (95% CI: 0.483, 0.844; joint p-value for the quadratic model = 0.0064).

Our observation of U-shaped estimated associations between rare regulatory allele burden and gene expression is consistent with the findings by Zhao et al. (2016), who are the investigators that developed the burden analysis approach on which our modified approach was based. These investigators analyzed associations of low-frequency (MAF < 0.05) promoter-proximal variants with peripheral blood transcript level for 472 genes (not focusing on SZ-associated genes), and identified an enrichment of these variants for both low and high expression (they replicated this finding in a supplemental analysis of MAF < 0.01 variants). They also observed that this pattern of enrichment was stronger for variants within 1kb downstream of the TSS as compared with variants within 1kb upstream of the TSS, which is consistent with our finding that U-shaped estimated associations were most pronounced when only analyzing 5'UTR variants (which are directly downstream of the TSS).

We also stratified the 64 genes with SZ-associated expression into genes with low versus high expression associated with SZ, and separately analyzed these two gene sets. The

resulting association estimates were rather different for the two gene sets, possibly reflecting a genuine difference in the distribution of rare regulatory allele burden for genes with low versus high expression associated with SZ. However, the estimates from these analyses were insufficiently precise to be confident in any real difference. Furthermore, these differences in estimated associations did not replicate when analyzing the larger sets of genes with low versus high dosage associated with SZ (gene sets generated by stratifying the combined set of 149 genes based on SZ-associated low expression or deletion versus SZ-associated high expression or duplication).

In addition, we analyzed gene sets created by filtering the combined set of 149 genes based on LoF and missense variant constraint metrics. These analyses yielded our strongest association estimates. In comparison with pLi ≥ 0.90 genes (highly intolerant to LoF mutations), analyzing genes with pLi < 0.10 (LoF-tolerant) tended to yield more pronounced U-shaped estimated associations between rare regulatory allele burden and gene expression. Application of missense constraint filters yielded similar results: we observed much stronger association estimates for genes extremely tolerant to missense variation (with rare allele odds by far greatest at the lowest expression levels) as compared with missense-intolerant genes. These results are consistent with prior findings by Lek et al. (2016)[125] that genes that are highly constrained for protein-truncating variants as well as genes highly constrained for missense variation are depleted for eQTLs (common variant associations with gene expression) as compared with genes exhibiting medium or low constraint for these variant classes. A plausible explanation for these findings is that rare regulatory variants with stronger effects on gene expression may have been eliminated to a certain extent from the intolerant genes through negative selection, since aberrant gene expression caused by such rare variants might have particularly detrimental consequences for these genes, which are known to be intolerant to alterations arising from missense and/or LoF mutations (note that pLi ≥ 0.90 means that the gene is highly likely to be haploinsufficient and therefore intolerant to a 50% reduction in

expression due to the loss of one gene copy). The remaining rare variants would then be less likely to strongly impact gene expression. On the other hand, genes with evidence of being tolerant or highly tolerant to such mutations might also be highly tolerant to rare regulatory variants with large effects on gene expression, with the result that such variants are less likely to be eliminated due to negative selective pressure. Our gene constraint findings may be supported by the observation by Zhao et al (2016)[27] that the association between rare promoter allele burden and gene expression was much stronger (more pronounced U-shape) for a set of genes not associated with disease than it was for a set of metabolic disease-related genes. These investigators noted that this difference might be explained by relaxed purifying selection on the former set of genes, which is perhaps less likely to include highly constrained genes as compared with the set of disease genes.

We performed sensitivity analyses using controls only, SZ cases only, and all samples without adjustment for case-control status, and observed that these analyses yielded estimated associations of rare regulatory allele burden and gene expression that were highly consistent with the estimates obtained from our main analyses (which analyzed all cases and controls combined, and adjusted for case-control status). This finding lends additional credibility to the results from our main analyses, and also supports the notion that an effect of rare regulatory variation on gene expression may not be different for those with and without SZ.

For this study, we examined a set of genes with LCL-based expression levels previously identified as associated with SZ, and our study also involved directly analyzing LCL-based expression levels. The use of LCLs to investigate SZ-associated gene expression is perhaps not ideal, as brain tissue is considered the most relevant tissue for SZ.[14] However, ascertainment of postmortem brain tissue or neuronal cell lines for a psychiatric study is often not feasible, particularly for a large genetic epidemiology study. As a result, LCLs have frequently been used in psychiatric studies as a substitute tissue. The relevance of LCLs is supported by research indicating that 35-80% of transcripts are expressed in both brain and

blood, with correlation between brain and blood gene expression levels estimated at r = 0.24-0.64.[14,126,127] In addition, research supports an important role for immune mechanisms in SZ, and LCLs are well-suited to investigate gene expression that may be involved in such processes.[14,16,19,128-130] Furthermore, the set of 64 genes with SZ-associated expression (based on LCLs) that we analyzed in this study are enriched for genes that are also expressed in the brain.[14] LCLs also provide the benefit of potentially reduced environmental influences on gene expression. Specifically, the process of generating LCLs results in a cell line that is thought to be more removed from the environmental influences and state aspects of the individual (e.g., disease status, medication usage), as compared with the original cell from which the LCL was derived.[16]

Our study focused exclusively on regulatory variants that are proximal to the gene. It is possible that including distal regulatory variants (e.g., those within enhancer, silencer and insulator sequences) may yield modified estimates of association between rare regulatory allele burden and gene expression. However, distal regulatory regions are presently less well defined in comparison with proximal regulatory regions, presenting challenges for identifying and studying variants located within these distal regulatory elements. In addition, there is evidence that variants proximal to genes tend to have larger effects than distal variants.[131]

The rare variant burden approach that we employed to examine relations of rare regulatory alleles and gene expression gains power for examining associations of rare variants and gene expression by analyzing sets of genes that have been grouped together. A consequence of this approach is that genes are not examined individually, and a unique pattern of association between rare regulatory alleles and gene expression for a given gene may be missed. With time, substantial increases in study sample sizes should provide sufficient power for analyzing genes individually; until then, a burden approach such as that which we used may be the best alternative.

In addition, the rare variant burden approach that we employed provides estimates of association that reflect the pattern of rare allele burden across gene expression levels, but does not inform as to why such patterns may exist. An analysis of genes yielding a rather weak U-shaped pattern of association between rare regulatory allele burden and gene expression may reflect little effect of any rare variation on expression for the gene set; alternatively, certain rare variants may have large effects on expression for these genes, with the weak U-shaped estimated association reflecting that such variants have been largely removed due to negative selective pressure. The latter possibility is illustrated by our different findings across levels of gene constraint. Thus, it is important to be cautious about interpreting a given set of results as representative of associations that would be estimated for any rare regulatory variation within the genes of interest. Supplemental approaches may be employed to gain a more accurate and complete understanding of the potential effects of rare regulatory variants on expression level for given genes.

Our rare variant burden approach was designed to examine associations between rare allele burden and gene expression bin; as such, it is able to provide suggestive evidence about whether rare regulatory alleles may be causing decreased and/or increased gene expression. However, the approach is inherently limited with respect to providing information about the magnitude of expression change (e.g., number of standard deviations in expression shifted) that might be caused by a rare allele. Alternative analytic methods, which do not assign expression values to bins, will be needed to examine this aspect of the association between rare regulatory variants and gene expression. Substantial sample size increases should eventually enable well-powered application of eQTL analysis methods to rare variants, which would allow quantification of the amount that expression levels differ for those with and without specific rare regulatory alleles. Future work might also include in vitro studies that examine the amount of change in gene expression corresponding to experimental manipulation of rare regulatory alleles.

An additional point of interest, which was beyond the scope of this project, is whether expression for the genes considered in our study (specifically, those identified to have SZ-associated expression levels) may in fact be mediators on potential causal pathways from rare regulatory variation to SZ. Additional analyses are needed to properly explore this possibility, including re-analyzing associations between gene expression and SZ while controlling for the rare regulatory variants of interest, to rule out the possibility that an observed association between gene expression and SZ is simply due to confounding by these rare regulatory variants.[132,133] Should future research support a causal pathway from rare regulatory variation to aberrant gene expression to SZ for genes among those which we have analyzed, it is possible that such knowledge could be leveraged in the effort to more effectively treat and prevent SZ.

In summary, most results from the analyses described in this chapter were consistent with a U-shaped pattern of association between rare regulatory variants and gene expression for genes with SZ-associated expression levels, whereby rare regulatory allele burden was greatest at the expression extremes. Despite employing a unique burden approach specifically designed to increase power for investigating associations of rare variants with gene expression, many of our estimates were rather imprecise, and no results were statistically significant following correction for multiple tests. Future studies involving more participants and/or analyzing more genes with SZ-associated expression will enable the precision needed to better estimate these associations.

SUPPLEMENT

Supplementary Table 4.1 (next page). Results from analyzing 64 genes with SZ-associated expression, using non-quadratic (no $\boldsymbol{Bin^2}$ term) and quadratic (includes $\boldsymbol{Bin^2}$ term) models, and applying various combinations of MAF, CADD, and region filters to the variants. In the model formulas, ***Y = 1*** indicates presence of a rare regulatory allele, and ***Bin*** represents expression bin number (these analyses used 25 expression bins, with 29 samples assigned to each bin). For the group of middle columns corresponding to the non-quadratic model, OR, CIs and P all correspond to ***Bin***. For the quadratic model, 'P: Joint' is the p-value resulting from a joint test of β_1 and β_2. In the 'Gene Region' column, 'Any reg' means variants within the promoter, 5'UTR or 3'UTR were included; 'Upstream reg' means only variants within the promoter or 5'UTR were included; 'Any UTR' means only variants within the 5'UTR or 3'UTR were included; 'Prom only' means only variants within the promoter region were included. The column 'Genes' shows the number of genes analyzed, which is less than 64 when variant filters removed all variants of interest from a gene. 'Rare sites' and 'Rare alleles' are the total number of rare variant sites and rare alleles, respectively, remaining after variant filters. MAF = minor allele frequency; CADD = Combined Annotation Dependent Depletion; OR = odds ratio; L 95% CI = lower bound for the 95% confidence interval; U 95% CI = upper bound for the 95% confidence interval; P = p-value; SE = standard error.

Dataset						Model: $logodds(Y=1) = \alpha + \beta_1 * Bin$				Model: $logodds(Y=1) = \alpha + \beta_1 * Bin + \beta_2 * Bin^2$						
MAF	Gene Region	CADD	Genes	Rare Sites	Rare Alleles	OR	L 95% CI	U 95% CI	P	β_1	SE(β_1)	P: Bin	β_2	SE(β_2)	P: Bin^2	P: Joint
< 0.01	Any reg	No filter	64	1,275	2,494	1.000	0.993	1.006	0.8731	-0.0184	0.0129	0.1550	6.90E-04	4.80E-04	0.1543	0.3599
< 0.01	Any reg	≥ 5	64	820	1,592	0.999	0.991	1.007	0.8090	-0.0302	0.0162	0.0629	1.13E-03	6.10E-04	0.0636	0.1766
< 0.01	Upstream reg	No filter	64	629	1,210	1.001	0.992	1.010	0.8314	-0.0149	0.0191	0.4358	6.10E-04	7.10E-04	0.3922	0.6792
< 0.01	Upstream reg	≥ 5	63	488	951	1.003	0.993	1.013	0.5669	-0.0346	0.0213	0.1041	1.44E-03	7.90E-04	0.0694	0.1662
< 0.01	Any UTR	No filter	62	856	1,744	0.999	0.992	1.006	0.7524	-0.0234	0.0150	0.1179	8.60E-04	5.60E-04	0.1257	0.2969
< 0.01	Any UTR	≥ 5	61	515	1,027	0.995	0.985	1.005	0.3064	-0.0453	0.0206	0.0278	1.55E-03	7.70E-04	0.0447	0.0810
< 0.01	Prom only	No filter	64	604	1,160	1.002	0.993	1.011	0.6575	-0.0160	0.0194	0.4107	6.90E-04	7.20E-04	0.3384	0.5747
< 0.01	Prom only	≥ 5	63	469	919	1.005	0.995	1.015	0.3574	-0.0332	0.0217	0.1271	1.45E-03	8.10E-04	0.0719	0.1321
< 0.01	5'UTR only	No filter	52	210	460	0.997	0.982	1.012	0.6670	-0.0456	0.0313	0.1456	1.63E-03	1.18E-03	0.1647	0.3512
< 0.01	5'UTR only	≥ 5	51	183	386	0.995	0.978	1.011	0.5188	-0.0810	0.0338	0.0169	2.92E-03	1.27E-03	0.0217	0.0613
< 0.01	3'UTR only	No filter	59	646	1,284	1.000	0.991	1.008	0.9152	-0.0207	0.0184	0.2620	7.80E-04	6.90E-04	0.2590	0.5280
< 0.01	3'UTR only	≥ 5	51	332	641	0.994	0.981	1.007	0.3909	-0.0488	0.0272	0.0728	1.67E-03	1.02E-03	0.1030	0.1864
< 0.001	Any reg	No filter	64	1,006	1,147	1.000	0.991	1.009	0.9693	-0.0304	0.0182	0.0962	1.16E-03	6.80E-04	0.0885	0.2373
< 0.001	Any reg	≥ 5	64	649	734	1.001	0.990	1.012	0.8961	-0.0500	0.0229	0.0289	1.95E-03	8.50E-04	0.0225	0.0763
< 0.001	Upstream reg	No filter	64	498	568	1.001	0.989	1.014	0.8211	-0.0435	0.0259	0.0924	1.73E-03	9.60E-04	0.0736	0.2009
< 0.001	Upstream reg	≥ 5	63	387	440	1.005	0.991	1.019	0.4880	-0.0491	0.0290	0.0906	2.07E-03	1.08E-03	0.0555	0.1292
< 0.001	Any UTR	No filter	61	664	760	0.997	0.987	1.007	0.5788	-0.0322	0.0216	0.1360	1.13E-03	8.10E-04	0.1628	0.3263
< 0.001	Any UTR	≥ 5	58	399	453	0.994	0.980	1.008	0.3680	-0.0648	0.0287	0.0241	2.27E-03	1.08E-03	0.0362	0.0772
< 0.001	Prom only	No filter	64	479	543	1.003	0.991	1.016	0.6073	-0.0404	0.0266	0.1280	1.68E-03	9.90E-04	0.0900	0.2122
< 0.001	Prom only	≥ 5	63	372	420	1.008	0.993	1.022	0.2992	-0.0450	0.0299	0.1328	2.00E-03	1.11E-03	0.0710	0.1172
< 0.001	5'UTR only	No filter	50	156	181	0.992	0.971	1.013	0.4494	-0.0778	0.0427	0.0685	2.71E-03	1.61E-03	0.0926	0.1871
< 0.001	5'UTR only	≥ 5	47	137	159	0.992	0.970	1.014	0.4792	-0.0945	0.0454	0.0376	3.36E-03	1.71E-03	0.0502	0.1190
< 0.001	3'UTR only	No filter	57	508	579	0.998	0.986	1.010	0.6879	-0.0228	0.0250	0.3613	7.90E-04	9.40E-04	0.4011	0.6498
< 0.001	3'UTR only	≥ 5	49	262	294	0.994	0.977	1.011	0.4810	-0.0548	0.0359	0.1274	1.88E-03	1.35E-03	0.1641	0.3009

Supplementary Table 4.2 (next page). Results from analyzing 17 genes with low expression associated with SZ, using non-quadratic (no $\boldsymbol{Bin^2}$ term) and quadratic (includes $\boldsymbol{Bin^2}$ term) models, and applying various combinations of MAF, CADD, and region filters to the variants. In the model formulas, $\boldsymbol{Y = 1}$ indicates presence of a rare regulatory allele, and ***Bin*** represents expression bin number (these analyses used 25 expression bins, with 29 samples assigned to each bin). For the group of middle columns corresponding to the non-quadratic model, OR, CIs and P all correspond to ***Bin***. For the quadratic model, 'P: Joint' is the p-value resulting from a joint test of β_1 and β_2. In the 'Gene Region' column, 'Any reg' means variants within the promoter, 5'UTR or 3'UTR were included; 'Upstream reg' means only variants within the promoter or 5'UTR were included; 'Any UTR' means only variants within the 5'UTR or 3'UTR were included; 'Prom only' means only variants within the promoter region were included. The column 'Genes' shows the number of genes analyzed, which is less than 17 when variant filters removed all variants of interest from a gene. 'Rare sites' and 'Rare alleles' are the total number of rare variant sites and rare alleles, respectively, remaining after variant filters. MAF = minor allele frequency; CADD = Combined Annotation Dependent Depletion; OR = odds ratio; L 95% CI = lower bound for the 95% confidence interval; U 95% CI = upper bound for the 95% confidence interval; P = p-value; SE = standard error.

Dataset						Model: $logodds(Y=1) = \alpha + \beta_1 * Bin$				Model: $logodds(Y=1) = \alpha + \beta_1 * Bin + \beta_2 * Bin^2$						
MAF	Gene Region	CADD	Genes	Rare Sites	Rare Alleles	OR	L 95% CI	U 95% CI	P	β_1	SE(β_1)	P: Bin	β_2	SE(β_2)	P: Bin^2	P: Joint
< 0.01	Any reg	No filter	17	332	631	1.006	0.994	1.017	0.3496	0.0043	0.0247	0.8622	5.00E-05	9.10E-04	0.9593	0.6450
< 0.01	Any reg	≥ 5	17	232	464	1.007	0.993	1.021	0.3206	0.0172	0.0298	0.5638	-3.90E-04	1.10E-03	0.7244	0.5737
< 0.01	Upstream reg	No filter	17	165	298	1.007	0.991	1.025	0.3862	-0.0036	0.0360	0.9201	4.20E-04	1.33E-03	0.7516	0.6536
< 0.01	Upstream reg	≥ 5	17	139	262	1.008	0.989	1.028	0.4006	0.0004	0.0407	0.9922	3.00E-04	1.50E-03	0.8445	0.6890
< 0.01	Any UTR	No filter	16	221	438	1.000	0.986	1.014	0.9665	-0.0072	0.0295	0.8081	2.60E-04	1.10E-03	0.8106	0.9708
< 0.01	Any UTR	≥ 5	16	144	304	1.000	0.983	1.018	0.9932	0.0013	0.0371	0.9728	-5.00E-05	1.38E-03	0.9737	0.9994
< 0.01	Prom only	No filter	17	162	295	1.007	0.990	1.025	0.4088	-0.0079	0.0365	0.8287	5.80E-04	1.35E-03	0.6699	0.6495
< 0.01	Prom only	≥ 5	17	137	260	1.007	0.988	1.027	0.4690	-0.0030	0.0411	0.9422	3.80E-04	1.52E-03	0.8005	0.7454
< 0.01	5'UTR only	No filter	12	54	105	0.985	0.954	1.018	0.3709	-0.1026	0.0639	0.1093	3.45E-03	2.43E-03	0.1569	0.2489
< 0.01	5'UTR only	≥ 5	12	51	102	0.989	0.957	1.023	0.5270	-0.1039	0.0667	0.1203	3.63E-03	2.52E-03	0.1514	0.2974
< 0.01	3'UTR only	No filter	15	167	333	1.006	0.990	1.022	0.4430	0.0034	0.0340	0.9198	1.10E-04	1.26E-03	0.9326	0.7426
< 0.01	3'UTR only	≥ 5	14	93	202	1.011	0.988	1.034	0.3461	0.0140	0.0484	0.7730	-1.20E-04	1.78E-03	0.9459	0.6399
< 0.001	Any reg	No filter	17	260	298	1.015	0.999	1.032	0.0733	0.0333	0.0357	0.3514	-6.90E-04	1.30E-03	0.5963	0.1739
< 0.001	Any reg	≥ 5	17	181	212	1.012	0.992	1.033	0.2412	0.0310	0.0438	0.4804	-7.20E-04	1.61E-03	0.6566	0.4551
< 0.001	Upstream reg	No filter	17	127	145	1.010	0.987	1.033	0.3863	0.0337	0.0493	0.4946	-9.00E-04	1.82E-03	0.6206	0.6070
< 0.001	Upstream reg	≥ 5	17	108	124	1.014	0.988	1.040	0.2840	0.0493	0.0565	0.3833	-1.33E-03	2.06E-03	0.5200	0.4558
< 0.001	Any UTR	No filter	16	173	199	1.015	0.995	1.036	0.1466	-0.0015	0.0433	0.9733	6.20E-04	1.58E-03	0.6971	0.3223
< 0.001	Any UTR	≥ 5	16	110	131	1.007	0.981	1.033	0.5974	-0.0314	0.0541	0.5620	1.46E-03	2.00E-03	0.4665	0.6692
< 0.001	Prom only	No filter	17	124	142	1.010	0.987	1.033	0.4061	0.0288	0.0501	0.5655	-7.20E-04	1.84E-03	0.6955	0.6554
< 0.001	Prom only	≥ 5	17	106	122	1.013	0.987	1.039	0.3352	0.0454	0.0569	0.4259	-1.23E-03	2.09E-03	0.5552	0.5259
< 0.001	5'UTR only	No filter	11	40	46	0.999	0.958	1.041	0.9445	-0.0978	0.0848	0.2494	3.71E-03	3.18E-03	0.2432	0.5122
< 0.001	5'UTR only	≥ 5	11	37	43	1.005	0.962	1.050	0.8274	-0.0978	0.0892	0.2737	3.92E-03	3.32E-03	0.2378	0.4938
< 0.001	3'UTR only	No filter	15	133	153	1.019	0.995	1.043	0.1197	0.0257	0.0512	0.6167	-2.60E-04	1.86E-03	0.8877	0.2935
< 0.001	3'UTR only	≥ 5	14	73	88	1.005	0.973	1.039	0.7521	-0.0021	0.0706	0.9769	2.80E-04	2.62E-03	0.9143	0.9459

Supplementary Table 4.3 (next page). Results from analyzing 39 genes with high expression associated with SZ, using non-quadratic (no $\boldsymbol{Bin^2}$ term) and quadratic (includes $\boldsymbol{Bin^2}$ term) models, and applying various combinations of MAF, CADD, and region filters to the variants. In the model formulas, $\boldsymbol{Y = 1}$ indicates presence of a rare regulatory allele, and ***Bin*** represents expression bin number (these analyses used 25 expression bins, with 29 samples assigned to each bin). For the group of middle columns corresponding to the non-quadratic model, OR, CIs and P all correspond to ***Bin***. For the quadratic model, 'P: Joint' is the p-value resulting from a joint test of β_1 and β_2. In the 'Gene Region' column, 'Any reg' means variants within the promoter, 5'UTR or 3'UTR were included; 'Upstream reg' means only variants within the promoter or 5'UTR were included; 'Any UTR' means only variants within the 5'UTR or 3'UTR were included; 'Prom only' means only variants within the promoter region were included. The column 'Genes' shows the number of genes analyzed, which is less than 39 when variant filters removed all variants of interest from a gene. 'Rare sites' and 'Rare alleles' are the total number of rare variant sites and rare alleles, respectively, remaining after variant filters. MAF = minor allele frequency; CADD = Combined Annotation Dependent Depletion; OR = odds ratio; L 95% CI = lower bound for the 95% confidence interval; U 95% CI = upper bound for the 95% confidence interval; P = p-value; SE = standard error.

Dataset						Model: $logodds(Y=1) = \alpha + \beta_1 * Bin$				Model: $logodds(Y=1) = \alpha + \beta_1 * Bin + \beta_2 * Bin^2$						
MAF	Gene Region	CADD	Genes	Rare Sites	Rare Alleles	OR	L 95% CI	U 95% CI	P	β_1	SE(β_1)	P: Bin	β_2	SE(β_2)	P: Bin^2	P: Joint
< 0.01	Any reg	No filter	39	832	1,705	0.999	0.992	1.007	0.8583	-0.0261	0.0157	0.0961	9.80E-04	5.90E-04	0.0952	0.2471
< 0.01	Any reg	≥ 5	39	509	1,017	0.997	0.987	1.007	0.5139	-0.0516	0.0205	0.0119	1.86E-03	7.70E-04	0.0153	0.0444
< 0.01	Upstream reg	No filter	39	394	811	0.998	0.987	1.010	0.7847	-0.0212	0.0237	0.3711	7.60E-04	8.80E-04	0.3936	0.6710
< 0.01	Upstream reg	≥ 5	38	290	599	1.000	0.988	1.013	0.9498	-0.0587	0.0267	0.0278	2.27E-03	1.00E-03	0.0226	0.0773
< 0.01	Any UTR	No filter	38	570	1,210	1.001	0.993	1.010	0.8150	-0.0233	0.0179	0.1916	9.40E-04	6.70E-04	0.1605	0.3660
< 0.01	Any UTR	≥ 5	37	329	665	0.994	0.981	1.006	0.3366	-0.0566	0.0259	0.0292	1.96E-03	9.80E-04	0.0453	0.0877
< 0.01	Prom only	No filter	39	374	770	0.999	0.988	1.011	0.9032	-0.0212	0.0242	0.3805	7.90E-04	9.10E-04	0.3828	0.6798
< 0.01	Prom only	≥ 5	38	275	575	1.003	0.990	1.016	0.6749	-0.0556	0.0275	0.0431	2.24E-03	1.02E-03	0.0288	0.0871
< 0.01	5'UTR only	No filter	32	132	316	1.000	0.982	1.019	0.9708	-0.0121	0.0386	0.7534	4.80E-04	1.44E-03	0.7394	0.9457
< 0.01	5'UTR only	≥ 5	31	110	247	0.995	0.975	1.016	0.6233	-0.0657	0.0424	0.1215	2.34E-03	1.59E-03	0.1416	0.3064
< 0.01	3'UTR only	No filter	38	438	894	1.001	0.990	1.011	0.9046	-0.0223	0.0225	0.3209	8.80E-04	8.40E-04	0.2927	0.5729
< 0.01	3'UTR only	≥ 5	31	219	418	0.991	0.975	1.007	0.2741	-0.0627	0.0336	0.0623	2.09E-03	1.27E-03	0.1004	0.1465
< 0.001	Any reg	No filter	39	654	747	0.997	0.986	1.008	0.5987	-0.0497	0.0229	0.0303	1.81E-03	8.60E-04	0.0360	0.0994
< 0.001	Any reg	≥ 5	39	402	448	0.999	0.985	1.013	0.8860	-0.0857	0.0294	0.0036	3.26E-03	1.10E-03	0.0031	0.0137
< 0.001	Upstream reg	No filter	39	313	357	1.001	0.985	1.017	0.8902	-0.0787	0.0332	0.0179	3.06E-03	1.24E-03	0.0135	0.0498
< 0.001	Upstream reg	≥ 5	38	232	261	1.005	0.987	1.023	0.6075	-0.1004	0.0374	0.0073	4.02E-03	1.39E-03	0.0039	0.0150
< 0.001	Any UTR	No filter	37	437	498	0.994	0.981	1.007	0.3763	-0.0345	0.0270	0.2016	1.11E-03	1.02E-03	0.2747	0.3743
< 0.001	Any UTR	≥ 5	34	252	278	0.992	0.974	1.010	0.3772	-0.0750	0.0374	0.0450	2.60E-03	1.41E-03	0.0657	0.1291
< 0.001	Prom only	No filter	39	299	341	1.003	0.987	1.020	0.7278	-0.0728	0.0340	0.0326	2.90E-03	1.27E-03	0.0222	0.0725
< 0.001	Prom only	≥ 5	38	221	249	1.008	0.989	1.027	0.4075	-0.0936	0.0387	0.0156	3.87E-03	1.43E-03	0.0071	0.0205
< 0.001	5'UTR only	No filter	31	96	108	0.996	0.969	1.023	0.7459	-0.0712	0.0557	0.2010	2.58E-03	2.09E-03	0.2173	0.4492
< 0.001	5'UTR only	≥ 5	28	82	91	0.992	0.963	1.022	0.6135	-0.0954	0.0599	0.1118	3.41E-03	2.26E-03	0.1318	0.2895
< 0.001	3'UTR only	No filter	36	341	390	0.993	0.979	1.008	0.3525	-0.0270	0.0305	0.3750	7.80E-04	1.15E-03	0.4967	0.5156
< 0.001	3'UTR only	≥ 5	29	170	187	0.992	0.971	1.014	0.4657	-0.0683	0.0445	0.1257	2.34E-03	1.68E-03	0.1638	0.2967

Supplementary Table 4.4. Results from analyzing combined set of 149 genes, including genes with SZ-associated expression and genes located within or near a SZ-associated large CNV interval, using non-quadratic (no Bin^2 term) and quadratic (includes Bin^2 term) models, and applying various combinations of MAF, CADD, and region filters to the variants. In the model formulas, ***Y = 1*** indicates presence of a rare regulatory allele, and ***Bin*** represents expression bin number (these analyses used 25 expression bins, with 29 samples assigned to each bin). For the group of middle columns corresponding to the non-quadratic model, OR, CIs and P all correspond to ***Bin***. For the quadratic model, 'P: Joint' is the p-value resulting from a joint test of β_1 and β_2. In the 'Gene Region' column, 'Any UTR' means variants within the 5'UTR or 3'UTR were included. The column 'Genes' shows the number of genes analyzed, which is less than 149 when variant filters removed all variants of interest from a gene. 'Rare sites' and 'Rare alleles' are the total number of rare variant sites and rare alleles, respectively, remaining after variant filters. MAF = minor allele frequency; CADD = Combined Annotation Dependent Depletion; OR = odds ratio; L 95% CI = lower bound for the 95% confidence interval; U 95% CI = upper bound for the 95% confidence interval; P = p-value; SE = standard error.

Dataset						Model: $logodds(Y=1) = \alpha + \beta_1 * Bin$				Model: $logodds(Y=1) = \alpha + \beta_1 * Bin + \beta_2 * Bin^2$						
MAF	Gene Region	CADD	Genes	Rare Sites	Rare Alleles	OR	L 95% CI	U 95% CI	P	β_1	SE(β_1)	P: Bin	β_2	SE(β_2)	P: Bin^2	P: Joint
< 0.01	Any UTR	No filter	149	1,872	3,929	0.999	0.994	1.004	0.6468	-0.0149	0.0105	0.1573	5.30E-04	3.90E-04	0.1795	0.3671
< 0.01	Any UTR	≥ 5	147	1,172	2,432	0.995	0.988	1.001	0.1251	-0.0349	0.0136	0.0107	1.15E-03	5.10E-04	0.0249	0.0255
< 0.01	5'UTR only	No filter	117	434	1,028	0.995	0.985	1.005	0.3500	-0.0650	0.0215	0.0024	2.32E-03	8.10E-04	0.0040	0.0109
< 0.01	5'UTR only	≥ 5	113	382	893	0.994	0.983	1.006	0.3347	-0.0746	0.0232	0.0013	2.67E-03	8.70E-04	0.0023	0.0064
< 0.01	3'UTR only	No filter	139	1,438	2,901	1.000	0.994	1.006	0.9479	-0.0002	0.0128	0.9858	0.00E+00	4.80E-04	0.9983	0.9979
< 0.01	3'UTR only	≥ 5	125	790	1,539	0.994	0.985	1.002	0.1531	-0.0205	0.0178	0.2495	5.50E-04	6.70E-04	0.4077	0.2567
< 0.001	Any UTR	No filter	148	1,438	1,707	0.998	0.991	1.005	0.5410	-0.0227	0.0147	0.1222	7.90E-04	5.50E-04	0.1502	0.2967
< 0.001	Any UTR	≥ 5	142	909	1,081	0.995	0.986	1.004	0.2899	-0.0453	0.0186	0.0146	1.57E-03	7.00E-04	0.0247	0.0473
< 0.001	5'UTR only	No filter	106	319	383	0.994	0.980	1.009	0.4386	-0.0821	0.0298	0.0060	2.96E-03	1.12E-03	0.0085	0.0249
< 0.001	5'UTR only	≥ 5	100	281	339	0.996	0.980	1.012	0.6079	-0.0763	0.0320	0.0170	2.79E-03	1.20E-03	0.0201	0.0618
< 0.001	3'UTR only	No filter	135	1,119	1,324	0.998	0.990	1.007	0.7076	-0.0087	0.0170	0.6080	2.80E-04	6.30E-04	0.6635	0.8481
< 0.001	3'UTR only	≥ 5	121	628	742	0.994	0.984	1.005	0.3172	-0.0348	0.0228	0.1261	1.13E-03	8.60E-04	0.1856	0.2549

Supplementary Table 4.5. Results from analyzing 72 genes with evidence of low dosage associated with SZ, including genes with low expression associated with SZ and genes located within or near a SZ-associated large deletion interval, using non-quadratic (no ***Bin²*** term) and quadratic (includes ***Bin²*** term) models, and applying various combinations of MAF, CADD, and region filters to the variants. In the model formulas, ***Y = 1*** indicates presence of a rare regulatory allele, and ***Bin*** represents expression bin number (these analyses used 25 expression bins, with 29 samples assigned to each bin). For the group of middle columns corresponding to the non-quadratic model, OR, CIs and P all correspond to ***Bin***. For the quadratic model, 'P: Joint' is the p-value resulting from a joint test of β_1 and β_2. In the 'Gene Region' column, 'Any UTR' means variants within the 5'UTR or 3'UTR were included. The column 'Genes' shows the number of genes analyzed, which is less than 72 when variant filters removed all variants of interest from a gene. 'Rare sites' and 'Rare alleles' are the total number of rare variant sites and rare alleles, respectively, remaining after variant filters. MAF = minor allele frequency; CADD = Combined Annotation Dependent Depletion; OR = odds ratio; L 95% CI = lower bound for the 95% confidence interval; U 95% CI = upper bound for the 95% confidence interval; P = p-value; SE = standard error.

Dataset						Model: $logodds(Y=1) = \alpha + \beta_1 * Bin$				Model: $logodds(Y=1) = \alpha + \beta_1 * Bin + \beta_2 * Bin^2$						
MAF	Gene Region	CADD	Genes	Rare Sites	Rare Alleles	OR	L 95% CI	U 95% CI	P	β_1	SE(β_1)	P: Bin	β_2	SE(β_2)	P: Bin^2	P: Joint
< 0.01	Any UTR	No filter	72	944	1,978	1.001	0.994	1.009	0.6994	-0.0232	0.0153	0.1307	9.40E-04	5.70E-04	0.0984	0.2392
< 0.01	Any UTR	≥ 5	71	609	1,302	1.000	0.991	1.009	0.9528	-0.0396	0.0189	0.0361	1.51E-03	7.10E-04	0.0321	0.1031
< 0.01	5'UTR only	No filter	56	205	470	0.994	0.978	1.011	0.4810	-0.1184	0.0328	0.0003	4.36E-03	1.23E-03	0.0004	0.0018
< 0.01	5'UTR only	≥ 5	54	184	428	0.997	0.980	1.014	0.7225	-0.1051	0.0351	0.0028	3.94E-03	1.32E-03	0.0029	0.0121
< 0.01	3'UTR only	No filter	67	739	1,508	1.004	0.995	1.012	0.3885	0.0019	0.0180	0.9175	7.00E-05	6.70E-04	0.9157	0.6857
< 0.01	3'UTR only	≥ 5	62	425	874	1.001	0.990	1.012	0.8930	-0.0126	0.0237	0.5937	5.10E-04	8.80E-04	0.5600	0.8368
< 0.001	Any UTR	No filter	72	719	872	1.005	0.995	1.014	0.3648	-0.0339	0.0206	0.0996	1.47E-03	7.60E-04	0.0547	0.1070
< 0.001	Any UTR	≥ 5	70	469	578	1.003	0.990	1.015	0.6748	-0.0557	0.0253	0.0277	2.24E-03	9.40E-04	0.0177	0.0575
< 0.001	5'UTR only	No filter	52	154	190	1.000	0.979	1.022	0.9761	-0.1125	0.0429	0.0088	4.34E-03	1.60E-03	0.0069	0.0284
< 0.001	5'UTR only	≥ 5	49	138	172	1.007	0.985	1.030	0.5205	-0.0888	0.0457	0.0520	3.66E-03	1.69E-03	0.0309	0.0835
< 0.001	3'UTR only	No filter	65	565	682	1.006	0.994	1.017	0.3327	-0.0146	0.0240	0.5431	7.70E-04	8.90E-04	0.3866	0.4312
< 0.001	3'UTR only	≥ 5	60	331	406	1.000	0.985	1.015	0.9840	-0.0444	0.0312	0.1548	1.70E-03	1.16E-03	0.1444	0.3492

Supplementary Table 4.6. Results from analyzing 69 genes with evidence of high dosage associated with SZ, including genes with high expression associated with SZ and genes located within or near a SZ-associated large duplication interval, using non-quadratic (no Bin^2 term) and quadratic (includes Bin^2 term) models, and applying various combinations of MAF, CADD, and region filters to the variants. In the model formulas, $Y = 1$ indicates presence of a rare regulatory allele, and Bin represents expression bin number (these analyses used 25 expression bins, with 29 samples assigned to each bin). For the group of middle columns corresponding to the non-quadratic model, OR, CIs and P all correspond to Bin. For the quadratic model, 'P: Joint' is the p-value resulting from a joint test of β_1 and β_2. In the 'Gene Region' column, 'Any UTR' means variants within the 5'UTR or 3'UTR were included. The column 'Genes' shows the number of genes analyzed, which is less than 69 when variant filters removed all variants of interest from a gene. 'Rare sites' and 'Rare alleles' are the total number of rare variant sites and rare alleles, respectively, remaining after variant filters. MAF = minor allele frequency; CADD = Combined Annotation Dependent Depletion; OR = odds ratio; L 95% CI = lower bound for the 95% confidence interval; U 95% CI = upper bound for the 95% confidence interval; P = p-value; SE = standard error.

Dataset						Model: $logodds(Y=1) = \alpha + \beta_1 * Bin$				Model: $logodds(Y=1) = \alpha + \beta_1 * Bin + \beta_2 * Bin^2$						
MAF	Gene Region	CADD	Genes	Rare Sites	Rare Alleles	OR	L 95% CI	U 95% CI	P	β_1	SE(β_1)	P: Bin	β_2	SE(β_2)	P: Bin^2	P: Joint
< 0.01	Any UTR	No filter	69	863	1,855	0.998	0.991	1.005	0.5043	-0.0007	0.0148	0.9600	-6.00E-05	5.50E-04	0.9093	0.7951
< 0.01	Any UTR	≥ 5	68	521	1,072	0.990	0.980	1.000	0.0453	-0.0226	0.0204	0.2683	4.90E-04	7.70E-04	0.5251	0.1101
< 0.01	5'UTR only	No filter	53	205	519	0.995	0.981	1.009	0.4832	-0.0061	0.0292	0.8358	4.00E-05	1.10E-03	0.9691	0.7814
< 0.01	5'UTR only	≥ 5	51	176	428	0.990	0.975	1.006	0.2264	-0.0382	0.0320	0.2337	1.11E-03	1.21E-03	0.3575	0.3160
< 0.01	3'UTR only	No filter	66	658	1,336	0.998	0.989	1.007	0.6341	0.0039	0.0187	0.8347	-2.30E-04	7.00E-04	0.7393	0.8448
< 0.01	3'UTR only	≥ 5	57	345	644	0.988	0.975	1.001	0.0618	-0.0206	0.0272	0.4487	3.20E-04	1.03E-03	0.7576	0.1664
< 0.001	Any UTR	No filter	68	665	772	0.993	0.983	1.004	0.2094	-0.0033	0.0220	0.8809	-1.30E-04	8.30E-04	0.8739	0.4490
< 0.001	Any UTR	≥ 5	64	403	459	0.989	0.976	1.003	0.1252	-0.0280	0.0288	0.3305	6.70E-04	1.09E-03	0.5386	0.2559
< 0.001	5'UTR only	No filter	46	145	166	0.992	0.971	1.014	0.4900	-0.0506	0.0452	0.2635	1.67E-03	1.71E-03	0.3290	0.4920
< 0.001	5'UTR only	≥ 5	43	125	142	0.988	0.965	1.012	0.3302	-0.0596	0.0489	0.2232	1.86E-03	1.85E-03	0.3147	0.3775
< 0.001	3'UTR only	No filter	64	520	606	0.993	0.982	1.005	0.2593	0.0076	0.0250	0.7608	-5.60E-04	9.40E-04	0.5528	0.4433
< 0.001	3'UTR only	≥ 5	55	278	317	0.990	0.974	1.006	0.2321	-0.0164	0.0344	0.6350	2.50E-04	1.30E-03	0.8504	0.4811

Supplementary Table 4.7. Results from analyzing 87 genes with pLi < 0.10, using non-quadratic (no Bin^2 term) and quadratic (includes Bin^2 term) models, and applying various combinations of MAF, CADD, and region filters to the variants. In the model formulas, $Y = 1$ indicates presence of a rare regulatory allele, and Bin represents expression bin number (these analyses used 25 expression bins, with 29 samples assigned to each bin). For the group of middle columns corresponding to the non-quadratic model, OR, CIs and P all correspond to Bin. For the quadratic model, 'P: Joint' is the p-value resulting from a joint test of β_1 and β_2. In the 'Gene Region' column, 'Any UTR' means variants within the 5'UTR or 3'UTR were included. The column 'Genes' shows the number of genes analyzed, which is less than 87 when variant filters removed all variants of interest from a gene. 'Rare sites' and 'Rare alleles' are the total number of rare variant sites and rare alleles, respectively, remaining after variant filters. MAF = minor allele frequency; CADD = Combined Annotation Dependent Depletion; OR = odds ratio; L 95% CI = lower bound for the 95% confidence interval; U 95% CI = upper bound for the 95% confidence interval; P = p-value; SE = standard error.

Dataset						Model: $logodds(Y = 1) = \alpha + \beta_1 * Bin$				Model: $logodds(Y = 1) = \alpha + \beta_1 * Bin + \beta_2 * Bin^2$						
MAF	Gene Region	CADD	Genes	Rare Sites	Rare Alleles	OR	L 95% CI	U 95% CI	P	β_1	SE(β_1)	P: Bin	β_2	SE(β_2)	P: Bin^2	P: Joint
< 0.01	Any UTR	No filter	87	854	1,793	0.994	0.986	1.001	0.0907	-0.0471	0.0153	0.0022	1.58E-03	5.80E-04	0.0063	0.0060
< 0.01	Any UTR	≥ 5	85	510	1,108	0.991	0.981	1.001	0.0908	-0.0697	0.0209	0.0009	2.37E-03	7.90E-04	0.0027	0.0029
< 0.01	5'UTR only	No filter	73	257	589	0.990	0.976	1.004	0.1763	-0.0861	0.0284	0.0025	2.98E-03	1.07E-03	0.0056	0.0093
< 0.01	5'UTR only	≥ 5	69	220	500	0.991	0.976	1.006	0.2331	-0.0948	0.0310	0.0023	3.33E-03	1.17E-03	0.0046	0.0096
< 0.01	3'UTR only	No filter	79	597	1,204	0.994	0.985	1.004	0.2464	-0.0277	0.0195	0.1567	8.60E-04	7.30E-04	0.2430	0.2600
< 0.01	3'UTR only	≥ 5	68	290	608	0.989	0.975	1.003	0.1355	-0.0475	0.0295	0.1071	1.43E-03	1.12E-03	0.2000	0.1470
< 0.001	Any UTR	No filter	86	654	769	0.992	0.982	1.002	0.1213	-0.0393	0.0215	0.0672	1.21E-03	8.10E-04	0.1354	0.1000
< 0.001	Any UTR	≥ 5	80	385	464	0.990	0.976	1.004	0.1490	-0.0643	0.0281	0.0219	2.12E-03	1.06E-03	0.0464	0.0502
< 0.001	5'UTR only	No filter	65	190	230	0.993	0.974	1.012	0.4689	-0.1096	0.0379	0.0039	3.98E-03	1.43E-03	0.0054	0.0176
< 0.001	5'UTR only	≥ 5	59	161	197	0.995	0.975	1.016	0.6532	-0.1110	0.0410	0.0069	4.11E-03	1.54E-03	0.0077	0.0283
< 0.001	3'UTR only	No filter	75	464	539	0.991	0.979	1.004	0.1660	-0.0138	0.0263	0.6004	1.90E-04	9.90E-04	0.8477	0.3761
< 0.001	3'UTR only	≥ 5	64	224	267	0.986	0.968	1.004	0.1180	-0.0330	0.0376	0.3804	7.20E-04	1.43E-03	0.6124	0.2596

Supplementary Table 4.8. Results from analyzing 33 genes with pLi ≥ 0.90, using non-quadratic (no Bin^2 term) and quadratic (includes Bin^2 term) models, and applying various combinations of MAF, CADD, and region filters to the variants. In the model formulas, ***Y = 1*** indicates presence of a rare regulatory allele, and ***Bin*** represents expression bin number (these analyses used 25 expression bins, with 29 samples assigned to each bin). For the group of middle columns corresponding to the non-quadratic model, OR, CIs and P all correspond to ***Bin***. For the quadratic model, 'P: Joint' is the p-value resulting from a joint test of β_1 and β_2. In the 'Gene Region' column, 'Any UTR' means variants within the 5'UTR or 3'UTR were included. The column 'Genes' shows the number of genes analyzed, which is less than 33 when variant filters removed all variants of interest from a gene. 'Rare sites' and 'Rare alleles' are the total number of rare variant sites and rare alleles, respectively, remaining after variant filters. MAF = minor allele frequency; CADD = Combined Annotation Dependent Depletion; OR = odds ratio; L 95% CI = lower bound for the 95% confidence interval; U 95% CI = upper bound for the 95% confidence interval; P = p-value; SE = standard error.

Dataset						Model: $logodds(Y=1) = \alpha + \beta_1 * Bin$				Model: $logodds(Y=1) = \alpha + \beta_1 * Bin + \beta_2 * Bin^2$						
MAF	Gene Region	CADD	Genes	Rare Sites	Rare Alleles	OR	L 95% CI	U 95% CI	P	β_1	SE(β_1)	P: Bin	β_2	SE(β_2)	P: Bin^2	P: Joint
< 0.01	Any UTR	No filter	33	584	1,247	1.001	0.992	1.009	0.8946	0.0118	0.0187	0.5285	-4.30E-04	7.00E-04	0.5373	0.8188
< 0.01	Any UTR	≥ 5	33	379	763	0.994	0.983	1.004	0.2434	-0.0176	0.0224	0.4331	4.30E-04	8.40E-04	0.6065	0.4437
< 0.01	5'UTR only	No filter	26	107	261	0.995	0.975	1.016	0.6484	-0.0315	0.0418	0.4521	1.04E-03	1.57E-03	0.5097	0.7259
< 0.01	5'UTR only	≥ 5	26	100	249	0.996	0.975	1.017	0.6810	-0.0373	0.0440	0.3974	1.27E-03	1.65E-03	0.4428	0.6859
< 0.01	3'UTR only	No filter	31	477	986	1.002	0.991	1.012	0.7464	0.0136	0.0224	0.5425	-4.60E-04	8.30E-04	0.5835	0.8161
< 0.01	3'UTR only	≥ 5	30	279	514	0.990	0.977	1.004	0.1671	-0.0255	0.0290	0.3783	6.10E-04	1.10E-03	0.5764	0.3292
< 0.001	Any UTR	No filter	33	447	536	0.997	0.984	1.010	0.6589	-0.0079	0.0274	0.7723	1.90E-04	1.03E-03	0.8507	0.8913
< 0.001	Any UTR	≥ 5	33	303	355	0.992	0.976	1.007	0.2954	-0.0390	0.0327	0.2344	1.19E-03	1.24E-03	0.3370	0.3666
< 0.001	5'UTR only	No filter	24	78	94	0.997	0.968	1.027	0.8435	-0.0013	0.0630	0.9834	-7.00E-05	2.36E-03	0.9780	0.9803
< 0.001	5'UTR only	≥ 5	24	72	88	1.000	0.969	1.031	0.9922	0.0012	0.0661	0.9854	-5.00E-05	2.47E-03	0.9830	0.9997
< 0.001	3'UTR only	No filter	31	369	442	0.997	0.983	1.012	0.6822	-0.0145	0.0306	0.6356	4.40E-04	1.15E-03	0.6997	0.8537
< 0.001	3'UTR only	≥ 5	30	231	267	0.988	0.970	1.007	0.2109	-0.0601	0.0385	0.1186	1.88E-03	1.46E-03	0.1980	0.2012

Supplementary Table 4.9. Results from analyzing 7 genes with extreme tolerance to missense variants, using non-quadratic (no Bin^2 term) and quadratic (includes Bin^2 term) models, and applying various combinations of MAF, CADD, and region filters to the variants. In the model formulas, $Y = 1$ indicates presence of a rare regulatory allele, and Bin represents expression bin number (these analyses used 25 expression bins, with 29 samples assigned to each bin). For the group of middle columns corresponding to the non-quadratic model, OR, CIs and P all correspond to Bin. For the quadratic model, 'P: Joint' is the p-value resulting from a joint test of β_1 and β_2. In the 'Gene Region' column, 'Any UTR' means variants within the 5'UTR or 3'UTR were included. The column 'Genes' shows the number of genes analyzed, which is less than 7 when variant filters removed all variants of interest from a gene. 'Rare sites' and 'Rare alleles' are the total number of rare variant sites and rare alleles, respectively, remaining after variant filters. MAF = minor allele frequency; CADD = Combined Annotation Dependent Depletion; OR = odds ratio; L 95% CI = lower bound for the 95% confidence interval; U 95% CI = upper bound for the 95% confidence interval; P = p-value; SE = standard error.

Dataset						Model: $logodds(Y = 1) = \alpha + \beta_1 * Bin$				Model: $logodds(Y = 1) = \alpha + \beta_1 * Bin + \beta_2 * Bin^2$						
MAF	Gene Region	CADD	Genes	Rare Sites	Rare Alleles	OR	L 95% CI	U 95% CI	P	β_1	SE(β_1)	P: Bin	β_2	SE(β_2)	P: Bin^2	P: Joint
< 0.01	Any UTR	No filter	7	99	208	0.964	0.941	0.988	0.0039	-0.0818	0.0476	0.0877	1.85E-03	1.86E-03	0.3232	0.0085
< 0.01	Any UTR	≥ 5	7	65	134	0.955	0.925	0.986	0.0048	-0.0901	0.0618	0.1468	1.80E-03	2.45E-03	0.4637	0.0127
< 0.01	5'UTR only	No filter	6	22	62	0.968	0.927	1.011	0.1424	-0.1215	0.0826	0.1435	3.57E-03	3.21E-03	0.2667	0.1737
< 0.01	5'UTR only	≥ 5	6	19	50	0.957	0.911	1.006	0.0874	-0.1885	0.0920	0.0422	5.89E-03	3.59E-03	0.1033	0.0574
< 0.01	3'UTR only	No filter	7	77	146	0.969	0.940	1.000	0.0496	-0.0557	0.0626	0.3749	9.80E-04	2.44E-03	0.6875	0.1330
< 0.01	3'UTR only	≥ 5	7	46	84	0.957	0.921	0.994	0.0229	-0.0291	0.0758	0.7019	-6.30E-04	3.02E-03	0.8356	0.0682
< 0.001	Any UTR	No filter	7	73	91	0.977	0.944	1.012	0.2037	0.0183	0.0734	0.8038	-1.64E-03	2.84E-03	0.5648	0.3729
< 0.001	Any UTR	≥ 5	7	51	69	0.974	0.938	1.012	0.1858	0.0807	0.0830	0.3319	-4.29E-03	3.24E-03	0.1871	0.1636
< 0.001	5'UTR only	No filter	4	16	22	0.997	0.941	1.055	0.9061	0.0697	0.1260	0.5814	-2.83E-03	4.73E-03	0.5509	0.8259
< 0.001	5'UTR only	≥ 5	4	14	20	0.995	0.935	1.058	0.8657	0.0045	0.1320	0.9731	-3.80E-04	4.96E-03	0.9389	0.9829
< 0.001	3'UTR only	No filter	7	57	69	0.968	0.928	1.009	0.1244	0.0093	0.0857	0.9137	-1.70E-03	3.36E-03	0.6137	0.2629
< 0.001	3'UTR only	≥ 5	7	37	49	0.962	0.919	1.008	0.1055	0.1343	0.1015	0.1874	-7.13E-03	4.09E-03	0.0834	0.0483

Supplementary Table 4.10. Results from analyzing 89 genes intolerant to missense variants, using non-quadratic (no Bin^2 term) and quadratic (includes Bin^2 term) models, and applying various combinations of MAF, CADD, and region filters to the variants. In the model formulas, $Y = 1$ indicates presence of a rare regulatory allele, and Bin represents expression bin number (these analyses used 25 expression bins, with 29 samples assigned to each bin). For the group of middle columns corresponding to the non-quadratic model, OR, CIs and P all correspond to Bin. For the quadratic model, 'P: Joint' is the p-value resulting from a joint test of β_1 and β_2. In the 'Gene Region' column, 'Any UTR' means variants within the 5'UTR or 3'UTR were included. The column 'Genes' shows the number of genes analyzed, which is less than 89 when variant filters removed all variants of interest from a gene. 'Rare sites' and 'Rare alleles' are the total number of rare variant sites and rare alleles, respectively, remaining after variant filters. MAF = minor allele frequency; CADD = Combined Annotation Dependent Depletion; OR = odds ratio; L 95% CI = lower bound for the 95% confidence interval; U 95% CI = upper bound for the 95% confidence interval; P = p-value; SE = standard error.

Dataset						Model: $logodds(Y = 1) = \alpha + \beta_1 * Bin$				Model: $logodds(Y = 1) = \alpha + \beta_1 * Bin + \beta_2 * Bin^2$						
MAF	Gene Region	CADD	Genes	Rare Sites	Rare Alleles	OR	L 95% CI	U 95% CI	P	β_1	SE(β_1)	P: Bin	β_2	SE(β_2)	P: Bin^2	P: Joint
< 0.01	Any UTR	No filter	89	1,210	2,599	1.001	0.995	1.007	0.6899	0.0006	0.0131	0.9644	3.00E-05	4.90E-04	0.9581	0.9222
< 0.01	Any UTR	≥ 5	88	780	1,639	0.996	0.989	1.004	0.3735	-0.0208	0.0163	0.2009	6.70E-04	6.10E-04	0.2736	0.3710
< 0.01	5'UTR only	No filter	70	263	637	1.000	0.987	1.013	0.9947	-0.0381	0.0276	0.1685	1.46E-03	1.03E-03	0.1568	0.3711
< 0.01	5'UTR only	≥ 5	68	235	568	0.998	0.984	1.012	0.7676	-0.0541	0.0295	0.0670	2.00E-03	1.11E-03	0.0701	0.1902
< 0.01	3'UTR only	No filter	83	947	1,962	1.001	0.994	1.009	0.7389	0.0064	0.0158	0.6840	-2.00E-04	5.90E-04	0.7360	0.8936
< 0.01	3'UTR only	≥ 5	76	545	1,071	0.995	0.985	1.005	0.2818	-0.0173	0.0210	0.4116	4.60E-04	7.90E-04	0.5644	0.4753
< 0.001	Any UTR	No filter	89	924	1,104	0.999	0.990	1.008	0.8240	-0.0261	0.0181	0.1491	9.70E-04	6.80E-04	0.1527	0.3536
< 0.001	Any UTR	≥ 5	84	604	713	0.995	0.984	1.006	0.3792	-0.0571	0.0226	0.0117	2.02E-03	8.50E-04	0.0176	0.0424
< 0.001	5'UTR only	No filter	66	190	226	0.994	0.975	1.013	0.5280	-0.0547	0.0394	0.1647	1.88E-03	1.48E-03	0.2038	0.3698
< 0.001	5'UTR only	≥ 5	61	170	203	0.994	0.974	1.014	0.5366	-0.0584	0.0418	0.1630	2.02E-03	1.58E-03	0.2006	0.3689
< 0.001	3'UTR only	No filter	81	734	878	1.000	0.991	1.010	0.9591	-0.0237	0.0206	0.2499	9.20E-04	7.70E-04	0.2310	0.4900
< 0.001	3'UTR only	≥ 5	74	434	510	0.996	0.983	1.009	0.5402	-0.0630	0.0272	0.0204	2.28E-03	1.02E-03	0.0256	0.0716

Supplementary Table 4.11 (next page). Results from analyzing 64 genes with SZ-associated expression, using controls only. Results are presented for the quadratic models only, as these provided the best fit for the data in our main analyses (which analyzed all 725 samples and adjusted for case-control status). We applied the same combinations of MAF, CADD, and region filters to the variants as were used when analyzing the 64 SZ-associated expression genes in our main analyses. We also applied the same weights (based on number of rare variant sites for each gene) as used in our main analyses. Last four columns show correlations between quadratic model betas estimated from controls only (presented in this table) versus those from our main analyses (presented in **Supplementary Table 4.1**), and correlations between quadratic model betas estimated from controls only versus SZ cases only (presented in **Supplementary Table 4.12**). In the model formula, ***Y = 1*** indicates presence of a rare regulatory allele, and ***Bin*** represents expression bin number (these analyses used 25 expression bins, with 14 samples assigned to each bin). For the quadratic model, 'P: Joint' is the p-value resulting from a joint test of β_1 and β_2. In the 'Gene Region' column, 'Any reg' means variants within the promoter, 5'UTR or 3'UTR were included; 'Upstream reg' means only variants within the promoter or 5'UTR were included; 'Any UTR' means only variants within the 5'UTR or 3'UTR were included; 'Prom only' means only variants within the promoter region were included. The column 'Genes' shows the number of genes analyzed, which is less than 64 when variant filters removed all variants of interest from a gene. 'Rare alleles' are the total number of rare alleles remaining after variant filters. MAF = minor allele frequency; CADD = Combined Annotation Dependent Depletion; OR = odds ratio; L 95% CI = lower bound for the 95% confidence interval; U 95% CI = upper bound for the 95% confidence interval; P = p-value; SE = standard error.

Dataset					Model: $logodds(Y=1) = \alpha + \beta_1 * Bin + \beta_2 * Bin^2$							Correlation between			
MAF	Gene Region	CADD	Genes	Rare Alleles	β_1	SE(β_1)	P: Bin	β_2	SE(β_2)	P: Bin^2	P: Joint	Controls only β_1 and main analysis β_1	Controls only β_2 and main analysis β_2	Controls only β_1 and cases only β_1	Controls only β_2 and cases only β_2
< 0.01	Any reg	No filter	64	1,194	-0.0187	0.0207	0.3662	0.0008	0.0008	0.2885	0.4977				
< 0.01	Any reg	≥ 5	62	748	-0.0310	0.0267	0.2447	0.0014	0.0010	0.1724	0.3180				
< 0.01	Upstream reg	No filter	64	583	-0.0232	0.0305	0.4465	0.0011	0.0011	0.3260	0.4505				
< 0.01	Upstream reg	≥ 5	61	451	-0.0356	0.0333	0.2846	0.0015	0.0012	0.2122	0.3904				
< 0.01	Any UTR	No filter	60	813	-0.0094	0.0246	0.7014	0.0004	0.0009	0.6815	0.9177				
< 0.01	Any UTR	≥ 5	55	470	-0.0342	0.0326	0.2942	0.0013	0.0012	0.2735	0.5514				
< 0.01	Prom only	No filter	64	550	-0.0277	0.0310	0.3719	0.0013	0.0012	0.2579	0.3685				
< 0.01	Prom only	≥ 5	61	429	-0.0357	0.0346	0.3016	0.0017	0.0013	0.1958	0.2850				
< 0.01	5'UTR only	No filter	45	202	-0.0607	0.0490	0.2164	0.0024	0.0018	0.1832	0.4072				
< 0.01	5'UTR only	≥ 5	43	173	-0.0910	0.0512	0.0754	0.0034	0.0019	0.0767	0.2101				
< 0.01	3'UTR only	No filter	53	611	0.0014	0.0275	0.9604	-0.0001	0.0010	0.9606	0.9988				
< 0.01	3'UTR only	≥ 5	41	297	-0.0226	0.0388	0.5608	0.0009	0.0015	0.5413	0.8300				
< 0.001	Any reg	No filter	64	569	-0.0324	0.0286	0.2584	0.0012	0.0011	0.2478	0.5159	**0.85**	**0.84**	**0.58**	**0.60**
< 0.001	Any reg	≥ 5	62	356	-0.0375	0.0360	0.2981	0.0016	0.0013	0.2408	0.4663				
< 0.001	Upstream reg	No filter	63	277	-0.0260	0.0417	0.5321	0.0013	0.0015	0.4163	0.5709				
< 0.001	Upstream reg	≥ 5	61	210	-0.0236	0.0462	0.6093	0.0014	0.0017	0.4228	0.3900				
< 0.001	Any UTR	No filter	57	364	-0.0367	0.0333	0.2706	0.0013	0.0013	0.3176	0.5282				
< 0.001	Any UTR	≥ 5	51	210	-0.0606	0.0440	0.1687	0.0021	0.0017	0.1956	0.3898				
< 0.001	Prom only	No filter	63	262	-0.0271	0.0424	0.5229	0.0014	0.0016	0.3835	0.4734				
< 0.001	Prom only	≥ 5	61	198	-0.0349	0.0483	0.4709	0.0018	0.0018	0.3143	0.3429				
< 0.001	5'UTR only	No filter	39	72	-0.1008	0.0689	0.1438	0.0039	0.0026	0.1362	0.3374				
< 0.001	5'UTR only	≥ 5	37	64	-0.1241	0.0713	0.0819	0.0047	0.0027	0.0765	0.2161				
< 0.001	3'UTR only	No filter	51	292	-0.0183	0.0374	0.6256	0.0005	0.0014	0.7159	0.8032				
< 0.001	3'UTR only	≥ 5	37	146	-0.0211	0.0541	0.6967	0.0006	0.0020	0.7536	0.8957				

Supplementary Table 4.12 (next page). Results from analyzing 64 genes with SZ-associated expression, using SZ cases only. Results are presented for the quadratic models only, as these provided the best fit for the data in our main analyses (which analyzed all 725 samples and adjusted for case-control status). We applied the same combinations of MAF, CADD, and region filters to the variants as were used when analyzing the 64 SZ-associated expression genes in our main analyses. We also applied the same weights (based on number of rare variant sites for each gene) as used in our main analyses. Last four columns show correlations between quadratic model betas estimated from SZ cases only (presented in this table) versus those from our main analyses (presented in **Supplementary Table 4.1**), and correlations between quadratic model betas estimated from SZ cases only versus controls only (presented in **Supplementary Table 4.11**). In the model formula, ***Y = 1*** indicates presence of a rare regulatory allele, and ***Bin*** represents expression bin number (these analyses used 25 expression bins, with 14 samples assigned to each bin). For the quadratic model, 'P: Joint' is the p-value resulting from a joint test of β_1 and β_2. In the 'Gene Region' column, 'Any reg' means variants within the promoter, 5'UTR or 3'UTR were included; 'Upstream reg' means only variants within the promoter or 5'UTR were included; 'Any UTR' means only variants within the 5'UTR or 3'UTR were included; 'Prom only' means only variants within the promoter region were included. The column 'Genes' shows the number of genes analyzed, which is less than 64 when variant filters removed all variants of interest from a gene. 'Rare alleles' are the total number of rare alleles remaining after variant filters. MAF = minor allele frequency; CADD = Combined Annotation Dependent Depletion; OR = odds ratio; L 95% CI = lower bound for the 95% confidence interval; U 95% CI = upper bound for the 95% confidence interval; P = p-value; SE = standard error.

Dataset					Model: $logodds(Y=1) = \alpha + \beta_1 * Bin + \beta_2 * Bin^2$							Correlation between			
MAF	Gene Region	CADD	Genes	Rare Alleles	β_1	SE(β_1)	P: Bin	β_2	SE(β_2)	P: Bin^2	P: Joint	Cases only β_1 and main analysis β_1	Cases only β_2 and main analysis β_2	Cases only β_1 and controls only β_1	Cases only β_2 and controls only β_2
< 0.01	Any reg	No filter	64	1,215	-0.0209	0.0187	0.2626	5.90E-04	7.00E-04	0.4008	0.3220				
< 0.01	Any reg	≥ 5	64	798	-0.0199	0.0244	0.4158	5.50E-04	9.20E-04	0.5518	0.5240				
< 0.01	Upstream reg	No filter	63	587	-0.0085	0.0289	0.7696	1.20E-04	1.09E-03	0.9136	0.7367				
< 0.01	Upstream reg	≥ 5	62	471	-0.0271	0.0339	0.4238	9.80E-04	1.27E-03	0.4412	0.7282				
< 0.01	Any UTR	No filter	62	872	-0.0335	0.0214	0.1174	1.12E-03	8.00E-04	0.1612	0.2654				
< 0.01	Any UTR	≥ 5	59	529	-0.0302	0.0291	0.2997	9.10E-04	1.10E-03	0.4050	0.4577				
< 0.01	Prom only	No filter	63	570	-0.0123	0.0296	0.6776	2.80E-04	1.11E-03	0.7977	0.7637				
< 0.01	Prom only	≥ 5	61	461	-0.0308	0.0341	0.3667	1.14E-03	1.28E-03	0.3714	0.6663				
< 0.01	5'UTR only	No filter	47	244	-0.0530	0.0438	0.2269	1.87E-03	1.65E-03	0.2574	0.4819				
< 0.01	5'UTR only	≥ 5	46	202	-0.0630	0.0515	0.2221	2.46E-03	1.92E-03	0.2014	0.4469				
< 0.01	3'UTR only	No filter	54	628	-0.0302	0.0253	0.2323	9.50E-04	9.50E-04	0.3159	0.4039				
< 0.01	3'UTR only	≥ 5	46	327	-0.0460	0.0354	0.1942	1.50E-03	1.33E-03	0.2598	0.3783				
< 0.001	Any reg	No filter	64	539	-0.0299	0.0277	0.2801	1.08E-03	1.04E-03	0.2964	0.5613	**0.80**	**0.85**	**0.58**	**0.60**
< 0.001	Any reg	≥ 5	63	357	-0.0630	0.0346	0.0685	2.25E-03	1.30E-03	0.0842	0.1943				
< 0.001	Upstream reg	No filter	60	269	-0.0659	0.0386	0.0880	2.38E-03	1.45E-03	0.1011	0.2400				
< 0.001	Upstream reg	≥ 5	57	213	-0.0750	0.0439	0.0879	2.65E-03	1.65E-03	0.1099	0.2351				
< 0.001	Any UTR	No filter	59	371	-0.0183	0.0334	0.5849	6.60E-04	1.25E-03	0.5947	0.8622				
< 0.001	Any UTR	≥ 5	52	231	-0.0440	0.0433	0.3095	1.45E-03	1.63E-03	0.3735	0.5580				
< 0.001	Prom only	No filter	60	259	-0.0701	0.0391	0.0734	2.60E-03	1.47E-03	0.0766	0.2059				
< 0.001	Prom only	≥ 5	56	205	-0.0799	0.0446	0.0733	2.94E-03	1.67E-03	0.0789	0.2078				
< 0.001	5'UTR only	No filter	40	101	-0.0950	0.0604	0.1162	3.27E-03	2.29E-03	0.1533	0.2809				
< 0.001	5'UTR only	≥ 5	37	87	-0.0895	0.0668	0.1805	3.05E-03	2.53E-03	0.2286	0.3898				
< 0.001	3'UTR only	No filter	48	270	-0.0199	0.0375	0.5953	6.60E-04	1.40E-03	0.6379	0.8548				
< 0.001	3'UTR only	≥ 5	41	144	-0.0615	0.0521	0.2381	2.26E-03	1.95E-03	0.2473	0.5044				

Supplementary Table 4.13 (next page). Results from analyzing 64 genes with SZ-associated expression, using all 725 samples (SZ cases and controls combined) and *not* adjusting for case-control status. Results are presented for the quadratic models only, as these provided the best fit for the data in our main analyses (which analyzed all 725 samples and adjusted for case-control status). We applied the same combinations of MAF, CADD, and region filters to the variants as were used when analyzing the 64 SZ-associated expression genes in our main analyses. We also applied the same weights (based on number of rare variant sites for each gene) as used in our main analyses. Last four columns show correlations between quadratic model betas estimated from all samples combined without adjustment for case-control status (presented in this table) versus those from our main analyses (presented in **Supplementary Table 4.1**), and correlations between quadratic model betas estimated from all samples combined without adjustment for case-control status versus controls only (presented in **Supplementary Table 4.11**). In the model formula, ***Y = 1*** indicates presence of a rare regulatory allele, and ***Bin*** represents expression bin number (these analyses used 25 expression bins, with 29 samples assigned to each bin). For the quadratic model, 'P: Joint' is the p-value resulting from a joint test of β_1 and β_2. In the 'Gene Region' column, 'Any reg' means variants within the promoter, 5'UTR or 3'UTR were included; 'Upstream reg' means only variants within the promoter or 5'UTR were included; 'Any UTR' means only variants within the 5'UTR or 3'UTR were included; 'Prom only' means only variants within the promoter region were included. The column 'Genes' shows the number of genes analyzed, which is less than 64 when variant filters removed all variants of interest from a gene. 'Rare alleles' are the total number of rare alleles remaining after variant filters. MAF = minor allele frequency; CADD = Combined Annotation Dependent Depletion; OR = odds ratio; L 95% CI = lower bound for the 95% confidence interval; U 95% CI = upper bound for the 95% confidence interval; P = p-value; SE = standard error.

Dataset					Model: $logodds(Y=1) = \alpha + \beta_1 * Bin + \beta_2 * Bin^2$							Correlation between			
MAF	Gene Region	CADD	Genes	Rare Alleles	β_1	SE(β_1)	P: Bin	β_2	SE(β_2)	P: Bin^2	P: Joint	Non-adjusted β_1 and main analysis β_1	Non-adjusted β_2 and main analysis β_2	Non-adjusted β_1 and controls only β_1	Non-adjusted β_2 and controls only β_2
< 0.01	Any reg	No filter	64	2,494	-0.0191	0.0131	0.1434	7.30E-04	4.90E-04	0.1368	0.3317				
< 0.01	Any reg	≥ 5	64	1,592	-0.0290	0.0169	0.0856	1.10E-03	6.30E-04	0.0815	0.2207				
< 0.01	Upstream reg	No filter	64	1,210	-0.0166	0.0191	0.3862	6.90E-04	7.10E-04	0.3360	0.6068				
< 0.01	Upstream reg	≥ 5	63	951	-0.0330	0.0215	0.1259	1.39E-03	8.00E-04	0.0831	0.1855				
< 0.01	Any UTR	No filter	62	1,744	-0.0233	0.0154	0.1290	8.60E-04	5.70E-04	0.1358	0.3178				
< 0.01	Any UTR	≥ 5	61	1,027	-0.0434	0.0210	0.0386	1.49E-03	7.90E-04	0.0583	0.1097				
< 0.01	Prom only	No filter	64	1,160	-0.0178	0.0195	0.3606	7.70E-04	7.20E-04	0.2860	0.5000				
< 0.01	Prom only	≥ 5	63	919	-0.0318	0.0220	0.1481	1.41E-03	8.20E-04	0.0836	0.1424				
< 0.01	5'UTR only	No filter	52	460	-0.0444	0.0312	0.1549	1.58E-03	1.17E-03	0.1789	0.3655				
< 0.01	5'UTR only	≥ 5	51	386	-0.0778	0.0337	0.0209	2.79E-03	1.27E-03	0.0278	0.0731				
< 0.01	3'UTR only	No filter	59	1,284	-0.0207	0.0189	0.2743	7.80E-04	7.10E-04	0.2717	0.5455				
< 0.01	3'UTR only	≥ 5	51	641	-0.0470	0.0275	0.0872	1.62E-03	1.03E-03	0.1171	0.2213				
< 0.001	Any reg	No filter	64	1,147	-0.0311	0.0183	0.0891	1.20E-03	6.80E-04	0.0781	0.2147	**0.997**	**0.996**	**0.86**	**0.84**
< 0.001	Any reg	≥ 5	64	734	-0.0495	0.0230	0.0318	1.95E-03	8.60E-04	0.0236	0.0781				
< 0.001	Upstream reg	No filter	64	568	-0.0464	0.0255	0.0692	1.85E-03	9.50E-04	0.0520	0.1488				
< 0.001	Upstream reg	≥ 5	63	440	-0.0499	0.0289	0.0837	2.11E-03	1.07E-03	0.0492	0.1125				
< 0.001	Any UTR	No filter	61	760	-0.0312	0.0219	0.1534	1.10E-03	8.20E-04	0.1804	0.3587				
< 0.001	Any UTR	≥ 5	58	453	-0.0635	0.0290	0.0285	2.22E-03	1.09E-03	0.0417	0.0899				
< 0.001	Prom only	No filter	64	543	-0.0436	0.0262	0.0964	1.81E-03	9.70E-04	0.0636	0.1552				
< 0.001	Prom only	≥ 5	63	420	-0.0461	0.0297	0.1212	2.05E-03	1.10E-03	0.0622	0.1000				
< 0.001	5'UTR only	No filter	50	181	-0.0785	0.0424	0.0645	2.72E-03	1.60E-03	0.0901	0.1750				
< 0.001	5'UTR only	≥ 5	47	159	-0.0957	0.0452	0.0343	3.38E-03	1.71E-03	0.0476	0.1089				
< 0.001	3'UTR only	No filter	57	579	-0.0211	0.0254	0.4059	7.40E-04	9.50E-04	0.4400	0.7028				
< 0.001	3'UTR only	≥ 5	49	294	-0.0519	0.0362	0.1517	1.79E-03	1.36E-03	0.1876	0.3492				

Supplementary Table 4.14 (next page)**.** Results from analyzing combined set of 149 genes using controls only. Genes had SZ-associated expression or were located within or near a SZ-associated large CNV interval. Results are presented for the quadratic models only, as these provided the best fit for the data in our main analyses (which analyzed all 725 samples and adjusted for case-control status). We applied the same combinations of MAF, CADD, and region filters to the variants as were used when analyzing the 149 genes using all 725 samples in our main analyses. We also applied the same weights (based on number of rare variant sites for each gene) as used in the main analyses. Last four columns show correlations between quadratic model betas estimated from controls only (presented in this table) versus those from our main analyses (presented in **Supplementary Table 4.4**), and correlations between quadratic model betas estimated from controls only versus SZ cases only (presented in **Supplementary Table 4.15**). In the model formula, ***Y = 1*** indicates presence of a rare regulatory allele, and ***Bin*** represents expression bin number (these analyses used 25 expression bins, with 14 samples assigned to each bin). For the quadratic model, 'P: Joint' is the p-value resulting from a joint test of β_1 and β_2. In the 'Gene Region' column, 'Any UTR' means variants within the 5'UTR or 3'UTR were included. The column 'Genes' shows the number of genes analyzed, which is less than 149 when variant filters removed all variants of interest from a gene. 'Rare alleles' are the total number of rare alleles remaining after variant filters. MAF = minor allele frequency; CADD = Combined Annotation Dependent Depletion; OR = odds ratio; L 95% CI = lower bound for the 95% confidence interval; U 95% CI = upper bound for the 95% confidence interval; P = p-value; SE = standard error.

Dataset					Model: $logodds(Y=1) = \alpha + \beta_1 * Bin + \beta_2 * Bin^2$							Correlation between			
MAF	Gene Region	CADD	Genes	Rare Alleles	β_1	SE(β_1)	P: Bin	β_2	SE(β_2)	P: Bin^2	P: Joint	Controls only β_1 and main analysis β_1	Controls only β_2 and main analysis β_2	Controls only β_1 and cases only β_1	Controls only β_2 and all cases only β_2
< 0.01	Any UTR	No filter	144	1,844	-0.0110	0.0162	0.4958	3.90E-04	6.00E-04	0.5159	0.7929				
< 0.01	Any UTR	≥ 5	136	1,144	-0.0298	0.0204	0.1449	1.04E-03	7.70E-04	0.1765	0.3383				
< 0.01	5'UTR only	No filter	102	487	-0.0535	0.0322	0.0969	2.16E-03	1.20E-03	0.0716	0.1890				
< 0.01	5'UTR only	≥ 5	96	426	-0.0633	0.0341	0.0631	2.44E-03	1.27E-03	0.0549	0.1634				
< 0.01	3'UTR only	No filter	127	1,357	0.0039	0.0188	0.8356	-1.80E-04	7.00E-04	0.7984	0.9543				
< 0.01	3'UTR only	≥ 5	107	718	-0.0142	0.0251	0.5722	2.20E-04	9.50E-04	0.8141	0.3744	**0.92**	**0.93**	**0.84**	**0.84**
< 0.001	Any UTR	No filter	140	819	-0.0383	0.0219	0.0804	1.42E-03	8.20E-04	0.0839	0.2192				
< 0.001	Any UTR	≥ 5	130	507	-0.0567	0.0280	0.0426	2.10E-03	1.05E-03	0.0457	0.1316				
< 0.001	5'UTR only	No filter	85	169	-0.0942	0.0463	0.0420	3.73E-03	1.73E-03	0.0309	0.0989				
< 0.001	5'UTR only	≥ 5	79	149	-0.1050	0.0485	0.0307	4.12E-03	1.81E-03	0.0229	0.0781				
< 0.001	3'UTR only	No filter	121	650	-0.0266	0.0248	0.2838	9.40E-04	9.30E-04	0.3139	0.5611				
< 0.001	3'UTR only	≥ 5	98	358	-0.0357	0.0331	0.2805	1.13E-03	1.25E-03	0.3630	0.4821				

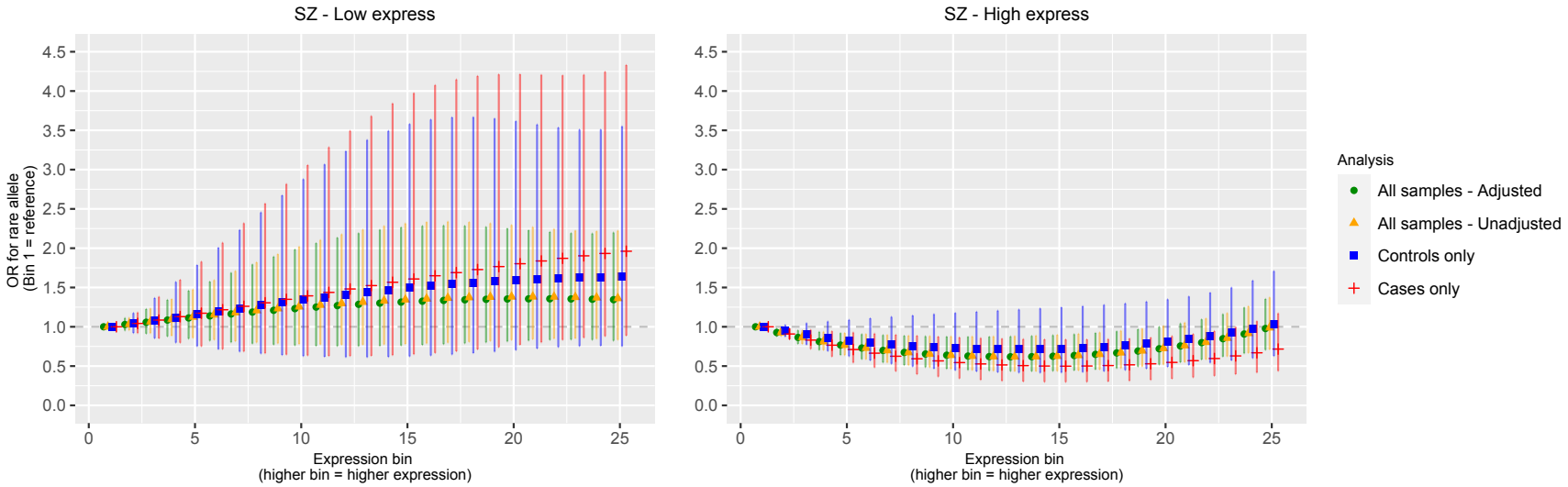
Supplementary Table 4.15 (next page)**.** Results from analyzing combined set of 149 genes using SZ cases only. Genes had SZ-associated expression or were located within or near a SZ-associated large CNV interval. Results are presented for the quadratic models only, as these provided the best fit for the data in our main analyses (which analyzed all 725 samples and adjusted for case-control status). We applied the same combinations of MAF, CADD, and region filters to the variants as were used when analyzing the 149 genes using all 725 samples in our main analyses. We also applied the same weights (based on number of rare variant sites for each gene) as used in the main analyses. Last four columns show correlations between quadratic model betas estimated from SZ cases only (presented in this table) versus those from our main analyses (presented in **Supplementary Table 4.4**), and correlations between quadratic model betas estimated from SZ cases only versus controls only (presented in **Supplementary Table 4.14**). In the model formula, ***Y = 1*** indicates presence of a rare regulatory allele, and ***Bin*** represents expression bin number (these analyses used 25 expression bins, with 14 samples assigned to each bin). For the quadratic model, 'P: Joint' is the p-value resulting from a joint test of β_1 and β_2. In the 'Gene Region' column, 'Any UTR' means variants within the 5'UTR or 3'UTR were included. The column 'Genes' shows the number of genes analyzed, which is less than 149 when variant filters removed all variants of interest from a gene. 'Rare alleles' are the total number of rare alleles remaining after variant filters. MAF = minor allele frequency; CADD = Combined Annotation Dependent Depletion; OR = odds ratio; L 95% CI = lower bound for the 95% confidence interval; U 95% CI = upper bound for the 95% confidence interval; P = p-value; SE = standard error.

Dataset					Model: $logodds(Y=1) = \alpha + \beta_1 * Bin + \beta_2 * Bin^2$							Correlation between			
MAF	Gene Region	CADD	Genes	Rare Alleles	β_1	SE(β_1)	P: Bin	β_2	SE(β_2)	P: Bin^2	P: Joint	Cases only β_1 and main analysis β_1	Cases only β_2 and main analysis β_2	Cases only β_1 and controls only β_1	Cases only β_2 and all controls only β_2
< 0.01	Any UTR	No filter	147	1,951	-0.0138	0.0154	0.3693	3.60E-04	5.80E-04	0.5380	0.3856				
< 0.01	Any UTR	≥ 5	141	1,215	-0.0205	0.0198	0.2989	5.60E-04	7.40E-04	0.4494	0.3445				
< 0.01	5'UTR only	No filter	108	511	-0.0725	0.0300	0.0158	2.51E-03	1.13E-03	0.0265	0.0516				
< 0.01	5'UTR only	≥ 5	105	441	-0.0647	0.0337	0.0552	2.37E-03	1.26E-03	0.0615	0.1649				
< 0.01	3'UTR only	No filter	126	1,440	-0.0080	0.0184	0.6657	1.30E-04	6.90E-04	0.8517	0.5751				
< 0.01	3'UTR only	≥ 5	107	774	-0.0164	0.0243	0.5013	4.50E-04	9.10E-04	0.6219	0.6445				
< 0.001	Any UTR	No filter	140	835	-0.0186	0.0225	0.4078	5.90E-04	8.40E-04	0.4859	0.6470	**0.93**	**0.93**	**0.84**	**0.84**
< 0.001	Any UTR	≥ 5	125	545	-0.0274	0.0278	0.3243	8.00E-04	1.05E-03	0.4467	0.4504				
< 0.001	5'UTR only	No filter	88	202	-0.1039	0.0422	0.0139	3.46E-03	1.61E-03	0.0316	0.0348				
< 0.001	5'UTR only	≥ 5	83	178	-0.0611	0.0459	0.1836	2.03E-03	1.73E-03	0.2413	0.3764				
< 0.001	3'UTR only	No filter	117	633	-0.0109	0.0258	0.6740	3.70E-04	9.60E-04	0.7045	0.9092				
< 0.001	3'UTR only	≥ 5	97	367	-0.0323	0.0333	0.3328	1.13E-03	1.25E-03	0.3676	0.6213				

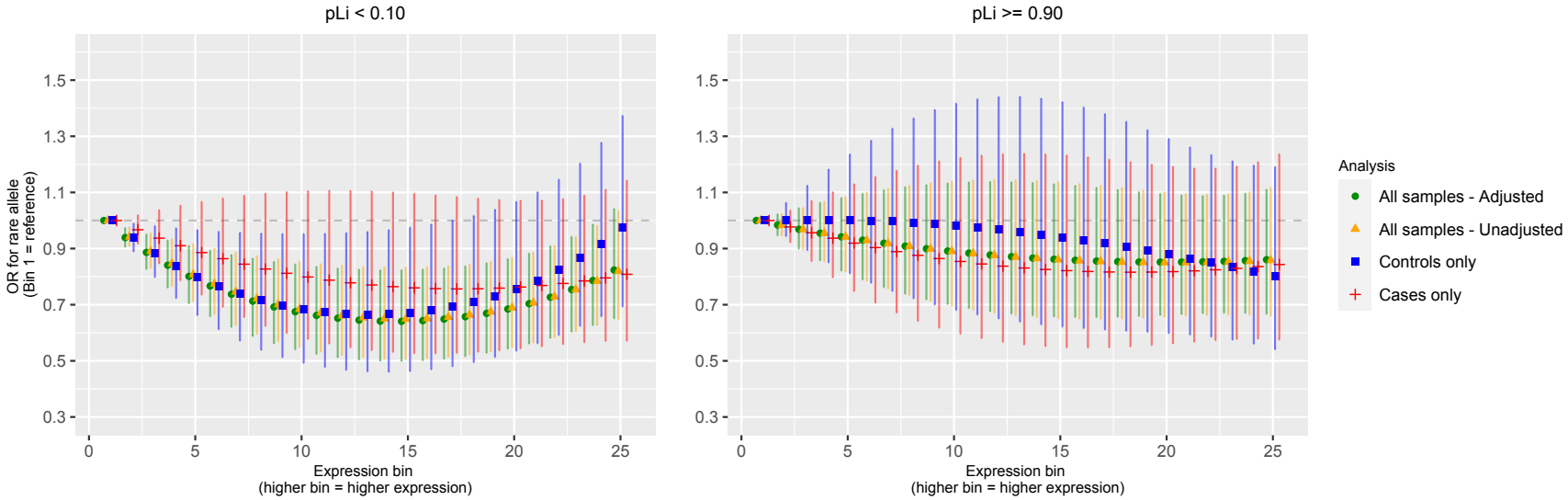
Supplementary Table 4.16 (next page). Results from analyzing combined set of 149 genes, using all 725 samples (SZ cases and controls combined) and *not* adjusting for case-control status. Genes had SZ-associated expression or were located within or near a SZ-associated large CNV interval. Results are presented for the quadratic models only, as these provided the best fit for the data in our main analyses (which analyzed all 725 samples and adjusted for case-control status). We applied the same combinations of MAF, CADD, and region filters to the variants as were used when analyzing the 149 genes using all 725 samples in our main analyses. We also applied the same weights (based on number of rare variant sites for each gene) as used in the main analyses. Last four columns show correlations between quadratic model betas estimated from all samples combined without adjustment for case-control status (presented in this table) versus those from our main analyses (presented in **Supplementary Table 4.4**), and correlations between quadratic model betas estimated from all samples combined without adjustment for case-control status versus controls only (presented in **Supplementary Table 4.14**). In the model formula, ***Y = 1*** indicates presence of a rare regulatory allele, and ***Bin*** represents expression bin number (these analyses used 25 expression bins, with 29 samples assigned to each bin). For the quadratic model, 'P: Joint' is the p-value resulting from a joint test of β_1 and β_2. In the 'Gene Region' column, 'Any UTR' means variants within the 5'UTR or 3'UTR were included. The column 'Genes' shows the number of genes analyzed, which is less than 149 when variant filters removed all variants of interest from a gene. 'Rare alleles' are the total number of rare alleles remaining after variant filters. MAF = minor allele frequency; CADD = Combined Annotation Dependent Depletion; OR = odds ratio; L 95% CI = lower bound for the 95% confidence interval; U 95% CI = upper bound for the 95% confidence interval; P = p-value; SE = standard error.

Dataset					Model: $logodds(Y=1) = \alpha + \beta_1 * Bin + \beta_2 * Bin^2$							Correlation between			
MAF	Gene Region	CADD	Genes	Rare Alleles	β_1	SE(β_1)	P: Bin	β_2	SE(β_2)	P: Bin^2	P: Joint	Non-adjusted β_1 and main analysis β_1	Non-adjusted β_2 and main analysis β_2	Non-adjusted β_1 and controls only β_1	Non-adjusted β_2 and all controls only β_2
< 0.01	Any UTR	No filter	149	3,929	-0.0151	0.0107	0.1574	5.30E-04	4.00E-04	0.1897	0.3587				
< 0.01	Any UTR	≥ 5	147	2,432	-0.0332	0.0138	0.0160	1.08E-03	5.20E-04	0.0365	0.0343				
< 0.01	5'UTR only	No filter	117	1,028	-0.0616	0.0211	0.0036	2.19E-03	7.90E-04	0.0059	0.0151				
< 0.01	5'UTR only	≥ 5	113	893	-0.0704	0.0229	0.0021	2.51E-03	8.60E-04	0.0036	0.0096				
< 0.01	3'UTR only	No filter	139	2,901	-0.0015	0.0130	0.9068	3.00E-05	4.90E-04	0.9533	0.9675				
< 0.01	3'UTR only	≥ 5	125	1,539	-0.0195	0.0178	0.2730	5.10E-04	6.70E-04	0.4484	0.2516				
< 0.001	Any UTR	No filter	148	1,707	-0.0232	0.0149	0.1195	8.00E-04	5.60E-04	0.1490	0.2893	**0.999**	**0.999**	**0.93**	**0.94**
< 0.001	Any UTR	≥ 5	142	1,081	-0.0447	0.0187	0.0168	1.54E-03	7.00E-04	0.0283	0.0528				
< 0.001	5'UTR only	No filter	106	383	-0.0809	0.0298	0.0066	2.91E-03	1.12E-03	0.0094	0.0271				
< 0.001	5'UTR only	≥ 5	100	339	-0.0746	0.0319	0.0194	2.73E-03	1.20E-03	0.0229	0.0691				
< 0.001	3'UTR only	No filter	135	1,324	-0.0097	0.0172	0.5710	3.10E-04	6.40E-04	0.6323	0.8159				
< 0.001	3'UTR only	≥ 5	97	367	-0.0323	0.0333	0.3328	1.13E-03	1.25E-03	0.3676	0.6213				

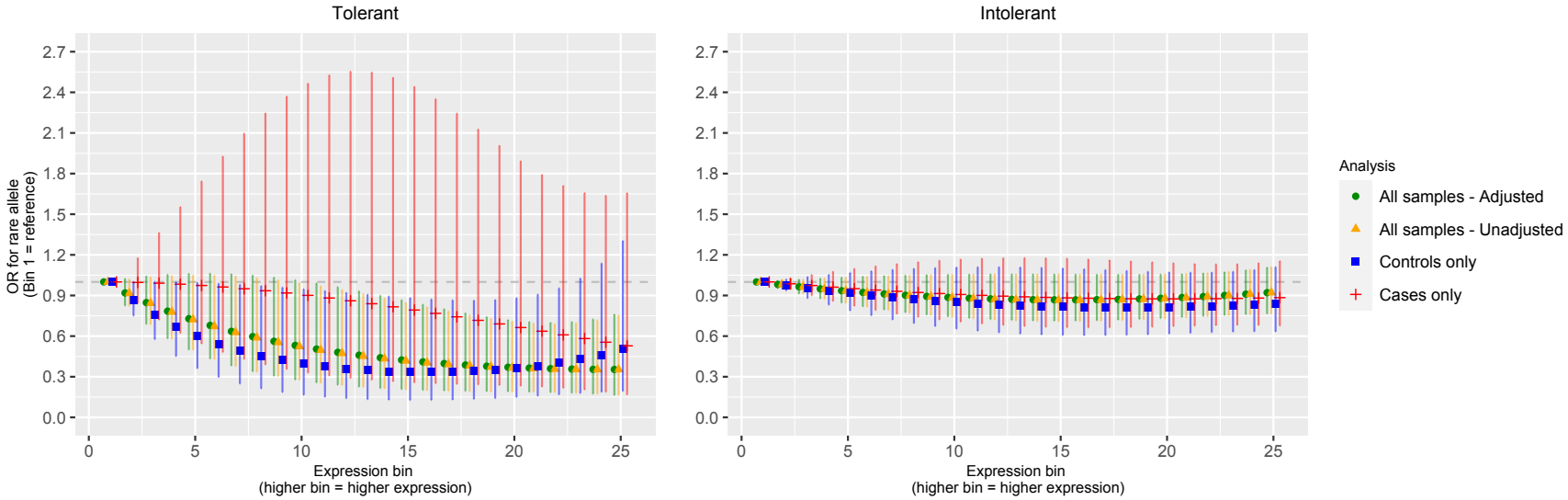
Supplementary Figure 4.1. Comparison of results when performing burden analyses for all samples with adjustment for case-control status ('All samples - Adjusted'; main analyses), all samples without adjustment for case-control status ('All samples - Unadjusted'), controls only, and cases only. Analyses were performed for up to 17 genes with low expression associated with SZ and up to 39 genes with high expression associated with SZ. These results are based on analyzing any regulatory variants (promoter, 5'UTR or 3'UTR variants) with MAF < 0.001 and CADD ≥ 5. Points are ORs comparing each expression bin to bin 1, and vertical lines are corresponding 95% CIs. ORs and 95% CIs were calculated using estimates from the corresponding quadratic regression models. Patterns of estimated associations, including apparent differences for genes with low versus high expression associated with SZ, are quite similar across the various analyses. Note that the somewhat more convergent results observed in the 'SZ - High express' plot, as compared with the 'SZ - Low express' plot, correspond to a larger gene set (and more rare allele observations) for analysis.



Supplementary Figure 4.2. Comparison of results when performing burden analyses for all samples with adjustment for case-control status ('All samples - Adjusted'; main analyses), all samples without adjustment for case-control status ('All samples - Unadjusted'), controls only, and cases only. Analyses were performed for up to 87 genes with pLi < 0.10 and up to 33 genes with pLi ≥ 0.90. These results are based on analyzing any UTR variants (5'UTR or 3'UTR variants) with MAF < 0.01 and CADD ≥ 5. Points are ORs comparing each expression bin to bin 1, and vertical lines are corresponding 95% CIs. ORs and 95% CIs were calculated using estimates from the corresponding quadratic regression models. Patterns of estimated associations, including apparent differences for genes with pLi < 0.10 versus pLi ≥ 0.90, are quite similar across the various analyses.



Supplementary Figure 4.3. Comparison of results when performing burden analyses for all samples with adjustment for case-control status ('All samples - Adjusted'; main analyses), all samples without adjustment for case-control status ('All samples - Unadjusted'), controls only, and cases only. Analyses were performed for up to 7 genes extremely tolerant to missense variation and up to 89 genes intolerant to missense variation. These results are based on analyzing any UTR variants (5'UTR or 3'UTR variants) with MAF < 0.01 and CADD ≥ 5. Points are ORs comparing each expression bin to bin 1, and vertical lines are corresponding 95% CIs. ORs and 95% CIs were calculated using estimates from the corresponding quadratic regression models. Patterns of estimated associations, including apparent differences for genes extremely tolerant versus intolerant to missense variation, are quite similar across the various analyses. Note that the more convergent results observed in the 'Intolerant' plot, as compared with the 'Tolerant' plot, correspond to a larger gene set (and more rare allele observations) for analysis.



Chapter 5:

Summary of Results, Future Research

This dissertation had the overarching goal of investigating genetic associations with particular phenotypes of interest, specifically by employing recently developed and cutting-edge genetic epidemiology analytic approaches that enable increased power for detecting associations. Such approaches can facilitate the examination of genetic associations that have historically presented challenges for study, and can do so without the need for increased sample size. In **Aim 1**, we employed PRS methods to optimally investigate a potential role for common (MAF > 1%) genetic variants in DS-associated AVSD. In **Aim 2**, we used data simulations to test the power of a cutting-edge multivariate analysis approach called GAMuT for identifying common variant associations with multivariate questionnaire data, and then applied GAMuT to real data to more powerfully examine genetic associations with multivariate psychiatric phenotypes. In **Aim 3**, we employed a modified version of a recently developed burden analysis method to increase power for investigating associations of rare (MAF < 1%) regulatory variants with gene expression for a set of genes enriched for having SZ-associated expression levels.

The **Aim 1** analyses were carried out to improve understanding of the potential contribution made by common variants to AVSD among those with DS. Prior investigations of the role of common variants in DS-associated AVSD have applied GWAS methods, examining associations of individual common variants across the genome with AVSD. These investigations have identified no robust signals of common variant associations; and given a maximum sample size of 210 DS-associated AVSD cases and 242 DS controls, they have been underpowered to identify all but the largest-effect common variants.[4] We employed an alternative strategy that enables a more powerful examination of the role of common variation in DS-associated AVSD: the PRS method. As opposed to individually analyzing each common variant, PRS methodology can be used to examine the collective contribution made by common variants all across the genome, with an aggregated common variant effect more readily detected than many smaller individual effects. We applied PRS methods to a case-control sample of 487 participants,

including 245 AVSD cases and 242 controls with structurally normal hearts, all of whom had DS. PRS were constructed using weights based on SNP effect sizes estimated in the largest GWAS of congenital heart defects available (2,594 cases and 5,159 controls; all without Down syndrome). We analyzed these genome-wide PRS for association with AVSD status, and found PRS to be associated with AVSD with odds ratios ranging from 1.2 to 1.3 per standard deviation increase in PRS, and with PRS explaining an estimated 1% of variance in outcome on the liability scale. Secondary analyses indicated that any additional contribution by common variants on the trisomic chromosome 21, over and above that of common variants on the other autosomes, was negligible.

While p-values for all PRS association estimates were > 0.05 following correction for multiple statistical tests, we interpret our PRS results as suggestive of at least a small role for common variants in DS-associated AVSD. Furthermore, considering the scientific literature on PRS as well as supplementary analyses that we performed, it is reasonable to expect that PRS constructed using weights from much larger and more phenotypically relevant GWAS may explain more of the variance in DS-associated AVSD, suggesting an even larger contribution by common variants. Thus, although our results are important in that they suggest common variants may indeed make some level of contribution to DS-associated AVSD, future studies are needed to quantify what the full extent of this contribution may be. Furthermore, confirmation of a common variation contribution through application of PRS methods would provide added motivation for continuing to grow DS-associated AVSD case-control sample sizes, with the goal of applying methods other than PRS which are able interrogate which particular common variants, proximal to which genes, seem to be making the greatest contributions (e.g., methods like the gene-based sequence kernel association test [SKAT]). Findings from such future studies of common variants, in combination with findings related to rare variant contributions to DS-associated AVSD, will help further our understanding of the biology underlying AVSD and

perhaps CHD more generally, with the possibility and hope of benefiting individuals both with and without DS.

The **Aim 2** analyses had the goals of 1) examining the power of GAMuT, a previously developed multivariate analysis method, specifically for identifying common variant associations with multivariate phenotype questionnaire data; and 2) applying GAMuT to identify common variant associations with multivariate psychiatric phenotypes assessed by the PSS and BDI questionnaires. This Aim 2 work was motivated by the recognition that psychiatric phenotypes are frequently analyzed by psychiatric genetics researchers in ways that may be suboptimal and lead to decreased ability to identify genetic associations. In particular, psychiatric disorders are syndromes involving the co-occurrence of multiple correlated yet discrete symptoms; yet this multivariate nature is hidden when the multivariate data are collapsed into a single univariate measure for analysis, as commonly occurs for the PSS and BDI. Such collapsing of the multivariate data into a single measure has the potential to greatly decrease power for identifying genetic associations, particularly in realistic circumstances in which a genetic variant may affect only a subset of the multivariate psychiatric questionnaire items. We examined the power of GAMuT to detect common variant associations under such realistic scenarios, in order to determine whether GAMuT provides a good alternative to the traditional approach of analyzing a univariate summary measure.

We applied GAMuT to simulated datasets, which involved simulated common variant data for the gene *LRFN5* (including 127 common variants), and multivariate BDI phenotypes that were simulated based on the common variant genotypes. Data were simulated under a variety of causal scenarios, including scenarios in which the causal variant affected 18/21, 12/21, or 6/21 BDI items. Across all scenarios, GAMuT exhibited good power for detecting the common variant associations, and in all instances showed a tendency to outperform the univariate approaches of KMR and standard linear regression with respect to power. The power differential between GAMuT and the univariate approaches was particularly pronounced when

the causal variant affected only 12/21 or 6/21 BDI items. For instance, when the causal variant was set to affect 12 BDI items, both KMR and linear regression showed less than 20% power for detecting all 127 causal common variants (each common variant was set as causal, one at a time); whereas GAMuT had greater than 80% power to detect 55 of the causal common variants and greater than 50% power to detect 96 of the causal variants (with power defined as the proportion of p-values < 0.001). We also confirmed that GAMuT properly preserves Type I error.

We then applied GAMuT to real data from the GTP, to examine associations of common variants with PSS and BDI phenotypes. GAMuT identified common variants within or near the *SIRPA* and *ZHX2* genes to be significantly associated with the PSS (avoidance-numbing subscale) and BDI, respectively. In comparison, univariate KMR and linear regression detected no study-wise significant associations. Follow-up examination of the GAMuT-identified association between *SIRPA* and the PSS showed that common variants within *SIRPA* were only associated with a subset of the PSS avoidance-numbing subscale items, thus illustrating GAMuT's ability to identify genetic associations for scenarios in which variants are associated with only a subset of multivariate phenotypes, while univariate analysis methods struggle to identify associations under such scenarios.

Our Aim 2 analyses support the multivariate GAMuT method as a powerful and computationally efficient means of detecting common variant associations with multivariate phenotype data, which displays substantial power advantages over methods that analyze a univariate summary of the multivariate phenotype data, particularly in situations where genetic variants are associated with only a subset of the multivariate phenotype items. Such situations would be expected to occur commonly for psychiatric disorders, which are heterogeneous conditions consisting of multiple correlated yet discrete symptoms that may be differentially affected by a genetic variant. With this in mind, application of GAMuT in future psychiatric genetics studies has the potential to facilitate the identification of robust common variant associations with psychiatric phenotypes, which have often evaded detection by traditional

analytic approaches. In addition, the GAMuT framework can be employed to investigate rare variant associations, and is also amenable to analyses of other omics data types like methylation data. We hope that the application of GAMuT in future psychiatric studies will help to further understanding of the genetic architecture of these disorders, and contribute to identifying genes for which variation is strongly associated with psychiatric phenotypes, which may enable a better understanding of the biological processes that underlie various psychiatric disorders.

Aim 3 had the goal of examining associations of rare regulatory variants with gene expression levels, for genes enriched for having SZ-associated expression levels. Numerous genes have been identified by previous studies to have expression levels that are associated with SZ.[14-16] It is possible that rare variants in regulatory sequences play a particularly important role in regulating the expression of these genes, but this has not been sufficiently examined, due to a combination of technological and analytical limitations. We were able to investigate these associations in a sample of 725 individuals, including SZ cases and controls analyzed in prior case-control studies. All individuals had both targeted DNA sequence data and RNA sequencing data for 160 genes, including 64 genes with SZ-associated expression levels and 96 genes located within or near SZ-associated large CNV intervals. We used a modified version of a recently developed burden approach specifically designed to increase power for examining rare variant associations with gene expression by considering gene sets rather than individual genes for analysis.[27]

Our analyses consistently yielded U-shaped patterns of estimated association between rare regulatory allele burden and gene expression, whereby rare regulatory alleles were most likely to be observed at the extremes (low and high) of gene expression. This observation was consistent with relevant previous literature.[27] We also observed a consistent tendency for the U-shaped estimated associations to be more pronounced (i.e., deeper U) when limiting analyses to the rarest variants (MAF < 0.001), variants more likely to be deleterious (CADD ≥ 5), and

when only considering variants in the 5'UTR (as compared with including promoter and/or 3'UTR variants). In addition, we found that estimated associations between rare regulatory variants and gene expression were comparatively weaker for genes intolerant to LoF or missense variation as compared with genes tolerant to these variant types, which, considering prior scientific literature,[125] we interpreted as possibly reflecting selection against variants with strong influences on expression for the highly constrained genes.

While none of our findings were statistically significant following correction for multiple testing, they may reflect genuine associations between rare regulatory variants and gene expression for genes with (or likely to have) SZ-associated expression, particularly considering consistencies between our findings and those of prior, non-SZ focused studies. Future studies that include larger sample sizes and/or analyze larger gene sets will have greater power for examining these associations. We note that we had planned to perform the same analyses in an independent set of 400 samples with microarray expression data, and then to meta-analyze the RNA sequencing sample and microarray sample results, which would have increased our power; but we ultimately excluded the microarray dataset due to concerns surrounding quality of the microarray expression data.

Future studies can also more fully sequence putative regulatory regions for all genes analyzed, including promoter regions (less than half of the full 149 genes we analyzed had undergone DNA sequencing upstream of the TSS); as well as distal regulatory elements including enhancers, silencers, and insulators, which we did not consider and which are currently more challenging to localize than the regions we analyzed. Such studies are needed to help provide a more complete picture of how different classes of rare regulatory variation may impact gene expression.

The burden approach which we employed for these analyses is informative with regard to the association between rare allele burden and gene expression, and can therefore provide suggestive evidence about whether rare regulatory alleles may be causing decreased and/or

increased gene expression. However, the approach is inherently limited with regard to providing information about the magnitude of expression change (e.g., number of standard deviations in expression shifted) that might be caused by a rare allele. Alternative analytic methods will be needed to examine this aspect of the association between rare regulatory variants and gene expression. Very large sample sizes should eventually enable successful application of eQTL analysis methods to rare variants, which would allow quantification of the amount to which expression levels are different for those with and without specific rare alleles. Future work might also include in vitro studies which examine the amount of change in gene expression corresponding with experimental manipulation of rare regulatory alleles.

Additional analyses are needed to understand whether expression for the genes considered in our study (specifically, those identified to have SZ-associated expression levels) may in fact be mediators on potential causal pathways from rare regulatory variation to SZ.[132,133] These additional analyses would need to include an examination of associations between gene expression and SZ that control for the rare regulatory variants of interest, to rule out the possibility that an observed association between gene expression and SZ is simply due to confounding by these rare regulatory variants. These additional analyses were beyond the scope of this dissertation, but are worthy of being the focus of future research. Should such future research support a causal pathway from rare regulatory variation to aberrant gene expression to SZ for particular genes, it is possible that this knowledge could be leveraged to develop more effective treatment and prevention approaches for SZ.

The three aims described in this dissertation involved the strategic application of particular genetic epidemiology analytic methods to increase power for investigating the genetic contributions to specific phenotypes. Prior examination of these genetic associations have tended to employ less-than-optimized analytic methods, limiting the ability to identify

associations that may exist. Our use of highly optimized approaches yielded results suggestive of genetic associations with the phenotypes we investigated, contributing new evidence of potential common and rare genetic variant effects on these phenotypes. Even with our power-optimizing analytic methods, however, many of our analyses remained underpowered, and would benefit from larger sample sizes (in addition to certain other modifications to help minimize potential biases). Nonetheless, we have generated findings that we believe represent important contributions to the field of genetic epidemiology, while at the same time emphasizing the importance of selecting analytic methods that are optimized for the objectives of one's research study.

References

1 Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet* **101**, 5-22, doi:10.1016/j.ajhg.2017.06.005 (2017).

2 The Schizophrenia Working Group of the Psychiatric Genomics Consortium. Mapping genomic loci prioritises genes and implicates synaptic biology in schizophrenia. *medRxiv*, doi:https://doi.org/10.1101/2020.09.12.20192922 (2020).

3 Psychiatric Genomics Consortium. Schizophrenia. *https://www.med.unc.edu/pgc/pgc-workgroups/schizophrenia/* (2021).

4 Ramachandran, D., Zeng, Z., Locke, A. E., Mulle, J. G., Bean, L. J., Rosser, T. C. *et al.* Genome-Wide Association Study of Down Syndrome-Associated Atrioventricular Septal Defects. *G3 (Bethesda)* **5**, 1961-1971, doi:10.1534/g3.115.019943 (2015).

5 Johnson, J. L. Genetic Association Study (GAS) Power Calculator. *https://csg.sph.umich.edu/abecasis/gas_power_calculator/* (2017).

6 American Psychiatric Association. *Diagnostic and statistical manual of mental disorders*. 5th edn, (American Psychiatric Publishing, 2013).

7 Wikipedia contributors. Syndrome. *https://en.wikipedia.org/wiki/Syndrome* (2021).

8 Foa, E. B., Riggs, D. S., Dancu, C. V. & Rothbaum, B. O. Reliability and validity of a brief instrument for assessing posttraumatic-stress-disorder. *J Trauma Stress* **6**, 459-473, doi:10.1007/Bf00974317 (1993).

9 Foa, E. B. & Tolin, D. F. Comparison of the PTSD Symptom Scale-Interview Version and the Clinician-Administered PTSD scale. *J Trauma Stress* **13**, 181-191, doi:10.1023/A:1007781909213 (2000).

10 Beck, A. T., Steer, R. A. & Brown, G. K. Manual for the Beck Depression Inventory-II. *San Antonio, TX: Psychological Corporation* (1996).

11 van der Sluis, S., Posthuma, D. & Dolan, C. V. TATES: efficient multivariate genotype-phenotype analysis for genome-wide association studies. *PLoS Genet* **9**, e1003235, doi:10.1371/journal.pgen.1003235 (2013).

12 Aschard, H., Vilhjalmsson, B. J., Greliche, N., Morange, P. E., Tregouet, D. A. & Kraft, P. Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies. *Am J Hum Genet* **94**, 662-676, doi:10.1016/j.ajhg.2014.03.016 (2014).

13 Broadaway, K. A., Cutler, D. J., Duncan, R., Moore, J. L., Ware, E. B., Jhun, M. A. *et al.* A Statistical Approach for Testing Cross-Phenotype Effects of Rare Variants. *Am J Hum Genet* **98**, 525-540, doi:10.1016/j.ajhg.2016.01.017 (2016).

14 Duan, J., Sanders, A. R., Moy, W., Drigalenko, E. I., Brown, E. C., Freda, J. *et al.* Transcriptome outlier analysis implicates schizophrenia susceptibility genes and enriches putatively functional rare genetic variants. *Hum Mol Genet* **24**, 4674-4685, doi:10.1093/hmg/ddv199 (2015).

15 Sanders, A. R., Drigalenko, E. I., Duan, J., Moy, W., Freda, J., Goring, H. H. H. *et al.* Transcriptome sequencing study implicates immune-related genes differentially

expressed in schizophrenia: new data and a meta-analysis. *Transl Psychiatry* **7**, e1093, doi:10.1038/tp.2017.47 (2017).

16 Sanders, A. R., Goring, H. H., Duan, J., Drigalenko, E. I., Moy, W., Freda, J. *et al.* Transcriptome study of differential expression in schizophrenia. *Hum Mol Genet* **22**, 5001-5014, doi:10.1093/hmg/ddt350 (2013).

17 Lichtenstein, P., Yip, B. H., Bjork, C., Pawitan, Y., Cannon, T. D., Sullivan, P. F. *et al.* Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. *Lancet* **373**, 234-239, doi:10.1016/S0140-6736(09)60072-6 (2009).

18 Sullivan, P. F., Kendler, K. S. & Neale, M. C. Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. *Arch Gen Psychiatry* **60**, 1187-1192, doi:10.1001/archpsyc.60.12.1187 (2003).

19 Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421-427, doi:10.1038/nature13595 (2014).

20 Marshall, C. R., Howrigan, D. P., Merico, D., Thiruvahindrapuram, B., Wu, W., Greer, D. S. *et al.* Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nat Genet* **49**, 27-35, doi:10.1038/ng.3725 (2017).

21 Rees, E., Walters, J. T., Georgieva, L., Isles, A. R., Chambert, K. D., Richards, A. L. *et al.* Analysis of copy number variations at 15 schizophrenia-associated loci. *Br J Psychiatry* **204**, 108-114, doi:10.1192/bjp.bp.113.131052 (2014).

22 Mulle, J. G. The 3q29 deletion confers >40-fold increase in risk for schizophrenia. *Mol Psychiatry* **20**, 1028-1029, doi:10.1038/mp.2015.76 (2015).

23 Mulle, J. G., Pulver, A. E., McGrath, J. A., Wolyniec, P. S., Dodd, A. F., Cutler, D. J. *et al.* Reciprocal duplication of the Williams-Beuren syndrome deletion on chromosome 7q11.23 is associated with schizophrenia. *Biol Psychiatry* **75**, 371-377, doi:10.1016/j.biopsych.2013.05.040 (2014).

24 GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* **550**, 204-213, doi:10.1038/nature24277 (2017).

25 Li, X., Kim, Y., Tsang, E. K., Davis, J. R., Damani, F. N., Chiang, C. *et al.* The impact of rare variation on gene expression across tissues. *Nature* **550**, 239-243, doi:10.1038/nature24267 (2017).

26 Montgomery, S. B., Lappalainen, T., Gutierrez-Arcelus, M. & Dermitzakis, E. T. Rare and common regulatory variation in population-scale sequenced human genomes. *PLoS Genet* **7**, e1002144, doi:10.1371/journal.pgen.1002144 (2011).

27 Zhao, J., Akinsanmi, I., Arafat, D., Cradick, T. J., Lee, C. M., Banskota, S. *et al.* A Burden of Rare Variants Associated with Extremes of Gene Expression in Human Peripheral Blood. *Am J Hum Genet* **98**, 299-309, doi:10.1016/j.ajhg.2015.12.023 (2016).

28 Zeng, Y., Wang, G., Yang, E., Ji, G., Brinkmeyer-Langford, C. L. & Cai, J. J. Aberrant gene expression in humans. *PLoS Genet* **11**, e1004942, doi:10.1371/journal.pgen.1004942 (2015).

29 Li, X., Battle, A., Karczewski, K. J., Zappala, Z., Knowles, D. A., Smith, K. S. *et al.* Transcriptome sequencing of a large human family identifies the impact of rare

noncoding variants. *Am J Hum Genet* **95**, 245-256, doi:10.1016/j.ajhg.2014.08.004 (2014).

30 Li, X. & Montgomery, S. B. Detection and impact of rare regulatory variants in human disease. *Front Genet* **4**, 67, doi:10.3389/fgene.2013.00067 (2013).

31 Freeman, S. B., Taft, L. F., Dooley, K. J., Allran, K., Sherman, S. L., Hassold, T. J. *et al.* Population-based study of congenital heart defects in Down syndrome. *Am J Med Genet* **80**, 213-217 (1998).

32 Reller, M. D., Strickland, M. J., Riehle-Colarusso, T., Mahle, W. T. & Correa, A. Prevalence of congenital heart defects in metropolitan Atlanta, 1998-2005. *J Pediatr* **153**, 807-813, doi:10.1016/j.jpeds.2008.05.059 (2008).

33 Hoffman, J. I. & Kaplan, S. The incidence of congenital heart disease. *J Am Coll Cardiol* **39**, 1890-1900, doi:10.1016/s0735-1097(02)01886-7 (2002).

34 Centers for Disease Control and Prevention. Facts about Atrioventricular Septal Defect (AVSD). https://www.cdc.gov/ncbddd/heartdefects/avsd.html (2020).

35 Mai, C. T., Isenburg, J. L., Canfield, M. A., Meyer, R. E., Correa, A., Alverson, C. J. *et al.* National population-based estimates for major birth defects, 2010-2014. *Birth Defects Res* **111**, 1420-1435, doi:10.1002/bdr2.1589 (2019).

36 Hartman, R. J., Riehle-Colarusso, T., Lin, A., Frias, J. L., Patel, S. S., Duwe, K. *et al.* Descriptive study of nonsyndromic atrioventricular septal defects in the National Birth Defects Prevention Study, 1997-2005. *Am J Med Genet A* **155A**, 555-564, doi:10.1002/ajmg.a.33874 (2011).

37 Ramachandran, D., Mulle, J. G., Locke, A. E., Bean, L. J., Rosser, T. C., Bose, P. *et al.* Contribution of copy-number variation to Down syndrome-associated atrioventricular septal defects. *Genet Med* **17**, 554-560, doi:10.1038/gim.2014.144 (2015).

38 Rambo-Martin, B. L., Mulle, J. G., Cutler, D. J., Bean, L. J. H., Rosser, T. C., Dooley, K. J. *et al.* Analysis of Copy Number Variants on Chromosome 21 in Down Syndrome-Associated Congenital Heart Defects. *G3 (Bethesda)* **8**, 105-111, doi:10.1534/g3.117.300366 (2018).

39 Khera, A. V., Chaffin, M., Aragam, K. G., Haas, M. E., Roselli, C., Choi, S. H. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet* **50**, 1219-1224, doi:10.1038/s41588-018-0183-z (2018).

40 Hoang, T. T., Goldmuntz, E., Roberts, A. E., Chung, W. K., Kline, J. K., Deanfield, J. E. *et al.* The Congenital Heart Disease Genetic Network Study: Cohort description. *PLoS One* **13**, e0191319, doi:10.1371/journal.pone.0191319 (2018).

41 Pediatric Cardiac Genomics Consortium, Gelb, B., Brueckner, M., Chung, W., Goldmuntz, E., Kaltman, J. *et al.* The Congenital Heart Disease Genetic Network Study: rationale, design, and early results. *Circ Res* **112**, 698-706, doi:10.1161/CIRCRESAHA.111.300297 (2013).

42 Johnston, H. R., Chopra, P., Wingo, T. S., Patel, V., International Consortium on Brain and Behavior in 22q11.2 Deletion Syndrome, Epstein, M. P. *et al.* PEMapper and PECaller provide a simplified approach to whole-genome sequencing. *Proc Natl Acad Sci U S A* **114**, E1923-E1932, doi:10.1073/pnas.1618065114 (2017).

43 Kotlar, A. V., Trevino, C. E., Zwick, M. E., Cutler, D. J. & Wingo, T. S. Bystro: rapid online variant annotation and natural-language filtering at whole-genome scale. *Genome Biol* **19**, 14, doi:10.1186/s13059-018-1387-3 (2018).

44 Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M. & Lee, J. J. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7, doi:10.1186/s13742-015-0047-8 (2015).

45 Purcell, S. & Chang, C. PLINK [v1.9b6.6]. www.cog-genomics.org/plink/1.9/.

46 R Core Team. R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria.* https://www.R-project.org (2017).

47 International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299-1320, doi:10.1038/nature04226 (2005).

48 Anderson, C. A., Pettersson, F. H., Clarke, G. M., Cardon, L. R., Morris, A. P. & Zondervan, K. T. Data quality control in genetic case-control association studies. *Nat Protoc* **5**, 1564-1573, doi:10.1038/nprot.2010.116 (2010).

49 Das, S., Forer, L., Schonherr, S., Sidore, C., Locke, A. E., Kwong, A. *et al.* Next-generation genotype imputation service and methods. *Nat Genet* **48**, 1284-1287, doi:10.1038/ng.3656 (2016).

50 Rayner, W. Script to check plink .bim files against HRC/1000G for strand, id names, positions, alleles, ref/alt assignment [v 4.2.9]. https://www.well.ox.ac.uk/~wrayner/tools/.

51 McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* **48**, 1279-1283, doi:10.1038/ng.3643 (2016).

52 Hinrichs, A. S., Karolchik, D., Baertsch, R., Barber, G. P., Bejerano, G., Clawson, H. *et al.* The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res* **34**, D590-598, doi:10.1093/nar/gkj144 (2006).

53 Cordell, H. J., Bentham, J., Topf, A., Zelenika, D., Heath, S., Mamasoula, C. *et al.* Genome-wide association study of multiple congenital heart disease phenotypes identifies a susceptibility locus for atrial septal defect at chromosome 4p16. *Nat Genet* **45**, 822-824, doi:10.1038/ng.2637 (2013).

54 Cordell, H. J., Topf, A., Mamasoula, C., Postma, A. V., Bentham, J., Zelenika, D. *et al.* Genome-wide association study identifies loci on 12q24 and 13q32 associated with tetralogy of Fallot. *Hum Mol Genet* **22**, 1473-1481, doi:10.1093/hmg/dds552 (2013).

55 Agopian, A. J., Goldmuntz, E., Hakonarson, H., Sewda, A., Taylor, D., Mitchell, L. E. *et al.* Genome-Wide Association Studies and Meta-Analyses for Congenital Heart Defects. *Circ Cardiovasc Genet* **10**, e001449, doi:10.1161/CIRCGENETICS.116.001449 (2017).

56 Magi, R. & Morris, A. P. GWAMA: software for genome-wide association meta-analysis. *BMC Bioinformatics* **11**, 288, doi:10.1186/1471-2105-11-288 (2010).

57 Choi, S. W. & O'Reilly, P. F. PRSice-2: Polygenic Risk Score software for biobank-scale data. *Gigascience* **8**, doi:10.1093/gigascience/giz082 (2019).

58 Conneely, K. N. & Boehnke, M. So many correlated tests, so little time! Rapid adjustment of P values for multiple correlated tests. *Am J Hum Genet* **81**, 1158-1168, doi:10.1086/522036 (2007).

59 Dudbridge, F. Power and predictive accuracy of polygenic risk scores. *PLoS Genet* **9**, e1003348, doi:10.1371/journal.pgen.1003348 (2013).

60 Uher, R. & Zwicker, A. Etiology in psychiatry: embracing the reality of poly-gene-environmental causation of mental illness. *World Psychiatry* **16**, 121-129, doi:10.1002/wps.20436 (2017).

61 Polderman, T. J., Benyamin, B., de Leeuw, C. A., Sullivan, P. F., van Bochoven, A., Visscher, P. M. *et al.* Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nat Genet* **47**, 702-709, doi:10.1038/ng.3285 (2015).

62 Posthuma, D. MaTCH: Meta-Analysis of Twin Correlations and Heritability. http://match.ctglab.nl/#/home.

63 Sartor, C. E., Grant, J. D., Lynskey, M. T., McCutcheon, V. V., Waldron, M., Statham, D. J. *et al.* Common heritable contributions to low-risk trauma, high-risk trauma, posttraumatic stress disorder, and major depression. *Arch Gen Psychiatry* **69**, 293-299, doi:10.1001/archgenpsychiatry.2011.1385 (2012).

64 Scherrer, J. F., Xian, H., Lyons, M. J., Goldberg, J., Eisen, S. A., True, W. R. *et al.* Posttraumatic stress disorder; combat exposure; and nicotine dependence, alcohol dependence, and major depression in male twins. *Compr Psychiatry* **49**, 297-304, doi:10.1016/j.comppsych.2007.11.001 (2008).

65 Stein, M. B., Jang, K. L., Taylor, S., Vernon, P. A. & Livesley, W. J. Genetic and environmental influences on trauma exposure and posttraumatic stress disorder symptoms: a twin study. *Am J Psychiatry* **159**, 1675-1681, doi:10.1176/appi.ajp.159.10.1675 (2002).

66 True, W. R., Rice, J., Eisen, S. A., Heath, A. C., Goldberg, J., Lyons, M. J. *et al.* A twin study of genetic and environmental contributions to liability for posttraumatic stress symptoms. *Arch Gen Psychiatry* **50**, 257-264 (1993).

67 Xian, H., Chantarujikapong, S. I., Scherrer, J. F., Eisen, S. A., Lyons, M. J., Goldberg, J. *et al.* Genetic and environmental influences on posttraumatic stress disorder, alcohol and drug dependence in twin pairs. *Drug Alcohol Depend* **61**, 95-102, doi:10.1016/s0376-8716(00)00127-7 (2000).

68 Pettersson, E., Lichtenstein, P., Larsson, H., Song, J., Attention Deficit/Hyperactivity Disorder Working Group of the iPSYCH-Broad-PGC Consortium, Autism Spectrum Disorder Working Group of the iPSYCH-Broad-PGC Consortium *et al.* Genetic influences on eight psychiatric disorders based on family data of 4 408 646 full and half-siblings, and genetic data of 333 748 cases and controls. *Psychol Med* **49**, 1166-1173, doi:10.1017/S0033291718002039 (2019).

69 Duncan, L. E., Ratanatharathorn, A., Aiello, A. E., Almli, L. M., Amstadter, A. B., Ashley-Koch, A. E. *et al.* Largest GWAS of PTSD (N=20 070) yields genetic overlap with schizophrenia and sex differences in heritability. *Mol Psychiatry* **23**, 666-673, doi:10.1038/mp.2017.77 (2018).

70 Clark, L. A., Cuthbert, B., Lewis-Fernandez, R., Narrow, W. E. & Reed, G. M. Three approaches to understanding and classifying mental disorder: ICD-11, DSM-5, and the National Institute of Mental Health's Research Domain Criteria (RDoC). *Psychol Sci Public Interest* **18**, 72-145, doi:10.1177/1529100617727266 (2017).

71 National Institute of Mental Health. About RDoC. https://www.nimh.nih.gov/research/research-funded-by-nimh/rdoc/about-rdoc (2021).

72 Wessel, J. & Schork, N. J. Generalized genomic distance-based regression methodology for multilocus association analysis. *Am J Hum Genet* **79**, 792-806, doi:10.1086/508346 (2006).

73 Wu, M. C., Kraft, P., Epstein, M. P., Taylor, D. M., Chanock, S. J., Hunter, D. J. *et al.* Powerful SNP-set analysis for case-control genome-wide association studies. *Am J Hum Genet* **86**, 929-942, doi:10.1016/j.ajhg.2010.05.002 (2010).

74 Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M. & Lin, X. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* **89**, 82-93, doi:10.1016/j.ajhg.2011.05.029 (2011).

75 Kwee, L. C., Liu, D., Lin, X., Ghosh, D. & Epstein, M. P. A powerful and flexible multilocus association test for quantitative traits. *Am J Hum Genet* **82**, 386-397, doi:10.1016/j.ajhg.2007.10.010 (2008).

76 Wray, N. R., Ripke, S., Mattheisen, M., Trzaskowski, M., Byrne, E. M., Abdellaoui, A. *et al.* Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat Genet* **50**, 668-681, doi:10.1038/s41588-018-0090-3 (2018).

77 Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S. *et al.* The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Curr Protoc Bioinformatics* **54**, 1 30 31-31 30 33, doi:10.1002/cpbi.5 (2016).

78 Su, Z., Marchini, J. & Donnelly, P. HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics* **27**, 2304-2305, doi:10.1093/bioinformatics/btr341 (2011).

79 Binder, E. B., Bradley, R. G., Liu, W., Epstein, M. P., Deveau, T. C., Mercer, K. B. *et al.* Association of FKBP5 polymorphisms and childhood abuse with risk of posttraumatic stress disorder symptoms in adults. *JAMA* **299**, 1291-1305, doi:10.1001/jama.299.11.1291 (2008).

80 Major Depressive Disorder Working Group of the Psychiatric GWAS Consortium. A mega-analysis of genome-wide association studies for major depressive disorder. *Mol Psychiatry* **18**, 497-511, doi:10.1038/mp.2012.21 (2013).

81 Psychiatric GWAS Consortium Bipolar Disorder Working Group. Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nat Genet* **43**, 977-983, doi:10.1038/ng.943 (2011).

82 Cross-Disorder Group of the Psychiatric Genomics Consortium. Genomic Relationships, Novel Loci, and Pleiotropic Mechanisms across Eight Psychiatric Disorders. *Cell* **179**, 1469-1482 e1411, doi:10.1016/j.cell.2019.11.020 (2019).

83 Fagerberg, L., Hallstrom, B. M., Oksvold, P., Kampf, C., Djureinovic, D., Odeberg, J. *et al.* Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol Cell Proteomics* **13**, 397-406, doi:10.1074/mcp.M113.035600 (2014).

84 Karpyak, V. M., Winham, S. J., Preuss, U. W., Zill, P., Cunningham, J. M., Walker, D. L. *et al.* Association of the PDYN gene with alcohol dependence and the propensity to drink in negative emotional states. *Int J Neuropsychopharmacol* **16**, 975-985, doi:10.1017/S1461145712001137 (2013).

85 Ventriglia, M., Bocchio Chiavetto, L., Bonvicini, C., Tura, G. B., Bignotti, S., Racagni, G. *et al.* Allelic variation in the human prodynorphin gene promoter and schizophrenia. *Neuropsychobiology* **46**, 17-21, doi:10.1159/000063571 (2002).

86 Zhang, C. S., Tan, Z., Lu, L., Wu, S. N., He, Y., Gu, N. F. *et al.* Polymorphism of Prodynorphin promoter is associated with schizophrenia in Chinese population. *Acta Pharmacol Sin* **25**, 1022-1026 (2004).

87 Kolsch, H., Wagner, M., Bilkei-Gorzo, A., Toliat, M. R., Pentzek, M., Fuchs, A. *et al.* Gene polymorphisms in prodynorphin (PDYN) are associated with episodic memory in the elderly. *J Neural Transm (Vienna)* **116**, 897-903, doi:10.1007/s00702-009-0238-5 (2009).

88 Zhang, N., Ouyang, T. H., Zhou, Q., Kang, H. C. & Zhu, S. Q. Prodynorphin gene promoter polymorphism and temporal lobe epilepsy: A meta-analysis. *J Huazhong Univ Sci Technolog Med Sci* **35**, 635-639, doi:10.1007/s11596-015-1482-6 (2015).

89 Williams, T. J., LaForge, K. S., Gordon, D., Bart, G., Kellogg, S., Ott, J. *et al.* Prodynorphin gene promoter repeat associated with cocaine/alcohol codependence. *Addict Biol* **12**, 496-502, doi:10.1111/j.1369-1600.2007.00069.x (2007).

90 Clarke, T. K., Ambrose-Lanci, L., Ferraro, T. N., Berrettini, W. H., Kampman, K. M., Dackis, C. A. *et al.* Genetic association analyses of PDYN polymorphisms with heroin and cocaine addiction. *Genes Brain Behav* **11**, 415-423, doi:10.1111/j.1601-183X.2012.00785.x (2012).

91 Dahl, J. P., Weller, A. E., Kampman, K. M., Oslin, D. W., Lohoff, F. W., Ferraro, T. N. *et al.* Confirmation of the association between a polymorphism in the promoter region of the prodynorphin gene and cocaine dependence. *Am J Med Genet B Neuropsychiatr Genet* **139B**, 106-108, doi:10.1002/ajmg.b.30238 (2005).

92 Yuferov, V., Ji, F., Nielsen, D. A., Levran, O., Ho, A., Morgello, S. *et al.* A functional haplotype implicated in vulnerability to develop cocaine dependence is associated with reduced PDYN expression in human brain. *Neuropsychopharmacology* **34**, 1185-1197, doi:10.1038/npp.2008.187 (2009).

93 Saify, K., Saadat, I. & Saadat, M. Association between VNTR polymorphism in promoter region of prodynorphin (PDYN) gene and heroin dependence. *Psychiatry Res* **219**, 690-692, doi:10.1016/j.psychres.2014.06.048 (2014).

94 Walker, S. & Scherer, S. W. Identification of candidate intergenic risk loci in autism spectrum disorder. *BMC Genomics* **14**, 499, doi:10.1186/1471-2164-14-499 (2013).

95 Kahn, R. S., Sommer, I. E., Murray, R. M., Meyer-Lindenberg, A., Weinberger, D. R., Cannon, T. D. *et al.* Schizophrenia. *Nat Rev Dis Primers* **1**, 15067, doi:10.1038/nrdp.2015.67 (2015).

96 National Human Genome Research Institute. The Cost of Sequencing a Human Genome. *https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost* (2020).

97 Jaffe, A. E., Straub, R. E., Shin, J. H., Tao, R., Gao, Y., Collado-Torres, L. *et al.* Developmental and genetic regulation of the human cortex transcriptome illuminate schizophrenia pathogenesis. *Nat Neurosci* **21**, 1117-1125, doi:10.1038/s41593-018-0197-y (2018).

98 Shi, J., Levinson, D. F., Duan, J., Sanders, A. R., Zheng, Y., Pe'er, I. *et al.* Common variants on chromosome 6p22.1 are associated with schizophrenia. *Nature* **460**, 753-757, doi:10.1038/nature08192 (2009).

99 Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alfoldi, J., Wang, Q. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434-443, doi:10.1038/s41586-020-2308-7 (2020).

100 Purcell, S. & Chang, C. PLINK [v2.0alpha2.3final]. www.cog-genomics.org/plink/2.0/.

101 Andrews, S. FastQC: a quality control tool for high throughput sequence data. *https://www.bioinformatics.babraham.ac.uk/projects/fastqc/* (2010).

102 Ewels, P., Magnusson, M., Lundin, S. & Kaller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047-3048, doi:10.1093/bioinformatics/btw354 (2016).

103 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120, doi:10.1093/bioinformatics/btu170 (2014).

104 Bushnell, B. BBMap: A Fast, Accurate, Splice-Aware Aligner. *https://www.osti.gov/biblio/1241166-bbmap-fast-accurate-splice-aware-aligner* (2014).

105 Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21, doi:10.1093/bioinformatics/bts635 (2013).

106 Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166-169, doi:10.1093/bioinformatics/btu638 (2015).

107 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550, doi:10.1186/s13059-014-0550-8 (2014).

108 Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140, doi:10.1093/bioinformatics/btp616 (2010).

109 Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**, 57-63, doi:10.1038/nrg2484 (2009).

110 Zhao, S., Fung-Leung, W. P., Bittner, A., Ngo, K. & Liu, X. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PLoS One* **9**, e78644, doi:10.1371/journal.pone.0078644 (2014).

111 Monsees, G. M., Tamimi, R. M. & Kraft, P. Genome-wide association scans for secondary traits using case-control samples. *Genet Epidemiol* **33**, 717-728, doi:10.1002/gepi.20424 (2009).

112 Lin, D. Y. & Zeng, D. Proper analysis of secondary phenotype data in case-control association studies. *Genet Epidemiol* **33**, 256-265, doi:10.1002/gepi.20377 (2009).

113 Chen, H. Y., Kittles, R. & Zhang, W. Bias correction to secondary trait analysis with case-control design. *Stat Med* **32**, 1494-1508, doi:10.1002/sim.5613 (2013).

114 Wang, J. & Shete, S. Estimation of odds ratios of genetic variants for the secondary phenotypes associated with primary diseases. *Genet Epidemiol* **35**, 190-200, doi:10.1002/gepi.20568 (2011).

115 Li, H., Gail, M., Berndt, S. & Chatterjee, N. Using Cases to Strengthen Inference on the Association between Single Nucleotide Polymorphisms and a Secondary Phenotype in

Genome-Wide Association Studies. *Genet Epidemiol* **34**, 427-433, doi:doi:10.1002/gepi.20495 (2010).

116 Lloyd-Jones, L. R., Holloway, A., McRae, A., Yang, J., Small, K., Zhao, J. *et al.* The Genetic Architecture of Gene Expression in Peripheral Blood. *Am J Hum Genet* **100**, 228-237, doi:10.1016/j.ajhg.2016.12.008 (2017).

117 Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat Protoc* **7**, 500-507, doi:10.1038/nprot.2011.457 (2012).

118 Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353-1358, doi:10.1093/bioinformatics/bts163 (2012).

119 Karolchik, D., Hinrichs, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* **32**, D493-496, doi:10.1093/nar/gkh103 (2004).

120 Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M. & Shendure, J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-315, doi:10.1038/ng.2892 (2014).

121 Richardson, D. B., Rzehak, P., Klenk, J. & Weiland, S. K. Analyses of case-control data for additional outcomes. *Epidemiology* **18**, 441-445, doi:10.1097/EDE.0b013e318060d25c (2007).

122 Jiang, Y., Scott, A. J. & Wild, C. J. Secondary analysis of case-control data. *Stat Med* **25**, 1323-1339, doi:10.1002/sim.2283 (2006).

123 Aleman, A., Kahn, R. S. & Selten, J. P. Sex differences in the risk of schizophrenia: evidence from meta-analysis. *Arch Gen Psychiatry* **60**, 565-571, doi:10.1001/archpsyc.60.6.565 (2003).

124 Dickson, H., Hedges, E. P., Ma, S. Y., Cullen, A. E., MacCabe, J. H., Kempton, M. J. *et al.* Academic achievement and schizophrenia: a systematic meta-analysis. *Psychol Med* **50**, 1949-1965, doi:10.1017/S0033291720002354 (2020).

125 Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-291, doi:10.1038/nature19057 (2016).

126 Tylee, D. S., Kawaguchi, D. M. & Glatt, S. J. On the outside, looking in: a review and evaluation of the comparability of blood and brain "-omes". *Am J Med Genet B Neuropsychiatr Genet* **162B**, 595-603, doi:10.1002/ajmg.b.32150 (2013).

127 Rollins, B., Martin, M. V., Morgan, L. & Vawter, M. P. Analysis of whole genome biomarker expression in blood and brain. *Am J Med Genet B Neuropsychiatr Genet* **153B**, 919-936, doi:10.1002/ajmg.b.31062 (2010).

128 Ripke, S., O'Dushlaine, C., Chambert, K., Moran, J. L., Kahler, A. K., Akterin, S. *et al.* Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat Genet* **45**, 1150-1159, doi:10.1038/ng.2742 (2013).

129 Eaton, W. W., Pedersen, M. G., Nielsen, P. R. & Mortensen, P. B. Autoimmune diseases, bipolar disorder, and non-affective psychosis. *Bipolar Disord* **12**, 638-646, doi:10.1111/j.1399-5618.2010.00853.x (2010).

130 Network and Pathway Analysis Subgroup of Psychiatric Genomics Consortium. Psychiatric genome-wide association study analyses implicate neuronal, immune and histone pathways. *Nat Neurosci* **18**, 199-209, doi:10.1038/nn.3922 (2015).

131 Goring, H. H., Curran, J. E., Johnson, M. P., Dyer, T. D., Charlesworth, J., Cole, S. A. *et al.* Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nat Genet* **39**, 1208-1216, doi:10.1038/ng2119 (2007).

132 Huang, Y. T., Vanderweele, T. J. & Lin, X. Joint Analysis of Snp and Gene Expression Data in Genetic Association Studies of Complex Diseases. *Ann Appl Stat* **8**, 352-376, doi:10.1214/13-AOAS690 (2014).

133 Valeri, L. & Vanderweele, T. J. Mediation analysis allowing for exposure-mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros. *Psychol Methods* **18**, 137-150, doi:10.1037/a0031034 (2013).