MULTI-OMIC DATA PROVIDE A MORE COMPLETE UNDERSTANDING OF THE AUTISTIC BRAIN

by

Shannon E. Ellis

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland

March, 2016

© 2016 Shannon E. Ellis
All Rights Reserved

ABSTRACT

Autism is a complex neurodevelopmental disorder characterized by persistent social deficits and restricted or repetitive patterns of behavior. Despite an established genetic basis of the disorder, efforts to elucidate the genetic underpinnings of the disorder and our understanding of its etiology remains incomplete. As such, we set out to study the effects downstream of genetic variation by studying alterations in both gene expression and DNA methylation (DNAm) in post-mortem brain samples collected from individuals affected with autism and controls. This work highlights that even when there is no primary genetic lesion detected, the autistic brain shows a characteristic pattern of upregulation at M2-activation state microglia genes, a state potentially driven by Type I interferon responses. Additionally, by combining transcriptomic data across autism and two related neuropsychiatric disorders, schizophrenia and bipolar disorder, we have garnered a better understanding of the relationship between these disorders, where genes differentially expressed in autism are concordantly differentially expressed in schizophrenia, but not in bipolar disorder. Finally, as gene expression is regulated, at least in part, by DNAm, we have characterized DNAm at cytosines across the genome and have detected hypermethylation at cytosines outside the commonly-studied CpG context, suggesting that autistic brains have slight increases at many CpH sites (where H=A,T, or C) throughout their genome. These sites are enriched in repetitive regions of the genome and regions containing human-specific CpGs, offering an insight into how this hypermethylation may be functioning mechanistically. Taken together, by studying the downstream effects of genetic variation, at the levels of DNAm and gene expression, we have moved toward a more complete understanding of the autistic brain.

Advisor: Dan E. Arking, Ph.D.

Reader: Jeff Leek, Ph.D.

PREFACE

I can only begin to express how thankful I am for the opportunity I have had to pursue my Ph.D. in the McKusick-Nathans Institute of Genetic Medicine. I will do my very best to express my gratitude here, but it will undoubtedly fall short of expressing just how thankful I truly am.

First and foremost, I want to thank my advisor, Dr. Dan Arking, for his unwavering support and incredible mentorship throughout graduate school. Having never coded before graduate school, I am thankful to Dan for taking a chance on me and spending the time to help make me a successful student. Dan has taught me how to generate, clean, and analyze data critically, creatively and thoroughly, for which I will be forever grateful. Additionally, Dan was always there to answer any questions I had and to help push me past each and every mental roadblock – whether that meant encouraging me to work through it on my own or supplying me with the tools to move past it. Last, I would be remiss not to mention Dan's ability to think critically and quickly, to act ethically, to manage effectively, and to present data clearly – all skills that I will try my very best to emulate throughout my own career.

Additionally, my thesis work was only accomplished thanks to contributions from a number of other people in Dan's lab. Dr. Simone Gupta, a former postdoc in the lab, was immeasurably helpful my first few years in the lab. Her patience when I joined the lab is worth its weight in gold. I cannot even begin to explain how much Simone taught me. On my first day she sat and explained what a server was to me, demonstrated how to transfer files, and taught me how to log on to a node without ever making me feel like I didn't know anything (which, in fact, I didn't). Since that was my level of computational understanding at the start of my PhD, I think it's fair to say that at least half of what I now know can be directly attributed to her teaching. Additionally, I was so lucky to gain Anna Moes, our longtime lab manager, as a friend and coworker when I joined the lab. Her impeccable organizational skills, attention to detail, and wonderful personality are rarely found in one individual. My work was also immensely influenced by Foram Ashar. She played a pivotal role in every single thing I accomplished in Dan's lab. From discussing problems I was struggling with to editing both code and everything I've ever written and everything in between, Foram was there to support, help, listen, and be the best labmate and friend for which I could have ever asked.

Finally, every member of the Arking lab during my time played a role in making me the scientist I am today and truly made coming into work an enjoyable experience day in and day out. Thank you for always being there to bounce ideas off of, to get ice cream when things got tough, and for being incredible scientists and even more amazing people.

Outside of the Arking lab, there are many people who have been essential along the way. First, I am incredibly thankful to each and every collaborator with whom I've had the pleasure of working. Specifically, the scientific input Andy West has had throughout all of my work is impossible to measure. Additionally, he has not only read every draft of every paper, offering helpful comments every single time, but also responded with speedy answers to every e-mail I've ever written to him, for which I am so thankful. Additionally, I hit the jackpot with the members of my thesis committee: Drs. Joel Bader, Dani Fallin, Jeff Leek, and Juan Troncoso. The reasons I am indebted to this group are manifold, but I'll highlight two. First, and foremost, each person offered a different skillset to jointly help me work through each and every issue I ran into along the way; however, secondly, they were all wonderful people to work with. Their criticism was always constructive and their feedback relevant. It was so nice to feel as though we were working as a team toward a common goal each time I scheduled a meeting. Finally, the human genetics program has provided me with a second family throughout graduate school. The supportive environment of the program is unique and indispensable. Thank you to Dr. Dave Valle, Dr. Kirby Smith, and Sandy Muscelli for the countless and tireless hours you all invest so that our program continues to thrive and so that us students always feel supported. And, thank you to all of the faculty members within the IGM who have played a role in my life these past six years.

As a graduate student, I was also fortunate to go through the Human Genetics program with the most amazing classmates. I never expected to become close friends with my classmates, but I am so glad my expectation was wrong. Working as a team made getting through our coursework together so much less painful. And, I am so thankful we had one another as we studied for our comprehensive exams. However, I am most thankful for the friendships we formed beyond class and lab. Memories of our trip to Florida, annual white water rafting trips, and seemingly never-ending debates will be memories that I cherish long past our time together at Hopkins.

I would also like to thank the mentors that helped me get to Hopkins. I struggle to find the words to explain what a huge impact Dr. Jeramia Ory, my undergraduate research mentor, has had on my career. He is a tough but incredible instructor, a supportive mentor, and a wonderful scientist. Not only did he offer me my first position in a research lab, but he helped ensure that I was given every opportunity going forward. He encouraged me to pursue summer research interests at other institutions, gave me opportunities to present at numerous conferences as an undergrad from a tiny liberal arts college, and made sure that I applied to top graduate programs. Through Jeramia's encouragement, I also had the great fortune of being mentored by Dr. Joseph Ayoob during a summer at Carnegie Mellon University. Joe is not only responsible for any pipetting chops I have (even if I may not use them very much these days) but is also the reason I applied to Hopkins in the first place, as he received his PhD from Hopkins and strongly encouraged that I apply. Joe is a wonderful mentor, a superb scientist, and a great friend. I certainly lucked out getting the chance to work with him and am thankful for all the support and advice he's given me since my summer at Carnegie Mellon.

Beyond science, I have no idea how to thank the amazing friends who helped me make it to Baltimore – a city that I now absolutely love – and those I have made since making the move. I feel so lucky to have friends that have known me since I was five years old, whom know me better than I know myself, whom know where I come from, and who completely understand how I became the person I am today. Additionally, college brought me friends that are so different from myself in so many ways but who share the same importance I place on family, who taught me how to truly enjoy life, and showed me just how big others' hearts can be, and for all of that I am so very grateful. Since moving to Baltimore, while Nicole and Foram started simply as classmates, they have become my best friends. Along with Tim Kelly and Courtney Woods, I am so excited to see where life takes each of you! And, of course, volleyball and softball have brought the most wonderful combination of intelligent and fun people into my life, and I cannot imagine my life without them.

Last, and of course not least, to the most amazing family. To my parents, Tricia Fleming and Anthony Ellis, and my sisters, Briana and MacKenzie Ellis, your support and love mean the world to me. Thank you for putting up with me and my crazy schedule, for always being there for me, and for loving me no matter what. And, thank you for each being so unique. I have been so fortunate to be part of this family and have loved watching each of you pursue your varied passions and have been inspired by the way in which you each approach life. Truthfully, the words do not exist in the English language to accurately convey how much you each mean to me, but thank you for everything.

TABLE OF CONTENTS

ABSTRACT ii

PREFACE iii

LIST OF TABLES 11

LIST OF FIGURES 13

CHAPTER 1: Introduction 1

- 1.1 The Genetic basis of autism 1
- 1.2 Gene expression in autism 2
- 1.3 Methylation in autism 3
- 1.4 The benefits of a multi-omics approach 4

CHAPTER 2: RNA-Seq Optimization with eQTL Gold Standards 5

- 2.1 Introduction 5
- 2.2 Methods 7
  - 2.2.1 Sample information 7
  - 2.2.2 Genotyping 7
  - 2.2.3 RNA-Sequencing 7
  - 2.2.4 Mapping 8
  - 2.2.5 Normalization 9
  - 2.2.6 Data decomposition 9
  - 2.2.7 Variant calling 9
  - 2.2.8 Simulated sample mixing experiment 10
  - 2.2.9 Assembling lists of previously identified eQTLs 10
  - 2.2.10 Covariate inclusion in eQTL analyses 11
  - 2.2.11 Detecting inflation in each data set 11
  - 2.2.12 Replication of previously identified eQTLs 11
  - 2.2.13 Differential gene expression analyses 12
  - 2.2.14 Availability of supporting data 12
- 2.3 Results 12
  - 2.3.1 Data normalization in RNA-Seq 12
  - 2.3.2 Identifying outliers in RNA-Seq data 17
  - 2.3.3 Reported brain eQTLs are reproducible in RNA-Seq data 21

2.3.4 Monitoring eQTL replication to gauge quality control measures....24
2.3.5 Independent RNA-Seq data set supports use of eQTL gold standards....26
2.4 Discussion....29
2.5 Conclusion....31
2.6 Tables....33
CHAPTER 3: Transcriptome Analysis Reveals Deregulation of Innate Immune Response Genes and Neuronal Activity-Dependent Genes in Autism....39
3.1 Introduction....39
3.2 Methods....40
3.2.1 Brain tissue samples....40
3.2.2 RNA library preparation and RNA sequencing....40
3.2.3 Mapping and gene summarization of data from RNA-Seq....41
3.2.4 Normalization of gene estimates....41
3.2.5 Quality assessment....41
3.2.6 Single gene differential expression analysis....42
3.2.7 Single exon differential expression analysis....45
3.2.8 Single gene - Gene Ontology and pathway enrichment analysis....45
3.2.9 Artifact corrected dataset....46
3.2.10 Combined co-expression analysis....47
3.2.11 Stratified co-expression analysis....48
3.2.12 Gene Ontology analysis....48
3.2.13 Gene list enrichment analysis....48
3.2.14 Compilation of genetic association genes....49
3.2.15 Other gene list compilations....50
3.3 Results....50
3.3.1 Sample Summary....50
3.3.2 Differential Gene Expression Analysis....50
3.3.3 Pathway and Functional Enrichment Analyses....54
3.4 Conclusions....60
3.5 Tables....61
CHAPTER 4: Transcriptome Analysis of Cortical Tissue Reveals Shared Sets of Down-Regulated Genes in Autism and Schizophrenia....97

4.1 Introduction 97
4.2 Methods...97
4.2.1 Autism sample information 97
4.2.2 Schizophrenia and bipolar disorder sample information 98
4.2.3 RNA-Seq, alignment & quality control 98
4.2.4 Gene expression estimation and normalization 98
4.2.5 Differential gene expression analysis (DGEA) 99
4.2.6 Null DGEA 99
4.2.7 Calculating genes differentially expressed across disorders 100
4.2.8 Calculating the correlation of DCEGs across phenotypes 100
4.2.9 Pathway analysis of DCEGs 100
4.2.10 Enrichment for genetic signal analysis 101
4.3 Results..102
4.3.1 Sample summary 102
4.3.2 Genes differentially expressed across SCZ, BPD, and AUT 102
4.3.3 Correlation in gene expression across SCZ, BPD, and AUT 106
4.3.4 Pathway enrichment analyses of genes differentially expressed across disorders 110
4.3.5 Cross-disorder DEGs enrichment in association signals 112
4.4. Conclusions 117
4.4.1 Results summary...117
4.4.2 Accounting for unknown covariates is critical in transcriptome analyses 117
4.4.3 Correlations in differential gene expression across disorders highlights similarities between AUT and SCZ...118
4.4.4 Differences in genetic variation not explained by overlapping gene expression profiles 119
4.5 Tables 121
CHAPTER 5: Exaggerated CpH Methylation in the Autism-Affected Brain 130
5.1 Introduction 130
5.2 Methods...131
5.2.1 Samples...131
5.2.2 RRBS library preparation...131
5.2.3 Alignment...132

5.2.4 Methylation estimation 132
5.2.5 Illumina 27K methylation array 132
5.2.6 Sample outlier removal....................133
5.2.7 Single site differential methylation analysis 134
5.2.8 Global altered methylation analysis 135
5.2.9 Lists of functional genomic categories 135
5.2.10 Functional enrichment testing....................136
5.2.11 Power calculation 136
5.3 Results....................136
5.4 Conclusions....................145
5.5 Tables....................146
REFERENCES....................149
APPENDIX 1: Using the human brain to understand “omics” in autism 166
APPENDIX 2: Analyses of gene activity can yield clues to roots of autism 169
CURRICULUM VITAE 172

LIST OF TABLES

CHAPTER 2:

Table 2.1: Sample Information. 33
Table 2.2: Summary of eQTL replication analyses carried out in brain samples. 33
Table 2.3: Summary of eQTL replication carried out in GTEx blood samples. 37
Table 2.4: Technical Artifacts are correlated with Independent Surrogate Variables. 37

CHAPTER 3:

Table 3.1: Sample information 61
Table 3.2: Gene Lists 65
Table 3.3: Enrichment analysis of 1964, 749, 494 , 185 genes differentially expressed at P < 0.05, P < 0.01, P < 0.005, P < 0.001 , respectively. 67
Table 3.4: Gene Set enrichment analysis to evaluate expression at the level of genetically associated genes 68
Table 3.5: Correlation (r, spearman correlation) of the ISV with sequencing artifacts. 69
Table 3.6: Correlation (r, spearman correlation) between the sequencing artifacts metrics. 70
Table 3.7: Correlation between the 12 module eigengenes (ME) in the network 71
Table 3.8: Association of co-expression modules with disease status 72
Table 3.9: mod1 and enrichment of Gene Ontology categories with FWER < 0.05 73
Table 3.10: mod2 and enrichment of Gene Ontology categories with FWER < 0.05. 74
Table 3.11: mod3 and enrichment of Gene Ontology categories with FWER < 0.05. 76
Table 3.12: mod4 and enrichment of Gene Ontology categories with FWER < 0.05. 77
Table 3.13: mod5 and enrichment of Gene Ontology categories with FWER < 0.05. 80
Table 3.14: mod6 and enrichment of Gene Ontology categories with FWER < 0.05. 82
Table 3.15: mod7 and enrichment of Gene Ontology categories with FWER < 0.05. 83
Table 3.16: mod8 and enrichment of Gene Ontology categories with FWER < 0.05. 85
Table 3.17: mod9 and enrichment of Gene Ontology categories with FWER < 0.05. 85
Table 3.18: mod10 and enrichment of Gene Ontology categories with FWER < 0.05. 86
Table 3.19: mod11 and enrichment of Gene Ontology categories with FWER < 0.05. 87
Table 3.20: Modules and the enrichment of gene sets 89
Table 3.21: Enrichment analysis of 749 genes differentially expressed at P < 0.01. 92
Table 3.22: Enrichment analysis of 1964 genes differentially expressed at P < 0.05. 93
Table 3.23: Enrichment analysis of 185 genes differentially expressed at P < 0.001. 94
Table 3.24: Enrichment analysis of 50 genes differentially expressed at P < 0.0001. 95
Table 3.25: Association of co-expression modules with disease status 96

CHAPTER 4:

Table 4.1: Autism sample information 121
Table 4.2: Schizophrenia and bipolar disorder sample information 123
Table 4.3: Sample Summary 126

Table 4.4: DAVID Pathway Analysis for cross-disorder DEGs 126
Table 4.5: DAVID Pathway Analysis for cross-disorder DEGs controlling for # of genes input 127
Table 4.6: Association Signal Enrichment Results across multiple Z-score cutoffs. 128
Table 4.7: Genes significantly differentially expressed across disorders 129

CHAPTER 5:

Table 5.1 Sample Information 146
Table 5.2 Top 10 differentially methylated CpG sites 148
Table 5.3 Top 10 differentially methylated CpH sites 148

LIST OF FIGURES

CHAPTER 2:

Figure 2.1 Data analysis pipeline for analysis of RNA-Seq data. 6
Figure 2.2: Comparing normalization methods.....14
Figure 2.3: Sample outliers identified and removed on a gene-by-gene basis improves robustness of differential gene expression analyses. 15
Figure 2.4: eQTL replication in brain data.....16
Figure 2.5: PCA identifies sample outliers.....19
Figure 2.6: Using discordance ratio (DR) to assess quality of RNA.Seq data 20
Figure 2.7: Simulated contamination of RNA-Seq reads.....22
Figure 2.8: Replication of previously-reported eQTLs.....23
Figure 2.9: QQ Plots for Data Decomposition Methods.....27
Figure 2.10: eQTL replication in blood data. 28

CHAPTER 3:

Figure 3.1: Transcriptome-wide top hits. 44
Figure 3.2: Single-gene expression analysis identifies two transcriptome-wide significantly differentially expressed genes between autism brains and control brains. 52
Figure 3.3: Weighted gene correlation network analysis (WGCNA) identifies 12 co-expression modules.....53
Figure 3.4: Gene co-expression module mod5 is associated with autism. 55
Figure 3.5: Visualization of the (a) mod5, (b) mod1, (c) mod2, and (d) mod6 modules.....58
Figure 3.6: ‘Synaptic Transmission’ GO genes in mod1, mod2 and mod6.....59

CHAPTER 4:

Figure 4.1: Differential Gene Expression Across AUT, SCZ, and BPD. A density plot for the cross three-disorder Z-scores ($Z_{AUT}*Z_{SCZ}*Z_{BPD}$) are plotted in black with the cutoff for transcriptome-wide significance highlighted in red ($p<0.05$, determined empirically by permutation). The two genes that meet transcriptome wide significance are labeled. 104
Figure 4.2: Correlation of Genes Differentially Expressed Across All Three Disorders. 105
Figure 4.3: Correlation of Cross-Disorder Differential Gene Expression.....107
Figure 4.4: Assessing the significance for Correlations of Cross-Disorder Transcriptomic Similarity. 108
Figure 4.5: Accounting for Unknown Covariates Affects Correlation. 109
Figure 4.6: GO Analysis of cross-disorder DEGsAUT-SCZ. 111
Figure 4.7: Enrichment of DEGs among GWAS signal. 113
Figure 4.8: Enrichment of DEGs among GWAS at a more permissive p-value cutoff ($p<0.1$). 114

Figure 4.9: Enrichment of DEGs among GWAS signal at a more stringent p-value cutoff (p<0.01). 115
Figure 4.10: Enrichment of DEGs among all genes (no gene based GWAS p-value cutoff imposed). 116

CHAPTER 5:

Figure 5.1 Correlation between mean methylation measured by RRBS and array. 133
Figure 5.2 Previously reported meQTLs after sample outlier removal. 134
Figure 5.3 Summary of CpG and CpH sites. 138
Figure 5.4 Single site differential methylation analysis. 139
Figure 5.5 Proportion of hyper- and hypo- methylated sites in the CpG and CpH contexts.140
Figure 5.6 Null distributions of hypermethylation in CpG and CpH contexts. 141
Figure 5.7 Functional genomic enrichment of hypermethylated CpH sites. 143
Figure 5.8 Power calculation. 144

CHAPTER 1: Introduction

1.1 The Genetic basis of autism

Autism is a complex neurodevelopmental disorder that develops in early childhood, continues throughout one's life, and is characterized by persistent social deficits and restricted or repetitive patterns of behavior[1]. In addition to this characteristic set of core symptoms, affected individuals often also display a subset of associated neurological and physical symptoms including, but not limited to, alterations in mood, language ability, sleep habits, and gastrointestinal function[2,3]. Taken together, affected individuals display an incredibly heterogeneous phenotype with approximately one in 68 individuals in the United States receiving an autism diagnosis[4]. However, despite its incredibly heterogeneous nature, the genetic basis of the disorder has been definitively established. Historically, family and twin studies established the heritable nature of the disorder[5,6] with more recent large cohort studies definitively establishing that more than half of the liability in autism can be attributed to genetic causes [7].

Specifically, the past decade has made a number of seminal discoveries about the genetic architecture of autism. SNP genotyping arrays have established an excess of inherited and *de novo* CNVs in cases relative to controls[8–10]. Further, exome sequencing analysis of simplex families has identified a number of rare *de novo* mutations found more frequently in affected individuals relative to controls and have estimated that there are 1,000 genes involved in autism[11–16]. On the other hand, genome wide association studies (GWAS) have yet to identify replicable common variants associated with autism[17]. However, it is important to keep in mind that these analyses remain underpowered. Given the recent findings in schizophrenia[18], in which 108 disease-associated loci were identified once suitable sample sizes were obtained to detect the small effect sizes of variants playing a role in schizophrenia, it is reasonable to infer that autism, a related disorder with overlapping genetic hits[19], will follow a similar path. Thus, common variants robustly associated with the autism will likely be identified as suitable sample sizes are obtained. Nevertheless, while yet incomplete, research has begun to identify genetic variation that plays a role in autism.

1.2 Gene expression in autism

As we have begun to understand genetic variation, it has become increasingly clear that DNA mutations are only the beginning. Understanding the downstream effects of genetic mutations is another important piece of the puzzle. Thus, as we've begun to catalog mutations at the level of DNA, we've simultaneously started to get a handle on what is altered at the level of gene expression.

This is particularly important because while the individual genes affected in autism may be numerous, there is a growing body of evidence supporting the fact that these alterations converge on a limited number of biological pathways including alterations in cortical development, synapse function, transcription and translation, chromatin modification, and microglial activation[20–25]. Ultimately, these findings suggest that, from a treatment standpoint, despite variable genetic causes, there may be a limited number of pathways that need to be targeted to begin to treat affected individuals[24]. And, thus, by understanding gene expression data as a whole we can obtain a more complete understanding of autism and how to approach treatment moving forward.

Early attempts to study the autistic transcriptome focused on assessing gene expression in lymphoblastoid cell lines or whole blood[26–30]. However, given the core neurodevelopment phenotypes associated with autism, there is little doubt that direct assessment of gene expression in brains is critical. Initial work to study gene expression within the primary affected tissue in autism, the brain, utilized gene expression microarrays to determine which genes and pathways are altered in affected individuals [20]. This work began to provide insight into the autistic brain as it identified two modules – or groups of co-expressed genes – that demonstrate altered gene expression profiles in the brains of affected individuals. Their findings suggest alterations in neuronal, immune genes, and glial markers are driving the differences detected. With these findings in hand, we carried out RNA-Sequencing, which enables digital transcriptome profiling at unprecedented resolution, in 104 post-mortem brains samples obtained across three cortical brain regions (BA10, BA19, and BA44/45) from individuals with autism and age-

and sex- matched controls to identify individual genes and biological pathways that are differentially expressed in the brains of individuals with autism.

In generating these data, it became increasingly clear that there was no well-defined path to move from raw data obtained from the prepared sequencing libraries to biological insights. While a few analysis pipelines for RNA-Sequencing data had been published, the uniqueness of each experiment did not lend itself to a single pipeline being adequately applicable across experiments. Therefore, to address this, we developed a method to utilize known relationships between an individual's genotype and his or her expression levels at a nearby gene – known as eQTLs, or expression quantitative trait loci – as gold standards to help make decisions during data analysis. This method assists analysts as they make decisions during analytical steps involving quality control, expression estimation, and statistical modeling of the data to ultimately provide confidence in the analyses carried out. This method was used to help identify transcriptomic differences that exist in the autistic brain relative to controls.

1.3 Methylation in autism

In addition to understanding gene expression alterations that exist in the autistic brain, understanding DNA methylation (DNAm), which directly regulates gene expression, is important for a number of reasons. First, prenatal life is a critical time for both DNAm and brain development[31], making DNAm an epigenetic mark of interest in autism. Further, imprinting errors, which are mechanistically driven by alterations in DNA methylation patterns, have been sufficient to cause neurodevelopmental problems, such as in the case of Angelman Syndrome [32]. Additionally, mutations in epigenetic effectors can result in human neurodevelopmental disorders, such as is the case in Fragile X Syndrome and Rett Syndrome[33,34]. Taken together, these data support investigating the role of altered DNAm in autism.

As such, researchers set out to investigate the role for alterations in CpG DNAm in brains from individuals with autism spectrum disorder (ASD), reporting a handful of regions demonstrating altered methylation in individuals with ASD relative to controls. These initial studies, however, were carried out on a limited number of samples and utilized methylation array technology at a limited number of sites [35,36]. Given the limited sample size and number of cytosines tested, we decided to carry out reduced representation bisulfite sequencing (RRBS) on the same post-mortem cortical brain samples from which we had previously acquired transcriptome data. RRBS data enabled us to characterize DNAm alterations occurring at cytosine-rich regions throughout the genome within both the classically-studied CpG context as well as the CpH (H=A, C, or T) context.

Differences in methylation levels between affected individuals and controls could then be detected across the genome, at each single site, and within regions of the genome. Portions of the genome demonstrating altered methylation levels within the autistic brain were then assigned to functional regions of the genome to suggest how altered methylation may be functioning within affected individuals.

1.4 The benefits of a multi-omics approach

The advantages to obtaining multiple levels of data from the same individual samples are manifold. First, sample acquisition is subject to human error with sample mix up an unavoidable risk. By obtaining genetic and expression information from the same individuals we were able to ensure the samples were derived from the same individuals across the various levels of data, thus ensuring our confidence in all results from subsequent analyses.

Further, while each level of data (genotyping, gene expression, and methylation) is interesting on its own, there is also a relationship between each level of data collected, enhancing the utility and information that can be garnered to ultimately move us toward a more complete understanding of the autistic brain. Specifically, genotype not only affects expression levels and methylation at expression quantitative trait loci (eQTLs) and methylation quantitative trait loci (meQTLs), respectively, but there is also a well-established relationship between increased methylation at a gene's promoter and decreased expression at that gene. Given these established relationships, more information can be gained by combining information across all analyses carried out to ultimately better understand genes and pathways playing a role in the autistic brain.

CHAPTER 2: RNA-Seq Optimization with eQTL Gold Standards

2.1 Introduction

The advent of RNA-Seq[37] and dramatic decrease in next-generation sequencing costs have led to numerous RNA-Seq studies in recent years. This revolutionary technique has enabled digital transcriptome profiling at unprecedented resolution that avoids many of the limitations inherent to the analog nature of microarray technolog[38,39]. However, despite numerous publications and the fact that RNA-Seq studies have supplanted microarrays as the gold standard for transcriptome analysis, it is not without its own inherent limitations.

Early concerns regarding library preparation, sequencing error, read mapping, and gene expression quantification have been resolved by a number of studies; however, there is no standardized approach for quality control and data adjustment of RNA-Seq data after the generation of gene expression estimates. Without an appropriate approach to data analysis, reproducibility of these studies remains limited[40]. Further, the unique designs of sequencing studies suggest that a single black box approach is unlikely to be uniformly optimal across all experiments. Thus, we propose an approach to address data cleaning, normalization, and adjustment in RNA-Seq data analysis (**Figure 2.1**). This pipeline is informed by best practices that we and others have developed for genome-wide association studies (GWAS) [41,42], which also suffered from similar sources of error prior to the development of optimized methods.

We demonstrate the applicability of our approach in 64 autism-affected and control brain samples. Specifically, our outlier detection method is based on utilizing the RNA-Seq gene expression estimates as well as DNA and RNA genotypes obtained from the same individual. Further, expression quantitative trait loci (eQTLs) are biologically meaningful loci at which gene expression is modified by genotype. Accordingly, we utilize replication of *cis*-eQTL data from two recently published brain studies[43,44] as a means to assess the integrity of sequencing data and appropriateness of data handling procedures. We replicate the findings from this eQTL analysis in an independently-generated RNA-Seq data set of 162 blood samples from the Genotype-Tissue Expression (GTEx) project[45]. Within the context of eQTL replication, we particularly highlight the need to identify and remove outlier samples in RNA-

Seq experiments and further corroborate the need to account for unknown sources of variation in high-throughput data[46]. While a number of publications have presented methods by which one can analyze RNA-Seq data (many of which are reviewed in[47]) and account for unknown covariates[48–51], the steps we present herein ultimately provide a straightforward approach that allows for more accurate approximation of gene expression values that can be confidently used in downstream disease-based comparisons.

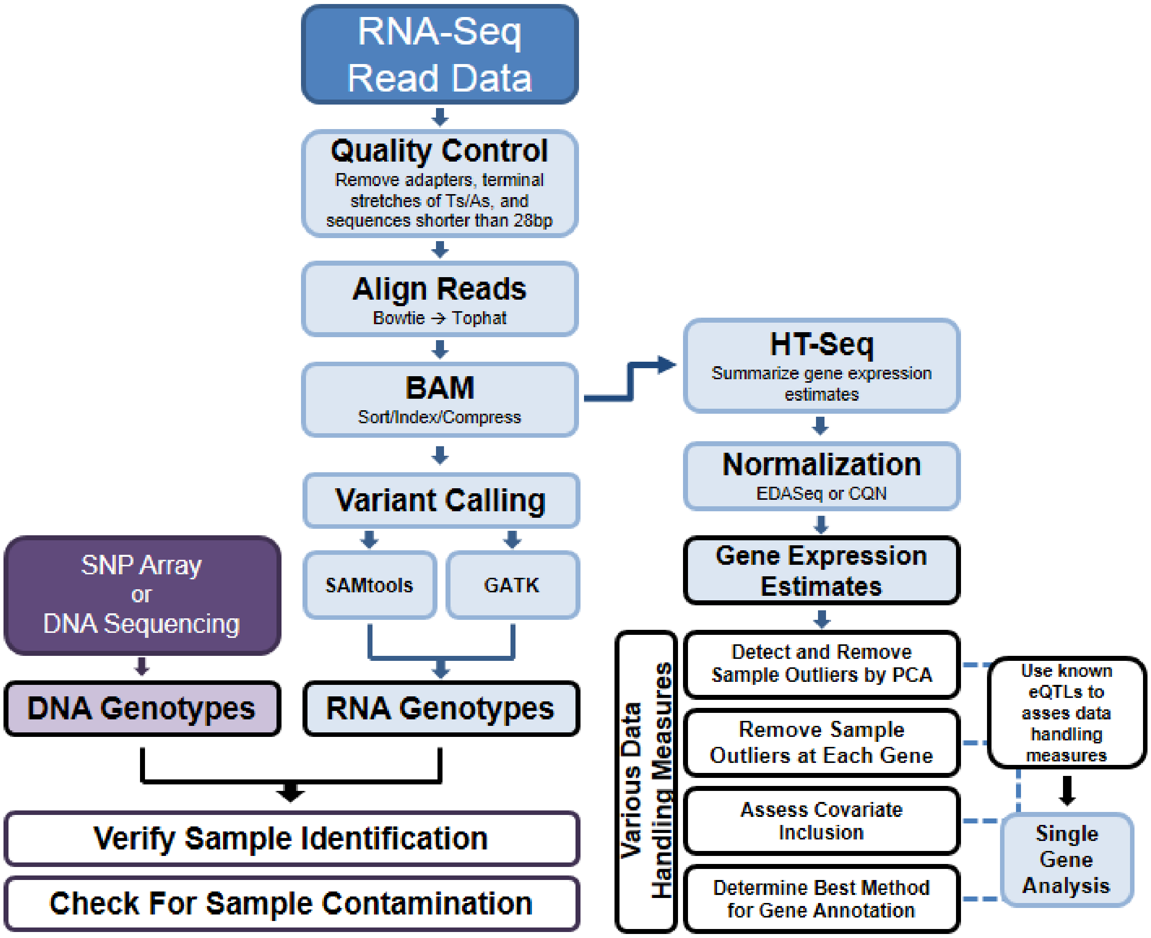

RNA-Seq Read Data
Quality Control
Remove adapters, terminal stretches of Ts/As, and sequences shorter than 28bp
Align Reads
Bowtie → Tophat
BAM
Sort/Index/Compress
Variant Calling
SAMtools
GATK
SNP Array or DNA Sequencing
DNA Genotypes
RNA Genotypes
Verify Sample Identification
Check For Sample Contamination
HT-Seq
Summarize gene expression estimates
Normalization
EDASeq or CQN
Gene Expression Estimates
Various Data Handling Measures
Detect and Remove Sample Outliers by PCA
Remove Sample Outliers at Each Gene
Assess Covariate Inclusion
Determine Best Method for Gene Annotation
Use known eQTLs to asses data handling measures
Single Gene Analysis

Figure 2.1 Data analysis pipeline for analysis of RNA-Seq data.
Blue boxes are data analyses carried out on RNA. Purple indicates DNA.

2.2 Methods

2.2.1 Sample information

Brain. Post-mortem brain samples were acquired through the Autism Tissue Program (http://www.atpportal.org), with samples originating from two different sites: the Harvard Brain Tissue Resource center and the NICHD Brain and Tissue Bank at the University of Maryland. Cortical tissue corresponding to Brodmann Area 19 (BA19) was sequenced in 40 controls and 25 autism-affected cases. Among this set of brains, the average age at time of death is similar between cases and controls (22.2 and 21.3 years, respectively), and there is no significant difference in cause of death between the two groups. One sample had fewer than 20,000 sequenced reads (average across all other samples was 109M reads) and was excluded. The resultant 64 samples were included for study. This study was approved by the IRB of The Johns Hopkins Hospital and conducted in accordance with institutional guidelines.

Blood. Sample data were acquired from the NHGRI GTEx project (phs000424.v3.p1)[45]. Whole blood RNA-Seq and genotyping data were available for 162 samples. This data set comprised of 103 males and 59 females with an average age of 49.7 years.

2.2.2 Genotyping

Brain. Each sample was genotyped at ~900,000 SNPs using the Affymetrix 6.0 array calling genotypes using the Birdsuite software package[52]. High quality genotyping was completed for all samples with an average call rate of 99.63% [range: 97.91 to 99.91].

Blood. The GTEx project used the Illumina Omni5 array for direct sample genotyping and subsequently imputed with IMPUTE2[53] using the 1000 Genome phase 1 release reference panel.

2.2.3 RNA-Sequencing

Brain. RNA-Seq libraries were prepared from 50 µg of total RNA from postmortem brain obtaining a fraction of purified polyadenylated (polyA) mRNA after two rounds of hybridization with oligo(dT) dynabeads. Standard quality control measures were employed using “no template controls”, “no ligase controls”, and “no adapter controls” in RNA-Seq library preparation. These samples did not demonstrate detectable product by PCR prior to sequencing. This process was followed by random fragmentation to avoid bias at the 3’ end of the transcript. First-strand cDNA synthesis was performed using random primers (Illumina) and SuperScriptII Reverse-Transcriptase (Invitrogen) followed by second strand cDNA synthesis using RNaseH and DNA Pol I (Illumina). Illumina adaptors were ligated to the purified, end-repaired and 3' adenylated cDNA and performed 200 bp size-selection of the final product by gel-excision, following the Illumina-recommended protocol. The 200 bp cDNA template molecules were amplified with the adaptor attached by PCR to create the final library. Each library was sequenced on a single lane of the Illumina's Hiseq 2000 to produce 100 base pair (bp) single-end reads.

Blood. RNA-Seq read count data was obtained from the GTEx project, which used a TruSeq library preparation protocol on poly-A selected mRNA to obtain 72 base paired-end sequencing from the Illumina Hiseq 2000.

2.2.4 Mapping

Brain. The number of total reads per lane varied from 26M to 202M, with a mean of 109M. We used in-house Python scripts to map the sequence reads to the genome (hg19) using Bowtie[54] followed by TopHat[55]. To improve mapping, reads were trimmed to remove stretches of terminal A’s or T’s (N=3-12) that occurred as a result of the polyA pulldown step. In addition, we removed contaminating adaptor sequences using a Python script, cutadapt (v0.09). Only uniquely mapped reads with a maximum of 3 mismatches were used to calculate gene expression values. Aligned reads were sorted, indexed and compressed into the BAM format for easy storage and usage in downstream analysis. The number of total mapped reads per lane varied from 2.7M to 84.2M, with a mean of 35M for the 57 samples used in the final analysis. The RNA-Seq reads were mapped to approximately 44,611 Ensembl genes (average 70% reads mapping per sample). For all analyses (save the case where we analyzed CDS only), we summarized these reads to all exons of genes based on the coordinates on the hg19/GRCh37 gene annotations provided from Ensembl using the python script HTSeq-count (intersection strict). For the CDS only analysis, HTSeq-count (intersection strict) was again used; however, we excluded reads that mapped to coordinates within the 5’ and 3’ UTRs for summarization. In both cases, regardless of quantification method, we then assessed summarized values on a gene-by-gene basis, removing samples whose gene expression values were more than three SD from the mean expression at each gene. After sample outlier removal, the final gene expression data set was pared down to include the

20,717 genes whose log2 gene expression estimates summed across all 57 samples totaled at least 100.

Blood. Mapping was carried out by the GTEx consortium[45]. Our data analysis of these data began with the mapped read count values.

2.2.5 Normalization

Subsequent to mapping, the gene count data was normalized to minimize biases due to gene-length, GC content, and sequencing depths. CQN normalization procedure was carried out with the recommended default setting. EDASeq normalization was completed using the full-quantile, within-lane GC-content normalization procedure as recommended[56].

2.2.6 Data decomposition

Data decomposition was performed on the log2 scale for those genes with at least ten gene-level counts across all samples. PCA was performed using the procedure implemented in the R function '*prcomp'.* SVA was performed on the matrix of the expression counts, after controlling for case-control status, age, sex and site using the 'sva' function implemented in the R package '*sva*'. ISVs were generated while protecting for case-control status using 'isvaFn' function in the '*isva'* package in R. We applied the unsupervised Bayesian factor analysis method implemented in Probabilistic estimation of expression residuals (PEER) on the count gene expression data[49,57]. PEER yields residual expression factors that can be used in for downstream analysis.

2.2.7 Variant calling

Variant Calling was completed using two different genotyping methods: SAMtools v0.1.12[58] and the Genome Analysis Toolkit v1.0 (GATK)[59]. SAMtools genotype calls were made for each sample individually using the recommended settings (http://samtools.sourceforge.net/mpileup.shtml); however, we excluded indels from these analyses and filtering was done in-house. Multi-sample GATK calls were made according to the suggested Unified Genotyper generic command line (http://www.broadinstitute.org/gatk/). The default settings were used except in the cases of standard minimum Phred-scaled confidence, which was increased to 60 to increase output of confident calls, and downsampling coverage, which was set to 250. We extracted genotypes from each method from the output files and assigned rsIDs (dbSNP build 132) using in-house scripts, keeping genotypes for which there were greater than twenty reads in downstream analyses. Genotypes both concordant across the two variant calling methods and present on the Affymetrix Genome-Wide Array 6.0 were used for downstream analyses.

2.2.8 Simulated sample mixing experiment

SNPs present on the Affymetrix Genome-Wide Array 6.0 that were called concordantly by both SAMtools and GATK were used in these analyses. The genome function in PLINK (v1.07) [60] was used for pair-wise comparisons to verify that, based on pair-wise IBS distance values (DST), the closest sample match for each RNA sample came from its corresponding DNA sample. Samples with low pair-wise concordance (IBS DST <0.89) were assessed further, computing each sample's Discordance Ratio (DR). A sample's DR can be calculated by taking the number of SNPs called homozygous at the DNA level but heterozygous at the RNA level divided by the total number of heterozygous RNA calls. Utilizing this metric, we simulated contamination at the RNA-Seq level by choosing eighteen high-confidence BAM files (DR<0.1) at random. The Picard (http://picard.sourceforge.net, v1.64) command DownsampleSam was then used to randomly sample a subset or reads from these BAM files. We combined RNA-Seq reads from these 18 samples in controlled ratios [10:90, 20:80, 30:70, 40:60, and 50:50] using samtools' [58] merge command. After controlled mixing of sequencing reads, we carried out variant calling and comparison back to DNA genotypes on these mixed samples as described above. The DR for each intentionally contaminated sample was calculated and compared the three samples in question to our intentionally contaminated subset to determine the level of sample contamination present in the three samples in question.

2.2.9 Assembling lists of previously identified eQTLs

Brain. We manually curated a list of brain SNPs and their associated genes from two recent publications[43,44]. These lists were generated from Table S6[44] and Tables S4 &S6[43] in the previous publications and included SNPs from the previous publications that had a proxy SNP on the Affymetrix Array 6.0 (r^2>0.90) as determined in SNAP with 1000G CEU as a reference population (http://www.broadinstitute.org/mpg/snap). Additionally, we retained eQTLs whose associated SNPs passed default filtering in PLINK, thus keeping SNPs with <10% missing and SNPs with a minor allele frequency > 0.01. We removed eQTLs whose associated genes were not present in our RNA-Seq data as well as duplicate SNP:gene pairs across the studies (defined as SNP:gene pairs with SNPs w/ r^2>0.8). Combining the lists from the two publications and performing the aforementioned filtering, resulted in a list of 909 eQTLs for study.

Blood. To test for known eQTLs in blood, we generated a list of 538 *cis*-eQTLs initially identified from a lymphoblastoid cell line[61]. From this data set we started with those *cis*-eQTLs with a q-value<0.01 in the previously published meta-analysis. Known eQTLs for which the genotyped SNP was present in the imputed GTEx genotype data and the gene was present in the GTEx RNA-Seq expression data were included. This resulted in 538 eQTLs for study.

2.2.10 Covariate inclusion in eQTL analyses

Covariates included in each analysis varied but included a subset or combination of known, unknown, and technical artifacts. The known covariates included were age, sex, and either sample collection site (Harvard or Maryland) in the brain data set or cohort (organ donor, postmortem, or surgical) in the blood data set. We utilized four data decomposition methods – independent surrogate variable analysis (ISVA), surrogate variable analysis (SVA), principal component analysis (PCA) and PEER[49,57] – to account for unknown covariates. We included percent coding bases, percent intronic bases, percent mRNA bases, median 3' bias, percent UTR bases, and AT dropout as the technical sequencing artifacts in our analyses. [See Picard documentation for further explanation of these artifacts, http://picard.sourceforge.net.]

2.2.11 Detecting inflation in each data set

To assess inflation of p-values, a genome-wide *cis*-eQTL analysis was carried out for each condition in the R package 'MatrixEQTL' (v1.6.1)[62]. eQTLs were detected by looking for *cis*-associations among all directly-genotyped SNPs and genome-wide RNA-Seq gene expression data. *cis*-associations were defined as SNP-gene associations in which the tested SNP was localized within 1Mb of either the 5' or the 3' end of the gene. From the p-value distribution of these analyses, the genome-wide inflation factor in each data (**Tables 2.2 & 2.3**) set was determined using the R package 'GenABEL'[63].

2.2.12 Replication of previously identified eQTLs

We utilized the curated list of 909 brain eQTLs and 538 LCL eQTLs to detect eQTLs in our brain and blood data sets, respectively. MatrixEQTL (v1.6.1)[62] was used to test for *cis*-

associations between the previously-reported SNP genotypes (or proxy SNPs) and corresponding gene expression estimates from the RNA-Seq data. *cis*-associations were defined as above. In each analysis, p-values were adjusted for inflation[64] using the inflation factor estimated from the genome-wide *cis*-eQTL analysis (**Table 2.2 and 2.3**). P-values from this analysis were used to obtain q-values using the R package 'qvalue'[65] keeping lambda constant at 0.50. Finally, as used previously[66], in order to assess eQTL replication, the π_1 statistic was calculated from the inflation-adjusted p-values using the 'qvalue' package. π_1, an estimate of the proportion of replicating eQTLs, is defined as $1-\pi_0$, where π_0 is the proportion of true null associations. These three statistics (p-value, q-value, and π_1) were used to assess the need for and success of each quality control step.

2.2.13 Differential gene expression analyses

A linear regression framework was utilized to identify differential gene expression between 36 controls and 21 cases with site, age, sex and ISVs as covariates.

2.2.14 Availability of supporting data

Genotyping and RNA-Sequencing data have been submitted to the NIH's National Database for Autism Research (NDARCOL0002034). Additional scripts developed for these analyses are available on the Arking lab website (www.arkinglab.org).

2.3 Results

2.3.1 Data normalization in RNA-Seq

Brain RNA-Seq data were generated from post-mortem cortical samples collected from Brodmann Area 19 (BA19) in 39 control and 25 autism-affected cases (**Table 2.1**). After estimating gene expression from the sequencing reads, two methods for data normalization were assessed: Exploratory Data Analysis and Normalization for RNA-Seq (EDASeq) [56] and Conditional Quantile Normalization (CQN) [67]. The normalized gene expression values from each algorithm demonstrated method-specific biases. Examining p-values from our covariate adjusted case-control analysis, we note that normalization by CQN leads to a marked increase in the test statistics for shorter and low GC content genes(gene length<1000 bp, GC content < 35%), a problem not observed with EDASeq (**Figure 2.2)**. On the other hand, genes with both lower gene expression estimates and the assignment of zero values by EDASeq led to an increase in outliers on a per-gene basis in our eQTL analyses (**Figure 2.3A**), whereas

CQN did a better job handling these genes (**Figure 2.3B)**. Further comparison by eQTL replication to assess the biologic reproducibility (discussed below) of these two normalization methods was performed with CQN slightly outperforming EDASeq (**Figure 2.4**). While one unified approach that directly addresses the limitations of each approach more effectively would improve results, we selected CQN for downstream analyses due to its slight improvement in eQTL replication. Nonetheless, we recommend that, until the presented issues are directly addressed, both methods be considered as part of an analysis pipeline.

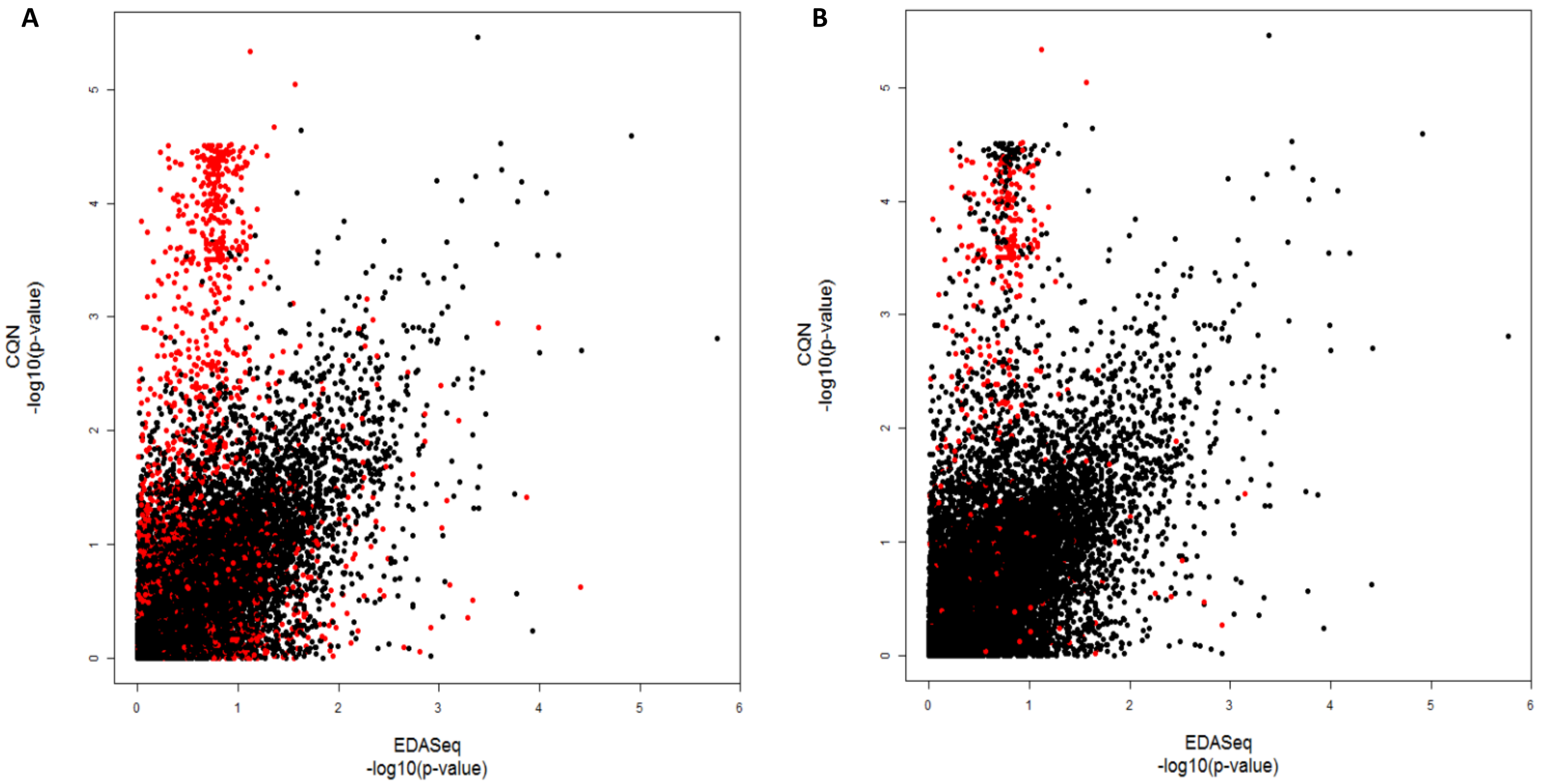
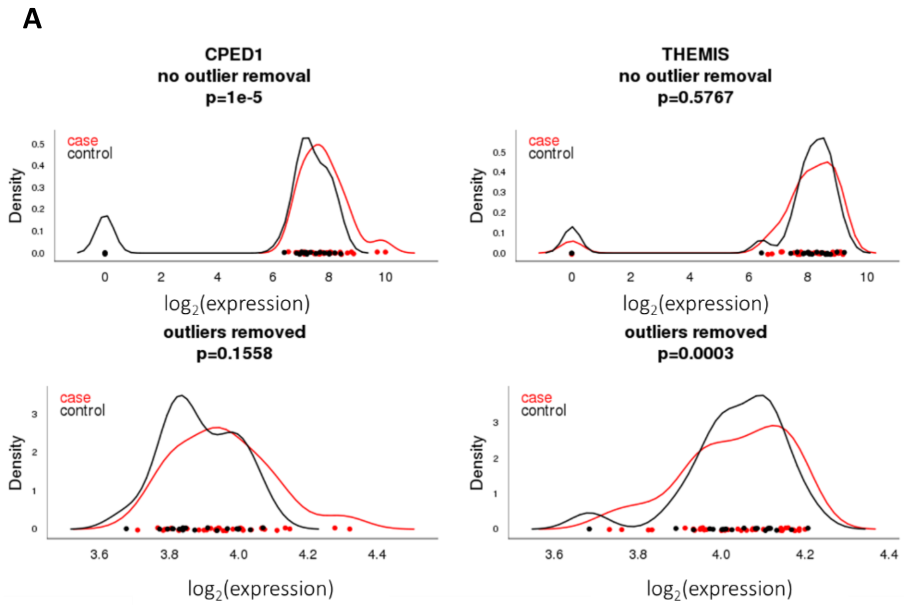


Figure 2.2: Comparing normalization methods.
(A) Scatterplot of case-control analysis (–log10) p-values from EDASeq (x-axis) vs CQN (y-axis). Genes with lengths less than 1000 base pairs are highlighted in red. (B) Scatterplot of case-control analysis (–log10) p-values from EDASeq (x-axis) vs CQN (y-axis), with genes with GC content < 35% are highlighted in red.



B

		EDASeq		CQN	
		Per gene outliers removed?			
		-	+	-	+
Gene	*CPED1*	1x10^{-5}	0.156	0.092	0.094
	THEMIS	0.577	3x10^{-4}	0.310	0.065
	RIC8B	0.925	0.197	0.988	0.017
	CHPF	0.022	0.636	0.048	0.992

Figure 2.3: Sample outliers identified and removed on a gene-by-gene basis improves robustness of differential gene expression analyses.

(A) EDASeq gene expression values are plotted. *CPED1,* a false positive, demonstrates a case where the presence of an outlier sample skews the differential gene expression analysis between cases and controls. Prior to outlier removal, the p-value for this gene was 1.04x10^{-5}; however, by removing the outlier sample, this gene is no longer differentially expressed between cases and controls. On the other hand, *THEMIS* shows the opposite trend in which the comparison demonstrates a more significant difference between cases and controls (p-value from 0.58 to 0.0003) after outliers are removed. (B) Differential gene expression p-values before and after per-gene outlier removal in EDASeq and CQN. *CPED1* and *THEMIS* summarize the data presented above in both EDASeq and CQN. In the CQN data, *RIC8B* and *CHPF* are comparable to *THEMIS* and *CPED* in the EDASeq analyses; however, the difference in p-value before and after per-gene outlier removal is less, due to CQN's better handling of outlier samples.

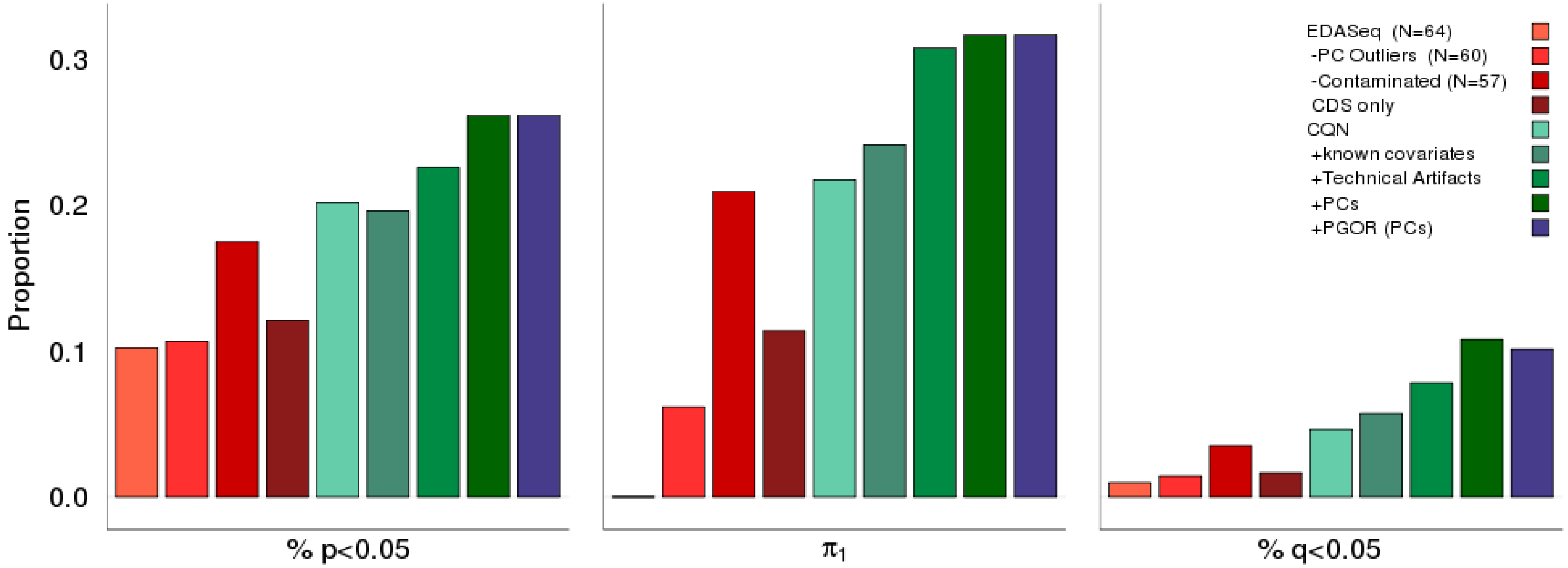


Figure 2.4: eQTL replication in brain data.

Normalization by EDASeq (red bars) demonstrates that sample outlier removal improves eQTL replication and that whole gene annotation provides improved gene expression estimation than coding sequence (CDS) only. CQN normalization (green bars) provides slightly improved eQTL replication over EDASeq normalization and demonstrates the necessity of covariate inclusion in eQTL replication, particularly highlighting the necessity of accounting for unknown covariates. Per-gene outlier removal (blue bars) does not hamper our ability to detect *cis*-eQTLs.

2.3.2 Identifying outliers in RNA-Seq data

In large sequencing studies, specific samples, for technical or biological reasons, can be recognized as outliers and should be removed from the study[68]. To identify outlier samples, whose global gene expression pattern is not explained by known covariates, we used Principal Component Analysis (PCA), investigating the first six principal components, which together explain ~60% of the variance in the brain data. Samples greater than three standard deviations (SD) from the mean in any of the first six principal components were deemed outliers and removed from analysis (N=4 or 6.3% of all samples) (**Figure 2.5**).

After sample-based outlier removal described above, it was apparent that, on a gene-by-gene basis, there were samples whose expression estimates differed greatly from the rest of the samples for that particular gene (**Figure 2.3**). Using a cut-off of three SD from the mean, 20.2% [7,027/34,738] of genes tested for differential expression between cases and controls had at least one sample flagged as an outlier for gene expression level. As these sample outliers are gene-specific, they suggest a clear artefactual origin, as opposed to a problem with the sample as a whole. Comparing the 50 most significantly differentially expressed genes between cases and controls before and after outlier removal, the lists differ at 60% [30/50] of the genes present, demonstrating that inaccurate results would be reported if gene-by gene outliers were not removed. To further ensure that this was indeed biologically sound, we assessed the validity of this approach using our eQTL analysis (discussed below).

After flagging outlier samples for removal in the brain data set, we obtained genotypes from both DNA and RNA. As a check on our data, we verified sample identity by comparing each RNA-Seq sample against all DNA samples. Pair-wise Identity by State (IBS) distances (DSTs) were calculated in PLINK with the expectation that DNA and RNA genotypes generated from the same individual should have a DST value approaching 1.0. In all samples, DNA genotypes best matched their corresponding RNA genotypes with a DST>0.83, indicating that our DNA and RNA samples were, in fact, from the same individual.

Despite correct identification of sample identity by IBS, three samples had borderline DST values (DSTs=0.83-0.89), warranting further investigation. These samples demonstrated an unexpected genotyping comparison profile such that all three showed an increased number of genotype calls deemed homozygous by DNA genotyping but called heterozygous at the

RNA level. As DNA genotyping by Affymetrix array has proven to be extremely accurate[69], an excess of sites where the DNA genotype indicates homozygosity but heterozygous calls are present at the RNA level indicates possible contamination. We quantify these occurrences in each sample using a metric we refer to as the Discordance Ratio (DR). For the majority of our samples, for which there is no suspected contamination, the DR approaches zero, with a value less than 0.2 indicating RNA-Seq data of sufficient quality for further analysis. The three samples in question had elevated DRs (0.32, 0.41, and 0.47), suggestive of sample cross-contamination (**Figure 2.6**).

To address the possibility of contamination, we conducted a mixing experiment where we combined high quality RNA-Seq samples (identified as having a DR<0.1) in controlled ratios. We carried out variant calling on these intentionally contaminated samples as had been previously carried out in the RNA-Seq data and calculated the DR for each. This ratio was then compared between the RNA-Seq samples in question and those from which mixing had been simulated. This comparison suggests that, for the three sample libraries in question, 370% of the RNA-Seq reads originated from a different sample (**Figure 2.7**). As reads from a foreign sample would lead to inaccurate gene expression estimates, we removed these samples from downstream analysis, resulting in a final data set of 57 samples, comprising 21 controls and 36 cases.

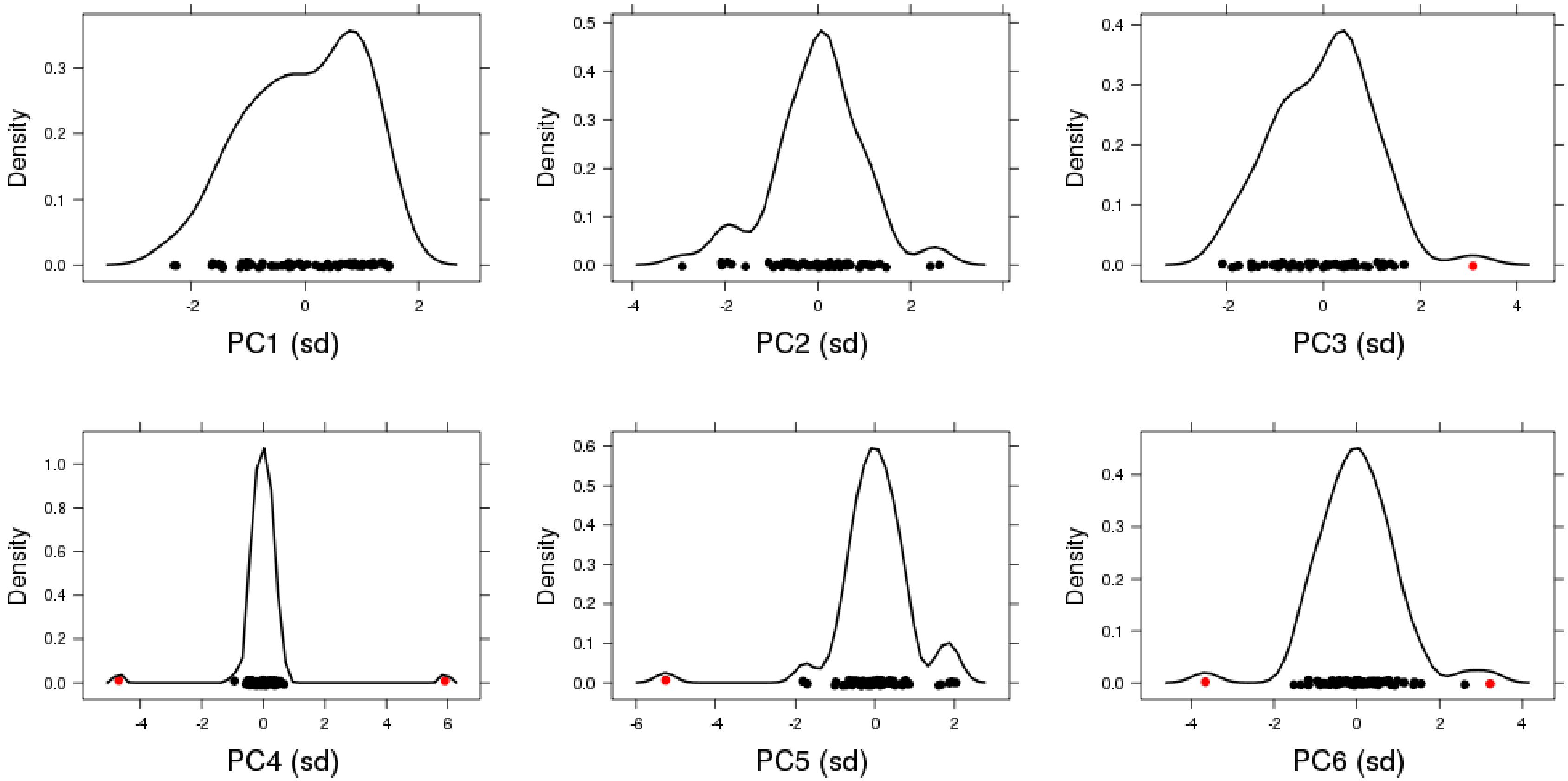


Figure 2.5: PCA identifies sample outliers.

The first six principle components (PCs) were assessed. Those samples whose gene expression profiles placed them greater than three standard deviations (sd) away from the mean of any of the first six PCs were identified (red) as sample outliers and removed from downstream analyses.

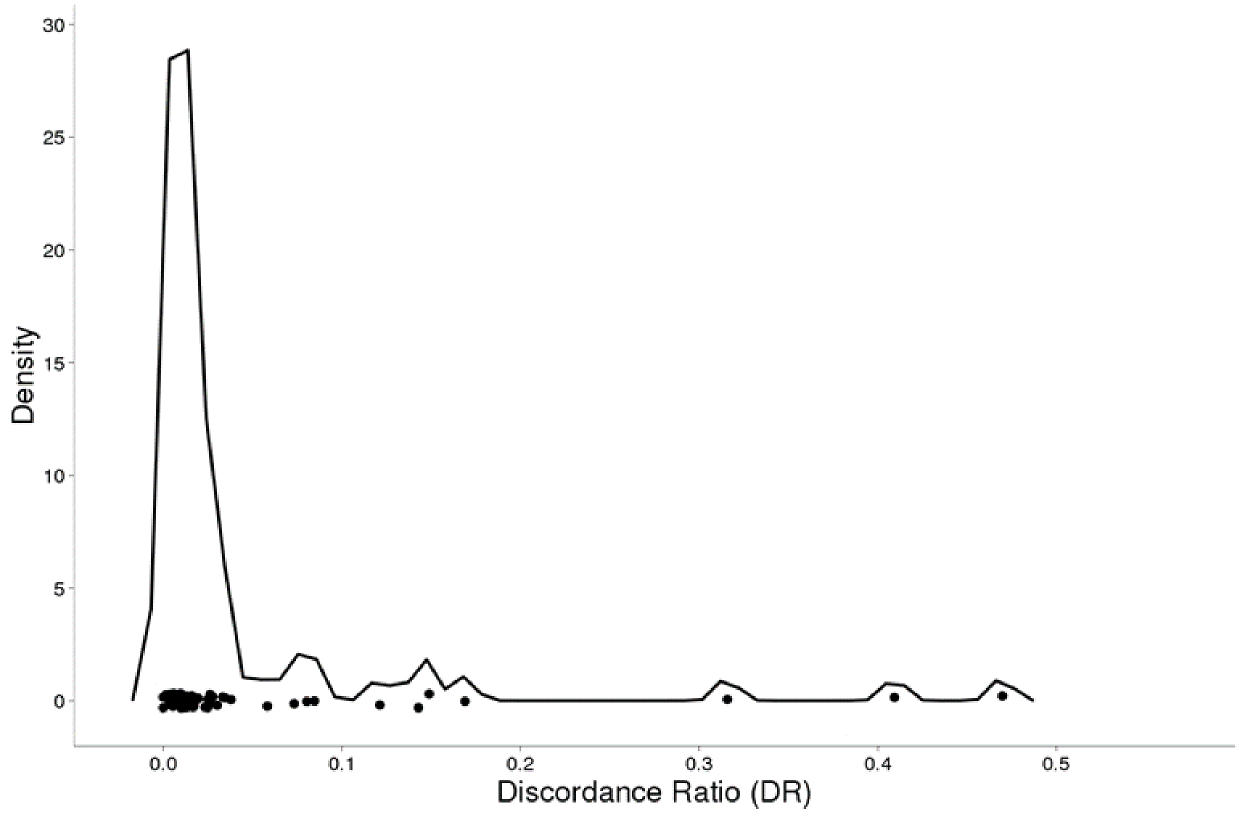


Figure 2.6: Using discordance ratio (DR) to assess quality of RNA.Seq data

Densityplot of RNA-Seq samples' discordance ratios (DR). A sample's DR can be calculated by taking the number of SNPs called homozygous at the DNA level but heterozygous at the RNA level divided by the total number of heterozygous RNA calls.

2.3.3 Reported brain eQTLs are reproducible in RNA-Seq data

Previously, surrogate measures of RNA quality (e.g., pH, post-mortem intervals, RIN values, etc.) have been used in an attempt to predict biologic validity, but none has been uniformly successful. Using published sets of brain eQTLs– regulatory genomic loci at which gene expression levels in the brain differ by genotype – we looked to recapitulate a number of the previously reported brain eQTLs in our gene expression data. We postulated that if we could replicate these eQTLs in our data, this would indicate that the use of post-mortem brain tissue may be representative of physiological conditions. We used a list of 909 *cis*-eQTLs generated from two recent studies that detected brain eQTLs in multiple disease populations across a number of brain regions [43,44]. Despite a smaller sample size and only one brain region under interrogation, we replicate 26.1% [237/909] of the tested associations (inflation-adjusted $p<0.05$) when age, sex, site and principal components are included as covariates (**Figure 2.4, Figure 2.8 & Table 2.2**).

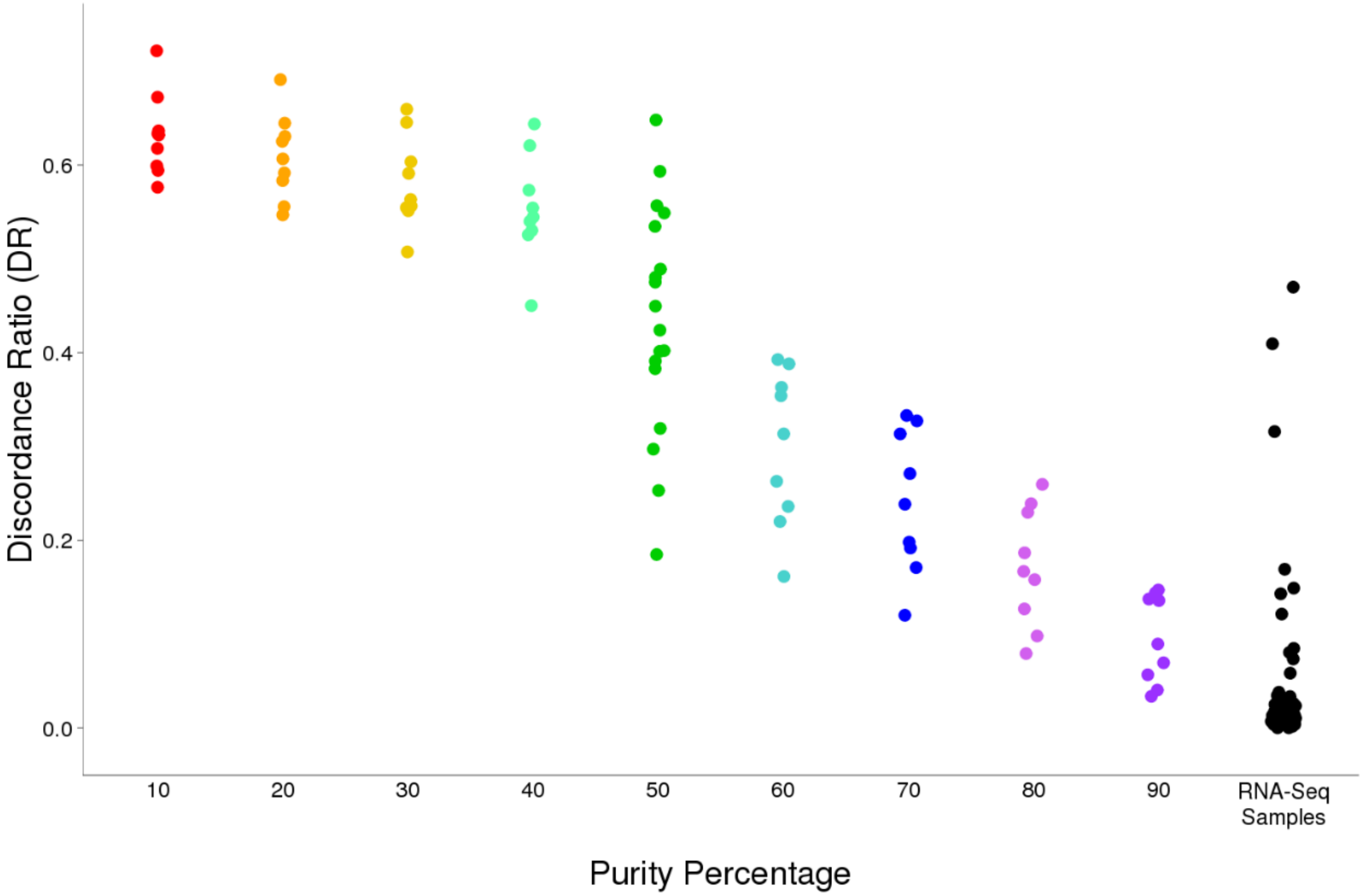


Figure 2.7: Simulated contamination of RNA-Seq reads.

Sample contamination was simulated by mixing RNA-Seq reads from two different samples in controlled ratios. These intentionally contaminated samples' RNA genotypes were then compared back to the DNA genotypes from which they were mixed. Specifically, a Purity Percentage of '10' indicates that 10% of the reads in that RNA genotype file were sampled from the DNA sample to which it was compared.

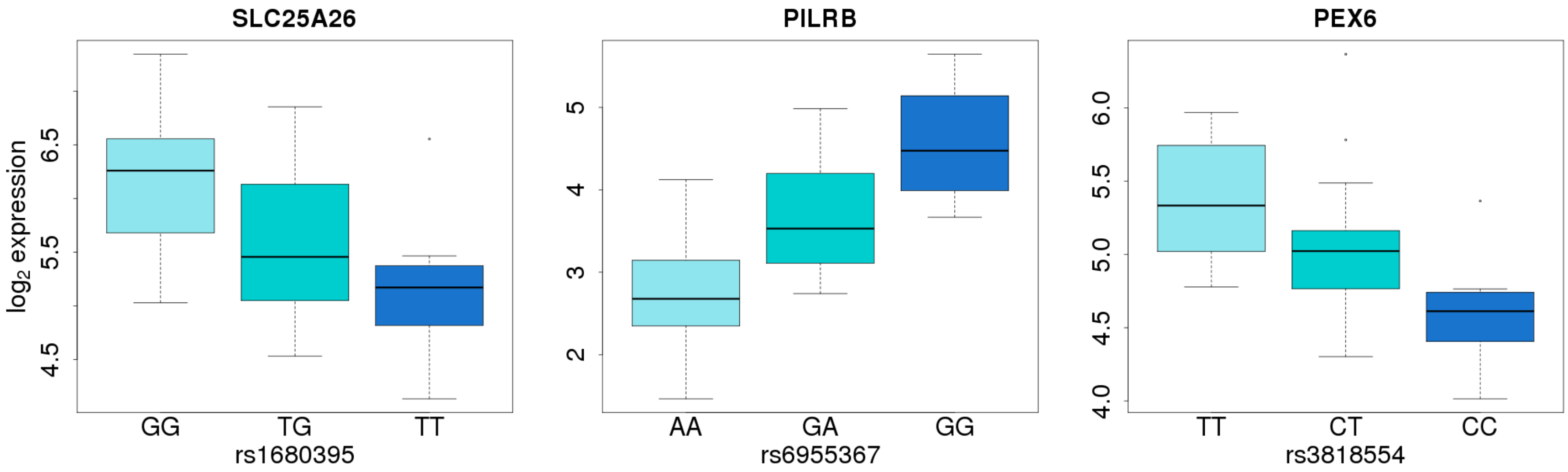


Figure 2.8: Replication of previously-reported eQTLs.

A representative set of three eQTLs are shown. Despite a smaller sample size and data from only one brain region, 26.1% of the previously published eQTLs are replicated in our combined dataset at $p<0.05$.

2.3.4 Monitoring eQTL replication to gauge quality control measures

We posit that if we are appropriately handling our data, known brain eQTLs should demonstrate improved association after each data correction step as well as an overall increase in the number of previously reported eQTLs that replicate. We have measured the ability to replicate known *cis*-eQTLs associations using three metrics: (1) the percentage of known eQTLs that replicate at $p<0.05$ after adjusting for genome-wide inflation (See Methods) (2) π_1, a statistic that estimates of the proportion of significant tests [65], and (3) the percentage of known eQTLs that replicate at $q<0.05$. When taken together, these three metrics offer a profile of the validity of each data handling step.

As part of the initial quality control, seven of the 64 samples (11% of total) were flagged as PCA outliers or contaminated samples, and removed. To assess the effect of sample removal, we compared eQTL replication in three data sets: (1) prior to outlier removal (N=64), (2) after dropping PC outliers (N=60), and (3) after dropping likely contaminated samples (N=57). Sample Outlier removal allows for the detection of 7.4% more known eQTLs $p<0.05$ and 3.5% more eQTLs $q<0.05$. Similarly, $\pi_1$estimates a dramatic increase in the proportion of replicating eQTLs from 0.00 to 0.209. These data indicate the necessity of removing suspect samples in these data (**Figure 2.4 & Table 2.2**).

We further utilized eQTL replication to determine the most appropriate model for gene annotation. There is evidence that suggests expression levels estimated from RNA-Seq data at the coding sequence (CDS) alone correspond better with qRT-PCR measurements than RNA-Seq estimates that include both the CDS and its untranslated regions (UTRs). However, recent RNA-Seq analyses have generally included gene annotation from the whole gene – that is the CDS and its UTRs – under the argument that gene annotation gains accuracy upon UTR inclusion [70]. To address this discrepancy in the literature, we compared these two gene annotation approaches by eQTL replication. The whole gene annotation clearly replicates known eQTLs better than the CDS alone (**Figure 2.4 & Table 2.2**) detecting 5% more known eQTLs at $p<0.05$ and 1.9% more at $q<0.05$. Replication, as measured by π_1 demonstrates an increase in this test statistic as well (0.114 in CDS, 0.209 in whole gene annotation). This improvement in eQTL detection offers support for the use of UTR inclusion in gene annotation in these data.

Similarly, eQTL replication was used to compare normalization methods. We note that when considering the overall number of known eQTLs detected, CQN replicates 2.7% more eQTLs ($p<0.05$) than does EDASeq (**Figure 2.4 & Table 2.2**), further supporting its use in analyzing gene expression in this data set.

Disease-based comparisons are frequently adjusted for known covariates (age, sex, etc.). However, comparative studies are also frequently plagued by unknown covariates, or confounders within the data that are not easily attributable to any recorded measurement [46,68]. These unknown covariates can be approximated through various data decomposition methods. We initially considered using PCA to accomplish this goal but observed that the first PC was correlated with both collection site (see Methods) and disease status, which often occurs whenever different sites have different fractions of cases and controls. As this could be a likely issue in many case-control study, limiting the utility of PCs in downstream analyses, we also considered Surrogate Variable Analysis (SVA)[50] and Independent Surrogate Variable Analysis (ISVA) [51], as these approaches allow for disease status to be protected during their generation. Lastly, we also considered utilizing PEER[49,57] to account for unknown covariates, as this method has been used and performed well in previous eQTL analyses [66]. In eQTL replication analyses, performance was comparable with ISVs, SVs, PEER and PCs detecting 25.1, 26.2, 26.9 and 26.1 percent of the previously reported eQTLs, respectively ($p<0.05$) (**Table 2.2**). Ultimately, however, to address the case-control confounding issue, we had to decide between ISV and SV usage. To do so, we tested both methods by assessing Q-Q Plots generated for disease-based comparisons. As the inclusion of SVs, but not ISVs, demonstrated inflated p-values in these analyses (**Figure 2.9**), we decided to move forward with ISVs to account for unknown covariates.

Finally, regarding covariate inclusion, we note that certain metrics for technical artifacts of sequencing (percent coding bases, percent intronic bases, percent mRNA bases, median 3' bias, percent UTR bases, and AT dropout) were correlated with specific ISVs (**Table 2.4**), suggesting that the unknown covariates detected by ISVA may simply be accounting for known technical artifacts of sequencing. We tested this possibility and demonstrate that, while including technical artifacts as covariates does improve eQTL detection over known covariates alone (2.4% increase at $p<0.05$, increase in π_1 from 0.217 to 0.308), both PCs and

ISVs perform even better, demonstrating a 5.9% and 4.9% increase at $p<0.05$, respectively, when compared to no covariate inclusion (**Figure 2.4 & Table 2.2**). These data ultimately support the inclusion of covariates, as captured by data decomposition methods, in downstream analyses suggesting that such methods are either (a) accounting for unknown covariates beyond technical sequencing artifacts or (b) appropriately weighting the effects of the technical artifacts amongst the ISVs/PCs generated.

As noted above, sample outliers were also identified on a per-gene basis and removed from analysis. To ensure that removing these outliers was biologically sound and that these outliers did not represent true measures of differential expression, we tested data sets where sample outliers were removed at each gene using our eQTL replication approach. While per gene outlier removal did not demonstrate a marked increase or decrease in eQTLs detected (**Figure 2.4 & Table 2.2**), the presence of outlier samples leads to a lack of robustness in the case-control analysis where single samples dramatically skewed the results (**Figure 2.4**). As per-gene outlier removal helped to stabilize the case-control analyses and did not hinder our ability to detect known eQTLs, we support its inclusion in RNA-Seq data analysis.

2.3.5 Independent RNA-Seq data set supports use of eQTL gold standards

To bolster the results of our brain RNA-Seq data set, we set out to replicate the main findings of our initial analysis in an independent RNA-Seq data set generated from a distinct tissue source. To do this, we used 162 blood samples from the GTEx project[45], for whom we had DNA genotypes as well as raw count data from RNA-Seq. In these data, four samples (2.5% of total) were identified as PC Outliers, using the same criteria as was used in the brain data. Sample outlier removal led to a slight decrease in the number of eQTLs detected (29.6% versus 28.3% at $p<0.05$); however, there was an increase in π_1 (0.374 to 0.387 after outlier removal) (**Figure 2.10 & Table 2.3**). Normalizing using CQN again led to an overall increase in eQTLs detected (3.5% increase at $p<0.05$) (**Figure 2.10 & Table 2.3**). In assessing covariate addition, a pattern similar to what was seen in the brain data was observed. While known covariates (age, sex, and cohort) in the brain data did not improve the eQTL detection, there was a similar improvement seen upon the addition of PCs to account for unknown covariates (9.3% increase when compared to the use of no covariates) (**Figure 2.10 & Table 2.3**). Again, per gene outlier removal does not hamper the ability to detect known eQTLs (**Figure 2.10 & Table 2.3**).

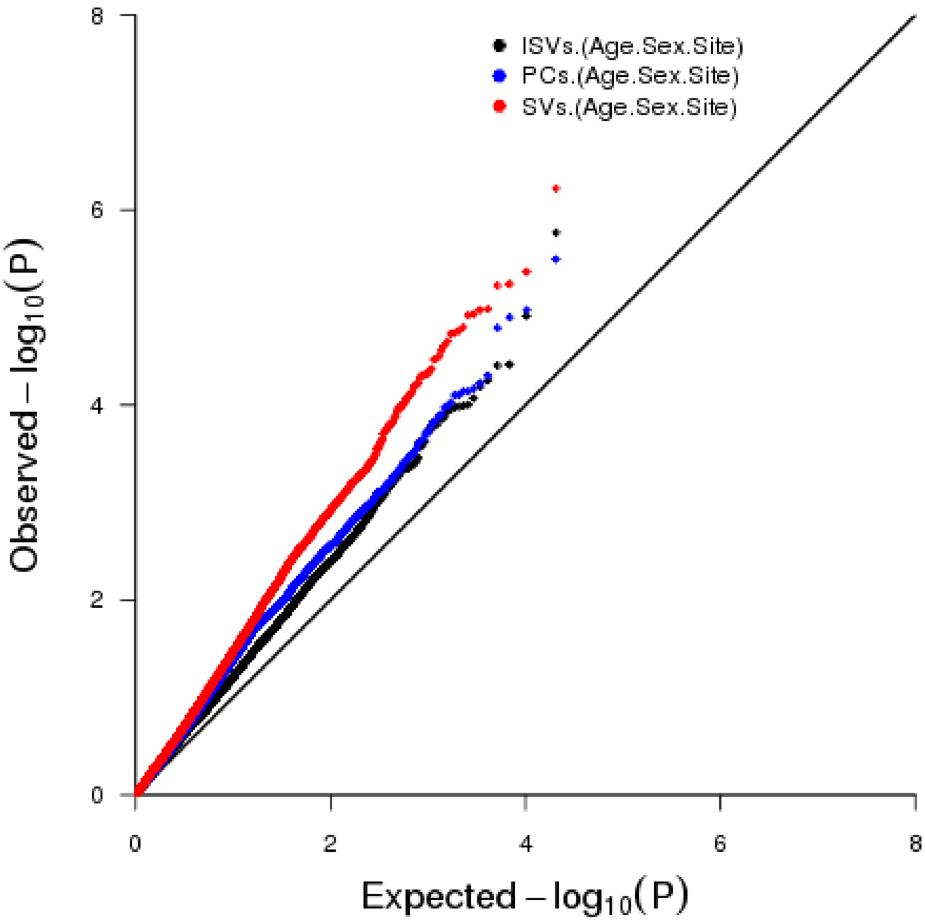


Figure 2.9: QQ Plots for Data Decomposition Methods.
The deviation from the expected in this QQ-Plot demonstrates inflation in these data, with SVA (red) demonstrating the most significant overinflation among the three methods of data decomposition. This finding supported using ISVs over SVs in our analyses.

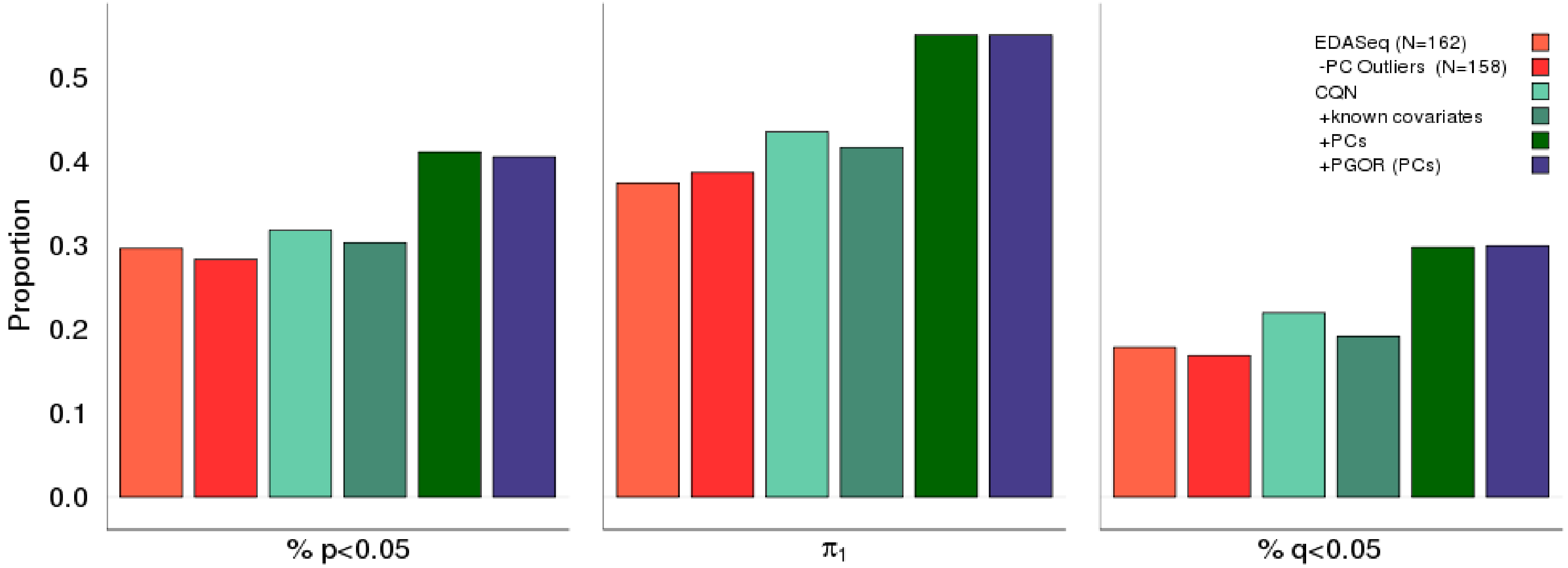


Figure 2.10: eQTL replication in blood data.

Colors correspond to the comparable analyses carried out in the brain data (Figure 2). Again, these data show that CQN (green bars) slightly improves eQTL detection over EDASeq (red bars) and that a considerable increase in eQTL detection is seen when unknown covariates are considered in the analysis.

2.4 Discussion

Just as it took more than ten years for the field to reach a consensus on the analysis of microarray data, RNA-Seq analysis has been undergoing a similar struggle since the first RNA-Seq publication. Since then, accurate library preparation, appropriate mapping of short sequencing reads, and correct estimation of gene expression values have been the primary focus. While many of the original experimental and data analysis hurdles have been addressed, a framework in which one can assess the quality control and data adjustment measures taken after obtaining accurate gene expression estimates was lacking. For experiments with RNA-Seq data, DNA genotypes and a list of tissue-appropriate eQTLs, we demonstrate that our approach can be easily employed to generate a clean set of expression estimates for downstream analyses (**Figure 2.1**). Specifically, we propose using the ability to replicate eQTLs as a biologically meaningful check on the integrity of the data and to help ensure that the data is being handled appropriately at each quality control and data adjustment step.

We demonstrate that upon generating normalized gene expression estimates, PCA can be utilized to identify global gene expression sample outliers and that DNA and RNA genotypes can be employed to verify sample identity and check for sample contamination. Using the replication of known eQTLs, we also demonstrate the importance of including both known and unknown covariates in downstream analyses. This result is consistent with expectation [68], illustrating the utility of eQTL replication as a simple approach to assess data handling measures and offering credence to its usage in additional comparisons. This eQTL replication approach was further employed to demonstrate that ISVs are not simply explaining known technical artifacts of next-generation sequencing, that gene annotation best replicates known biology when UTRs are included in gene annotation, and that we do not overtly lose power to detect known eQTLs when reducing the sample size by removing suspect samples.

Of the data cleaning steps employed, we highlight the importance of removing individuals on a per-gene basis, as this is not a standard quality control step. Indeed, either due to low coverage (leading to zero counts) or undetected PCR duplication (leading to an overabundance of counts), a sample may exhibit gene expression values vastly different than the rest of the samples, and as a result, should be removed from analysis at that particular gene prior to ISV/PC generation. We note that this process is more important when using EDASeq than when CQN is used for normalization, as quantile normalization produces fewer outlier values. In the brain data set, when EDASeq was employed, 42.5% of genes had at least one sample flagged as an outlier (**Figure 2.3A**), whereas the nature of quantile normalization produced outliers in only 20.2% of genes tested. Nevertheless, with both methods, the differential gene expression analysis becomes more robust upon the removal of individual samples present in the data skewing the results (**Figure 2.3B**). While these sample-specific outlier genes could certainly reflect an insertion-deletion event or reflect another genetic variation in these individual samples, our goal is to maximize one's ability to find eQTLs and assess overall data handling measures, and as such, these individuals should be removed from analysis. Given that our ability to detect eQTLs is not hampered and that differential gene expression analysis demonstrates improved robustness upon per-gene outlier removal, we argue that this novel outlier identification approach be incorporated in future RNA-Seq expression studies.

The utilization of two distinct RNA-Seq data sets – one generated from brain (N=64) and another independently-generated from blood (N=162) – helps to demonstrate the main findings of this work. In both brain and blood data sets, eQTL replication was improved with the use of CQN for data normalization and further improved upon the addition of PCs as covariates. Further, in both data sets, removing per-gene outliers did not hamper the ability to detect known eQTLs. However, there are data handling measures imposed on the data that were more important in the brain data set than in the blood data set, likely reflecting the fact that the brain data was plagued by a smaller sample size and that the sequencing data was generally of overall lower quality due to degradation of the starting material. Reflecting the sample size difference, overall replication was higher across the board in the blood data set. Further, the degraded nature of the starting brain RNA-Seq material was reflected in the need for extensive processing of the sequencing reads due to poor library quality. Accordingly, sample outlier removal proved essential in the brain data set, but did not make a huge difference in the blood data. While the blood data set was less sensitive to the presence of outliers, this work demonstrates that despite the use of a degraded starting product, biologically meaningful data was still generated from the brain data and that careful data analysis can augment the information garnered from a limited data set. These distinctions between the two independent data sets furthers the point that each RNA-Seq experiment is unique and carries its own limitations, but eQTL replication can be used to guide one's analysis pipeline.

Finally, it is important to note that our eQTL approach was more helpful in some comparisons than others. While this approach can greatly help to guide one's analysis, there will be cases where the choice is not so obvious and further steps will need to be taken to assess one's data processing. For example, when comparing the three data decomposition methods (PCA, SVA, and ISVA) (**Tables 2.2 & 2.3**), the answer was unclear, as all methods do a similar job accounting for the unknown covariates. Thus, the choice between the methods was based on additional criteria with PCs being excluded due to confounding within the first PC between sample collection site and disease status and SVs due to their overinflated p-values in differential gene expression analysis (**Figure 2.9**). Additionally, we note that there are several caveats to the use of data decomposition methods. First, when dealing with small sample sizes, including a large number of covariates can lead to overfitting of the data. Second, data decomposition to account for unknown covariates will minimize the ability to detect global differences in gene expression, which may be correlated with one or more of the eigenvectors (e.g. comparisons across tissues would by necessity not incorporate PCs). In addition to this approach not being applicable for all comparisons, we note that RNA-Seq remains an imperfect measure of gene expression. Technical and analytical limitations remain. Cell type heterogeneity and the need for cDNA generation currently result in unavoidable biases in data generation. While single cell RNA-Seq and direct RNA sequencing methods will address these issues, any improvement that further reduces bias in library construction will lead to more accurate gene estimate values, allowing for further protocol improvement. Additionally, improvements in mapping algorithms, normalization procedures, and gene estimate quantification will also aid in reproducibility.

2.5 Conclusion

In recent years, RNA-Sequencing (RNA-Seq) experiments have moved the forefront of the transcriptomics field to become the gold standard approach for the study of genome-wide gene expression. While this period has led to protocols that aim to optimize library preparation and computational methods that aid in improved mapping and accurate gene expression estimation, a method to assess downstream data handling approaches was lacking. Here, we offer a framework that utilizes DNA genotypes and RNA-Seq data along with previously published eQTLs to assess possible sample contamination and assess the biological validity of each data analysis step to ultimately enable confident downstream analyses.

2.6 Tables

Table 2.1: Sample Information.

FID	Diagnosis	Sex	Age (y)	Total Mapped Reads	Reason Removed from RNA-Seq Analysis?
AN16641	Autism	M	9	21,584,033	
AN00493	Autism	M	27	4,475,274	
AN00764	Autism	M	20	14,019,291	
AN08792	Autism	M	30	8,226,165	Sample Contamination
AN08873	Autism	M	5	16,702,649	Sample Contamination
AN19511	Autism	M	8	15,225,859	
AN01570	Autism	F	18	29,208,145	
AN09730	Autism	M	22	28,008,384	Sample Contamination
AN17777	Autism	F	49	27,039,497	
AN12457	Autism	F	29	36,768,708	
AN11989	Autism	M	30	34,359,536	
AN13872	Autism	F	5	35,951,442	
AN17678	Autism	M	11	14,071,311	
AN04682	Autism	M	15	3,687,486	
AN03632	Autism	F	49	44,055,322	
AN09714	Autism	M	60	42,351,515	
AN10606	Control	M	56	30,737,028	PC Outlier
AN16665	Control	M	36	19,701,091	
AN01357	Control	M	42	11,514,063	
AN17425	Control	M	16	46,259,333	
AN14368	Control	M	22	2,691,397	
AN15566	Control	F	32	14,500,662	

AN13295	Control	M	56	39,663,314	
UMB797	Autism	M	9	28,573,237	
UMB1349	Autism	M	5	24,697,811	
UMB1638	Autism	F	20	15,048,023	
UMB4231	Autism	M	8	28,320,026	
UMB4721	Autism	M	8	8,185,563	
UMB4999	Autism	M	20	31,898,006	
AN16641	Autism	M	9	21,584,033	
AN00493	Autism	M	27	4,475,274	
AN00764	Autism	M	20	14,019,291	
AN08792	Autism	M	30	8,226,165	Sample Contamination
AN08873	Autism	M	5	16,702,649	Sample Contamination
AN19511	Autism	M	8	15,225,859	
AN01570	Autism	F	18	29,208,145	
AN09730	Autism	M	22	28,008,384	Sample Contamination
AN17777	Autism	F	49	27,039,497	
AN12457	Autism	F	29	36,768,708	
AN11989	Autism	M	30	34,359,536	
AN13872	Autism	F	5	35,951,442	
AN17678	Autism	M	11	14,071,311	
AN04682	Autism	M	15	3,687,486	
AN03632	Autism	F	49	44,055,322	
AN09714	Autism	M	60	42,351,515	
AN10606	Control	M	56	30,737,028	PC Outlier
AN16665	Control	M	36	19,701,091	
AN01357	Control	M	42	11,514,063	
AN17425	Control	M	16	46,259,333	

AN14368	Control	M	22	2,691,397	
AN15566	Control	F	32	14,500,662	
AN13295	Control	M	56	39,663,314	
UMB797	Autism	M	9	28,573,237	
UMB1349	Autism	M	5	24,697,811	
UMB1638	Autism	F	20	15,048,023	
UMB4231	Autism	M	8	28,320,026	
UMB4721	Autism	M	8	8,185,563	
UMB4999	Autism	M	20	31,898,006	

Table 2.2: Summary of eQTL replication analyses carried out in brain samples.

EDASeq	64	WG	No	no covariates	10.2%	0.000	1.0%	1.412
	60				10.7%	0.062	1.4%	1.462
	57				17.5%	0.209	3.5%	0.988
		CDS			12.1%	0.114	1.7%	1.137
CQN		WG			20.2%	0.217	4.6%	0.968
				known covariates (age.sex.site)	19.6%	0.241	5.7%	0.966
				Technical Artifacts	22.6%	0.308	7.8%	0.946
				PCs	26.1%	0.316	10.8%	0.981
				ISVs	25.1%	0.327	10.1%	0.965
				SVs	26.2%	0.290	9.9%	0.960
				PEER	26.9%	0.321	10.0%	0.955
			Yes	PCs	26.1%	0.316	10.1%	0.972
				ISVs	25.7%	0.303	10.3%	0.952
				PEER	25.7%	0.279	11.5%	0.968

Table 2.3: Summary of eQTL replication carried out in GTEx blood samples.

Normalization Method	Sample Size (N)	Gene Annotation	Per Gene Outliers Removed?	Covariates	% detected $p<0.05$*	π_1*	% detected $q<0.05$*	Inflation factor (λ)*
EDASeq	162	WG	No	no covariates	29.6%	0.374	17.8%	1.182
	158				28.3%	0.387	16.9%	1.188
CQN					31.8%	0.435	21.9%	1.192
				known covariates (age.sex.cohort)	30.3%	0.416	19.1%	1.305
				PCs	41.1%	0.550	29.7%	1.297
			Yes		40.5%	0.550	29.9%	1.268
				PEER	41.1%	0.550	32.3%	1.269

Table 2.4: Technical Artifacts are correlated with Independent Surrogate Variables. Coefficients greater than 0.45 are bold for emphasis.

Technical Artifact	**Correlation Coefficient (r)**							
	ISV1	**ISV2**	**ISV3**	**ISV4**	**ISV5**	**ISV6**	**ISV7**	**ISV8**
percent coding bases	0.34	-0.28	0.41	**0.78**	-0.36	**-0.54**	**-0.51**	**0.5**
percent intronic bases	-0.03	0.38	**-0.55**	**-0.66**	**0.61**	**0.59**	0.23	**-0.56**
percent mRNA bases	0.41	-0.25	**0.45**	**0.48**	**-0.61**	**-0.63**	-0.22	**0.46**
median 3' bias	-0.19	0.18	-0.25	**-0.79**	0.09	0.31	**0.57**	-0.32
percent UTR bases	0.19	0	0.15	-0.34	**-0.48**	-0.24	0.39	0.02
AT dropout	**0.64**	-0.14	0.05	0.19	-0.03	-0.26	**-0.74**	0.18

CHAPTER 3: Transcriptome Analysis Reveals Deregulation of Innate Immune Response Genes and Neuronal Activity-Dependent Genes in Autism

3.1 Introduction

Recent studies to elucidate the molecular basis of autism have largely focused on genetic approaches, including both genome-wide association studies[17] and whole-exome sequencing[11–14], to identify both inherited and de novo variation contributing to autism. A major mode of action for genetic variation is through altered gene expression, so direct analysis of gene expression in a disease-relevant tissue is a complementary approach to genetic studies. Despite the extreme genetic heterogeneity observed in autism, it is possible that common downstream mechanisms may be altered[20]. Thus, there has been an effort to use transcriptomics to identify and dissect molecular pathways that may be altered in ASD.

Given low tissue sample availability in autism research, efforts have focused on assessing gene expression in lymphoblastoid cell lines or whole blood[26–30]. However, given the core neurodevelopment phenotypes associated with autism, there is little doubt that direct assessment of gene expression in brains may be critical. Indeed, Voineagu et. al recently utilized co-expressed gene networks from RNA-Sequencing (RNA-Seq) carried out in 19 autism brains and 17 controls to identify a set of co-expressed neuronal genes enriched for known autism susceptibility genes as well as a set of co-expressed genes enriched for both immune genes and glial markers[20].

In the current study, we present results from the largest RNA sequencing of autism brains effort to date that allows for new insights into the etiology of autism. We find clear differences in the transcriptome between control and ASD cortical brains. Using co-expression network analysis, we demonstrate that autism brains are specifically enriched for “activated” M2 microglial and “immune response” genes. Remarkably, the M2 microglial module is strongly negatively correlated with one of two differentially expressed neuronal modules, highlighting the interplay between innate immunity and neuronal activity in the etiology of ASD.

3.2 Methods

3.2.1 Brain tissue samples

Brain tissue: Frozen brain samples were acquired through the Autism Tissue Program (http://www.atpportal.org), with samples originating from two different sites: the Harvard Brain Tissue Resource center and the NICHD Brain and Tissue Bank at the University of Maryland. Tissue was obtained post-mortem and written informed consent was obtained from next-of-kin or a legal guardian. This work was approved by the IRB of The Johns Hopkins Hospital and University of Alabama at Birmingham and was conducted in accordance with institutional guidelines. The brain samples were dissected to obtain the cerebral cortex Brodmann area (BA) 19, anterior prefrontal cortex (BA10) and a part of the frontal cortex (BA44). Multiple cortical tissues corresponding to BA19, BA10 and BA44, were sequenced in 62, 14, and 28 samples, respectively, resulting in a total of 57 (40 unique individuals) control and 47 (32 unique individuals) autism samples. The average age at time of death of the 40 control and 32 autism individuals was similar (cases median age = 20 yrs, controls median age = 17 yrs), and there was no significant difference in cause of death between the two groups. **Table 3.1** contains the details of the corresponding subject phenotypes and additional characteristics.

3.2.2 RNA library preparation and RNA sequencing

RNA-Seq libraries were prepared from 50 µg of total RNA from postmortem brain tissue extracted with Trizol reagent according to manufacturer's protocol (Invitrogen). The TruSeq RNAseq kit (Illumina) was used with minor modifications as follows. Total RNA pools were subjected to two rounds of hybridization and elution with oligo(dT) dynabeads (Invitrogen) to obtain purified polyadenylated (polyA) RNA. After mRNA selection, samples were randomly fragmented to minimize bias at the 3' end of the transcript. First-strand cDNA synthesis was performed using random primers (Illumina) and SuperScriptII Reverse-Transcriptase (Invitrogen) followed by second strand cDNA synthesis using RNaseH and DNA polymerase I (Illumina). Illumina supplied adaptors (TruSeq kit) were ligated to the purified, end-repaired and 3' adenylated cDNA, and we performed manual 200 bp size-selection of the final product by gel-excision. The 200 bp cDNA template molecules were then amplified by PCR to create the final library. Quality control measures during library amplification included PCR from reactions with no template, from libraries made with no ligase (hence no adaptors), and finally from libraries with no adaptor oligonucleotides included in the ligase reaction. In these cases, the library failed to amplify, thereby ensuring specificity of the expected product for each run. Each library was evaluated for uniformity on a 2100 Bioanalyzer (Agilent) prior to sequencing on a single lane of Illumina's HiSeq 2000 to produce 100 base pair (bp) single-end reads. Each sequencing run included samples randomized by sex, collection site, and case-control status.

3.2.3 Mapping and gene summarization of data from RNA-Seq

The sequenced reads for each sample were obtained as fastq files for 110 samples. To improve mapping, reads were trimmed to remove stretches of terminal A's or T's (N=3-12) and contaminating adaptor sequences using a Python script, cutadapt (v1.2.1)[71]. The sequenced reads were mapped using Tophat2[55,72]. Only uniquely mapped reads with a maximum of 3 mismatches were used to estimate gene counts. The RNA-Seq reads were mapped to a set of sequences derived from the Genome Reference Consortium Human build 37 (GRCh37) assembly, recommended by the 1000 Genomes Project[73]. Gene expression estimates were made for approximately 48,260 of the total 62,069 reported Ensembl gene annotations (GRCh37 or Human release 70), recommended by Kim et al.[72], using the python script 'HTSeq-count' (model type - intersection strict, http://www-huber.embl.de/users/anders/HTSeq/)[74].

3.2.4 Normalization of gene estimates

Subsequent to mapping, the gene count data was normalized for within and between lane biases (e.g. GC content) and sequencing depth by methods implemented in Conditional Quantile Normalization (CQN)[67] and Exploratory Data Analysis and Normalization for RNA-Seq (EDASeq)[56], using the default settings for each method. We present the EDASeq-normalized data, and for a detailed discussion about the differences between EDASeq and CQN, see Ellis *et al*.[75].

We assessed summarized values on a per-gene basis, removing gene estimates for samples whose gene expression values were more than three standard deviations (SD) from the mean expression of each gene (per-gene outlier), as these outliers are artefactual in origin[75].

3.2.5 Quality assessment

Picard (http://picard.sourceforge.net, v1.87) command-line tools 'CollectRnaSeqMetrics' and 'CollectGcBiasMetrics' were used to provide RNA-Seq summary statistics. Six samples with low gene coverage (> 20% of the 48,260 genes had zero coverage) were dropped from all downstream analyses, resulting in 104 samples. In addition, to detect global sample outliers due to technical or biological reasons, we used principal component analysis (PCA) and identified a subset of 2,582 genes with at least 10 reads per sample using the '*prcomp*' function in the stats package in R (http://www.R-project.org/). All 104 samples were within three SD of the mean of the first six principal components, which together explained ~55% of the variance[75].

3.2.6 Single gene differential expression analysis

After normalization and outlier removal, independent surrogate variables (ISVs)[51] were generated on a subset of 2,500 genes with at least 10 read coverage in each sample. Data decomposition was performed on the log2 scale for the 2,500 genes. ISVs were generated while protecting for case-control status using the '*isvaFn*' function in the 'isva' package in R. Differential gene analysis was performed using a subset of 13,262 genes that had at least 3 reads per sample across 90% of the samples.

A linear mixed regression framework was utilized to identify differential gene expression between 57 controls and 47 cases. To remove unwanted sources of variation while protecting differences due to the primary variable of interest (case-control status), site of sample collection, age, sex, brain region and ISVs were included as fixed effects. . Additionally, the model included a random intercept term to account for the correlation of gene expression estimated from multiple brain regions obtained from the same individual.

Permutation testing was used to estimate the threshold for transcriptome-wide significant differential expression (EDASeq, 400 permutations, $P = 4.76 \times 1^{7}$). We reiterate that the samples were obtained from two collection sites, and to estimate the threshold for transcriptome-wide significant differential expression, we permutated the case-control status within each site, maintaining the same phenotype for multiple samples (i.e. brain regions) derived from a single individual.

We assessed the possibility of confounding in the expression of the two differentially expressed genes by investigating the expression stratified by sample collection site (**Figure 3.1a,d**). The sequencing coverage of *MAL* was calculated across the 4 exons and for 21 exons for *C11orf30* using 'coverageBed' from bedtools[76] (**Figure 3.1b,e**). We also investigated for the expression of *MAL* and *C11orf30* during development and across different brain regions from Brainspan (http://hbatlas.org/pages/hbtd).

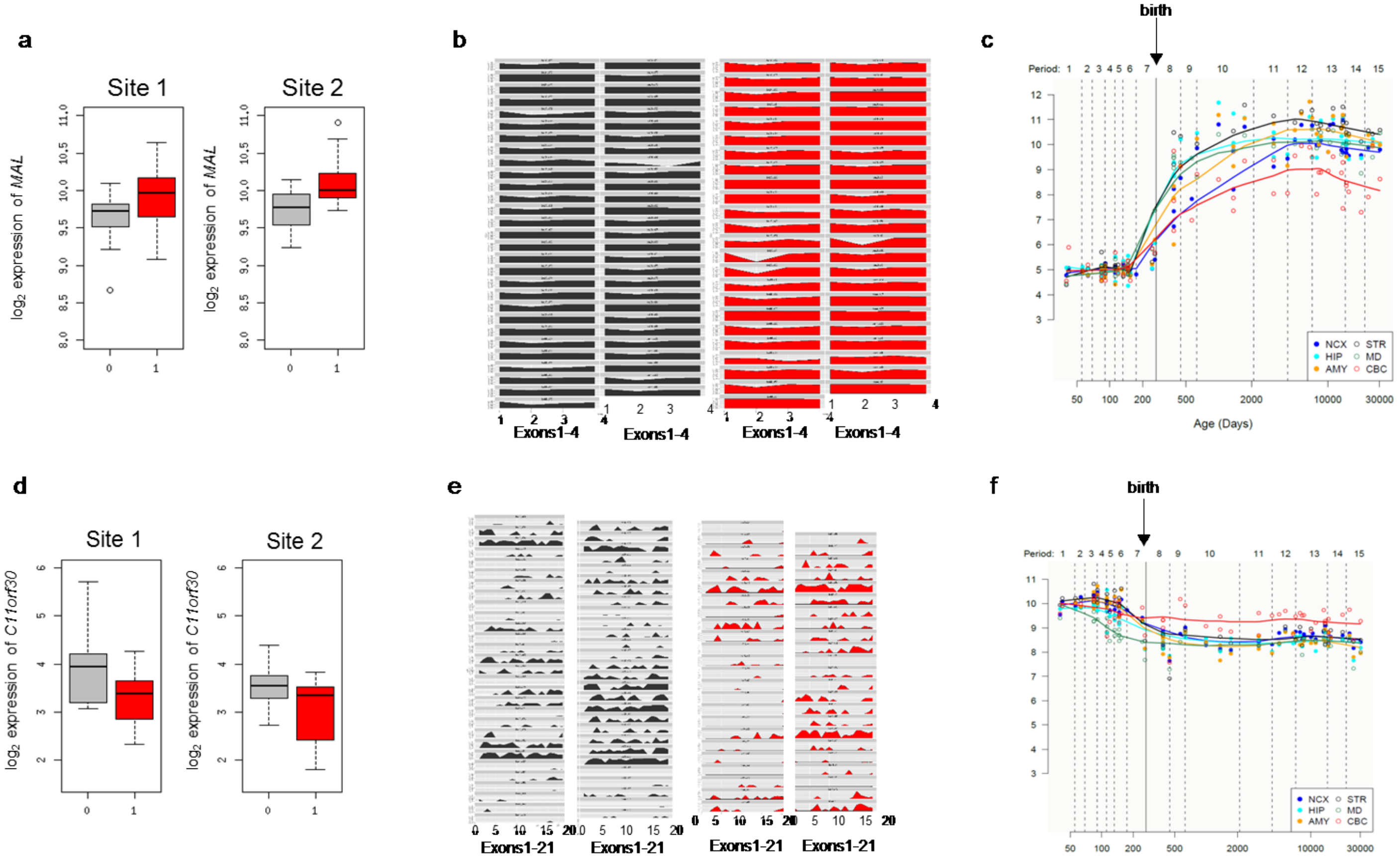


Figure 3.1: Transcriptome-wide top hits.

(**a**,**d**) Expression levels of *MAL* (**a**) and *C11orf30* (**d**) in controls (grey) and cases (red) stratified by collection site demonstrate that the association with autism is consistent across collections sites. (**b**,**e**) A continuous plot with the coverage on a scale of 0-1 in all the 57 controls (grey) and 47 cases (red) across the 4 exons for *MAL* (**b**) and 21 exons for *C11orf 30* (**e**). (**c**,**f**) Expression of *MAL* (**c**) and *C11orf30* (**f**) through human brain development from post-conception week (PCW) 4 to 82 years (age on the x-axis and $\log_2$ expression of the gene across the different regions of the brain on the y-axis). The expression of *MAL* increases after birth in all brain regions, whereas *C11orf30* decreases after birth in all brain regions other than the amygdala (data from brainspan, http://www.brainspan.org).

3.2.7 Single exon differential expression analysis

Exon-level estimates were obtained using the ‘count.py’ script from DEXSeq[77] for each of the 104 samples. Exons with more than three reads across 90 percent of all samples were included for analysis. These 21,310 exons were modeled utilizing a linear mixed regression framework to identify differential exon usage between the 57 controls and 47 cases. Site of sample collection, age, sex, brain region and ISVs (generated from the single gene-level analysis) were included as covariates to account for unknown confounding factors as fixed effects. Additionally, the model included a random intercept term to account for the correlation of gene expression estimated from multiple brain regions obtained from the same individual.

3.2.8 Single gene - Gene Ontology and pathway enrichment analysis

To determine a common functional relationship among the top differentially expressed genes we tested for the enrichment of biological processes using Gene Ontology annotations (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2go.gz). The number of genes differentially expressed at $P < 0.05$, $P < 0.01$, $P < 0.001$ and $P < 0.0001$ were 1,964, 749, 185 and 50 respectively. For each P value cutoff, we generated 2,000 random gene sets of equal size (e.g. 1,964 for $P < 0.05$) and performed the same enrichment analysis as on the original dataset. Minimum P values for each enrichment analysis were stored. The 0.05 family-wise error rate (FWER) was then calculated to estimate false positives by setting the 100th (out of 2,000) best P value as the threshold for a true discovery.

We also used an alternate method for pathway enrichment analysis for the identification of common functional categories represented by GO and curated gene sets. The pathways include all the pathways in GO and curated gene sets, which can be downloaded from MsigDB (http://www.broadinstitute.org/gsea/msigdb/collections.jsp#C1). The P value of each gene was determined from a linear mixed model. We then mapped these P values to non-negative z-scores(z_i), where Pr(z>{z-score_{gene}}) = { P value_{gene}}/2.0 (Equation 1), assuming the P values were two-tailed. Then for each pathway, we calculated the P value for a one-sided t-test of *z_i in the pathway ≥. z_i not in the pathway*. We refer to this as the “pathway enrichment test”.

To account for false positives, we first generated 100 sets of balanced permutations, where for each permutation, the permuted case and control groups contain equal number of case and control samples from the original data set. Then, for each permuted dataset, we did the same gene set enrichment test as we did with the original data and stored the *P* values for each pathway in each permuted dataset. We then extract the best *P* value for each permutation, ranked them, and set the 5th best *P* value as the threshold for a true discovery. This gives us a 0.05 FWER.

We also tested whether the genes associated with the risk of autism and intellectual disability (ID) were differentially expressed. Gene lists are in **Table 3.2**, and details on the gene lists are provided below**.** The enrichment of the gene sets categories in the differentially expressed genes was tested based on the hypergeometric distribution model. Four lists of differentially expressed genes at *P* values < 0.001, < 0.005, < 0.01 and < 0.05, with 185, 494, 749 and 1,964 genes, respectively, were generated. The percentage of genes in each gene-set and *P* value corresponding to FWER = 0.05 are tabulated in **Table 3.3**. We performed the "pathway enrichment test" using Equation 1 that does not rely upon defining a differentially expressed set of genes, broadly looking for differential expression (without a *P* value cut-off) among genetically associated genes (**Table 3.4**). We obtained the *P* values at FWER < 0.05 (100 permutations) as described above.

3.2.9 Artifact corrected dataset

For the single gene differential expression analysis we used the ISVs to correct for both technical and biological confounders in the expression data. For the differential co-expression analysis, we identified various sequencing artifacts, computed by Picard command-line tools, which were largely confounding the expression data on the EDASeq-normalized genes (**Table 3.5**). The correlations between the various sequencing artifacts are provided in **Table 3.6**. We used a multivariate linear regression model to correct the gene expression estimates of sequencing artifacts (SA), collection site (CS), sex (S), age (A), and brain region (BR), yielding an artifact corrected (AC) dataset. The sequencing artifacts used in the model were 'percent coding bases', 'percent utr bases', 'percent intronic bases', 'percent intergenic bases', 'median CV coverage', 'median 5' to 3' bias', 'aligned reads' and 'AT dropout'. The correction formula was (Equation 2) as follows assuming we are correcting for only two sequencing artifacts:

$$AC = GeneExpression - \left(\beta_{SA1} * \left(SA1 - mean(SA1)\right)\right) - \left(\beta_{SA2} * \left(SA2 - \right.\right.$$

$$mean(SA2))\Big) - \big(\beta_A * (A - meanA))\big) - (\beta_{CS} * CS) - (\beta_S * S) - (\beta_{BR19} * BR19) - (\beta_{BR10} * BR10) - (\beta_{BR44} * BR44)$$

Equation 2

3.2.10 Combined co-expression analysis

We investigated the entire AC dataset to obtain gene sets or modules that were differentially co-expressed between autism cases and control brains using weighted gene correlation network analysis (WGCNA)[78]. We used WGCNA's "signed" co-expression measure to construct the interconnected gene modules to track the sign of the co-expression information[78]. Pearson's correlations were calculated between 13,443 genes in the 104 samples. The WGCNA method transforms the correlation values to an adjacency matrix using a power function. This power function is selected based on a fit to scale-free topology, and a threshold of 9 (scale-free R^2 of 0.7) was chosen in this study. This power function weights the network by transforming the pair-wise correlation values and computing pairwise topological overlap (TO) between genes[78]. TO is a measure of connection strength between genes. Genes with high TO are clustered into co-expression modules. Each group of interconnected genes is co-expressed and the module is represented by the module eigengene (the first principal component of the module, ME). The connectivity of every gene in every module is represented by correlation to the ME, kME. In this study, this intramodule strength (kME) was ≥ 0.45 for all the modules.

Once the co-expression modules were created, they were numerically labeled by module size, with mod1 denoting the largest module. The co-expression analysis on 13,443 genes from the AC data identified 12 modules, with each module being represented by its first principal component or eigengene (e.g ME1) for each sample. We tested the association of each eigengene with case-control status using a univariate linear mixed regression model, with a random intercept term to account for the correlation among multiple samples derived from the same individual. No additional variables are included in the analysis, since the modules were constructed using the AC data (described above). The multiple test correction threshold using the Bonferroni method was 4.0 x 1^3. The permutation threshold, $P < 2.0 \times 1^3$, was determined by permuting case-control labels (n = 100 permutations) for the eigengene values and re-running the regression analysis. The 5th lowest *P* value was deemed as the study-wide empirical threshold for $P < 0.05$.

Most of the 13,443 genes were clustered into mutually exclusive co-expressing modules. However, 5,075 genes were assigned into the predefined mod0, which is reserved for non-module genes. The Pearson's correlation between modules is shown in **Table 3.7**, and the robustness of correlations was assessed using bootstrap with replacement analysis.

3.2.11 Stratified co-expression analysis

We investigated the global similarity in transcriptome organization in the autism case and control brains by constructing signed networks for the autism case and control brains separately (power function threshold of 9). The construction of the signed networks separately for cases and controls identified 18 modules for each. We utilized the module preservation statistic Zsummary[79], described in the 'modulePreservation' R function implemented in WGCNA, to assess the overlap in network modules obtained from the autism case and control brain datasets. The Zsummary statistic takes into account the overlap in module membership, the density (mean connectivity) and connectivity (sum of connections) patterns of modules. We adopted the recommended significance thresholds: Zsummary < 2 implies no evidence for module preservation, 2 > Zsummary < 10 implies weak to moderate evidence, and Zsummary > 10 implies strong evidence for module preservation. Using the recommended thresholds, we clearly observe that all 18 modules are conserved between the case and control brains (Zsummary > 2, **Table 3.8**).

3.2.12 Gene Ontology analysis

We functionally annotated the 12 modules with Gene Ontology (GO) terms. The enrichment of the GO terms in each of the 12 modules was evaluated based on the hypergeometric test. In order to account for false positives, twelve random modules of the same size were generated 2,000 times, the hypergeometic test was carried out, and the 0.05 FWER was calculated. We have tabulated the GO Term enrichment for each module at FWER = 0.05 in **Tables 3.9-3.19**.

3.2.13 Gene list enrichment analysis

To help provide insight into the interpretation of the gene expression data we compiled gene sets that have been either implicated in ASD[14] or have been designated as markers for specific cell types[80–82]. The main lists in this study are provided in **Table 3.2**, along with sources for each list (gene lists are available at www.arkinglab.org/resources). We present the enrichment of each module, with significance calculated based on the hypergeometric model and implemented in the GeneMerge software package[83]. To account for false positives, the 0.05 FWER was calculated as described above. The enrichment analysis with each module's *P* value corresponding to FWER = 0.05 are tabulated in **Table 3.20**.

3.2.14 Compilation of genetic association genes

As previously discussed with the single gene analysis, the genes associated with autism and intellectual disability were taken from numerous sources presented in **Table 3.2**. The genetic association is presented as independent but not mutually exclusive lists: **1)** 155 genes (**ASD SFARI 2012**) was compiled by Parikshak *et. al.* and is a manually curated set of candidate genes implicated by common variant association, candidate gene studies, genes within ASD-associated CNV, and, to a lesser extent, syndromic forms of ASD. This list from the Simons Foundation Autism Research Initiative Autism (SFARI) AutDB was restricted to genes with strong genetic evidence by also filtering by the category S (syndromic) and evidence levels 1-4 (1 = high confidence, 4 = minimal evidence). The ASD SFARI 2012 list excludes any exome sequencing-implicated RDNV genes; **2)** 235 genes (**ASD SFARI 2014**) from the SFARI AutDB database[84] (accessed July, 2014). The list was restricted to genes with strong genetic evidence by filtering by the category S (syndromic) and evidence levels 1-4 (1 = high confidence, 4 = minimal evidence).; **3)** 197 genes (**ASD SFARI 2014 CV**) are a subset of the ASD SFARI 2014 after removing the 896 genes with rare *de novo* variant from the 4 whole-exome sequencing publications[11,12,14,15] ; **4)** Pinto et al. [85] compiled a list of 124 genes (**ASD [Pinto]**) that have been implicated in ASD and was updated from a list provided by Betancur in 2011[86]. All of the 124 genes have also been implicated in ID. Only autosomal (AD) or X-linked (XL) genes were included. The genes and loci were included only if there was independent evidence from other studies[46]; **5)** 896 **rare *de novo*** variants (RDNV) associated with autism were compiled by Parikshak *et al*. from four whole exome sequencing publications [11,12,14,15]. **6)** Steinberg *et al*.[87] compiled a list of genes disrupted by *de novo* nonsense, frameshift, or splice-site point mutations in autism probands that were obtained from Iossifov et al.[14] (59 genes; referred to as "**I-exomes**") and three other recent studies by Sanders et al.[12], O'Roak et al.[11], and Neale et al.[15] (65 genes combined from all three; referred to as "**SON-exomes**"); **7)** A list of genes disrupted by breakpoints of balanced chromosomal abnormalities (BCAs) observed in individuals with ASD was obtained from Talkowski et al.[88] (32 genes; referred to as "**T-BCAs**").

3.2.15 Other gene list compilations

Voineagu et al. identified two co-expression gene modules that were dysregulated in postmortem ASD brains, **asdM12** (a neuronal module, enriched for ASD associated genes) and **asdM16** (enriched with astrocyte, activated microglial markers, with functional annotation immune response, but no enrichment for ASD associated genes)[20].

Fragile X mental retardation protein (FMRP) and its interacting partners (**FMRP interacting**), implicated with translational regulation of synaptic proteins[89] and shown to be enriched with diverse class of ASD variants[90]**.** Additionally, Steinberg *et al*.[87] demonstrated that the 832 FMRP interacting partners, particularly in two modules (represented as **FMRP 1** and **FMRP 2**), exhibit differential temporal expression: genes in FMRP 1 tend to be specifically upregulated during fetal development, whereas genes in FMRP 2 were generally upregulated in adolescence and adulthood. Numerous cell type markers were tested, as presented in **Table 3.2** and **Table 3.20.** Finally, Uddin et al. identified 3,955 exons mapping to 1,744 genes with high expression in the brain and a low burden of rare mutations, and designate these as 'brain-critical exons'[91].

3.3 Results

3.3.1 Sample Summary

Transcriptomes from 104 human brain cortical tissue samples were resolved using next generation RNA sequencing technology at single gene resolution and through co-expressing gene clusters or modules. Multiple cortical tissues corresponding to Brodmann Area 19 (BA19), Brodmann Area 10 (BA10) and Brodmann Area 44 (BA44) were sequenced in 62, 14, and 28 samples, respectively, resulting in a total of 57 (40 unique individuals) control and 47 (32 unique individuals) autism samples (**Table 3.1** and see[75]).

3.3.2 Differential Gene Expression Analysis

Differential gene expression was estimated between the 57 controls and 47 cases, with sample collection site, age, sex, brain region and independent surrogate variables (ISVs) as fixed effects in a linear mixed regression model. In total, 13,262 genes with at least 3 reads per sample across 90% of the samples were tested, and two transcriptome-wide significant differentially expressed genes associated with autism were identified (**Figure 3.2a**).

The most significant differentially expressed gene was *Myelin And Lymphocyte Protein* (*MAL)* (*P* = 2.16 x 1^{7}), (**Figure 3.2b** and **Figure 3.1a-c**). *MAL*, along with other myelination genes, has previously been reported to show altered expression in patients with psychiatric disorders[92]. The second differentially expressed gene was *C11orf30* (*EMSY)* (*P* = 3.29 ×1^{7}), (**Figure 3.1c** and **Figure 3.2d-f**). This gene has been implicated in chromatin modification, DNA repair, and transcriptional regulation, and previous GWASs have linked *C11orf30* to inflammatory and malignant diseases[93,94]. We also performed analyses at the exon level, testing 21,310 exons for differential exon expression; however, no differences could be identified in autism cases and controls after correction for multiple testing.

Next we asked whether the top differentially expressed genes from the single gene analysis shared common pathways or functional categories. We tested for the enrichment of biological processes using Gene Ontology (GO) annotations and MsigDB curated gene sets. No gene-set enrichment with family-wise error rate (FWER) ≤ 0.05 was observed (**Tables 3.21-3.24**).

For common genetic variation, altered gene expression is a major mode of action[95,96]. We therefore tested whether genes previously associated with autism through genetic analyses are enriched for altered gene expression. To identify genes underlying susceptibility to autism, we utilized a list of expertly curated genes developed by the Simons Foundation for Autism Research (SFARI)[84]. In addition to the SFARI list of genes, we also integrated genes associated with rare *de novo* variation (RDNV) and genes involved in intellectual disability (ID) compiled from four published whole-exome sequencing studies[11,12,14,15] and review articles[21] (**Table 3.2**). Overall, we find comparable expression of these genes in autism and control brain tissue (**Table 3.3** and **Table 3.4**). While these gene lists are not comprehensive and only reflect the current understanding of the genetic basis of autism, the lack of enrichment for genes known to harbor genetic signal for autism in altered gene expression suggests the potential for non-overlapping mechanisms between genetic and transcriptomic determinants of autism. However, we do find modest enrichment for a broader set of genes containing 'brain-critical exons', which have high gene expression in brain and low rare burden of rare mutations, and have been proposed to represent autism candidate genes[91] (**Table 3.3**).

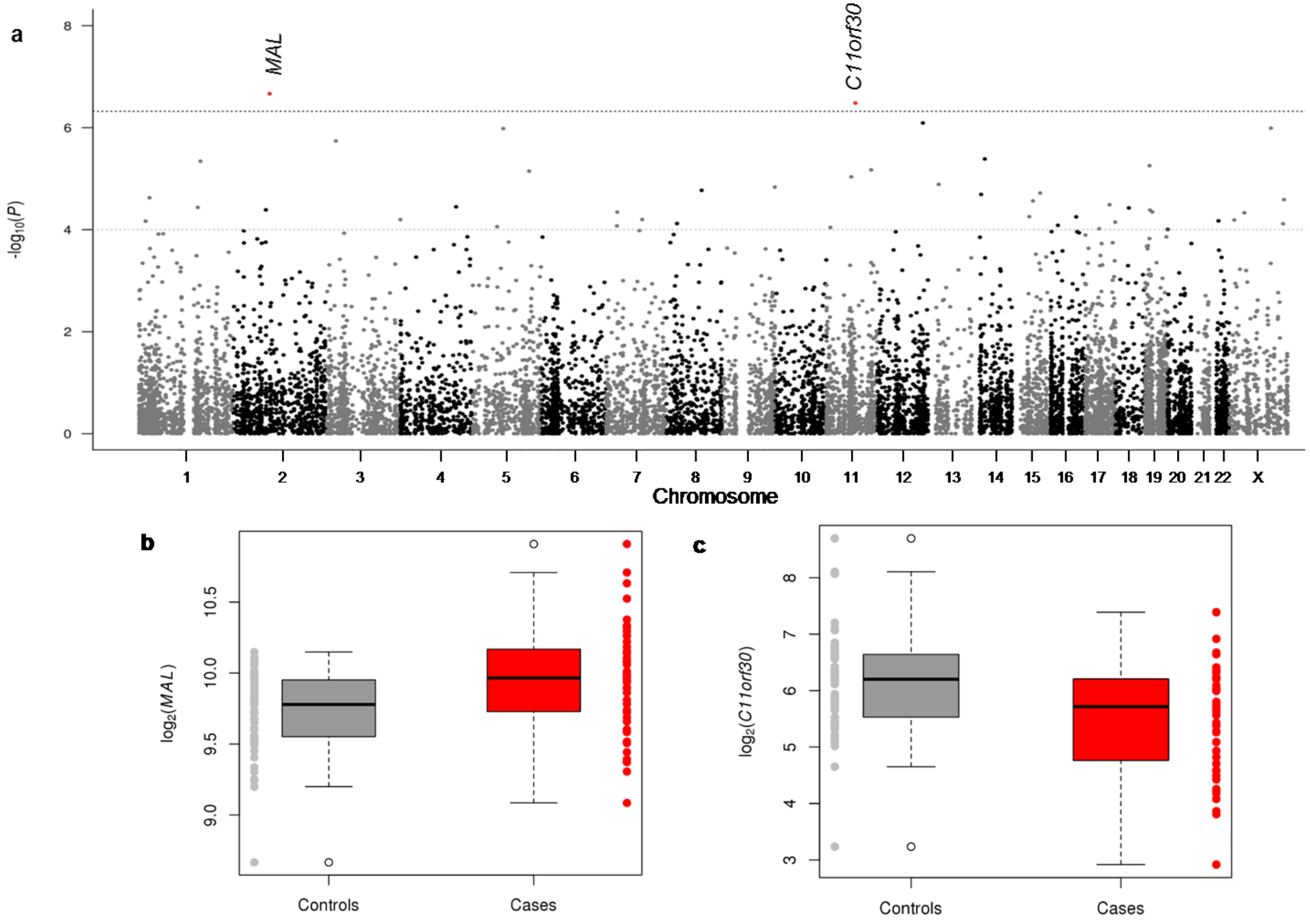


Figure 3.2: Single-gene expression analysis identifies two transcriptome-wide significantly differentially expressed genes between autism brains and control brains.

(**a**) Manhattan plot for 13,262 expressed genes. The threshold for transcriptome-wide significance was calculated based on 400 permutations ($P < 4.76 \times 10^{-7}$) and is indicated by the dotted gray line. (**b, c**) Boxplot of gene expression in 57 controls (gray) and 47 cases (red), indicating a 1.2-fold increase for *MAL*, and a 0.6-fold decrease for *C11orf30* in cases relative to controls.

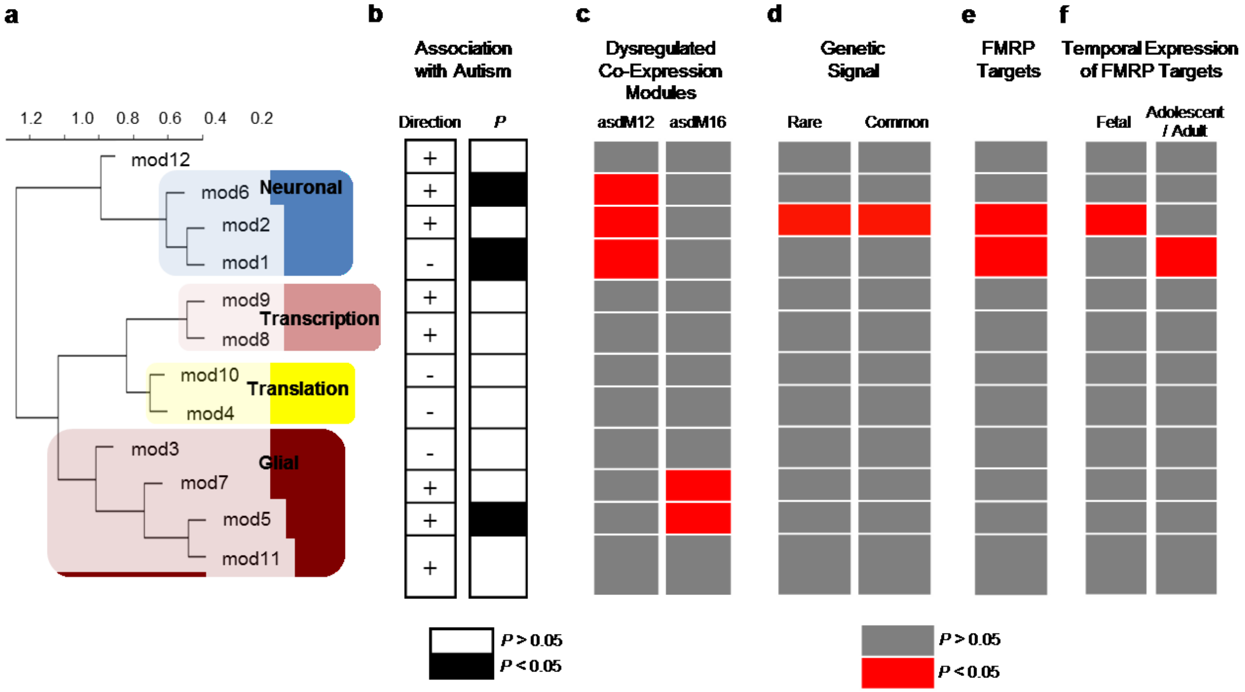


Figure 3.3: Weighted gene correlation network analysis (WGCNA) identifies 12 co-expression modules.

(**a**) Dendogram of 12 co-expressed modules, with major cell type/function enrichment noted (**Table 3.20** and **Tables 3.9-3.19**). mod12 was not significantly enriched for any cell type or GO terms. (**b**) Disease association of each module, represented by each module's first principal component (eigengene). Three co-expression modules are associated with autism with nominal significance (black), with mod5 significant after multi-test correction ($P < 9.64 \times 10^{-4}$) (**Table 3.25**). A positive (+) sign indicates upregulated gene expression in autism cases. (**c-f**) Enrichment analysis for gene lists compiled from the literature (**Table 3.2** and **Table 3.20**). (**c**) Individual dysregulated co-expression modules in autism brains are captured by multiple co-expression modules in the current study, allowing for refinement of the signal associated with autism. (**d**) Genes with known common and rare *de novo* variants associated with autism are enriched only for mod2, which does not show differential expression. (**e**) FMRP targets are enriched in neuronal co-expression modules. (**f**) FMRP targets were split into fetal and adult/adolescent expression patterns, and are captured by different co-expression modules. Red, $P < 0.05$; grey, $P > 0.05$.

3.3.3 Pathway and Functional Enrichment Analyses

In addition to the single gene analyses, we applied weighted gene correlation network analysis[93] (WGCNA) to identify discrete gene modules based on co-expression between genes. Considerable overlap was observed in the modules constructed separately from cases and controls, indicating that overall organization of transcript co-expression is conserved between autism and control brains. Therefore, we applied WGCNA to construct networks derived from the entire dataset of 104 samples, adjusted for sequencing artifacts, age, sex, collection site, and brain region, identifying 12 co-expressed modules. We tested the association of each module, represented by its corresponding first principal component or module eigengene (ME), with case-control status using a linear mixed regression framework (**Table 3.25** and **Figure 3.3b**). Three of the twelve modules were differentially co-expressed ($P < 0.005$), with mod5 ($P = 9.64 \text{ x } 1^{4}$) exceeding the multi-test correction threshold ($P_{\text{permutated}} < 0.002$, $P_{\text{Bonferroni}} < 0.004$) (**Figure 3.4a,b**). Mod5 comprised 759 genes (**Figure 3.3a**) with enrichment for M2-microglial cell states ($P_{\text{hypergeometric}} = 1.22 \text{ x } 1^{39}$) (**Table 3.20** and **Figure 3.4c**) and the GO term 'Type I Interferon pathway' ($P_{\text{hypergeometric}} = 1.19 \text{ x } 1^{20}$) (**Table 3.13** and **Figure 3.4d**). Type I Interferon responses in the brain are classically attributed to viral infections that can produce M1 activation states in microglia[97]. Accordingly, mod5 also shows enrichment for GO terms "defense response to virus" ($P_{\text{hypergeometric}} = 6.83 \text{ x } 1^{17}$) and "cytokine-mediated signaling pathway" ($P_{\text{hypergeometric}} = 6.31 \text{ x } 1^{16}$) (**Table 3.13**). In opposition to M1 activated microglia, M2 responses are responsible for mediating anti-inflammatory remediation responses to damage caused by viral infections. M2 microglial cells also secrete BDNF, increase the production of neural progenitor cells (NPC), and promote myelination[98–100]. These data provide support for a mechanistic connection for viral-infection hypotheses[101] for autism with neural over-growth hypotheses[102] through the novel identification of exaggerated M2 activation states in autism brain tissue.

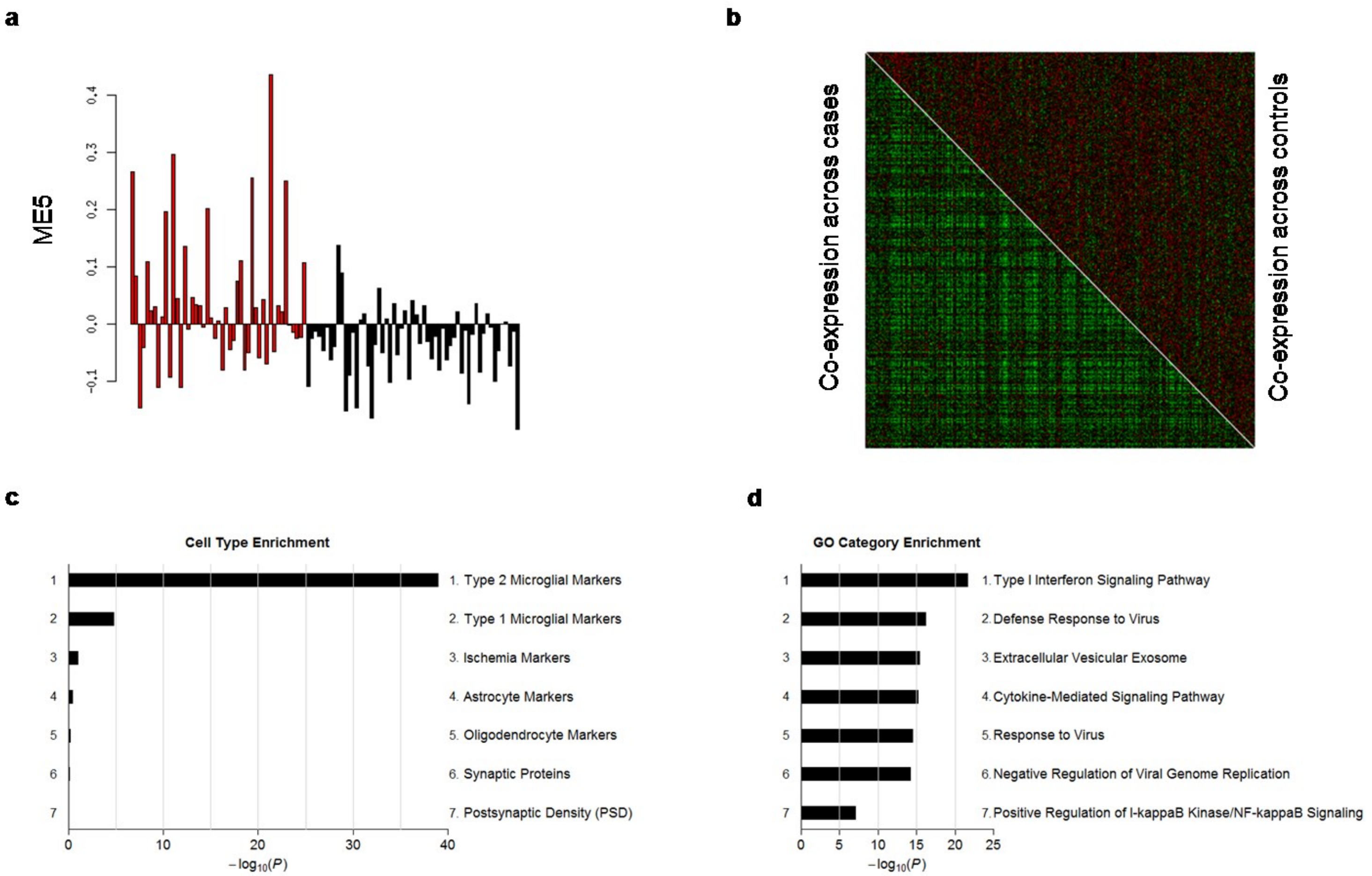


Figure 3.4: **Gene co-expression module mod5 is associated with autism.**

(**a**) The module eigengene (ME) of mod5 is upregulated in autism cases (red) compared to controls (black). (**b**) Heatmap of mod5 co-expression for 759 genes, stratified by disease status, showing greater co-expression between cases (bottom left) compared to controls (upper right). (**c**) mod5 is significantly enriched for genes associated with M2-microglial cell states (**Table 3.20**). (**d**) mod5 is significantly enriched for GO terms related to immune response (**Table 3.13**).

Voineagu *et al.* previously reported a co-expression module dysregulated in autism brains, termed asdM16, enriched in astrocytes and microglia-expressed genes[20]. To better understand the functional implications of asdM16 in autism, we looked for asdM16 signal enrichment amongst the modules generated utilizing our substantially larger data set (**Table 3.20** and **Figure 3.3b**). Two modules—mod5 ($P_{\text{hypergeometric}}$ = 9.3 x 1^{59}, described above) and mod7 ($P_{\text{hypergeometric}}$ = 1.45 x 1^{89})—were enriched for asdM16 signal. However, mod7 is not differentially expressed with respect to autism (**Table 3.25** and **Figure 3.3b**) and accounts for the astrocyte markers ($P_{\text{hypergeometric}}$ = 1.65 x 1^{75}) (**Table 3.20**), while mod5 is differentially expressed (P = 9.64 x 1^{4}). By substantially increasing the sample size and number of genes evaluated, we are able to accurately pinpoint the relevant signal from the previously-reported asdM16 module as coming from M2-state microglial cells and immunogenic responses (type I interferon responses) (**Table 3.13** and **Figure 3.4d**), and not from astrocytes. To our knowledge, M2 activation state responses have not previously been attributed to the pathogenesis of autism.

We also identified three distinct modules (mod1, mod2, and mod6) (**Figure 3.5b-d**) enriched for neuronal markers that contain genes with the shared GO term, 'synaptic transmission', all of which showed enrichment for an additional co-expression module reported to be dysregulated in autism, asdM12 (**Tables 3.9, 3.10, 3.14, 3.20** and **Figure 3.3c**). Two of the three modules—mod1, down-regulated in autism, and mod6, up-regulated in autism—were nominally differentially co-expressed between the autism and control brain samples (P < 0.005) (**Table 3.25** and **Figure 3.3b**). mod1 contains synaptic transmission genes enriched in GABA-related ion channel activity, whereas mod6 contains genes enriched in peptide and hormone signaling (**Figure 3.6**).

Previous studies have identified an enrichment of genetic association signals in genes selectively expressed in neurons[20,85,91,103]. Here, we find that mod2 was enriched for both common ($P_{\text{hypergeometric}}$ = 2.49 x 1^{06}) and rare classes of autism genetic variants ($P_{\text{hypergeometric}}$ = 4.29 x 1^{4}) but comparably expressed between cases and controls (**Table 3.20, Table 3.25,** and **Figure 3.3d**). That neuronal genes genetically associated with autism do not appear to have altered expression (mod2), coupled with the observation that neuronal genes without genetic signal do appear to be differentially expressed (mod1 and mod6), suggests that autism-associated differentially expressed genes are separable from genetic determinants of autism. Corroborating this idea, a recent study of gene networks in coronary artery disease has shown that genes at the center of the networks, referred to as ‘key drivers’, were largely not GWAS signal genes, suggesting that key regulatory genes may not harbor common inherited variation due to natural selection[104].

Direct evidence for the role of Fragile X mental retardation protein (FMRP) in autism has been provided by Darnell et al. These authors reported that many of the protein interacting partners of FMRP harbor autism-spectrum disease (ASD)-associated common variants[89]. Similarly, Iossifov and colleagues reported an enrichment of ASD RDNVs in FMRP targets[14]. We therefore investigated whether FMRP targets were enriched in any of the co-expression modules detected in autism brain tissue. We report a 20% enrichment of FMRP targets in one of the differentially co-expressed neuronal modules, mod1 ($P_{\text{hypergeometric}} = 1.80 \text{ x } 1^{10}$), and the non-differentially co-expressed neuronal module, mod2, which showed a substantially stronger enrichment of 39% of FMRP targets ($P_{\text{hypergeometric}} = 7.38 \text{ x } 1^{110}$) (**Table 3.20** and **Figure 3.3e**).

Recently, Steinberg et al. organized the FMRP target genes into distinct temporally expressed subpopulations affected by different classes of genetic variation associated with ASD[87]. Based on this classification, we found that mod1 was enriched for FMRP targets expressed in the synapse during adolescence and adulthood ($P_{\text{hypergeometric}} = 4.66 \text{ x } 1^{4}$), while mod2 was enriched for the FMRP targets in the modules that were expressed during fetal development ($P_{\text{hypergeometric}} = 3.32 \text{ x } 1^{4}$) (**Table 3.20** and **Figure 3.3f**). Thus, we again find evidence that the genetic signal is stronger in the non-differentially expressed module (mod2), with a 2-fold enrichment for FMRP targets compared to the differentially expressed mod6, while no enrichment was observed for mod1. Incorporating the temporal data leads to a hypothesis that one important mechanism of action at the neuronal level is that primary mutations may occur in genes important in fetal development (captured by mod2), and altered expression of those genes would not be captured in the current study, where the youngest individual was 2 years of age. These mutations may lead to developmental changes reflected in adolescent and adult expressed genes showing differential expression between cases and controls (mod1 and mod6).

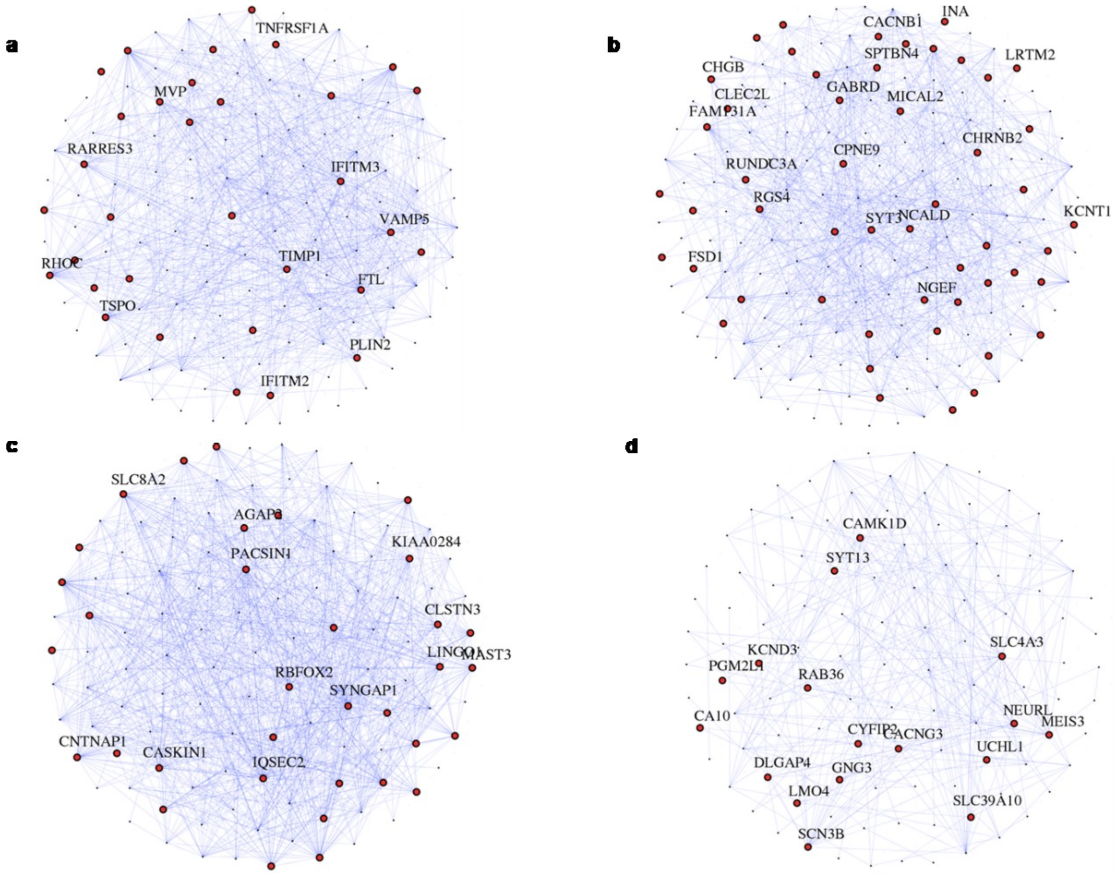
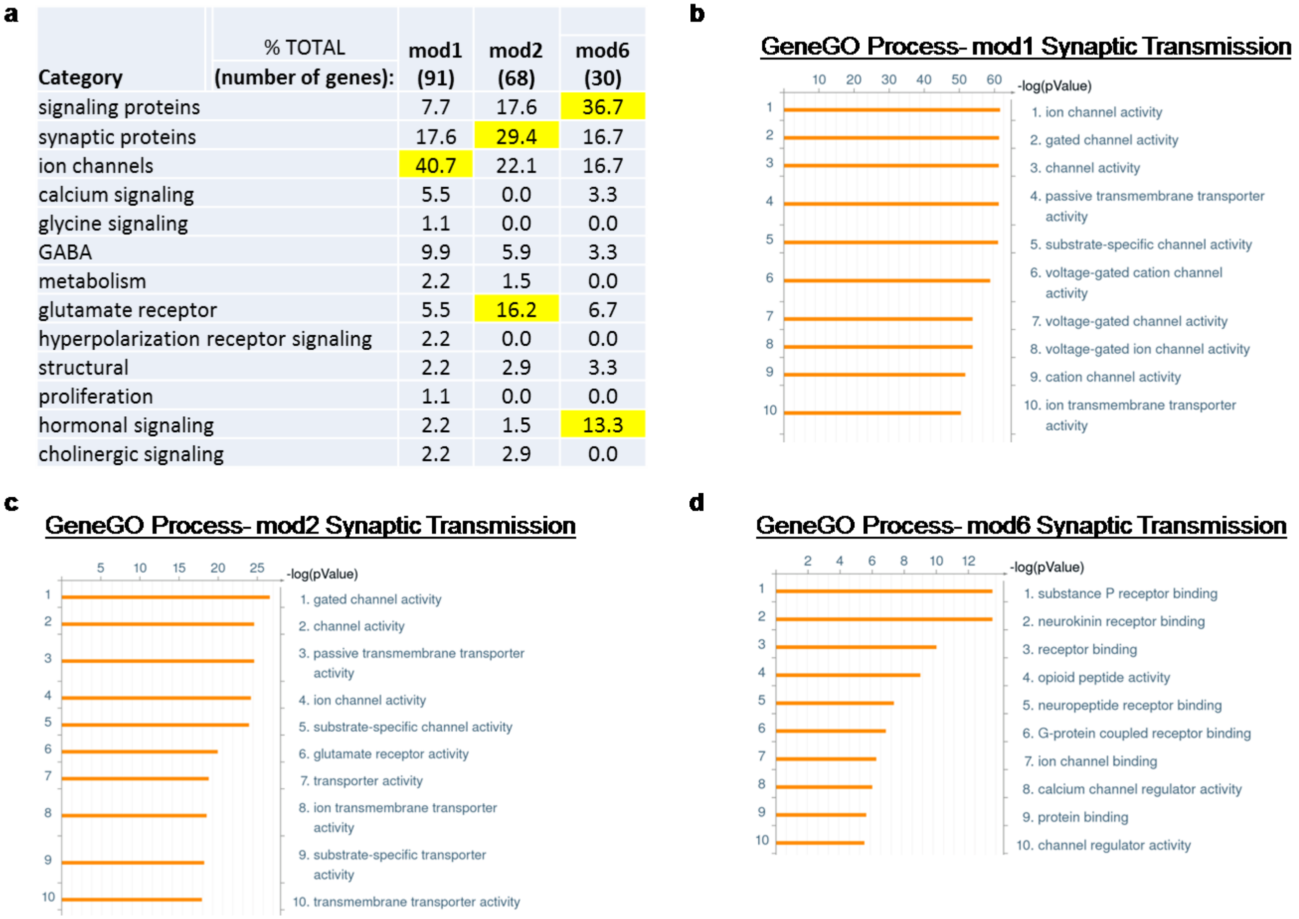


Figure 3.5: Visualization of the (a) mod5, (b) mod1, (c) mod2, and (d) mod6 modules.

The top 150 connections for each module are represented as nodes. Genes with the highest correlation with the module eigengene value are represented by large node sizes.



a

Category	% TOTAL (number of genes):	mod1 (91)	mod2 (68)	mod6 (30)
signaling proteins		7.7	17.6	36.7
synaptic proteins		17.6	29.4	16.7
ion channels		40.7	22.1	16.7
calcium signaling		5.5	0.0	3.3
glycine signaling		1.1	0.0	0.0
GABA		9.9	5.9	3.3
metabolism		2.2	1.5	0.0
glutamate receptor		5.5	16.2	6.7
hyperpolarization receptor signaling		2.2	0.0	0.0
structural		2.2	2.9	3.3
proliferation		1.1	0.0	0.0
hormonal signaling		2.2	1.5	13.3
cholinergic signaling		2.2	2.9	0.0

b
GeneGO Process- mod1 Synaptic Transmission
10 20 30 40 50 60
-log(pValue)
1 2 3 4 5 6 7 8 9 10
1. ion channel activity
2. gated channel activity
3. channel activity
4. passive transmembrane transporter activity
5. substrate-specific channel activity
6. voltage-gated cation channel activity
7. voltage-gated channel activity
8. voltage-gated ion channel activity
9. cation channel activity
10. ion transmembrane transporter activity
c
GeneGO Process- mod2 Synaptic Transmission
5 10 15 20 25
-log(pValue)
1 2 3 4 5 6 7 8 9 10
1. gated channel activity
2. channel activity
3. passive transmembrane transporter activity
4. ion channel activity
5. substrate-specific channel activity
6. glutamate receptor activity
7. transporter activity
8. ion transmembrane transporter activity
9. substrate-specific transporter activity
10. transmembrane transporter activity
d
GeneGO Process- mod6 Synaptic Transmission
2 4 6 8 10 12
-log(pValue)
1 2 3 4 5 6 7 8 9 10
1. substance P receptor binding
2. neurokinin receptor binding
3. receptor binding
4. opioid peptide activity
5. neuropeptide receptor binding
6. G-protein coupled receptor binding
7. ion channel binding
8. calcium channel regulator activity
9. protein binding
10. channel regulator activity

Figure 3.6: 'Synaptic Transmission' GO genes in mod1, mod2 and mod6.
(**a**) Categorical assignments of genes within the 'Synaptic Transmission' GO term of mod1, mod2 and mod6. (**b**) Enrichment of ion channel activity (GABA-related) in mod1. (**c**) Enrichment of synaptic and glutamate receptor signaling in mod2. (**d**) Enrichment in peptide and hormone signaling in mod6. mod1 is down-regulated in autism, mod2 and mod6 are upregulated in autism (Table 3.25 and Fig. 3.3).

3.4 Conclusions

In this study we provide transcriptomic evidence for type-I interferon and M2-activation state abnormalities in autism that may lead to a variety of pathologic and phenotypic consequences. We further note that there is a strong negative correlation between two differentially co-expressed modules, mod5 (activated M2-state microglia genes) and mod1 (synaptic transmission genes) (r = -0.92, **Table 3.7**). Recently, microglia have been identified as cells capable of restoring neural function in the ASD-model *MECP2* knockout mice [105]. We observe, for the first time, that M2-activation state microglia genes, in particular, are altered in autism, potentially driven by type I interferon responses. This process may drive changes in NPC proliferation and connectivity with resultant altered activity-dependent neural expression profiles in post-natal development [106,107]. The linkage of this pathway to autism may lead to more accurate and predictive models of idiopathic disease that might contribute to the identification of effective therapeutic approaches.

3.5 Tables

Table 3.1: Sample information

Brain Number	**Old Brain Bank ID**	**Sample Code**	**Site**	**Diagnosis**	**Sex**	**Age**	**Ethnicity**	PMI **(hrs)**	**Sample Excluded**
1	B-4925	AN16641	Harvard	Autism	Male	9	W	27	
2	B-5000	AN00493	Harvard	Autism	Male	27	W	8.3	
3	B-5144	AN00764	Harvard	Autism	Male	20	W	23.7	
4	B-5173	AN08792	Harvard	Autism	Male	30	W	20.3	
6	B-5505	AN01227	Harvard	Autism	Male	82	W	24.67	
7	B-5562	AN14613	Harvard	Autism	Male	39	W	22.75	
8	B-5569	AN08873	Harvard	Autism	Male	5	W	25.5	
9	B-5666	AN19511	Harvard	Autism	Male	8	W	22.2	
11	B-6184	AN01570	Harvard	Autism	Female	18	W	6.75	
12	B-6294	AN17138	Harvard	15qdup	Male	16	A	NA	YES
13	B-6337	AN09730	Harvard	Autism	Male	22	W	25	
14	B-6399	AN03345	Harvard	Autism	Male	2	W	4	
15	B-6469	AN17777	Harvard	Autism	Female	49	NA	16.33	
16	B-6640	AN12457	Harvard	Autism	Female	29	W	17.83	
17	B-6677	AN11989	Harvard	Autism	Male	30	W	16.06	
19	B-6756	AN07591	Harvard	Seizures	Male	16	NA	22	YES
21	B-6856	AN14829	Harvard	15qdup	Female	26	NA	28.67	YES
22	B-6994	AN08166	Harvard	Autism	Male	28	NA	43.25	
23	B-7002	AN13872	Harvard	Autism	Female	5	NA	33	
24	B-7014	AN09402	Harvard	15qdup	Male	11	W	10.5	YES
25	B-7078	AN17678	Harvard	Autism	Male	11	W	NA	
26	B-7079	AN04682	Harvard	Autism	Male	15	NA	23.23	
27	B-7085	AN03632	Harvard	Autism	Female	49	NA	21.08	

28	B-7090	AN09714	Harvard	Autism	Male	60	NA	26.5	
29	B-7109	AN17254	Harvard	Autism	Male	51	NA	22.16	
30	B-4756	AN10606	Harvard	control	Male	56	NA	23	YES
31	B-4786	AN16665	Harvard	control	Male	36	W	20	
32	B-4981	AN01357	Harvard	control	Male	42	W	18.33	
33	B-5386	AN02583	Harvard	control	Male	68	NA	16.58	
34	B-5813	AN01410	Harvard	control	Male	41	NA	27.17	
35	B-6004	AN15240	Harvard	control	Female	36	NA	18.08	
36	B-6076	AN08677	Harvard	control	Male	38	NA	25.47	
37	B-6078	AN07176	Harvard	control	Male	21	NA	29.91	
38	B-6207	AN17425	Harvard	control	Male	16	NA	26.16	
39	B-6221	AN14368	Harvard	control	Male	22	NA	24.2	
40	B-6316	AN15566	Harvard	control	Female	32	NA	28.92	
41	B-6860	AN13295	Harvard	control	Male	56	NA	22.12	
42	797	UMB797	Maryland	Autism	Male	9	W	13	
43	1182	UMB1182	Maryland	Autism	Female	9	AA	24	YES
44	1349	UMB1349	Maryland	Autism	Male	5	W	39	
45	1638	UMB1638	Maryland	Autism	Female	20	W	50	
46	4231	UMB4231	Maryland	Autism	Male	8	AA	12	
47	4721	UMB4721	Maryland	Autism	Male	8	AA	16	
48	4849	UMB4849	Maryland	Autism	Male	7	AA	20	
49	4899	UMB4899	Maryland	Autism	Male	14	W	9	YES
50	4999	UMB4999	Maryland	Autism	Male	20	W	14	
51	4671	UMB4671	Maryland	Autism	Female	4	AA	13	
52	451	UMB451	Maryland	control	Male	4	W	15	
53	497	UMB497	Maryland	control	Male	12	W	16	
54	662	UMB662	Maryland	control	Female	12	W	18	YES
55	1185	UMB1185	Maryland	control	Male	4	W	17	
56	1377	UMB1377	Maryland	control	Female	5	W	20	

57	1500	UMB1500	Maryland	control	Male	6	W	18	
58	1674	UMB1674	Maryland	control	Male	NA	NA	NA	
60	4670	UMB4670	Maryland	control	Male	4	W	17	
61	4898	UMB4898	Maryland	control	Male	7	W	12	
62	1323	UMB1323	Maryland	control	Male	16	W	25	
63	1409	UMB1409	Maryland	control	Male	18	W	6	
64	1429	UMB1429	Maryland	control	Male	18	W	9	
65	1465	UMB1465	Maryland	control	Male	17	W	4	
66	1322	UMB1322	Maryland	control	Male	16	W	25	
67	1541	UMB1541	Maryland	control	Female	20	W	19	
68	1543	UMB1543	Maryland	control	Male	17	W	22	
69	1571	UMB1571	Maryland	control	Female	18	W	8	
70	1584	UMB1584	Maryland	control	Female	18	W	15	
71	1712	UMB1712	Maryland	control	Male	20	W	8	
72	1790	UMB1790	Maryland	control	Male	13	W	18	
73	1796	UMB1796	Maryland	control	Male	16	W	16	
74	1823	UMB1823	Maryland	control	Male	15	W	18	
75	1841	UMB1841	Maryland	control	Male	19	W	14	
76	1843	UMB1843	Maryland	control	Female	15	W	9	
77	1862	UMB1862	Maryland	control	Male	20	W	6	YES
78	1908	UMB1908	Maryland	control	Male	13	W	13	
79	1944	UMB1944	Maryland	control	Female	16	W	20	
80	4590	UMB4590	Maryland	control	Male	20	W	19	
81	4591	UMB4591	Maryland	control	Female	16	W	14	
82	4669	UMB4669	Maryland	control	Male	16	W	16	
83	4724	UMB4724	Maryland	control	Female	16	W	15	YES
84	4727	UMB4727	Maryland	control	Male	20	W	5	
85	4728	UMB4728	Maryland	control	Male	17	W	23	
86	AN01093	AN01093	Harvard	Autism	Male	56	NA	19	

87	AN03935	AN03935	Harvard	15qdup	Male	20	NA	28	YES
88	AN06420	AN06420	Harvard	Autism	Male	39	NA	14	
89	AN16115	AN16115	Harvard	Autism	Female	11	NA	13	

Table 3.2: Gene Lists

Gene categories	Source	Description
Neuronal Markers	Cahoy et al.[80]	Supplementary Table 6 in Cahoy et al.[80]
Oligodendrocyte Markers	Cahoy et al.[80]	Supplementary Table 5 in Cahoy et al.[80]
Astrocyte Markers	Cahoy et al.[80]	Supplementary Table 4 in Cahoy et al.[80]
Type 1 Microglial Markers	userListEnrichment function in WGCNA[93]	userListEnrichment is a pre-defined or user-defined collections of brain- and blood-related lists curated from the literature part of the WGCNA package. This specific list of cell class genes was compiled by Miller et al., 2010[81]
Type 2 Microglial Markers	userListEnrichment function in WGCNA[93]	userListEnrichment is a pre-defined or user-defined collections of brain- and blood-related lists curated from the literature part of the WGCNA package. This specific list of cell class genes was compiled by Miller et al., 2010[81]
Ischemia Markers	Nagata et al.[90]	
Synaptic Proteins	Bayes et al.[5]	PMID: 21170055
Postsynaptic Density (PSD)	Bayes et al.	PMID: 21170055
ASD SFARI 2012	Parikshak et al.[21]	manually curated list of genes with strong genetic evidence (SFARI category S and evidence levels 1-4); includes genes harboring CNVs associated with autism, genes implicated by candidate gene studies, and syndromic autism genes, excludes RDNV genes
ASD SFARI 2014	(https://gene.sfari.org /autdb/)[84]	
ASD SFARI	(https://gene.sfari.org	subset of ASD SFARI 2012 after removing genes with RDNVs

2014 CV	/autdb)[84]	
ASD (Pinto)	Pinto et al[85] (Table S6a)	genes implicated in both ASD and ID
Rare de novo	Parikshak et al.[21]	RDNVs associated with autism
I-exomes	Steinberg et al.[87]	genes disrupted by *de novo,* nonsense, frameshift, or splice-site point mutations in autism probands
SON-exomes	Steinberg et al. [87]	genes disrupted by *de novo,* nonsense, frameshift, or splice-site point mutations in autism probands
AGP	Steinberg et al. [87]	rare deletion CNVs in 561 autism probands with "strict ASD"
San	Steinberg et al. [87]	rare deletion CNVs in 1124 autism probands with "strict ASD"
asdM12	Voineagu at al.[20]	a neuronal module, enriched for ASD-associated genes
asdM16	Voineagu at al. [20]	astrocyte and activated microglial maker module, with functional annotation immune response but no enrichment for ASD-associated genes
FMRP interacting	Parikshak et al. [21]	original source of this list is Darnell et al.[12]
FMRP 1	Steinberg et al. [87]	a module upregulated during fetal development obtained after clustering 832 FMRP targets in human brain transcriptome (Brainspan)
FMRP 2	Steinberg et al. [87]	a module upregulated during adolescence and adulthood obtained after clustering 832 FMRP targets in human brain transcriptome (Brainspan)
ID 2009	Parikshak et al. [21]	intellectual disability genes
ID (Pinto)	Pinto et al[85] (Table S6c)	genes implicated in ID, but not yet in ASD
Brain Critical	Uddin et al.[91]	1,744 genes containing exons with high brain expression and low burden of rare mutations
T-BCAs	Steinberg et al. [87]	genes with breakpoints of balanced abnormalities

Table 3.3: Enrichment analysis of 1964, 749, 494 , 185 genes differentially expressed at P < 0.05, P < 0.01, P < 0.005, P < 0.001 , respectively. P value and Ratio (number of genes in Study group/Number of genes transcriptome-wide, in parentheses) for each gene set.

Differentially expressed genes (*P* value threshold for FWER =0.05)	**Number of Genes**	**ASD SFARI 2012**	**ASD SFARI 2014**	**ASD SFARI 2014 CV**	**ASD (Pinto)**	**Rare *de novo***	**I-exomes**	**SON-exomes**	**AGP**	**San**	**ID 2009**	**ID (Pinto)**	**Brain Critical**	**T-BCAs**
Genes differentially expressed at *P* < 0.05 (1.65e-04)	1,964	0.63 (0.16)	0.56 (0.14)	0.47 (0.15)	0.17 (0.18)	0.56 (0.14)	0.54 (0.16)	NA	0.51 (0.16)	0.49 (0.15)	0.30 (0.15)	0.95 (0.11)	0.06 (0.16)	NA
Genes differentially expressed at *P* < 0.01 (5.94e-04)	749	0.10 (0.08)	0.13 (0.07)	0.06 (0.09)	0.32 (0.06)	0.46 (0.05)	NA	NA	NA	0.84 (0.03)	0.71 (0.05)	0.99 (0.02)	1.00E-03 (0.07)	NA
Genes differentially expressed at *P* < 0.005 (4.40e-04)	494	0.24 (0.05)	0.12 (0.05)	0.09 (0.06)	0.25 (0.05)	0.43 (0.04)	NA	NA	NA	0.64 (0.03)	0.60 (0.03)	0.99 (0.009)	3.00E-04 (0.05)	NA
Genes differentially expressed at *P* < 0.001 (3.87e-04)	185	0.47 (0.01)	0.71 (0.01)	0.60 (0.013)	0.48 (0.01)	0.98 (0.006)	NA	NA	NA	NA	0.75 (0.01)	NA	3.70E-03 (0.02)	NA

NA indicates that modules contain one or fewer genes in the category being tested.

Table 3.4: Gene Set enrichment analysis to evaluate expression at the level of genetically associated genes.
Reported P values calculated by a one-sided t-test for z in each gene set ≥ z score not in the gene set. The single gene expression P were mapped to Z score using equation 1 (see Methods). At FWER < 0.05, P value threshold after 100 permutations was 0.005.

Gene Sets	***P* value**
ASD SFARI 2012	0.68
ASD SFARI 2014	0.89
ASD SFARI 2014 CV	0.84
ASD (Pinto)	0.83
Rare *de novo*	0.73
I-exomes	0.98
SON-exomes	0.73
AGP	0.57
San	0.25
ID 2009	0.36
ID (Pinto)	0.01
Brain Critical	0.99
T-BCAs	0.63

Table 3.5: Correlation (r, spearman correlation) of the ISV with sequencing artifacts.
Highlighted cells have r +/- > 0.6.

ISVs	% CODING BASES	% UTR BASES	% INTRONIC BASES	% INTERGENIC BASES	% mRNA BASES	% USABLE BASES	MEDIAN CV COVERAGE	MEDIAN 5' BIAS	MEDIAN 3' BIAS	MEDIAN 5' TO 3' BIAS	ALIGNED READS	AT DROPOUT	GC DROPOUT
ISV1	-0.36	-0.01	0.36	0.09	-0.23	-0.23	0.34	-0.38	0.27	-0.38	-0.02	-0.12	-0.02
ISV2	-0.54	-0.27	0.65	0.21	-0.47	-0.47	0.65	-0.58	0.37	-0.52	-0.26	-0.06	0.17
ISV3	0.09	0.02	0.02	-0.06	0.06	0.06	0.01	0.09	-0.11	0.11	0.03	-0.07	-0.04
ISV4	0.49	0.05	-0.48	-0.32	0.46	0.46	-0.28	0.39	-0.37	0.41	0.03	0.17	-0.01
ISV5	-0.57	0.26	0.33	0.6	-0.62	-0.62	-0.04	-0.33	0.44	-0.4	0.32	-0.76	-0.1
ISV6	0.51	0	-0.46	-0.51	0.58	0.58	-0.17	0.32	-0.34	0.35	-0.04	0.32	0.11
ISV7	-0.53	-0.1	0.66	0.14	-0.38	-0.38	0.34	-0.54	0.56	-0.59	-0.1	-0.07	0.08
ISV8	0.38	-0.69	0.15	-0.35	0.16	0.16	0.32	0.26	-0.55	0.4	-0.62	0.69	-0.02
ISV9	-0.72	0.06	0.5	0.14	-0.35	-0.35	0.61	-0.85	0.81	-0.89	-0.02	-0.16	0.38
ISV10	0.15	0.51	-0.22	0.27	-0.07	-0.07	-0.51	0.29	-0.21	0.26	0.49	-0.28	-0.39
ISV11	-0.4	-0.01	0.36	0.4	-0.46	-0.46	0.1	-0.3	0.26	-0.31	0.03	-0.36	-0.13

Table 3.6: Correlation (r, spearman correlation) between the sequencing artifacts metrics.
The cells highlighted have an r +/- > 0.80.

	% CODING BASES	% UTR BASES	% INTRONIC BASES	% INTERGENIC BASES	% mRNA BASES	% USABLE BASES	MEDIAN CV COVERAGE	MEDIAN 5' BIAS	MEDIAN 3' BIAS	MEDIAN 5' TO 3' BIAS	ALIGNED READS	AT DROPOUT	GC DROPOUT
% CODING BASES	1	-0.11	-0.64	-0.67	**0.83**	**0.83**	-0.36	0.78	**-0.88**	**0.88**	-0.2	0.63	-0.18
% UTR BASES	-0.11	1	-0.12	-0.45	0.39	0.39	0.33	-0.44	0.4	-0.45	0.01	0.14	0.41
% INTRONIC BASES	-0.64	-0.12	1	0.3	-0.64	-0.64	0.49	-0.51	0.54	-0.56	-0.16	-0.21	0.03
% INTERGENIC BASES	-0.67	-0.45	0.3	**1**	**-0.89**	**-0.89**	-0.17	-0.24	0.41	-0.34	0.4	**-0.8**	-0.25
% mRNA BASES	**0.83**	0.39	-0.64	**-0.89**	**1**	**1**	-0.12	0.45	-0.55	0.54	-0.21	0.66	0.11
% USABLE BASES	**0.83**	0.39	-0.64	**-0.89**	**1**	**1**	-0.12	0.45	-0.55	0.54	-0.21	0.66	0.11
MEDIAN CV COVERAGE	-0.36	0.33	0.49	-0.17	-0.12	-0.12	1	-0.66	0.29	-0.52	-0.51	0.29	0.48
MEDIAN 5' BIAS	0.78	-0.44	-0.51	-0.24	0.45	0.45	-0.66	1	-0.74	**0.94**	0.04	0.27	-0.43
MEDIAN 3' BIAS	**-0.88**	0.4	0.54	0.41	-0.55	-0.55	0.29	-0.74	1	**-0.91**	0.26	-0.54	0.25
MEDIAN 5' TO 3' BIAS	**0.88**	-0.45	-0.56	-0.34	0.54	0.54	-0.52	**0.94**	**-0.91**	**1**	-0.11	0.41	-0.35
ALIGNED READS	-0.2	0.01	-0.16	0.4	-0.21	-0.21	-0.51	0.04	0.26	-0.11	**1**	-0.58	-0.05
AT DROPOUT	0.63	0.14	-0.21	**-0.8**	0.66	0.66	0.29	0.27	-0.54	0.41	-0.58	1	0.16
GC DROPOUT	-0.18	0.41	0.03	-0.25	0.11	0.11	0.48	-0.43	0.25	-0.35	-0.05	0.16	1

	ME12	**ME6**	**ME2**	**ME1**	**ME9**	**ME8**	**ME10**	**ME4**	**ME3**	**ME7**	**ME5**	**ME11**
ME12	1.00	0.10	0.36	0.17	-0.19	-0.11	-0.19	-0.46	-0.04	-0.24	-0.14	-0.12
ME6	0.10	1.00	0.59	0.60	-0.11	0.57	0.54	-0.27	**-0.82**	-0.42	-0.47	-0.25
ME2	0.36	0.59	1.00	0.71	-0.61	-0.11	-0.10	-0.85	-0.30	-0.50	-0.65	-0.43
ME1	0.17	0.60	0.71	1.00	-0.23	0.26	0.37	-0.49	-0.41	-0.66	**-0.92**	-0.62
ME9	-0.19	-0.11	-0.61	-0.23	1.00	0.71	0.19	0.49	-0.01	0.21	0.25	0.08
ME8	-0.11	0.57	-0.11	0.26	0.71	1.00	0.54	0.21	-0.60	-0.15	-0.15	-0.15
ME10	-0.19	0.54	-0.10	0.37	0.19	0.54	1.00	0.50	-0.67	-0.21	-0.28	-0.15
ME4	-0.46	-0.27	-0.85	-0.49	0.49	0.21	0.50	1.00	0.00	0.39	0.49	0.37
ME3	-0.04	-0.82	-0.30	-0.41	-0.01	-0.60	-0.67	0.00	1.00	0.28	0.22	0.04
ME7	-0.24	-0.42	-0.50	-0.66	0.21	-0.15	-0.21	0.39	0.28	1.00	0.66	0.27
ME5	-0.14	-0.47	-0.65	**-0.92**	0.25	-0.15	-0.28	0.49	0.22	0.66	1.00	0.72
ME11	-0.12	-0.25	-0.43	-0.62	0.08	-0.15	-0.15	0.37	0.04	0.27	0.72	1.00

Table 3.7: Correlation between the 12 module eigengenes (ME) in the network

Table 3.8: Association of co-expression modules with disease status

Module	Module Size	$Z_{summary}$
mod7	584	25.0
mod2	1000	34.4
mod3	1000	27.3
mod13	248	4.6
mod5	622	25.9
mod11	339	9.9
mod0	121	6.2
mod14	140	4.4
mod15	99	5.9
mod9	420	13.4
mod16	233	15.1
mod8	432	10.2
mod10	374	19.2
mod6	602	21.8
mod17	260	13.9
mod12	312	13.5
mod1	1000	30.5
mod4	1000	19.2

Table 3.9: mod1 and enrichment of Gene Ontology categories with FWER < 0.05.
The P value threshold after 2000 permutations was 4.80e-05

GMRG Term	**Description**	***P* value (hypergeometric test)**	**Ratio (number of genes in Study group/Number of genes transcriptome-wide)**
GO:0007268	synaptic transmission	3.74E-19	0.32
GO:0008076	voltage-gated potassium channel complex	2.05E-12	0.57
GO:0030054	cell junction	8.83E-12	0.26
GO:0030672	synaptic vesicle membrane	1.05E-10	0.55
GO:0045211	postsynaptic membrane	1.13E-09	0.31
GO:0030426	growth cone	3.96E-09	0.37
GO:0005251	delayed rectifier potassium channel activity	9.38E-09	0.63
GO:0005249	voltage-gated potassium channel activity	8.68E-08	0.56
GO:0006813	potassium ion transport	1.91E-07	0.40
GO:0050796	regulation of insulin secretion	2.53E-07	0.38
GO:0061202	clathrin-sculpted gamma-aminobutyric acid transport vesicle membrane	3.03E-06	0.88
GO:0071805	potassium ion transmembrane transport	5.36E-06	0.48
GO:0001975	response to amphetamine	6.67E-06	0.59
GO:0030425	dendrite	1.13E-05	0.23
GO:0005886	plasma membrane	1.20E-05	0.15
GO:0008021	synaptic vesicle	2.60E-05	0.30
GO:0005516	calmodulin binding	3.79E-05	0.25

Table 3.10: mod2 and enrichment of Gene Ontology categories with FWER < 0.05.
The P value threshold after 2000 permutations was 6.37e-05.

GMRG Term	Description	*P* value (hypergeometric test)	Ratio (number of genes in Study group/Number of genes transcriptome-wide)
GO:0030054	cell junction	4.14E-15	0.25
GO:0005886	plasma membrane	5.90E-14	0.15
GO:0007268	synaptic transmission	3.49E-12	0.24
GO:0045211	postsynaptic membrane	4.74E-12	0.30
GO:0016021	integral component of membrane	1.02E-10	0.14
GO:0007411	axon guidance	3.25E-09	0.22
GO:0030425	dendrite	3.45E-08	0.23
GO:0043197	dendritic spine	8.76E-08	0.33
GO:0034220	ion transmembrane transport	3.15E-07	0.27
GO:0007156	homophilic cell adhesion	1.36E-06	0.31
GO:0000139	Golgi membrane	1.78E-06	0.18
GO:0030424	axon	1.96E-06	0.23
GO:0043005	neuron projection	2.15E-06	0.24
GO:0048813	dendrite morphogenesis	2.39E-06	0.43
GO:0035235	ionotropic glutamate receptor signaling pathway	2.56E-06	0.50
GO:0043025	neuronal cell body	2.60E-06	0.21
GO:0045202	synapse	4.47E-06	0.24
GO:0004674	protein serine/threonine kinase activity	6.04E-06	0.19
GO:0014069	postsynaptic density	9.15E-06	0.26
GO:0042734	presynaptic membrane	2.57E-05	0.33
GO:0035249	synaptic transmission, glutamatergic	3.43E-05	0.38
GO:0070509	calcium ion import	3.75E-05	0.47

GO:0007155	cell adhesion	3.80E-05	0.18
GO:0007626	locomotory behavior	5.28E-05	0.30
GO:0006112	energy reserve metabolic process	5.33E-05	0.26
GO:0032320	positive regulation of Ras GTPase activity	6.35E-05	0.50

Table 3.11: mod3 and enrichment of Gene Ontology categories with FWER < 0.05.
The P value threshold after 2000 permutations was 4.70e-05.

GMRG Term	Description	*P* value (hypergeometric)	Ratio (number of genes in Study group/Number of genes transcriptome-wide)
GO:0016021	integral component of membrane	1.38E-09	0.11
GO:0005886	plasma membrane	3.13E-09	0.12
GO:0070062	extracellular vesicular exosome	2.38E-07	0.12
GO:0042552	myelination	5.62E-06	0.36
GO:0009986	cell surface	8.67E-06	0.17
GO:0005887	integral component of plasma membrane	1.72E-05	0.13
GO:0006897	endocytosis	1.88E-05	0.22
GO:0016324	apical plasma membrane	3.17E-05	0.20
GO:0019911	structural constituent of myelin sheath	4.36E-05	1.00
GO:0042246	tissue regeneration	4.52E-05	0.60

Table 3.12: mod4 and enrichment of Gene Ontology categories with FWER < 0.05.
The P value threshold after 2000 permutations was 4.69e-05.

GMRG Term	Description	*P* value (hypergeometric)	Ratio (number of genes in Study group/Number of genes transcriptome-wide)
GO:0006414	translational elongation	2.65E-78	0.85
GO:0019083	viral transcription	7.74E-77	0.89
GO:0006415	translational termination	2.64E-74	0.85
GO:0003735	structural constituent of ribosome	1.09E-69	0.60
GO:0006614	SRP-dependent cotranslational protein targeting to membrane	1.20E-67	0.71
GO:0006413	translational initiation	1.38E-65	0.66
GO:0019058	viral life cycle	2.50E-62	0.65
GO:0000184	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	9.44E-60	0.64
GO:0005840	ribosome	1.47E-53	0.55
GO:0016070	RNA metabolic process	1.19E-51	0.39
GO:0016071	mRNA metabolic process	1.63E-49	0.40
GO:0022625	cytosolic large ribosomal subunit	2.72E-42	0.82
GO:0010467	gene expression	2.78E-35	0.20
GO:0016032	viral process	3.95E-35	0.23
GO:0044267	cellular protein metabolic process	7.89E-35	0.24
GO:0022627	cytosolic small ribosomal subunit	2.19E-34	0.89
GO:0044822	poly(A) RNA binding	3.45E-23	0.14
GO:0022904	respiratory electron transport chain	8.22E-21	0.41
GO:0015935	small ribosomal subunit	3.78E-20	0.77
GO:0005739	mitochondrion	9.88E-15	0.12

GO:0044237	cellular metabolic process	1.53E-14	0.28
GO:0005829	cytosol	2.11E-14	0.10
GO:0003723	RNA binding	3.94E-13	0.16
GO:0005743	mitochondrial inner membrane	1.61E-12	0.18
GO:0005747	mitochondrial respiratory chain complex I	1.49E-09	0.39
GO:0004129	cytochrome-c oxidase activity	6.10E-09	0.52
GO:0008137	NADH dehydrogenase (ubiquinone) activity	6.66E-09	0.39
GO:0005689	U12-type spliceosomal complex	2.09E-08	0.48
GO:0070062	extracellular vesicular exosome	2.77E-08	0.10
GO:0000398	mRNA splicing, via spliceosome	2.79E-08	0.18
GO:0002479	antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP-dependent	5.34E-08	0.28
GO:0005687	U4 snRNP	8.17E-08	0.78
GO:0006120	mitochondrial electron transport, NADH to ubiquinone	9.63E-08	0.35
GO:0042590	antigen processing and presentation of exogenous peptide antigen via MHC class I	1.16E-07	0.27
GO:0019843	rRNA binding	2.88E-07	0.43
GO:0051436	negative regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle	9.38E-07	0.26
GO:0005759	mitochondrial matrix	1.36E-06	0.15
GO:0042274	ribosomal small subunit biogenesis	1.54E-06	0.58
GO:0042273	ribosomal large subunit biogenesis	1.54E-06	0.58
GO:0000502	proteasome complex	1.64E-06	0.26
GO:0030529	ribonucleoprotein complex	1.91E-06	0.19
GO:0006521	regulation of cellular amino acid metabolic process	2.19E-06	0.28
GO:0051437	positive regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle	2.35E-06	0.24
GO:0051439	regulation of ubiquitin-protein ligase activity involved in mitotic cell	2.92E-06	0.24

	cycle		
GO:0002474	antigen processing and presentation of peptide antigen via MHC class I	3.48E-06	0.21
GO:0071013	catalytic step 2 spliceosome	3.48E-06	0.21
GO:0000028	ribosomal small subunit assembly	4.15E-06	0.83
GO:0031145	anaphase-promoting complex-dependent proteasomal ubiquitin-dependent protein catabolic process	4.43E-06	0.23
GO:0042776	mitochondrial ATP synthesis coupled proton transport	5.99E-06	0.50
GO:0005685	U1 snRNP	5.99E-06	0.50
GO:0006364	rRNA processing	6.26E-06	0.20
GO:0034709	methylosome	7.32E-06	0.60
GO:1902600	hydrogen ion transmembrane transport	9.85E-06	0.40
GO:0006977	DNA damage response, signal transduction by p53 class mediator resulting in cell cycle arrest	9.85E-06	0.23
GO:0034719	SMN-Sm protein complex	1.07E-05	0.47
GO:0005762	mitochondrial large ribosomal subunit	1.07E-05	0.47
GO:0044281	small molecule metabolic process	1.40E-05	0.09
GO:0005753	mitochondrial proton-transporting ATP synthase complex	4.50E-05	0.38

Table 3.13: mod5 and enrichment of Gene Ontology categories with FWER < 0.05.
The P value threshold after 2000 permutations was 5.81e-05.

GMRG Term	**Description**	***P* value (hypergeometric)**	**Ratio (number of genes in Study group/Number of genes transcriptome-wide)**
GO:0060337	type I interferon signaling pathway	2.58E-22	0.66
GO:0051607	defense response to virus	6.83E-17	0.37
GO:0070062	extracellular vesicular exosome	4.70E-16	0.12
GO:0019221	cytokine-mediated signaling pathway	6.31E-16	0.27
GO:0009615	response to virus	3.54E-15	0.38
GO:0045071	negative regulation of viral genome replication	7.43E-15	0.67
GO:0043123	positive regulation of I-kappaB kinase/NF-kappaB signaling	9.93E-08	0.21
GO:0060333	interferon-gamma-mediated signaling pathway	1.64E-07	0.36
GO:0042470	melanosome	1.75E-07	0.23
GO:0034341	response to interferon-gamma	5.81E-07	0.53
GO:0031012	extracellular matrix	1.46E-06	0.20
GO:0045087	innate immune response	3.22E-06	0.12
GO:0035456	response to interferon-beta	3.83E-06	0.83
GO:0034097	response to cytokine	5.28E-06	0.30
GO:0030198	extracellular matrix organization	6.26E-06	0.15
GO:0005789	endoplasmic reticulum membrane	8.00E-06	0.11
GO:0005783	endoplasmic reticulum	1.08E-05	0.11
GO:0034340	response to type I interferon	1.15E-05	1.00
GO:0006955	immune response	1.60E-05	0.17
GO:0043066	negative regulation of apoptotic process	1.65E-05	0.12
GO:0005764	lysosome	1.66E-05	0.15

GO:0005576	extracellular region	1.97E-05	0.10
GO:0005829	cytosol	3.09E-05	0.77
GO:0004859	phospholipase inhibitor activity	3.24E-05	0.62
GO:0019060	intracellular transport of viral protein in host cell	5.50E-05	0.08

Table 3.14: mod6 and enrichment of Gene Ontology categories with FWER < 0.05.
The P value threshold after 2000 permutations was 4.15e-05

GMRG Term	Description	*P* value (hypergeometric)	Ratio (number of genes in Study group/Number of genes transcriptome-wide)
GO:0007268	synaptic transmission	6.37E-05	0.11

Table 3.15: mod7 and enrichment of Gene Ontology categories with FWER < 0.05.
The P value threshold after 2000 permutations was 4.55e-05.

GMRG Term	**Description**	***P* value (hypergeometric)**	**Ratio (number of genes in Study group/Number of genes transcriptome-wide)**
GO:0044281	small molecule metabolic process	2.00E-10	0.09
GO:0005615	extracellular space	2.46E-08	0.11
GO:0030198	extracellular matrix organization	5.50E-08	0.15
GO:0005576	extracellular region	1.11E-07	0.10
GO:0071276	cellular response to cadmium ion	1.15E-07	0.64
GO:0031012	extracellular matrix	1.54E-07	0.19
GO:0005886	plasma membrane	2.38E-07	0.07
GO:0007601	visual perception	3.40E-07	0.20
GO:0008201	heparin binding	5.80E-07	0.21
GO:0030165	PDZ domain binding	1.22E-06	0.20
GO:0016491	oxidoreductase activity	1.39E-06	0.16
GO:0035019	somatic stem cell maintenance	2.64E-06	0.32
GO:0034641	cellular nitrogen compound metabolic process	5.15E-06	0.14
GO:0006635	fatty acid beta-oxidation	6.81E-06	0.29
GO:0001523	retinoid metabolic process	7.09E-06	0.33
GO:0005759	mitochondrial matrix	7.54E-06	0.12
GO:0071294	cellular response to zinc ion	2.14E-05	0.56
GO:0070371	ERK1 and ERK2 cascade	2.14E-05	0.56
GO:0044255	cellular lipid metabolic process	2.32E-05	0.14
GO:0005578	proteinaceous extracellular matrix	2.87E-05	0.16
GO:0017134	fibroblast growth factor binding	3.16E-05	0.40

GO:0007603	phototransduction, visible light	3.28E-05	0.24
GO:0005887	integral component of plasma membrane	3.38E-05	0.09
GO:0043235	receptor complex	3.67E-05	0.17
GO:0070062	extracellular vesicular exosome	3.97E-05	0.07

Table 3.16: mod8 and enrichment of Gene Ontology categories with FWER < 0.05.
The P value threshold after 2000 permutations was 4.83e-05

GMRG Term	Description	*P* value (hypergeometric)	Ratio (number of genes in Study group/Number of genes transcriptome-wide)
GO:0044822	poly(A) RNA binding	9.22E-12	0.08
GO:0005730	nucleolus	2.79E-08	0.06
GO:0006364	rRNA processing	1.52E-05	0.14

Table 3.17: mod9 and enrichment of Gene Ontology categories with FWER < 0.05.
The P value threshold after 2000 permutations was 2.74e-05

GMRG Term	Description	*P* value (hypergeometric)	Ratio (number of genes in Study group/Number of genes transcriptome-wide)
GO:0044822	poly(A) RNA binding	1.01E-16	0.08
GO:0005730	nucleolus	1.90E-14	0.06
GO:0005634	nucleus	8.44E-09	0.04
GO:0008380	RNA splicing	1.41E-07	0.10
GO:0006351	transcription, DNA-templated	8.78E-07	0.05
GO:0003682	chromatin binding	1.22E-05	0.08
GO:0006397	mRNA processing	1.44E-05	0.10
GO:0035845	photoreceptor cell outer segment organization	2.34E-05	1.00

Table 3.18: mod10 and enrichment of Gene Ontology categories with FWER < 0.05.
The P value threshold after 2000 permutations was 4.17e-05

GMRG Term	**Description**	***P* value (hypergeometric)**	**Ratio (number of genes in Study group/Number of genes transcriptome-wide)**
GO:0005739	mitochondrion	3.64E-12	0.05
GO:0005743	mitochondrial inner membrane	1.71E-11	0.10
GO:0022904	respiratory electron transport chain	9.29E-10	0.17
GO:0005747	mitochondrial respiratory chain complex I	2.90E-09	0.26
GO:0044237	cellular metabolic process	2.76E-08	0.12
GO:0008137	NADH dehydrogenase (ubiquinone) activity	3.01E-08	0.25
GO:0006120	mitochondrial electron transport, NADH to ubiquinone	3.90E-08	0.24
GO:0015450	P-P-bond-hydrolysis-driven protein transmembrane transporter activity	2.42E-06	0.67
GO:0006626	protein targeting to mitochondrion	2.94E-06	0.18
GO:0005744	mitochondrial inner membrane presequence translocase complex	1.10E-05	0.50

Table 3.19: mod11 and enrichment of Gene Ontology categories with FWER < 0.05.
The P value threshold after 2000 permutations was 3.59e-05

GMRG Term	Description	*P* value (hypergeometric)	Ratio (number of genes in Study group/Number of genes transcriptome-wide)
GO:0005886	plasma membrane	1.08E-15	0.04
GO:0045087	innate immune response	1.91E-15	0.09
GO:0006954	inflammatory response	2.59E-13	0.15
GO:0006955	immune response	1.21E-12	0.17
GO:0005576	extracellular region	1.49E-10	0.06
GO:0007165	signal transduction	6.22E-09	0.05
GO:0007596	blood coagulation	8.09E-09	0.08
GO:0009897	external side of plasma membrane	1.29E-08	0.16
GO:0030168	platelet activation	4.02E-08	0.11
GO:0050776	regulation of immune response	4.92E-08	0.27
GO:0007159	leukocyte cell-cell adhesion	9.83E-08	0.43
GO:0070022	transforming growth factor beta receptor homodimeric complex	1.13E-07	1.00
GO:0070062	extracellular vesicular exosome	1.72E-07	0.04
GO:0005615	extracellular space	1.74E-07	0.06
GO:0004896	cytokine receptor activity	2.43E-07	0.56
GO:0007229	integrin-mediated signaling pathway	3.30E-07	0.18
GO:0072562	blood microparticle	1.70E-06	0.17
GO:0006956	complement activation	2.34E-06	0.38
GO:0030593	neutrophil chemotaxis	2.87E-06	0.26
GO:0006958	complement activation, classical pathway	3.58E-06	0.36
GO:0002576	platelet degranulation	5.17E-06	0.15

GO:0050900	leukocyte migration	5.86E-06	0.13
GO:0002283	neutrophil activation involved in immune response	6.20E-06	1.00
GO:0090197	positive regulation of chemokine secretion	6.20E-06	1.00
GO:0050853	B cell receptor signaling pathway	7.58E-06	0.31
GO:0004872	receptor activity	9.93E-06	0.09
GO:0043011	myeloid dendritic cell differentiation	1.32E-05	0.44
GO:0002224	toll-like receptor signaling pathway	2.13E-05	0.11
GO:0019864	IgG binding	2.45E-05	0.75
GO:0042803	protein homodimerization activity	2.59E-05	0.05
GO:0060333	interferon-gamma-mediated signaling pathway	2.70E-05	0.18
GO:0030670	phagocytic vesicle membrane	3.23E-05	0.18
GO:0030890	positive regulation of B cell proliferation	3.28E-05	0.24

Table 3.20: Modules and the enrichment of gene sets.
All P values in red are FWER < 0.05, determined after 2000 permutations

Modules **(*P* value threshold for FWER =0.05)**	**mod5** **(1.79E-03)**	**mod11** **(2.03E-03)**	**mod7** **(1.75E-03)**	**mod3** **(2.01E-03)**	**mod2** **(1.70E-03)**	**mod6** **(1.63E-03)**	**mod1** **(2.13E-03)**	**mod9** **(2.05-03)**	**mod8** **(2.21E-03)**	**mod10** **(1.89E-03)**	**mod4** **(1.79E-03)**	**mod12** **(2.28E-03)**
Module Size	759	238	597	1059	1319	667	1646	385	465	280	824	129
Autism Association (univariate lme)	9.64E-04	5.84E-02	1.08E-01	7.61E-02	8.72E-01	6.39E-03	8.29E-03	2.55E-01	3.64E-02	6.47E-01	7.64E-01	1.43E-01
Neuronal Markers	NA	NA	NA	9.90E-01	**5.22E-10**	**5.40E-07**	**1.89E-34**	NA	9.70E-01	NA	NA	NA
Oligodendrocyte Markers	6.85E-01	NA	NA	**5.38E-42**	9.64E-01	NA	1.00E+00	NA	NA	NA	NA	NA
Astrocyte Markers	4.02E-01	NA	**1.65E-75**	9.33E-01	NA	9.80E-01	NA	NA	9.20E-01	NA	9.90E-01	NA
Type 1 Microglial Markers	**1.73E-05**	**3.38E-84**	NA	NA	NA	NA	NA	NA	NA	NA	8.70E-01	NA
Type 2 Microglial Markers	**1.22E-39**	**3.39E-04**	6.30E-01	9.54E-01	NA	NA	NA	NA	NA	NA	9.20E-01	NA
Ischemia Markers	1.08E-01	NA	NA	NA	NA	NA	4.54E-01	NA	NA	NA	NA	NA

Synaptic Proteins	7.62E-01	NA	5.10E-01	7.67E-01	**7.50E-06**	7.40E-01	**2.95E-08**	NA	2.45E-01	NA	8.70E-01	NA
Postsynaptic Density (PSD)	9.90E-01	9.98E-01	8.00E-01	8.08E-01	**1.19E-26**	1.00E-03	**2.67E-15**	9.90E-01	5.80E-01	9.03E-01	7.00E-02	9.90E-01
ASD SFARI 2012	8.00E-01	NA	6.11E-01	8.37E-01	**4.08E-04**	1.10E-02	1.69E-01	6.40E-01	3.60E-01	4.15E-01	9.70E-01	NA
ASD SFARI 2014	7.62E-01	8.64E-01	5.37E-01	9.10E-01	**2.49E-06**	2.00E-02	2.01E-01	9.45E-01	1.10E-01	9.00E-01	9.90E-01	NA
ASD SFARI 2014 CV	7.62E-01	NA	6.71E-01	6.59E-01	**1.26E-04**	1.00E-02	1.93E-01	7.98E-01	1.35E-01	8.00E-01	9.70E-01	NA
ASD (Pinto)	9.60E-01	NA	7.80E-01	3.44E-01	**1.84E-04**	6.80E-01	8.60E-01	1.17E-01	2.00E-03	4.20E-01	9.10E-01	NA
Rare *de novo*	9.00E-01	2.21E-01	8.46E-01	6.21E-02	**4.29E-04**	6.10E-01	1.24E-01	5.00E-02	9.08E-01	9.99E-01	9.90E-01	NA
I-exomes	NA	NA	NA	2.80E-01	3.82E-01	NA	4.88E-01	NA	NA	NA	NA	NA
SON-exomes	NA	NA	NA	NA	3.49E-02	NA	NA	NA	NA	NA	NA	NA
AGP	NA	NA	NA	NA	**6.61E-04**	3.70E-01	1.05E-02	NA	5.00E-02	NA	NA	NA

San	NA	NA	NA	2.10E-01	**8.05E-07**	3.40E-01	1.26E-02	NA	6.10E-01	NA	NA	NA
asdM12	NA	4.63E-03	NA	NA	**1.74E-10**	1.10E-03	**3.01E-102**	NA	5.98E-01	9.00E-01	NA	4.70E-01
asdM16	**9.38E-59**	NA	**1.45E-89**	8.62E-01	NA	NA	NA	9.90E-01	NA	NA	7.00E-03	NA
FMRP interacting	9.90E-01	1.00E+00	1.00E+00	2.18E-01	**7.38E-110**	3.10E-01	**1.80E-10**	9.70E-01	2.00E-02	NA	NA	9.64E-01
FMRP 1	5.99E-01	NA	NA	7.77E-01	**3.32E-04**	NA	9.34E-01	2.55E-01	3.20E-01	NA	NA	NA
FMRP 2	NA	NA	NA	NA	5.33E-02	NA	**4.66E-04**	NA	NA	NA	NA	NA
ID 2009	2.10E-01	9.69E-01	3.18E-01	3.59E-02	1.08E-02	7.50E-01	9.75E-01	5.60E-01	7.60E-01	5.16E-01	5.10E-01	5.58E-01
ID (Pinto)	2.20E-01	9.19E-01	6.90E-01	2.66E-01	1.25E-01	9.20E-01	4.77E-01	3.06E-01	3.50E-01	8.33E-01	9.90E-01	4.20E-01
Brain Critical	9.90E-01	NA	9.90E-01	7.40E-01	**3.56E-50**	**7.37E-04**	**4.33E-39**	3.10E-01	3.90E-01	9.90E-01	9.90E-01	NA
T-BCAs	NA	NA	NA	NA	NA	NA	NA	NA	2.00E-02	NA	NA	NA

NA indicates that modules contain one or fewer genes in the category being tested.

Table 3.21: Enrichment analysis of 749 genes differentially expressed at P < 0.01.
Top 15 Gene Ontology terms with P value (hypergeometric test) < 0.05. The P value threshold after 2000 permutations was 2.76e-05. No functional annotations exceeded a FWER < 0.05

GMRG Term	Description	*P* value (hypergeometric test)
GO:0035456	response to interferon-beta	1.38E-04
GO:1990247	N6-methyladenosine-containing RNA binding	1.79E-04
GO:0035455	response to interferon-alpha	1.01E-03
GO:0033690	positive regulation of osteoblast proliferation	1.65E-03
GO:0046597	negative regulation of viral entry into host cell	1.65E-03
GO:0055088	lipid homeostasis	2.42E-03
GO:1901214	regulation of neuron death	3.15E-03
GO:2000505	regulation of energy homeostasis	3.15E-03
GO:0008061	chitin binding	3.19E-03
GO:0009440	cyanate catabolic process	3.19E-03
GO:0010868	negative regulation of triglyceride biosynthetic process	3.19E-03
GO:0021740	principal sensory nucleus of trigeminal nerve development	3.19E-03
GO:0021978	telencephalon regionalization	3.19E-03
GO:0035873	lactate transmembrane transport	3.19E-03
GO:0060221	retinal rod cell differentiation	3.19E-03

Table 3.22: Enrichment analysis of 1964 genes differentially expressed at $P < 0.05$.
Reporting top 15 Gene Ontology Terms with P value (hypergeometric test) < 0.05. The P value threshold after 2000 permutations was 4.25e-05. No functional annotations exceeded a FWER < 0.05

GMRG Term	Description	*P* value (hypergeometric test)
GO:1901214	regulation of neuron death	3.74E-04
GO:0005246	calcium channel regulator activity	5.55E-04
GO:0008021	synaptic vesicle	1.15E-03
GO:0005267	potassium channel activity	1.83E-03
GO:0021772	olfactory bulb development	2.04E-03
GO:0032389	MutLalpha complex	2.12E-03
GO:0071204	histone pre-mRNA 3'end processing complex	2.12E-03
GO:0035435	phosphate ion transmembrane transport	2.12E-03
GO:0060666	dichotomous subdivision of terminal units involved in salivary gland branching	2.12E-03
GO:0048539	bone marrow development	3.25E-03
GO:0051224	negative regulation of protein transport	3.25E-03
GO:1990247	N6-methyladenosine-containing RNA binding	3.25E-03
GO:0045600	positive regulation of fat cell differentiation	3.85E-03
GO:0008543	fibroblast growth factor receptor signaling pathway	5.09E-03
GO:0035455	response to interferon-alpha	5.31E-03

Table 3.23: Enrichment analysis of 185 genes differentially expressed at $P < 0.001$.
Top 15 Gene Ontology terms with P value (hypergeometric test) < 0.05. The P value threshold after 2000 permutations was 3.42e-05. No functional annotations exceeded a FWER < 0.05

GMRG Term	Description	*P* value (hypergeometric test)
GO:1990247	N6-methyladenosine-containing RNA binding	5.88E-04
GO:0017091	AU-rich element binding	7.11E-04
GO:0086006	voltage-gated sodium channel activity involved in cardiac muscle cell action potential	1.16E-03
GO:0060371	regulation of atrial cardiac muscle cell membrane depolarization	1.16E-03
GO:0046872	metal ion binding	1.41E-03
GO:0086002	cardiac muscle cell action potential involved in contraction	2.86E-03
GO:2000009	negative regulation of protein localization to cell surface	2.86E-03
GO:0086012	membrane depolarization during cardiac muscle cell action potential	2.86E-03
GO:0031301	integral component of organelle membrane	2.86E-03
GO:0007399	nervous system development	3.67E-03
GO:0003730	mRNA 3'-UTR binding	3.97E-03
GO:0030166	proteoglycan biosynthetic process	5.24E-03
GO:0031588	AMP-activated protein kinase complex	5.24E-03
GO:0072358	cardiovascular system development	5.24E-03
GO:0001518	voltage-gated sodium channel complex	6.67E-03

Table 3.24: Enrichment analysis of 50 genes differentially expressed at P < 0.0001.
Top 15 Gene Ontology terms with P value (hypergeometric test) < 0.05. The P value threshold after 2000 permutations was 4.51e-05. No functional annotations exceeded a FWER < 0.05

GMRG Term	Description	*P* value (hypergeometric test)
GO:0043101	purine-containing compound salvage	9.34E-04
GO:0007409	axonogenesis	2.07E-03
GO:0006144	purine nucleobase metabolic process	5.89E-03
GO:0005070	SH3/SH2 adaptor activity	6.68E-03
GO:0009967	positive regulation of signal transduction	7.52E-03
GO:0030154	cell differentiation	8.13E-03
GO:0035725	sodium ion transmembrane transport	9.33E-03
GO:0007399	nervous system development	9.55E-03
GO:0045444	fat cell differentiation	9.81E-03
GO:0046872	metal ion binding	9.82E-03
GO:0009790	embryo development	2.23E-02
GO:0055086	nucleobase-containing small molecule metabolic process	2.51E-02
GO:0006811	ion transport	4.68E-02
GO:0005874	microtubule	5.13E-02
GO:0014069	postsynaptic density	5.45E-02

Table 3.25: Association of co-expression modules with disease status

Modules	Module Size	*P* value
mod5	759	9.64E-04
mod11	238	5.84E-02
mod7	597	1.08E-01
mod3	1059	7.61E-02
mod2	1319	8.72E-01
mod6	667	6.39E-03
mod1	1646	8.29E-03
mod9	385	2.55E-01
mod8	465	3.64E-02
mod10	280	6.47E-01
mod4	824	7.64E-01
mod12	129	1.43E-01

CHAPTER 4: Transcriptome Analysis of Cortical Tissue Reveals Shared Sets of Down-Regulated Genes in Autism and Schizophrenia

4.1 Introduction

The aggregation of psychiatric conditions and symptoms in families has long been recognized[19,108–111] with more recent genetic analyses suggesting overlap between a number of disorders[19,112–115]. Recent studies considering SNP-based genetic correlation demonstrated marked correlation between schizophrenia (SCZ) and bipolar disorder (BPD) and to a lesser extent between SCZ and autism spectrum disorder (ASD)[19], suggesting shared genetic etiologies. However, due to limited brain tissue availability, there have been fewer studies at the level of gene expression. We and others hypothesize that gene expression studies may begin to unravel how genetic correlations may functionally overlap in neuropsychiatric disorders.

In a recent publication, Zhao et. al suggested that SCZ and BPD show concordant differential gene expression (R=0.28) and that the genes contributing to this overlap are enriched for genetic association signal in both SCZ and BPD while highlighting several biological pathways [116]. Two separate recent studies of gene expression in autism (AUT) have resolved gene expression changes related to altered synaptic and neuronal signaling as well as immunological differences in autism-affected brains [20,117]. In particular, a marked increase was observed in gene expression related to alternative activation of the innate immune system, or the M2 response in autism-affected brains, relative to controls[117].

Here we set out to analyze RNA sequencing (RNA-Seq) data in combination from AUT, SCZ and BPD to identify cross-disorder transcriptomic relationships. We highlight the highly correlated nature of the SCZ and AUT transcriptomes, which together demonstrate a downregulation of genes involved in neurotransmission and synapse regulation across the two disorders.

4.2 Methods

4.2.1 Autism sample information

RNA-Seq for 104 cortical brain tissue samples across three brain regions (BA10, BA19, BA44/45), comprising 57 samples from 40 control subjects and 47 samples from 32 autism

(AUT) subjects was previously carried out[117]. We note that, as in the initial publication of these data[117], AUT samples harboring CNVs recurrent in autism spectrum disorder have not been included in these analyses. Details related to samples, sequencing, quality control, and informatics can be found in Gupta et. al[117] and are summarized in **Table 4.1**.

4.2.2 Schizophrenia and bipolar disorder sample information

RNA-Seq data was obtained from the Stanley Medical Research Institute (SMRI, http://www.stanleyresearch.org/) consisting of eighty-two (31 SCZ, 25 BPD and 26 controls) anterior cingulate cortex (BA24) samples. Detailed sequencing information can be found in Zhao et. al[116]. Sample information for those included in this analysis can be found in **Table 4.2**.

4.2.3 RNA-Seq, alignment & quality control

Sequencing, alignment, quality control and gene expression estimation for the AUT samples were carried out as previously described[117]. The reads from both the AUT and SMRI sequencing were subjected to a common analysis pipeline[117] in which quality control of raw sequences included removing both polyA stretches and adaptor sequence contamination using a Python script, 'cutadapt' (v1.2.1)[71]. Sequences were then aligned to the Genome Reference Consortium Human build 37 (GRCh37/hg19) assembly using TopHat2[55,72] allowing for only uniquely aligned sequences with fewer than three mismatches to align.

4.2.4 Gene expression estimation and normalization

Gene count estimates were obtained for 62,069 Ensembl gene annotations (GRCh37/hg19) using HTSeq (http://www-huber.embl.de/users/anders/HTSeq/) under an intersection-strict model. Of these, 8,856 genes with at least 10 reads across 75 percent of the SMRI samples were then normalized for gene length and GC content using Conditional Quantile Normalization (CQN)[67]. In the AUT samples, the 13,262 genes previously included for analysis [117] were normalized for gene length and GC content using CQN. Outliers were then removed from the CQN normalized gene expression estimates on a per-gene basis as described previously[118]. In either data set, any sample whose gene expression value was more than 2.7 standard deviations (sd) from the mean of the gene expression was excluded from analysis at that particular gene prior to linear modeling.

4.2.5 Differential gene expression analysis (DGEA)

Due to the unique experimental design in which multiple brain regions were sequenced from the same individual, AUT gene expression estimates were fit using a linear mixed effects model, with subject ID included as a random intercept term, and case-control status as the primary variable of interest. Age, sex, site of sample collection, brain region and twelve surrogate variables (SVs)[50] were included as fixed effects in the model to account for known and unknown covariates. SVs function to remove batch effects and sources of noise in gene expression data by adjusting for unknown or unmodeled sources of variation and are therefore included for analysis[50].

SCZ and BPD RNA-Seq data were analyzed using standard linear regression with case-control status as the primary variable of interest. The known covariates to which we had access and that were included in the analysis by Zhao et. al[116] (age, sex, cumulative antipsychotic use, brain pH, and postmortem interval (PMI)) were incorporated into the model here along with SVs to account for unknown sources of variation.

Because the SCZ and BPD cases share controls, two separate differential gene expression analyses were performed. For the comparison to AUT, all cases (SCZ or BPD) and all controls from the SMRI dataset were included in the analysis. Alternatively, when SCZ and BPD were to be compared directly, we employed a strategy similar to how these data were handled previously, in which controls were divided randomly in half [116]. One set of controls was then compared to the SCZ cases while the other set of controls was compared to the BPD cases. This procedure was carried out 100 times for each cross-disorder comparison and the Z-scores (β/se) were recorded for each gene for each run. The median Z-score for each gene across these 100 runs was then used for analyses comparing SCZ to BPD.

4.2.6 Null DGEA

To obtain a null set of differential gene expression values, each of the analyses in the previous section was carried out modeling the data exactly as described above save for the permutation of case-control status. In AUT datasets, case-control status was randomized between samples from the same collection sites, as described previously[117]. To minimize the possibility of reporting false-positive findings, one-thousand null permutations were utilized to determine significance.

4.2.7 Calculating genes differentially expressed across disorders

To determine which genes were differentially expressed across disorders, Z-scores were multiplied across each of the three disorder comparisons (Z_{SCZ}* Z_{BPD}, Z_{SCZ}* Z_{AUT}, Z_{BPD}* Z_{AUT}). Genes with large cross-disorder Z-scores were considered to be differentially expressed across disorders, with significance determined by permutation. For each cross-disorder comparison, the most extreme cross-disorder Z-score for each of these 1000 null permutations was recorded. Of these values, the cross-disorder cutoff for significance (defined at $p<0.05$) to determine which genes were differentially expressed across disorders was determined by taking the value for which only 5% of the null values were more extreme.

To determine differentially and concordantly expressed genes (DCEGs) common to all three disorders, Z-scores were multiplied for the 2,895 genes with Z-scores in the same direction across all three disorders (Z_{AUT}* Z_{SCZ}* Z_{BPD}). As SCZ and BPD are directly compared in the analysis, split-control generated Z-scores for SCZ and BPD were utilized to account for the shared control samples. To assess significance, the same analysis was carried out with 1000 null permutations as described above.

4.2.8 Calculating the correlation of DCEGs across phenotypes

Pearson's correlation coefficient (R) was calculated for the Z-scores from each disorder comparison (SCZ-AUT, SCZ-BPD, BPD-AUT) to assess the similarity of genes differentially expressed across disorders. To determine the significance of this correlation, Pearson's correlation coefficient was calculated after testing each of the 1000 null permutations.

4.2.9 Pathway analysis of DCEGs

Pathway enrichment analysis was carried out on genes differentially expressed across disorders. GO gene sets were downloaded from MsigDB (1466 gene sets, http://www.broadinstitute.org/gsea/msigdb/collections.jsp#C5). For each gene and across all three disease comparisons, Z-scores were summed across disorders using Stouffer's method [119] and pathways were tested for enrichment. Cross-disorder Z-scores were calculated, such that:

$$Z_{AUT\text{-}SCZ} = (Z_{AUT} + Z_{SCZ})/\mathrm{sqrt}(k)$$

where *k* is the number of comparisons made (here, *k*=2). A one-sided t-test was used to compare Z-scores between genes in the pathway and genes not in the pathway. To assess significance for each pathway in each of the cross-disorder comparisons, the absolute value of cross-disorder Z-scores for those genes in the pathway were compared to the absolute value of cross-disorder Z-scores for those genes not in the pathway using a one-sided t-test under the alternative hypothesis that Z-scores in the pathway were enriched for significant Z-scores relative to the genes not in the pathway. The 1,285 GO categories for which we had gene expression data for at least five genes in the pathway were included for analysis. Significance was determined empirically by permutation for each cross-disorder comparison ($1.51x1^4$ for AUT-SCZ, $1.72x1^4$ for AUT-BPD, and $4.25x1^6$ for SCZ-BPD).

As a complementary approach, we utilized two open source programs for pathway analysis: WebGestalt (v2, http://bioinfo.vanderbilt.edu/webgestalt/) [120,121] to run a Gene Ontology (GO) analysis[122,123] and DAVID[124] (v6.7, https://david.ncifcrf.gov/) for functional pathway analysis. As the input for these approaches requires gene lists, we input genes that were differentially expressed (absolute value(Z-score) > 2.2) in both disorders of the comparison: 1) SCZ-AUT (191 genes), 2) BPD-AUT (38 genes), and 3) SCZ-BPD (16 genes).

GO analysis used a hypergeometric test for enrichment utilizing the Benjamini-Hochberg method [125] for multiple test correction. GO categories whose adjusted p-values < 0.001 were considered to be statistically significantly enriched. For DAVID, gene lists were uploaded and a 'Functional Annotation Chart' was generated using default settings. Functional categories whose Bonferroni-adjusted p-value<0.05 were reported as significant.

To ensure that results from these analyses were not biased by the different number of genes input into the pathway analysis, we also carried out the GO and DAVID analyses described above with a fixed number of 191 genes from each cross-disorder comparison.

4.2.10 Enrichment for genetic signal analysis

GWAS results were downloaded from the Psychiatric Genetic Consortium (PGC, http://www.med.unc.edu/pgc/) for autism, bipolar disorder, and schizophrenia.

Gene-based p-values were computed on the summary data for each disorder using FAST (v1.8) [126] for the 8,856 genes included in the cross disorder DGEA.

The following tests were conducted allowing for up to one million permutations: 'logistic-minsnp-gene-perm' and 'logistic-gwis-perm'. LD was calculated on the fly using LD computed from HapMap Phase 1 CEU imputation data. Default settings were used, aside from the following: the flank parameter (region on either side of the gene for investigation) was set to 15kb, phenotype variance was estimated at 0.01, and a maf-cutoff of 0.01 was used. Sample sizes were estimated to be approximately 1.5x the number of cases used in each individual analysis. Accordingly, the sample sizes used as input were 10,000, 15,000 and 20,000 for AUT, BPD, and SCZ respectively. Downstream gene-based p-values were compiled such that the GWiS[127] p-value was used for all genes assigned p-values not equal to one (signifying that no permutations were carried out in GWiS). Otherwise, the more permissive minSNP-P p-value was assigned to the gene. The minSNP-P simply uses the best single SNP p-value within the gene, calculates a gene-based p-value by permutation test within each gene, and assigns that p-value to the gene[127].

To test for enrichment of genetic signal, we first took suggestive genes (gene-based p<0.05) for each individual GWAS (SCZ, BPD, and AUT) and compared these to p-values from the DGEA. Data were plotted in a QQ-plot among 100 null permutations to look for enrichment relative to the null data. To ensure that this analysis was not a reflection of the gene-based p-value restriction imposed on the data, a more permissive (p<0.1) and more restrictive (p<0.01) GWAS cutoff were used and the same enrichment analysis carried out.

4.3 Results

4.3.1 Sample summary

Of the 105 samples in the SMRI array collection, 82 cortical brain samples (BA24) were sequenced and included for analysis (31 SCZ, 25 BPD, and 26 controls). To accompany these data, 104 AUT samples from three cortical brain regions (BA10, BA19, BA44/45) were included for analysis, composed of 57 control and 47 AUT samples. A summary of sample statistics are provided in **Table 4.3** with detailed sample information in **Tables 4.1** and **4.2** for AUT and SMRI data, respectively. Further sample information can be found in the original publications [116,117].

4.3.2 Genes differentially expressed across SCZ, BPD, and AUT

Nine genes were differentially expressed (p<0.05) in both SCZ and AUT. None were significant when comparing BPD to SCZ, and one gene reached significance in the AUT-BPD comparison

(**Table 4.2**). We note that the single gene differentially expressed between AUT-BPD, *IQSEC3*, is significant in both AUT-SCZ and AUT-BPD comparisons. The relatively large Z-scores in SCZ (Z=-3.59) and BPD (Z=-3.46) suggest this result is not simply driven by the altered gene expression in AUT alone.

Differentially expressed genes (DEGs) across all three disorders were identified in a joint analysis of genes whose direction of effect was consistent across all three disorders ($Z_{AUT}*Z_{SCZ}*Z_{BPD}$). Two genes, *IQSEC3* (Z=-35.45, p=0.001) and *COPS7A* (Z=-22.52, p=0.017), are transcriptome-wide significant ($p<0.05$, absolute value ($Z_{AUT}*Z_{SCZ}*Z_{BPD}$) > 19.56), indicating a common role for altered gene expression of these genes across all three neuropsychiatric disorders (**Figure 4.1**). We note that these two genes, *IQSEC3* and *COPS7A*, are syntenic (12p13.33 and 12p13.31, respectively) with their expression being markedly correlated in both the SMRI and AUT data sets (R=0.41 and R=0.70, respectively) (**Figure 4.2**).

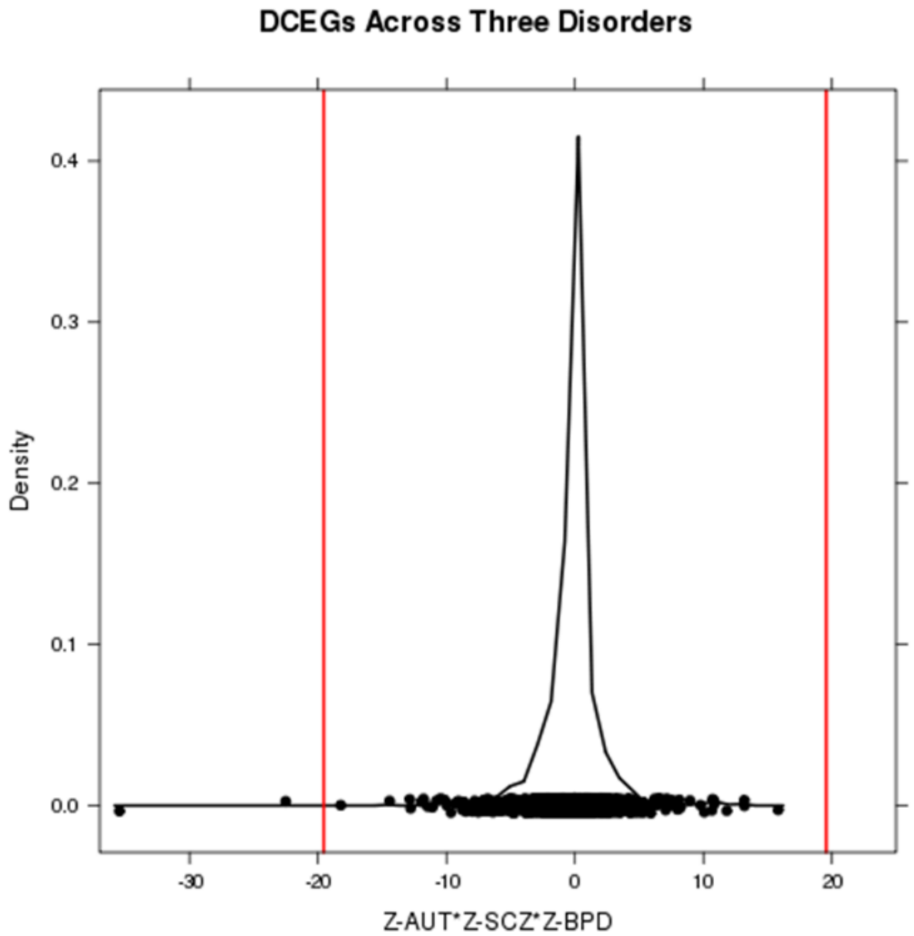


Figure 4.1: Differential Gene Expression Across AUT, SCZ, and BPD. A density plot for the cross three-disorder Z-scores (Z_{AUT}*Z_{SCZ}*Z_{BPD}) are plotted in black with the cutoff for transcriptome-wide significance highlighted in red ($p<0.05$, determined empirically by permutation). The two genes that meet transcriptome wide significance are labeled.

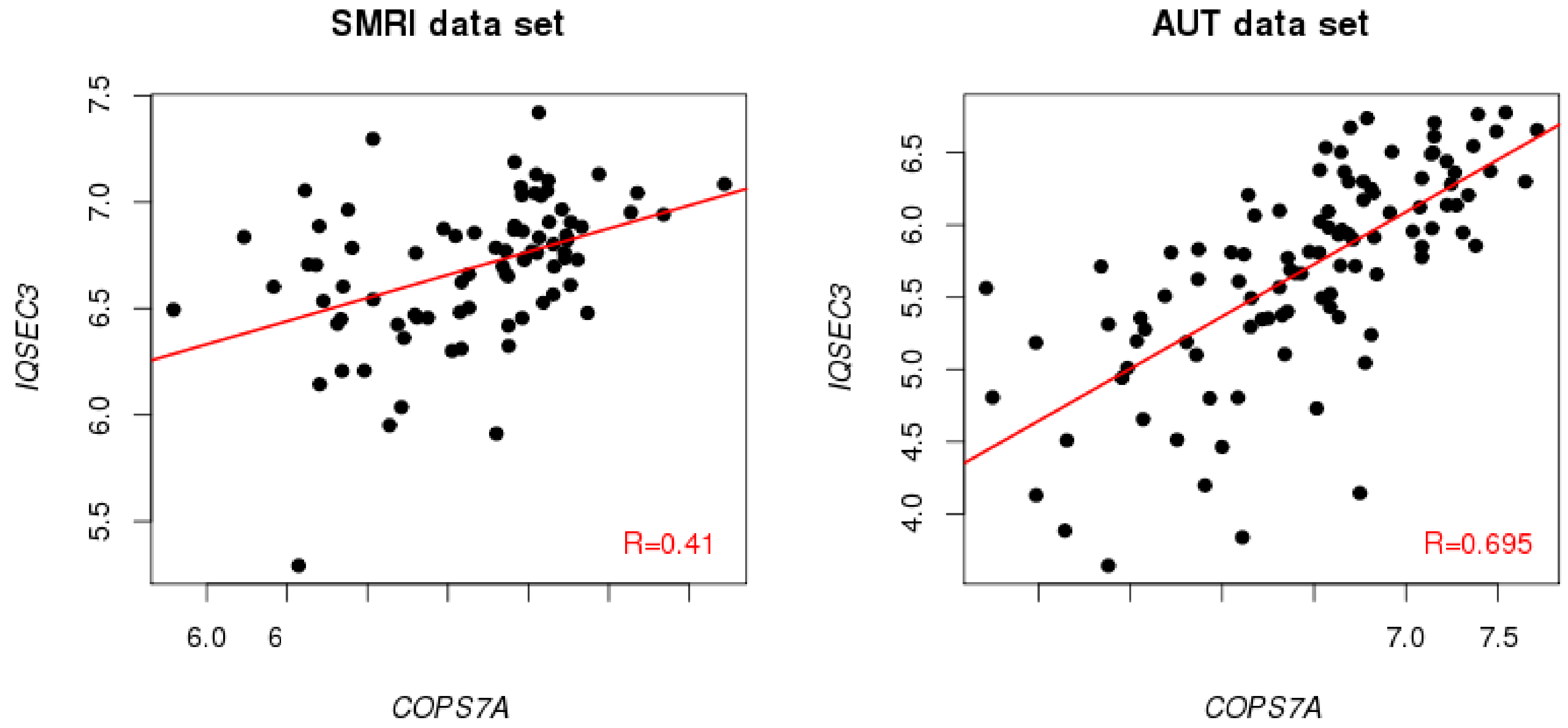


Figure 4.2: Correlation of Genes Differentially Expressed Across All Three Disorders.
The gene expression of *IQSEC* and *COPS7A* in (a) the SCZ and BPD data from the SMRI and (b) the AUT data are plotted. Pearson's correlation coefficient (R) is in red.

4.3.3 Correlation in gene expression across SCZ, BPD, and AUT

The transcriptomic relationship across disorders and correlation of test-statistics (Z-scores) was investigated. SCZ-AUT demonstrated the most significant correlation (R=0.298, p<0.001). SCZ-BPD also demonstrated a positive correlation (R=0.11). This level of correlation was neither significant (p=0.41) nor as high as previously reported (R=0.28)[116]. Similarly, the correlation between AUT and BPD was minimal and did not differ significantly from the null (R=0.06, p=0.25). (**Figure 4.3** and **Figure 4.4**).

To explore the discrepancy between the correlation reported here for SCZ and BPD and that previously reported, we carried out the same analysis without the inclusion of surrogate variables (SVs) in the model. The failure to include unknown covariates in the model led to a marked increase in the correlation between SCZ and BPD (R=0.50), suggesting that the previously reported correlation between these disorders may have been influenced by hidden structure in the data. (**Figure 4.5**).

.

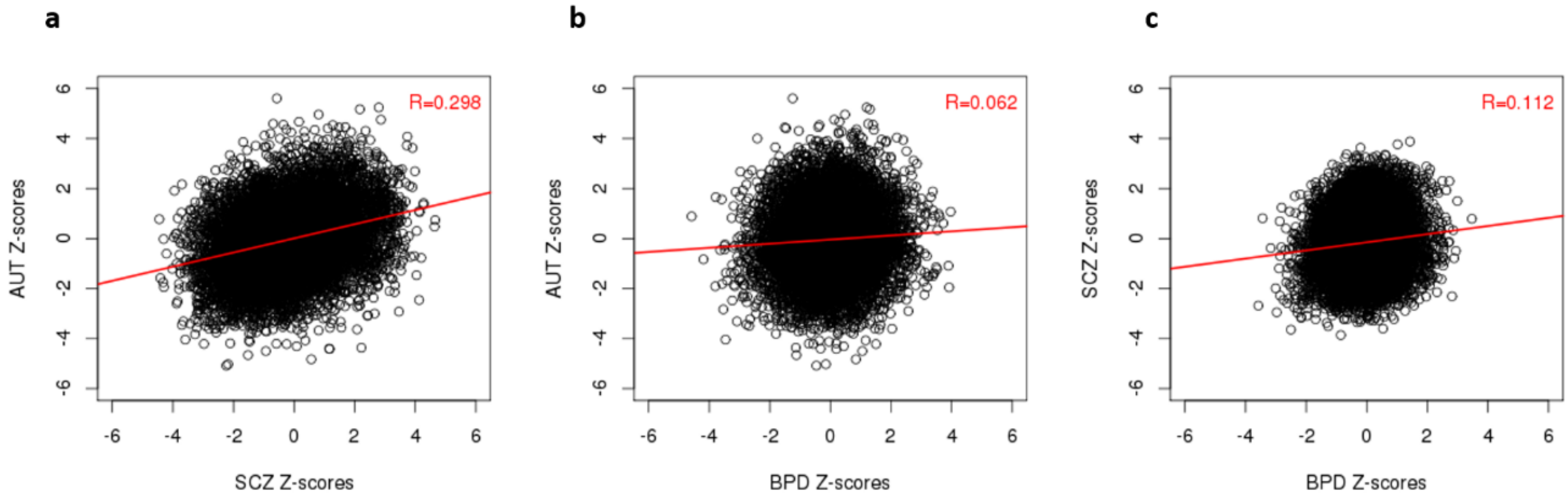


Figure 4.3: Correlation of Cross-Disorder Differential Gene Expression.
Z-scores for each cross-disorder comparison ((a) AUT-SCZ (b) AUT-BPD (c) SCZ-BPD) are plotted. The best fit line is in red. Pearson's Correlation Coefficient (R) is included on the graph, quantifying the level of correlation between the transcriptomes of each cross-disorder comparison.

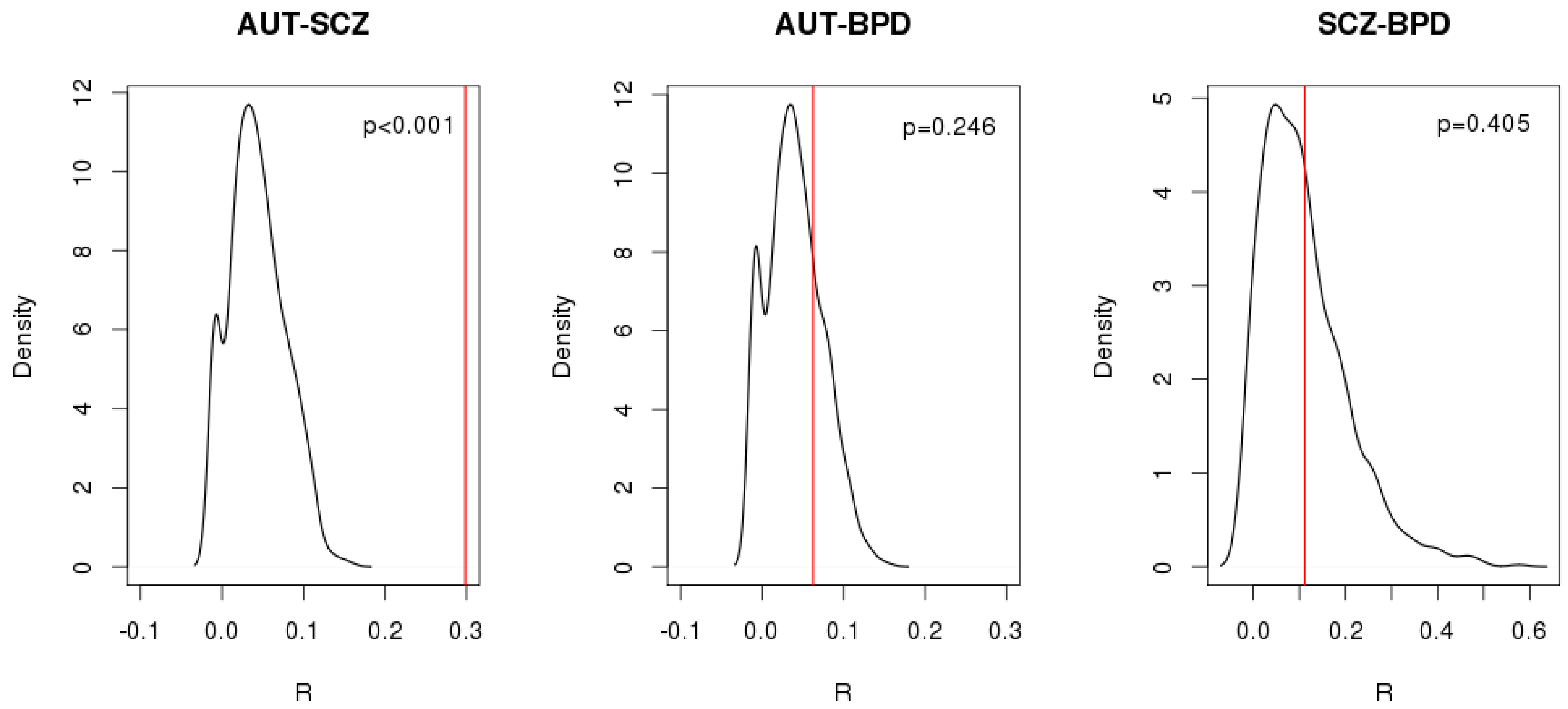


Figure 4.4: Assessing the significance for Correlations of Cross-Disorder Transcriptomic Similarity.
For each cross-disorder comparison, density plots for the correlations of the 1000 null permutations are plotted in black. The cross-disorder correlation derived from the data are plotted in red. (a) The correlation between AUT and SCZ is more extreme than the correlation in any of the 1000 null permutations ($p<0.001$). (b) The correlation between differential gene expression AUT and BPD is not significant relative to the null correlations ($p=0.246$). (c) The correlation between SCZ and BPD is similarly not significant ($p=0.405$).

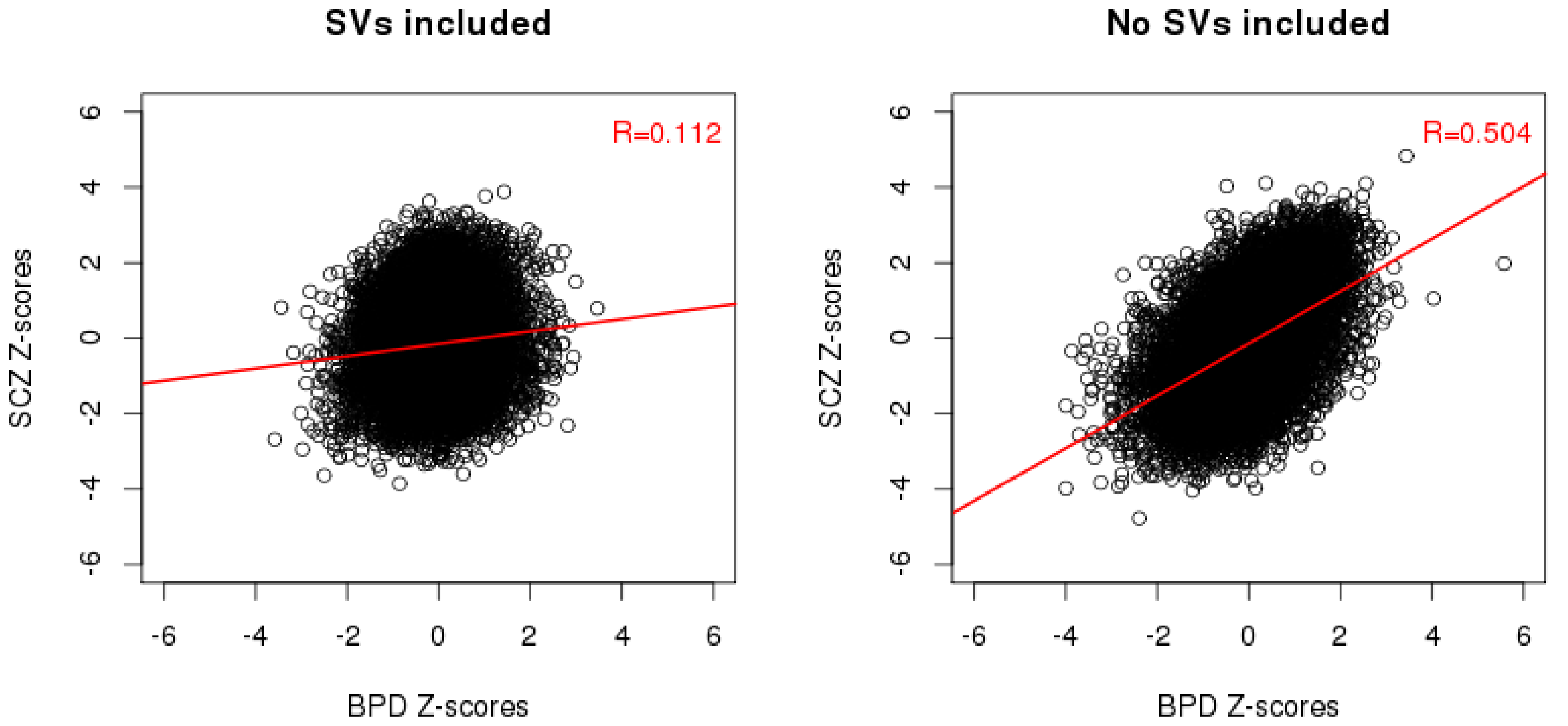


Figure 4.5: Accounting for Unknown Covariates Affects Correlation.

The correlation between differential gene expression in SCZ and BPD reported in this paper in which the linear model included SVs to account for unknown covariates (a) relative to an analysis in which these covariates were not included (b). The lack of SV inclusion in the linear model to detect differential gene expression leads to an artificially inflated correlation.

4.3.4 Pathway enrichment analyses of genes differentially expressed across disorders

Combined pathway analysis utilizing lists of genes differentially expressed across disorders (absolute value(Z-score)>2.2 in both disorders) was carried out using both Gene Ontology (GO) enrichment and DAVID pathway analysis. For this analysis, 191 DEGs for AUT-SCZ, 38 for AUT-BPD, and 16 for SCZ-BPD met these criteria. DAVID pathway analysis highlighted the role of neuron projection development ($p_{Bonferroni}$=0.012) in those genes differentially expressed in both AUT and SCZ (**Table 4.4**). Similarly, when these genes were characterized by GO, there was a clear abundance of altered gene expression in neuronal and synapse-related GOs (**Figure 4.6**). Further, when these $DEGs_{AUT-SCZ}$ genes were split up into those either concordantly up- or down-regulated in both disorders, 106 genes differentially downregulated in both disorders were driving the GO enrichments, with no contribution from the 69 genes upregulated in both disorders. As for AUT-BPD comparisons, there were no enrichments detected for any gene ontologies and the only emergent DAVID pathway was genes related to phosphoproteins ($p_{Bonferroni}=1.2x1^{4}$) (**Table 4.4**). Similarly, no GO or DAVD pathways were found to be significant for $DEGs_{SCZ-BPD}$. Substantially similar results were observed when the number of genes from each cross-disorder comparison input into the pathway analysis was fixed rather than imposing a Z-score cutoff (**Table 4.5**). Finally, we found that the number of cross-disorder discordant DEGs (upregulated in one disorder but downregulated in the other) differs across the three comparisons, such that there are fewer discordant cross-disorder DEGs (16/191, 8.4%) in the comparison between SCZ and AUT than in the comparison between AUT and BPD (76/191, 39.8%) or between SCZ and BPD (38/191, 19.9%), further supporting the transcriptomic similarities between AUT and SCZ.

Traditional pathway analysis requires a significance cutoff for the gene input for analysis. To avoid a potential bias by choosing an arbitrary cutoff, we used a Z-score based approach (see Methods) and identified gene enrichment of DCEGs common to all three disorder comparisons using the GO data from MSigDB. Three GO pathways – each of which indicated some enrichment for altered gene expression in transporter genes – were enriched for DEGs in both AUT and SCZ. No pathways were study-wide significant in the other two disorder comparisons.

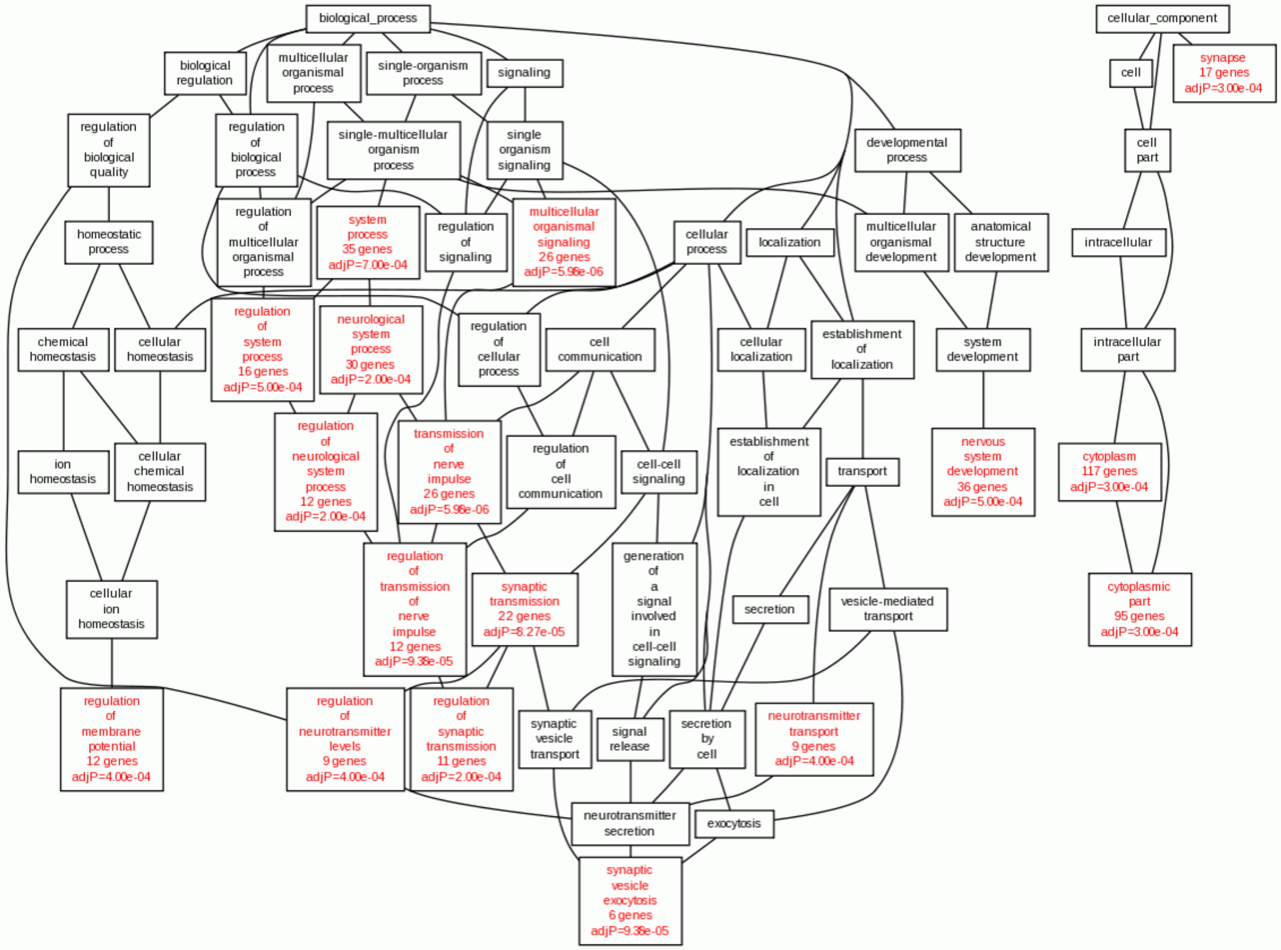

biological_process
cellular_component
biological regulation
multicellular organismal process
single-organism process
signaling
cell
synapse 17 genes adjP=3.00e-04
regulation of biological quality
regulation of biological process
single-multicellular organism process
single organism signaling
developmental process
cell part
homeostatic process
regulation of multicellular organismal process
system process 35 genes adjP=7.00e-04
regulation of signaling
multicellular organismal signaling 26 genes adjP=5.96e-06
cellular process
localization
multicellular organismal development
anatomical structure development
intracellular
chemical homeostasis
cellular homeostasis
regulation of system process 16 genes adjP=5.00e-04
neurological system process 30 genes adjP=2.00e-04
regulation of cellular process
cell communication
cellular localization
establishment of localization
system development
intracellular part
ion homeostasis
cellular chemical homeostasis
regulation of neurological system process 12 genes adjP=2.00e-04
transmission of nerve impulse 26 genes adjP=5.96e-06
regulation of cell communication
cell-cell signaling
establishment of localization in cell
transport
nervous system development 36 genes adjP=5.00e-04
cytoplasm 117 genes adjP=3.00e-04
cellular ion homeostasis
regulation of transmission of nerve impulse 12 genes adjP=9.38e-05
synaptic transmission 22 genes adjP=8.27e-05
generation of a signal involved in cell-cell signaling
secretion
vesicle-mediated transport
cytoplasmic part 95 genes adjP=3.00e-04
regulation of membrane potential 12 genes adjP=4.00e-04
regulation of neurotransmitter levels 9 genes adjP=4.00e-04
regulation of synaptic transmission 11 genes adjP=2.00e-04
synaptic vesicle transport
signal release
secretion by cell
neurotransmitter transport 9 genes adjP=4.00e-04
neurotransmitter secretion
exocytosis
synaptic vesicle exocytosis 6 genes adjP=9.38e-05

Figure 4.6: GO Analysis of cross-disorder DEGsAUT-SCZ.
Genes differentially expressed in both AUT and SCZ (absolute(Z-score) > 2.2) were analyzed for ontological enrichment of biological processes, developmental processes, and cellular component. Ontological categories with at least five genes and an adjusted p-value < 0.001 are highlighted in red. This tree highlights the role of nerve impulse transmission, synaptic transmission, and neurotransmitter transport in those genes differentially expressed in both AUT and SCZ.

4.3.5 Cross-disorder DEGs enrichment in association signals

To test whether genes differentially expressed across disorders were enriched for genetic associations, we compared cross-disorder DGEA results to gene-level GWAS results. We first directly compared gene-based GWAS p-values ($p<0.05$) from each individual GWAS (AUT, SCZ, BPD) to p-values from the cross-disorder differential gene expression analysis (AUT-SCZ, AUT-BPD, SCZ-BPD). No comparison was identified that would suggest any enrichment in signal overlap with respect to the null (**Figure 4.7**). Three additional p-value cutoffs ($p<0.1$, $p<0.01$, $p<1$) demonstrated that neither these null findings nor the inflation seen are a function of the gene-based p-value cutoff imposed on the data (**Figures 4.8, 4.9,** and **4.10, Table 4.6**). Likewise, there were no enrichments for cross-disorder DEGs seen in these analyses relative to the null. Finally, LOF variants have recently been reported in a number of AUT studies[11,12,14–16,128]; however, Gupta et. al demonstrated that these gene expression data are neither enriched for the findings from the exome studies nor for Structural Variants (SVs) [117]. Accordingly, these lists of variants have not been included in these analyses.

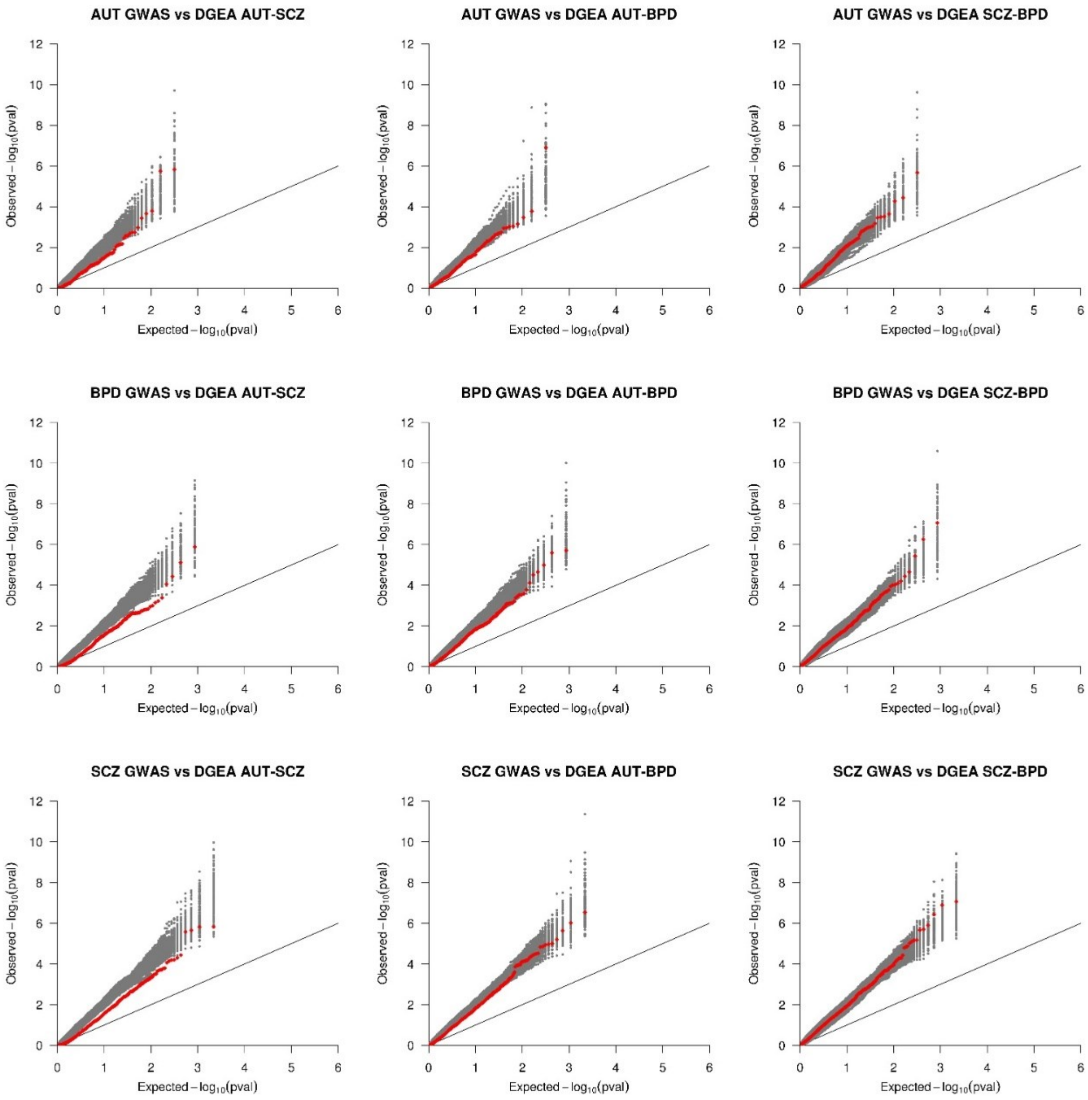


Figure 4.7: Enrichment of DEGs among GWAS signal.

QQ plots assess enrichment of differential gene expression signal (red) among suggestive GWAS results ($p<0.05$). Data for 100 null permutations are plotted in gray. Each row corresponds to GWAS data from a separate disorder (AUT, BPD, SCZ from top to bottom) and each column a different cross-disorder comparison (AUT-SCZ, AUT-BPD, and SCZ-BPD from left to right).

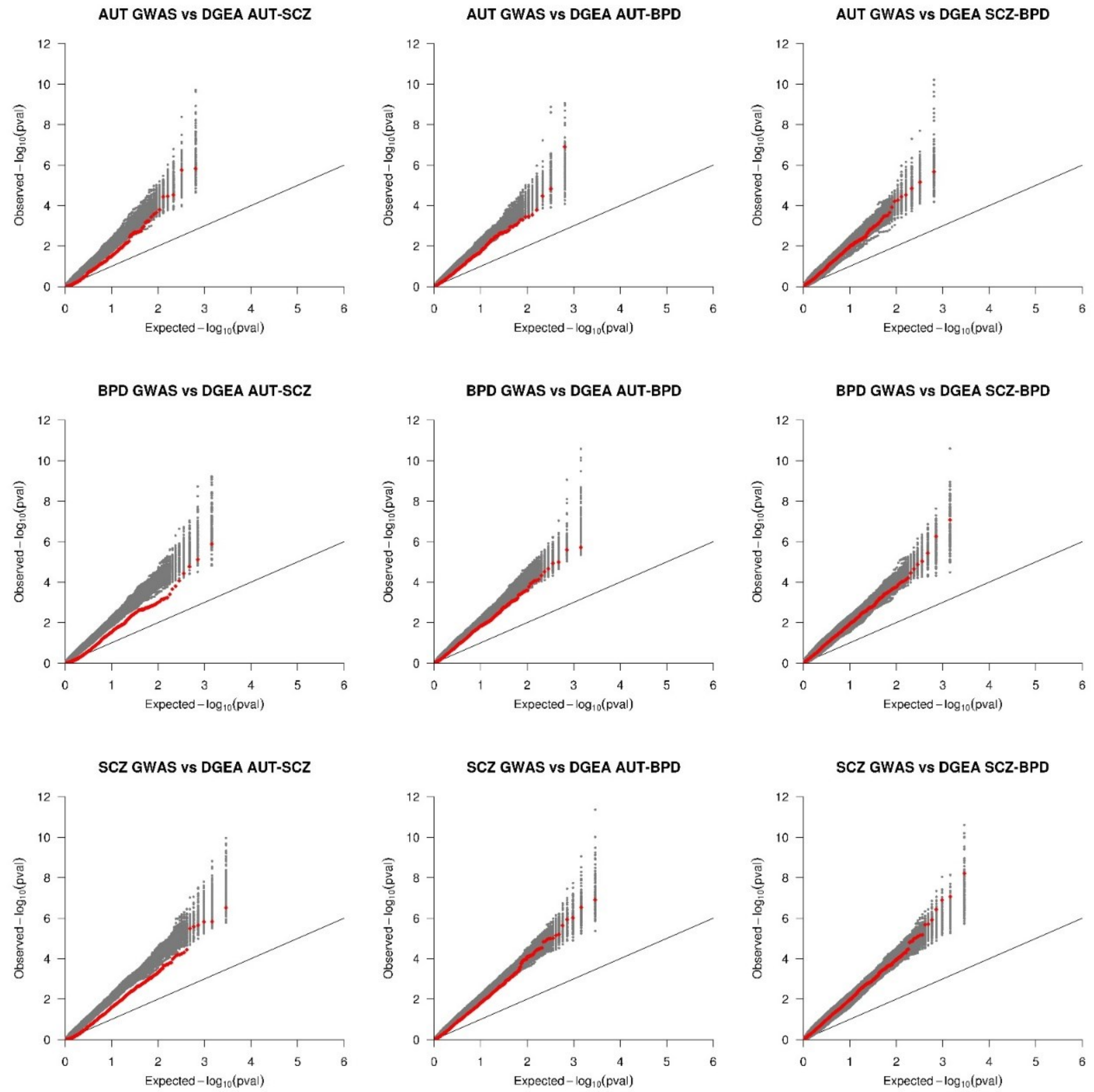


Figure 4.8: Enrichment of DEGs among GWAS at a more permissive p-value cutoff (p<0.1). QQ plots assess enrichment of differential gene expression signal (red) among suggestive GWAS results (p<0.1). Data for 100 null permutations are plotted in gray. Each row corresponds to GWAS data from a separate disorder (AUT, BPD, SCZ from top to bottom) and each column a different cross-disorder comparison (AUT-SCZ, AUT-BPD, and SCZ-BPD from left to right).

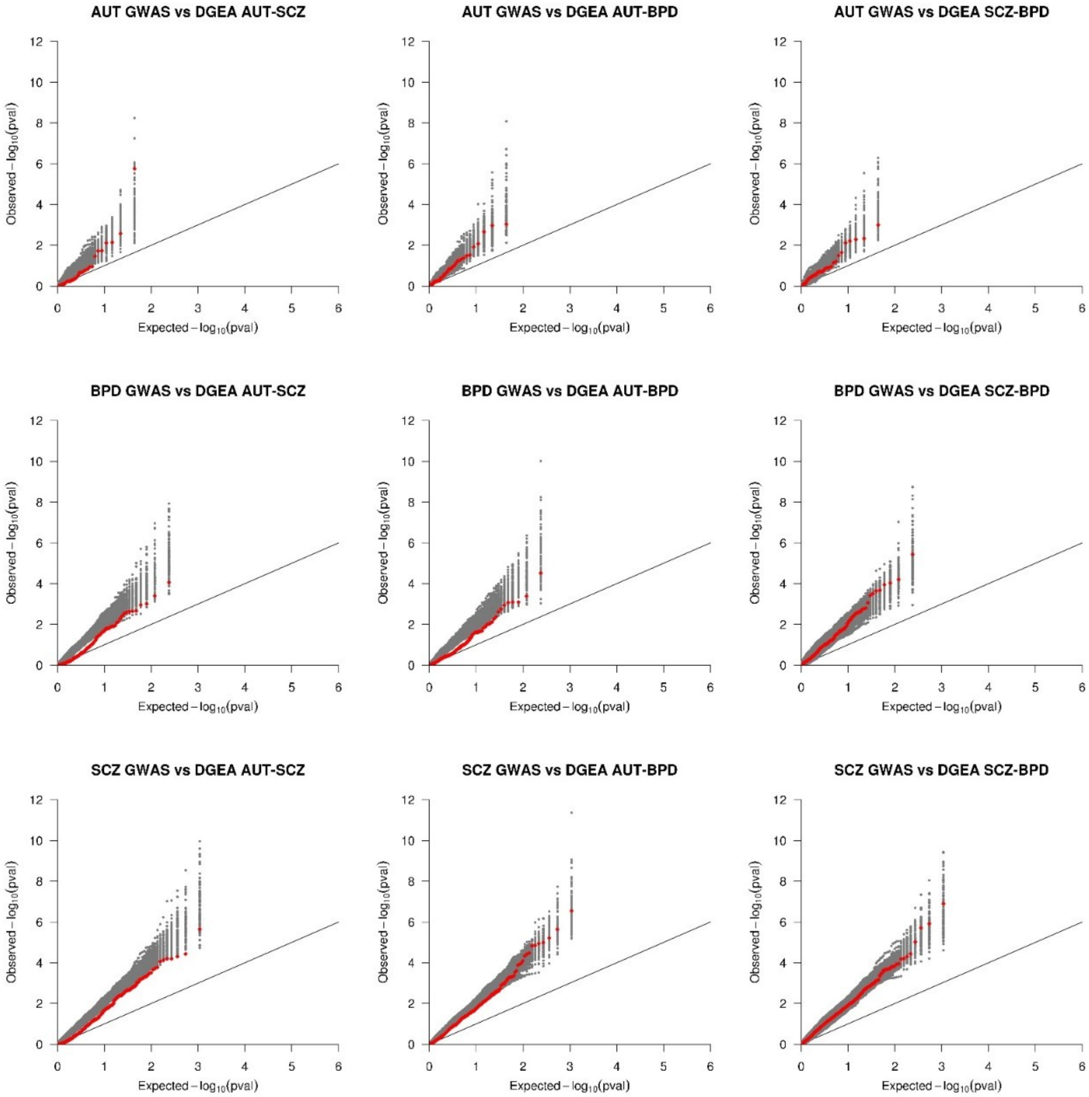


Figure 4.9: Enrichment of DEGs among GWAS signal at a more stringent p-value cutoff (p<0.01).

QQ plots assess enrichment of differential gene expression signal (red) among suggestive GWAS results (p<0.01). Data for 100 null permutations are plotted in gray. Each row corresponds to GWAS data from a separate disorder (AUT, BPD, SCZ from top to bottom) and each column a different cross-disorder comparison (AUT-SCZ, AUT-BPD, and SCZ-BPD from left to right).

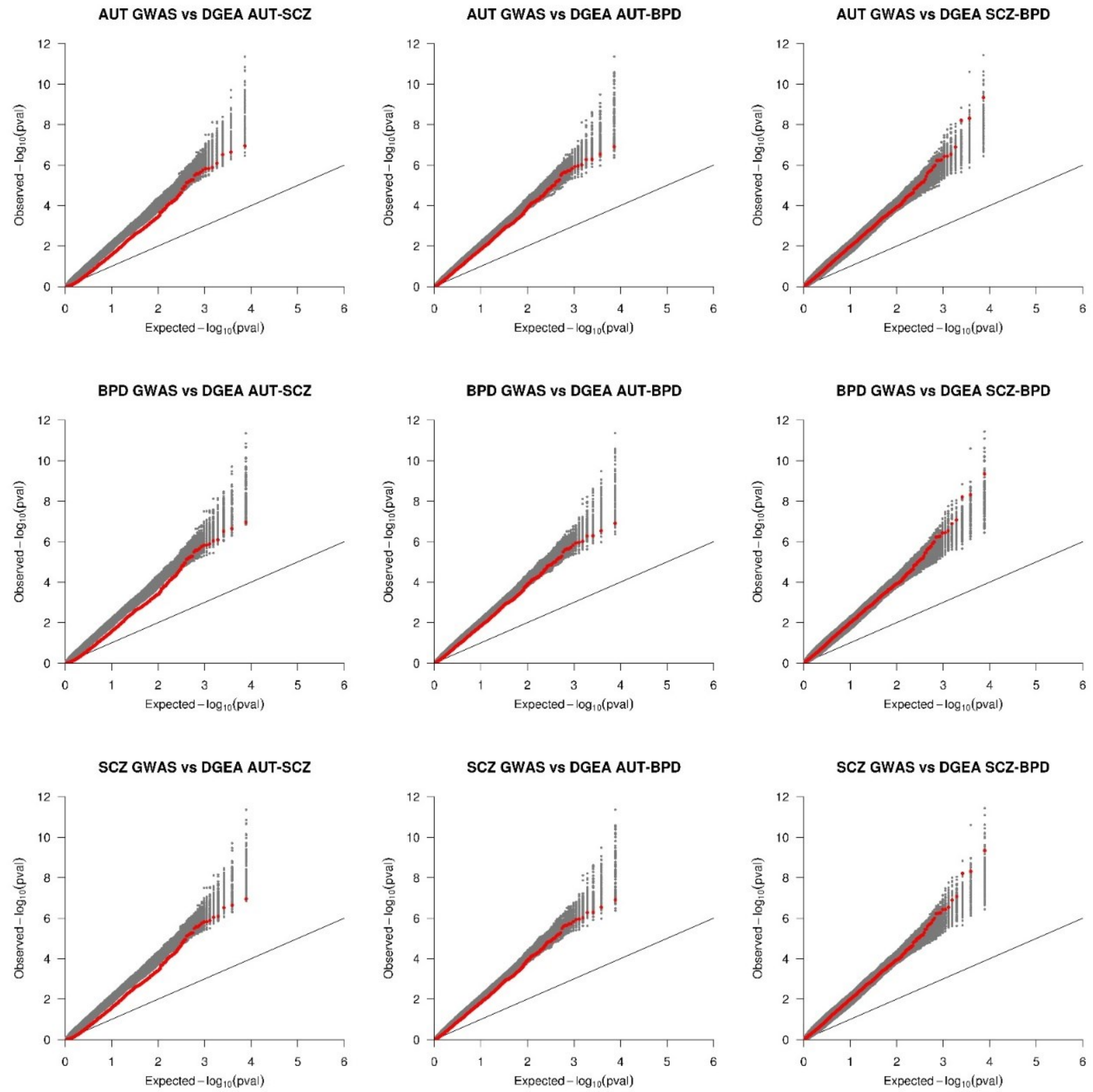


Figure 4.10: Enrichment of DEGs among all genes (no gene based GWAS p-value cutoff imposed).

QQ plots demonstrate that inflation of the test-statistic is present in the data regardless of gene-based p-value cut off. Data for 100 null permutations are plotted in gray. Each row corresponds to GWAS data from a separate disorder (AUT, BPD, SCZ from top to bottom) and each column a different cross-disorder comparison (AUT-SCZ, AUT-BPD, and SCZ-BPD from left to right).

4.4. Conclusions

4.4.1 Results summary

To our knowledge, this is the first study to combine next-generation sequencing gene expression analyses across AUT, SCZ, and BDP to assess the transcriptomic relationship and how gene expression relates to GWAS findings. We report that, at the transcriptome level, AUT and SCZ demonstrate a highly overlapping gene expression profile. The cross-disorder DEGs between AUT and SCZ highlight a shared relationship in synapse and projection formation, suggesting a role for neuronal development underlying the correlation. Further, despite the lack of global significant differential transcriptomic correlation between either BPD and SCZ or AUT and BPD, we highlight two genes, *IQSEC3* and *COPS7A*, for their consistent downregulation across all three disorders and support further investigation into these specific genes' expression and function to better understand their role in neuropsychiatric disorders. Finally, we report that the genes differentially expressed across disorders were not enriched in genetic association signals for AUT, SCZ or BPD.

4.4.2 Accounting for unknown covariates is critical in transcriptome analyses

In large gene expression studies, variation that confounds results can be introduced at any step despite a tremendous amount of effort to standardize approaches [129–132]. Fortunately, Surrogate Variable Analysis (SVA) can help to address this by accounting for unknown covariates within large genomic data sets[50]. Importantly, in these analyses, we demonstrate that failure to account for unknown sources of variation leads to an artificially inflated correlation between SCZ and BPD (R=0.50, **Figure 4.5**). The previously reported correlation between these two disorders (R=0.28) falls between the value reported herein as the correlation between SCZ and BPD (R=0.11, **Figure 4.3** & and **Figure 4.4**) and the correlation reported when unknown covariates fail to be considered (R=0.50). The previously reported correlation between the transcriptomes of SCZ and BPD was likely artificially inflated due to these unknown covariates. We note that the remaining discrepancy between our analysis without SVs included in the linear model and that previously reported is likely due to the fact that our linear model did not include all covariates included in the previous analysis; however, as we did not have access to a number of the technical covariates for the SMRI samples (cDNA concentration, RNA integrity number, or batch number), we were unable to directly test this hypothesis.

4.4.3 Correlations in differential gene expression across disorders highlights similarities between AUT and SCZ

After modeling the data for each individual-disorder comparison relative to their controls, the cross-disorder comparison demonstrated that SCZ and AUT share a similarly altered transcriptome ($p<0.001$), whereas AUT-BPD and SCZ-BPD ($p=0.25$ and $p=0.41$, respectively) do not show a significant correlation (**Figure 4.3, Figure 4.4**). We note that the lack of significant correlation between BDP-SCZ in our analysis is in conflict with a previous report [116], and is likely due to our inclusion of SVs to account for unknown sources of variation, suggesting that the previously reported analysis of these data is overstated (see Supplemental Discussion). Further, while these data do not directly support transcriptomic overlap between SCZ and BPD, this is likely reflective of the shared control design of the experiment. This experimental design results in a smaller effective sample size and a study underpowered to assess overlap between these two disorders. Given the genetic relationship between these disorders (where SCZ-BPD > AUT-SCZ > AUT-BPD)[19] , future work utilizing a larger sample for analysis may likely demonstrate a shared transcriptomic profile between SCZ and BPD; however, these data do not.

Analyzing the pathways in which DEGs in both SCZ and AUT were involved, we found that the genes differentially expressed in AUT and SCZ were enriched for neuron projection development ($p=0.012$, **Table 4.4**). Additionally, there was a clear enrichment for genes involved in synaptic and neuronal processes. The other two non-significant cross-disorder comparisons (AUT-BPD and SCZ-BPD) failed to demonstrate any enrichment for biological process ontology, even when controlling for the number of cross-disorder DEGs, further supporting the conclusion that differential transcriptomic correlation is biologically relevant between SCZ and AUT but is not observed in the other two cross-disorder comparisons. When the DEGs across AUT and SCZ were broken down into those concordantly upregulated versus those concordantly downregulated, the enrichment in GO was only present in those genes concordantly downregulated, suggesting that these synaptic and neuronal alterations were a result of decreased brain expression in both disorders.

Finally, in assessing which specific genes were differentially expressed across disorders, we identified *IQSEC3* and *COPS7A* as differentially expressed in all three disorders (**Table 4.7**).

IQSEC3 (*KIAA1110)* is a protein coding gene that has been shown to be specifically expressed in human adult brain with particularly high levels in the human cortex[133]. IQSEC3 has been suggested to act as a guanine exchange factor for ARF1 in endocytosis[133], and ARF1 critically regulates actin dynamics in neurons and synaptic strength and plasticity, potentially aligning with pathways previously implicated in AUT, SCZ and BPD. *COPS7A* is expressed broadly across tissues[45], and encodes part of the COP9 signalosome, a multi-subunit protease with a role in regulating the ubiquitin-proteasome pathway[134].

4.4.4 Differences in genetic variation not explained by overlapping gene expression profiles

We report no enrichment for significant cross-disorder DEGs among GWAS signal in any of the comparisons (**Figure 4.7**) relative to the null. These findings suggest either that 1) alterations at the genetic level do not largely manifest themselves in altered gene expression concordantly across these disorders, or 2) that primary genetic defects do not result in altered gene expression across disorders at the time points measured but could, perhaps, alter gene expression at other time points, such as during development, or 3) the effects of these genetic perturbations are small and that increased sample sizes will be required to detect these slight differences in cross-disorder altered gene expression. Regardless, large differences in gene expression across these disorders appear to be independent of known genetic variation in each of these disorders.

There were a number of limitations associated with our observations. As the analyses combine data across two studies with notable design differences in each (shared controls in the SMRI data, multiple brain regions from the same individual in the AUT data, limited ability to detect lowly expressed genes, and comparison of different cortical brain regions), there was certainly variation unrelated to disease state introduced into the differential gene expression analyses. However, we have controlled for this to the best of our ability by accounting for unknown covariates in all analyses and by determining all levels of significance relative to null permutations. While we have controlled for the differences in experimental design in our analysis, we note that the reported overlap in AUT and SCZ was significant ($p<0.001$) despite the fact that different cortical brain regions were studied in the two data sets. Due to this limitation, we hypothesize that our observed correlation between AUT and SCZ may underestimate the true transcriptomic correlation and that the similarities may be even more pronounced between AUT and SCZ had the same brain regions been studied. Similarly, sequencing depths in these data sets are lower than many RNA-Seq data sets currently being published. Thus, while lowly-expressed genes are not well-estimated here, their omission from analysis would only lead to false negatives – or genes missing from overlap. This does not detract for the findings, herein, but simply acknowledges that some genes may not be included in the analysis, herein. Conversely, we acknowledge that our power to detect correlation between SCZ and BPD is limited due to the smaller effective sample size, a consequence of the shared control design of the experiment and that, given a larger sample size, transcriptomic correlation between these two disorders may likely become evident and reflective of the known genetic relationships[19].

With future studies employing larger sample sizes and more powerful characterizations, we will gain a better understanding of the transcriptomic relationships that are common and disparate among neuropsychiatric disorders. Besides providing context for how the altered genetic landscape of each disorder affects the brain, we hope that identification of common aspects underlying susceptibility might be novel targets to therapeutically address the underlying pathogenic mechanisms.

4.5 Tables

Table 4.1: Autism sample information

Sample Code	Sample ID	Dx	Sex	Age	Site (Brain Bank)	RIN	PMI (hrs)	Brain Region
AN03345	s14	Autism	M	2	Harvard	2.2	4	BA10
AN12457	s16	Autism	F	29	Harvard	3.7	17.83	BA10
AN13872	s23	Autism	F	5	Harvard	7	33	BA10
UMB1571	s69	Control	F	18	Maryland	NA	8	BA10
UMB1712	s71	Control	M	20	Maryland	5.6	8	BA10
UMB1790	s72	Control	M	13	Maryland	6.4	18	BA10
UMB1796	s73	Control	M	16	Maryland	4.2	16	BA10
UMB1823	s74	Control	M	15	Maryland	1.7	18	BA10
UMB1841	s75	Control	M	19	Maryland	2.4	14	BA10
UMB1944	s79	Control	F	16	Maryland	NA	20	BA10
AN08873	s8	Autism	M	5	Harvard	5.7	25.5	BA10
UMB4728	s85	Control	M	17	Maryland	6	23	BA10
AN01093	s86	Autism	M	56	Harvard	2.3	19	BA10
AN06420	s88	Autism	M	39	Harvard	6.5	14	BA10
AN16641	s1	Autism	M	9	Harvard	6.4	27	BA19
AN01570	s11	Autism	F	18	Harvard	6.7	6.75	BA19
AN17777	s15	Autism	F	49	Harvard	2.2	16.33	BA19
AN12457	s16	Autism	F	29	Harvard	3.7	17.83	BA19
AN11989	s17	Autism	M	30	Harvard	4.7	16.06	BA19
AN00493	s2	Autism	M	27	Harvard	3.2	8.3	BA19
AN13872	s23	Autism	F	5	Harvard	7	33	BA19
AN17678	s25	Autism	M	11	Harvard	2.7	NA	BA19
AN04682	s26	Autism	M	15	Harvard	NA	23.23	BA19
AN03632	s27	Autism	F	49	Harvard	5.5	21.08	BA19
AN09714	s28	Autism	M	60	Harvard	5.2	26.5	BA19
AN00764	s3	Autism	M	20	Harvard	NA	23.7	BA19
AN16665	s31	Control	M	36	Harvard	2.2	20	BA19
AN01357	s32	Control	M	42	Harvard	2.5	18.33	BA19
AN02583	s33	Control	M	68	Harvard	3.5	16.58	BA19
AN01410	s34	Control	M	41	Harvard	2.1	27.17	BA19
AN15240	s35	Control	F	36	Harvard	5.7	18.08	BA19
AN08677	s36	Control	M	38	Harvard	3.2	25.47	BA19
AN07176	s37	Control	M	21	Harvard	2.3	29.91	BA19
AN17425	s38	Control	M	16	Harvard	3	26.16	BA19
AN14368	s39	Control	M	22	Harvard	2.1	24.2	BA19
AN15566	s40	Control	F	32	Harvard	4.6	28.92	BA19

AN13295	s41	Control	M	56	Harvard	6	22.12	BA19
UMB797	s42	Autism	M	9	Maryland	6.9	13	BA19
UMB1349	s44	Autism	M	5	Maryland	5.1	39	BA19
UMB1638	s45	Autism	F	20	Maryland	3.2	50	BA19
UMB4231	s46	Autism	M	8	Maryland	4.5	12	BA19
UMB4721	s47	Autism	M	8	Maryland	2.3	16	BA19
UMB4999	s50	Autism	M	20	Maryland	4.4	14	BA19
UMB4671	s51	Autism	F	4	Maryland	4.4	13	BA19
UMB451	s52	Control	M	4	Maryland	3	15	BA19
UMB497	s53	Control	M	12	Maryland	4.8	16	BA19
UMB1185	s55	Control	M	4	Maryland	2.6	17	BA19
UMB1377	s56	Control	F	5	Maryland	4	20	BA19
UMB1674	s58	Control	M	8	Maryland	5	NA	BA19
AN01227	s6	Autism	M	82	Harvard	5.2	24.67	BA19
UMB4898	s61	Control	M	7	Maryland	6.2	12	BA19
UMB1323	s62	Control	M	16	Maryland	6.1	25	BA19
UMB1409	s63	Control	M	18	Maryland	6.5	6	BA19
UMB1429	s64	Control	M	18	Maryland	5.2	9	BA19
UMB1322	s66	Control	M	16	Maryland	5.5	25	BA19
UMB1541	s67	Control	F	20	Maryland	NA	19	BA19
UMB1543	s68	Control	M	17	Maryland	5.6	22	BA19
UMB1571	s69	Control	F	18	Maryland	NA	8	BA19
AN14613	s7	Autism	M	39	Harvard	6.8	22.75	BA19
UMB1584	s70	Control	F	18	Maryland	5.5	15	BA19
UMB1712	s71	Control	M	20	Maryland	5.6	8	BA19
UMB1790	s72	Control	M	13	Maryland	6.4	18	BA19
UMB1796	s73	Control	M	16	Maryland	4.2	16	BA19
UMB1823	s74	Control	M	15	Maryland	1.7	18	BA19
UMB1841	s75	Control	M	19	Maryland	2.4	14	BA19
UMB1843	s76	Control	F	15	Maryland	2.6	9	BA19
UMB1908	s78	Control	M	13	Maryland	NA	13	BA19
UMB1944	s79	Control	F	16	Maryland	NA	20	BA19
UMB4590	s80	Control	M	20	Maryland	6.3	19	BA19
UMB4591	s81	Control	F	16	Maryland	6.3	14	BA19
UMB4669	s82	Control	M	16	Maryland	2.4	16	BA19
UMB4727	s84	Control	M	20	Maryland	6.4	5	BA19
UMB4728	s85	Control	M	17	Maryland	6	23	BA19
AN06420	s88	Autism	M	39	Harvard	6.5	14	BA19
AN16115	s89	Autism	F	11	Harvard	NA	13	BA19
AN19511	s9	Autism	M	8	Harvard	5.8	22.2	BA19
AN16641	s1	Autism	M	9	Harvard	6.4	27	BA44/45

AN09730	s13	Autism	M	22	Harvard	3	25	BA44/45
AN03345	s14	Autism	M	2	Harvard	2.2	4	BA44/45
AN11989	s17	Autism	M	30	Harvard	4.7	16.06	BA44/45
AN08166	s22	Autism	M	28	Harvard	6.7	43.25	BA44/45
AN03632	s27	Autism	F	49	Harvard	5.5	21.08	BA44/45
AN09714	s28	Autism	M	60	Harvard	5.2	26.5	BA44/45
AN17254	s29	Autism	M	51	Harvard	2.9	22.16	BA44/45
AN00764	s3	Autism	M	20	Harvard	NA	23.7	BA44/45
AN07176	s37	Control	M	21	Harvard	2.3	29.91	BA44/45
AN14368	s39	Control	M	22	Harvard	2.1	24.2	BA44/45
AN08792	s4	Autism	M	30	Harvard	5.4	20.3	BA44/45
UMB1349	s44	Autism	M	5	Maryland	5.1	39	BA44/45
UMB1638	s45	Autism	F	20	Maryland	3.2	50	BA44/45
UMB4721	s47	Autism	M	8	Maryland	2.3	16	BA44/45
UMB4849	s48	Autism	M	7	Maryland	3.9	20	BA44/45
UMB4999	s50	Autism	M	20	Maryland	4.4	14	BA44/45
UMB451	s52	Control	M	4	Maryland	3	15	BA44/45
UMB1185	s55	Control	M	4	Maryland	2.6	17	BA44/45
UMB1377	s56	Control	F	5	Maryland	4	20	BA44/45
UMB1674	s58	Control	M	8	Maryland	5	NA	BA44/45
UMB4670	s60	Control	M	4	Maryland	5	17	BA44/45
UMB1409	s63	Control	M	18	Maryland	6.5	6	BA44/45
UMB1429	s64	Control	M	18	Maryland	5.2	9	BA44/45
UMB1465	s65	Control	M	17	Maryland	5.9	4	BA44/45
UMB1841	s75	Control	M	19	Maryland	2.4	14	BA44/45
AN08873	s8	Autism	M	5	Harvard	5.7	25.5	BA44/45
AN19511	s9	Autism	M	8	Harvard	5.8	22.2	BA44/45

Table 4.2: Schizophrenia and bipolar disorder sample information

Stanley ID	Dx	Sex	Age	PMI (hrs)	Brain pH
A1	Bipolar	M	29	48	6.39
A10	Schizophrenia	M	40	34	6.18
A100	Schizophrenia	F	59	38	6.93
A101	Schizophrenia	M	52	16	6.52
A102	Bipolar	M	48	23	6.9
A104	Control	M	47	36	6.57
A12	Schizophrenia	M	19	28	6.73
A14	Bipolar	F	48	18	6.5
A16	Bipolar	M	42	32	6.65
A17	Schizophrenia	F	53	13	6.49

A18	Bipolar	M	35	35	6.3
A19	Control	M	49	46	6.5
A2	Bipolar	M	29	60	6.7
A20	Bipolar	F	59	53	6.2
A21	Bipolar	M	54	44	6.5
A22	Schizophrenia	M	37	30	6.8
A23	Bipolar	F	35	17	6.1
A24	Control	M	53	9	6.4
A25	Schizophrenia	M	52	10	6.1
A26	Schizophrenia	M	24	15	6.2
A27	Control	M	37	13	6.5
A29	Control	M	51	31	6.7
A3	Schizophrenia	M	43	26	6.42
A32	Bipolar	F	42	49	6.65
A33	Control	F	38	33	6
A34	Bipolar	F	58	34	6.5
A35	Control	F	38	28	6.7
A37	Schizophrenia	M	39	80	6.6
A38	Control	M	59	47	6.8
A39	Schizophrenia	M	33	29	6.5
A4	Bipolar	M	45	28	6.35
A40	Schizophrenia	M	50	9	6.2
A41	Schizophrenia	M	43	18	6.3
A42	Bipolar	M	64	16	6.1
A43	Control	M	35	52	6.7
A44	Schizophrenia	F	32	36	6.8
A45	Schizophrenia	M	35	47	6.4
A46	Bipolar	M	59	84	6.65
A47	Schizophrenia	M	44	32	6.67
A49	Control	M	34	22	6.48
A51	Control	M	47	21	6.81
A52	Control	M	45	21	6.94
A54	Control	M	42	37	6.91
A55	Schizophrenia	M	47	13	6.3
A57	Bipolar	M	51	23	6.67
A58	Bipolar	F	63	32	6.97
A6	Bipolar	F	29	62	6.74
A60	Control	M	45	18	6.81
A62	Bipolar	F	56	26	6.58
A63	Bipolar	F	43	39	6.74
A64	Bipolar	M	35	22	6.58

A65	Control	M	49	23	6.93
A66	Schizophrenia	M	45	35	6.66
A67	Control	M	32	24	7.03
A68	Schizophrenia	F	36	27	6.49
A69	Bipolar	F	50	62	6.51
A7	Schizophrenia	M	31	33	6.2
A70	Control	M	55	31	6.7
A71	Control	F	49	45	6.72
A72	Bipolar	F	49	38	6.39
A73	Schizophrenia	M	54	38	6.17
A75	Schizophrenia	F	54	42	6.65
A78	Schizophrenia	F	44	26	6.58
A79	Control	M	48	31	6.86
A8	Bipolar	M	44	19	6.74
A81	Schizophrenia	M	50	30	6.47
A83	Control	M	32	13	6.57
A84	Control	M	47	11	6.6
A85	Schizophrenia	M	38	35	6.69
A86	Control	M	46	31	6.67
A87	Schizophrenia	M	41	54	6.18
A88	Schizophrenia	M	43	65	6.67
A89	Bipolar	F	43	57	5.92
A90	Control	M	40	38	6.67
A91	Control	M	51	22	6.71
A92	Schizophrenia	M	42	26	6.19
A93	Schizophrenia	F	47	35	6.5
A94	Schizophrenia	M	42	19	6.48
A95	Control	M	31	11	6.13
A97	Schizophrenia	M	46	30	6.72
A98	Bipolar	M	56	23	6.07
A99	Control	F	39	58	6.46

Table 4.3: Sample Summary

	N	Unique Individuals	Mean Age (years)	Sex	
				F	M
AUT					
CTL	57	40	20	12	33
AUT	47	32	24	9	18
TOTAL	104	72	22	21	51
SMRI					
CTL	26	26	44	4	22
BPD	25	25	47	12	13
SCZ	31	31	42	7	24
TOTAL	82	82	44	23	59

Abbreviations: AUT, autism; BPD, bipolar disorder; SCZ, schizophrenia; CTL, control

Table 4.4: DAVID Pathway Analysis for cross-disorder DEGs

	TOTAL	UP	DOWN	DISCORDANT	DAVID Pathways
AUT-SCZ	191	69	106	0	neuron projection development (p=0.012)
AUT-BPD	38	8	19	11	phosphoprotein (p=$1.2x1^{4}$)
SCZ-BPD	16	2	13	1	--

Abbreviations: AUT, autism; BPD, bipolar disorder; SCZ, schizophrenia

Table 4.5: DAVID Pathway Analysis for cross-disorder DEGs controlling for # of genes input

	controlling for # of genes						
	TOTAL	**UP**	**DOWN**	**DISCORDANT**	**DAVID Pathways (ALL)**	**DAVID (UP)**	**DAVID (DOWN)**
AUT-SCZ	191	69	106	16	neuron projection development	phosphoprotein	transport, cytoplasm, syntaxin binding, SNARE binding
AUT-BPD	191	45	67	76	phosphoprotein, enzyme binding, cytosol, cytoplasm, acetylation	NONE	cytosol, synapse, phosphoprotein, GTPase regulator activity, nucleoside-triphosphate regulator activity, cell junction
SCZ-BPD	191	55	98	38	--	NONE	NONE

Table 4.6: Association Signal Enrichment Results across multiple Z-score cutoffs.

GWAS	Z-score cutoff				BACKGROUND (null)					
		$Z_{AUT}*Z_{SCZ}$	$Z_{AUT}*Z_{BPD}$	$Z_{SCZ}*Z_{BPD}$	$Z_{AUT}*Z_{SCZ}$	$Z_{AUT}*Z_{BPD}$	$Z_{SCZ}*Z_{BPD}$	$Z_{AUT}*Z_{SCZ}$	$Z_{AUT}*Z_{BPD}$	$Z_{SCZ}*Z_{BPD}$
AUT	3	0	0	0	0	0	0	1	1	1
	2.7	1	0	0	3	0	0	0.775	1	1
	2.5	1	0	0	4	2	0	0.946	1	1
	2.2	8	0	0	15	0	0	0.357	1	1
	2	14	2	0	16	3	2	0.264	0.826	1
BPD	3	0	0	0	0	0	0	1	1	1
	2.7	7	1	0	4	1	0	0.111	0.46	1
	2.5	10	1	0	9	2	0	0.295	0.804	1
	2.2	23	2	1	20	6	4	0.17	0.891	0.81
	2	33	7	5	24	10	3	0.381	0.711	0.216
SCZ	3	4	1	0	5	1	0	0.35	0.442	1
	2.7	13	1	0	10	1	0	0.232	0.838	1
	2.5	26	4	1	20	3	1	0.082	0.613	0.679
	2.2	54	11	4	56	11	3	0.155	0.33	0.581
	2	89	22	11	73	24	10	0.079	0.368	0.226

Table 4.7: Genes significantly differentially expressed across disorders

	# Sig. Genes	**Cross-Disorder Sig. Cutoff**	**Ensembl Gene IDs**	**Gene Name**	**chr**	**$Z_{cross\text{-}disorder}$**	**Z_{AUT}**	**Z_{SCZ}**	**Z_{BPD}**
AUT-SCZ	9	12.42	ENSG00000106261	*ZKSCAN1*	7	15.24	4.08	3.74	0.68
			ENSG00000172005	*MAL*	2	14.66	5.24	2.80	1.20
			ENSG00000120645	*IQSEC3*	12	14.53	-4.04	-3.59	-3.46
			ENSG00000046653	*GPM6B*	X	14.16	3.62	3.91	0.26
			ENSG00000167191	*GPRC5B*	16	13.85	3.72	3.72	0.78
			ENSG00000129521	*EGLN3*	14	13.62	4.76	2.86	0.13
			ENSG00000164068	*RNF123*	3	12.81	-3.46	-3.71	-0.99
			ENSG00000134780	*DAGLA*	11	12.54	-4.22	-2.98	-0.57
			ENSG00000183597	*TANGO2*	22	12.53	-3.83	-3.27	0.25
AUT-BPD	1	12.29	ENSG00000120645	*IQSEC3*	12	14.00	-4.04	-3.59	-3.46
SCZ-BPD	0	21.71	--	--	--	--	--	--	--
AUT-SCZ-BPD	2	19.56	ENSG00000120645	*IQSEC3*	12	-35.45	-4.04	-2.95	-2.97
			ENSG00000111652	*COPS7A*	12	-22.52	-3.31	-3.14	-2.17

Abbreviations: AUT, autism; BPD, bipolar disorder; SCZ, schizophrenia; Sig., significant; chr, chromosome; Z, Z-Score

CHAPTER 5: Exaggerated CpH Methylation in the Autism-Affected Brain

5.1 Introduction

Autism is a heritable neurodevelopmental disorder affecting one in 68 individuals in the United States[135]. Recent genetic studies have identified a handful of genes that contribute to autism[136–140] and gene expression studies have begun to unravel how altered gene expression manifests within the autistic brain[141,142]; however, the majority of risk remains unexplained. In addition to genetic causes, epigenetic mechanisms have been proposed to play an important role in the development of the disorder. This hypothesis was initially supported by three lines of evidence. First, direct alterations in epigenetic pathways can dramatically alter early embryonic and neonatal neurodevelopment in the same critical periods as autism-associated changes in the brain [31] Second, mutations in indirect epigenetic effectors can result in autism-spectrum and related disorders, such as Rett syndrome[143], Fragile X syndrome[144], and Angelman syndrome[145]. Finally, deficiencies in DNA methylation (DNAm), historically studied in CpG islands in gene promoters as an indicator of transcriptional repression, have previously been implicated in autism[146–148].

Initial studies of methylation in autism were limited by the number of sites investigated, a lack of dynamic range in microarrays, the number of samples available for study, and the prioritization of DNA that was procured from cell lines and tissue other than the brain. To gain a more complete picture of altered DNAm in autism, we carried out Reduced Representation Bisulfite Sequencing (RRBS) in 71 post-mortem cortical brain samples (BA19) at single nucleotide resolution with a quantitative measurement of DNAm across CpG-dense regions of the genome[149,150]. RRBS, in addition to querying methylation at more sites than the previously-used Infinium HumanMethylation450 array (Illumina)[35,36], enables measurement of methylation at cytosines outside of the classically studied CpG context. While CpH methylation (mCH, where H=A,C, or T) is rare in most tissues, it accumulates in DNA in human and mouse brain postnatally, ultimately reaching levels similar to that of CpG methylation (mCG) in brain DNA [151–153]. In contrast to mCG which remains largely unchanged during postnatal development, mCH accumulation correlates with synaptogenesis and increases especially during the first few years of life[151,152], a time period of particular interest in autism. Thus, we used post-mortem cortical brains samples to characterize CpG and CpH methylation in autism affected brain tissue and compared this to matched neurologically normal control brain tissue.

5.2 Methods

5.2.1 Samples

Samples were acquired through the Autism Tissue Program (which has since joined with the Autism Brain Net, https://autismbrainnet.org/). Post-mortem, frozen brain samples from the cerebral cortex Brodmann area (BA) 19 were collected at two different brain banks: the Harvard Brain Tissue Resource Center and the NICHD Brain and Tissue Bank at the University of Maryland with written informed consent having been obtained from next-of-kin or a legal guardian. Work herein was both approved by the IRB of The Johns Hopkins Hospital and University of Alabama at Birmingham and conducted in accordance with institutional guidelines.

5.2.2 RRBS library preparation

Seventy-one samples were prepared for reduced representation bisulfite sequencing (RRBS). RRBS libraries were prepared using 100 ng of genomic DNA. gDNA was first digested with MspI making cuts exclusively at methylated cytosines. 3' A-overhangs were created and filled in with Klenow Fragments. DNA was then purified using the Qiagen MinElute Kit. Methylated ilAdap PE adapters (Illumina) were ligated to purified gDNA. Fragment size selection (105-185bp) was carried out by gel extraction on a 2.5% NuSieve GTG agarose gel (Lonza). DNA was purified using Qiaquick Gel extraction Kit eluting DNA in elution bugger pre-warmed to 55 degrees Celsius. Bisulfite treatment was performed using the ZymoResearch EZ DNA Methylation Gold Kit following manufacturer's instructions; however, we eluted with 20µl M-Elution buffer. Bisulfite-treated DNA was cleaned up using EpiTect spin columns. Samples were PCR amplified (using the following primers: AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC*T and CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC*T; *=phosphorothioate bond) and size selection was carried out on a 3% Metaphor Agarose Gel to ensure that fragments of the correct size (175-275bp) were amplified. PCR product was cleaned up using the Qiagen minElute column, eluting with elution buffer warmed to 55 degrees Celsius. Each sample (10nM) was sequenced in a single lane on the Illumina HiSeq2000 to produce 50bp single end reads.

5.2.3 Alignment

Adaptor sequences were removed and reads shorter than 20 bp were excluded using Trim Galore! (v0.2.8). Remaining reads were aligned using Bismark (v0.7.7)[154], a methylation-specific wrapper for Bowtie[54] allowing for one mismatch and setting the seed substring length to 24.

5.2.4 Methylation estimation

Two separate analyses were carried out based on cytosine context; one for cytosines in the CpG context and a separate analysis for all other cytosines in the genome (CpH). Thus, samfiles for every sample and each of the two contexts were formatted for input into the R package 'methylKit'[155] (v0.9.5) using in-house scripts. Reads were filtered in methylKit based on read count discarding bases with coverage below 10X as well as those with coverage greater than the 99.9th percentile of coverage in each sample to remove reads suffering from PCR bias. Data were normalized based on median coverage and methylation percentage estimated using 'normalizeCoverage' and 'percMethylation', respectively within methylKit.

5.2.5 Illumina 27K methylation array

To independently verify methylation estimates from RRBS, CpG methylation was also analyzed in 71 cortical brain samples using the HumanMethylation27 BeadChip. These samples comprised 41 controls and 30 autism cases. Data were generated as described previously[156]. Normalized β–values were used for analysis. For comparison to RRBS data, mean methylation was quantified for the 1,249 CpGs that directly overlapping between the two platforms. Sites directly measured by both platforms had highly correlated measures of mean methylation (R^2=0.92), offering confidence in the measurements acquired by RRBS (**Figure 5.1**).

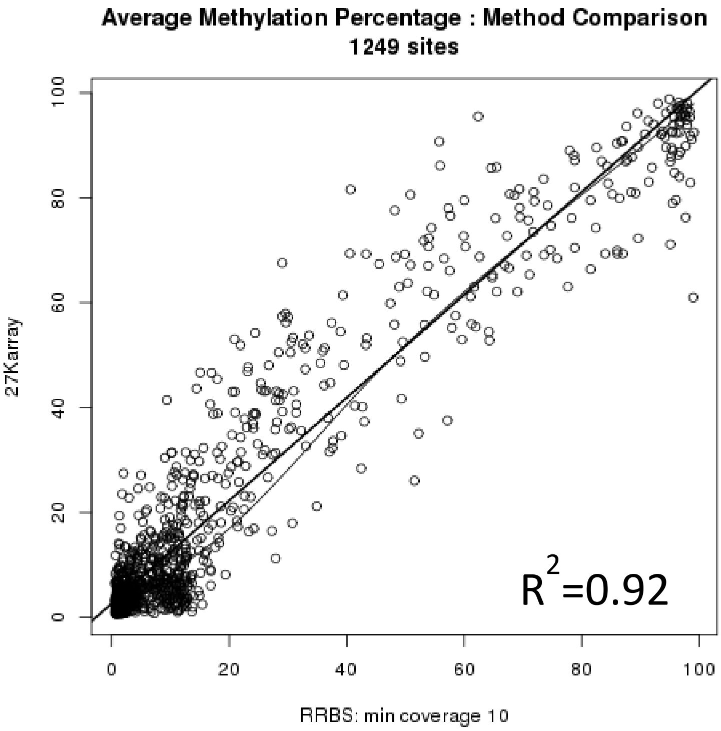


Figure 5.1 Correlation between mean methylation measured by RRBS and array.
The 1,249 CpGs directly measured by both RRBS and Illumina's 27K methylation array demonstrate highly correlated mean methylation values across samples (R^2=0.92).

5.2.6 Sample outlier removal

Four samples were excluded from analysis upon initial diagnostics as their profiles indicated failed library preparation or failed sequencing. Two were removed due to the fact that nearly all (>99%) of their cytosines were methylated after alignment and methylation estimation. A third sample was removed because its CpG methylation percentage distribution was not bimodal. The fourth sample was removed because its read coverage distribution did not match the expected distribution.

After identifying samples that failed library preparation and/or sequencing, remaining sample outliers were identified using surrogate variable analysis (SVA).[50] Ten surrogate variables (SVs) were generated using methylation estimates from CpG sites with data across all samples (254,824 CpGs). Samples greater than four standard deviations away from the mean in any of the SVs generated were identified as sample outliers. This process was carried out iteratively, and after each round of sample outlier removal, the percentage of known brain meQTLs[157,158] detected was quantified using a method previously developed for RNA-

Sequencing data[75] to guide data analysis. After each round of sample outlier removal, *cis* meQTLs (1Mb) were detected at SNPs and CpGs present in both the previously reported meQTL studies and the brain data using high quality genotype data described previously for these samples[75]. meQTLs were detected using MatrixEQTL[159] with age, sex, site and SVs included as covariates, and the percentage of known meQTLs was recorded. This process enabled us to confidently move forward with 63 samples, including 29 autistic cases and 34 controls, as this sample size maximized the percentage of known meQTLs detected, in all downstream analyses (**Figure 5.2**).

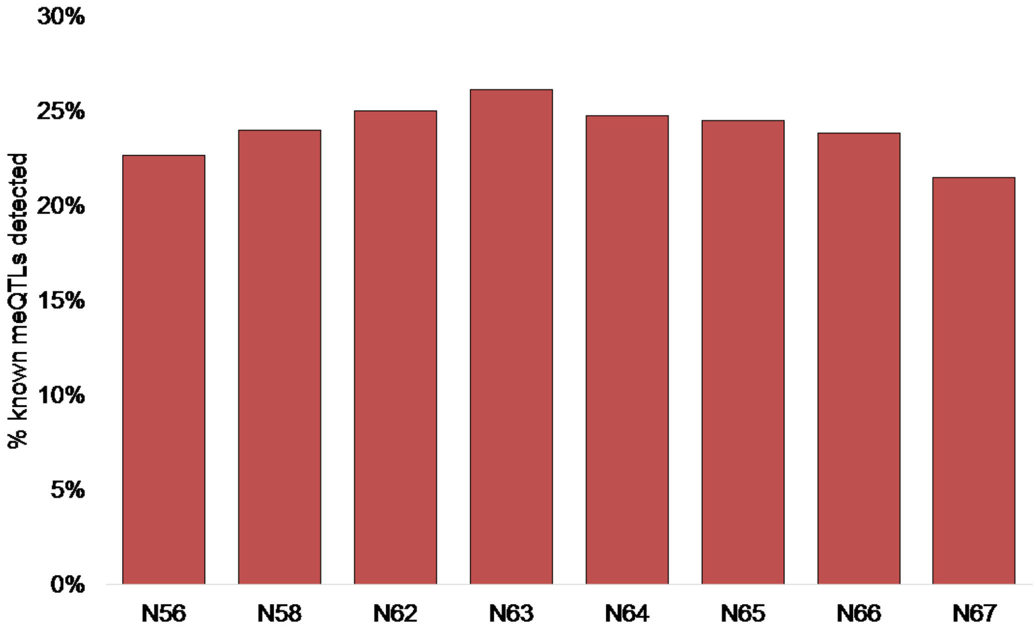


Figure 5.2 Previously reported meQTLs after sample outlier removal.
After removal of samples that failed sequencing, 63 samples remained for analysis. SVA was then utilized to identify sample outliers. After each round of outlier identification, the percentage of previously-reported meQTLs (y-axis) detected in the remaining samples (sample size indicated on the x-axis) was calculated. meQTL detection was maximized with 63 samples. These samples were included in downstream analyses.

5.2.7 Single site differential methylation analysis

Methylation outliers at each single site were defined as any sample greater than three standard deviations away from the mean methylation at that site and removed. Only variant sites were included for analysis, excluding the 25% least variable sites from analysis. Single site differential methylation was then carried out on each site using the '*lmFit*' function in the ‘limma’ R package[160]. For all cytosines, case-control status was regressed on methylation percentage with age, sex, site, and ten SVs included as covariates (‘full model’). Ten SVs were generated using methylation data from all variant sites with data across all samples utilizing the “irw” method and protecting case-control status. Additionally, as read coverage impacts our confidence in methylation estimates, the log10 of read coverage at each site was included as weights in the model.

Statistical significance was determined by residual bootstrapping, again using ‘limma’. For each bootstrap, the full model (described above) was fit and residuals recorded. A null model, in which the variable of interest (here, case-control status) was excluded, was also fit. The residuals from the full model were resampled with replacement, randomizing the sample order. ‘Pseudonull’ data were then generating adjusting the fits from the null model with the resampled residuals from the full model. These pseudonull methylation values were then substituted as the outcome variable into the full model, generating a null set of p-values. These p-values were collected for each of the 1,000 bootstraps to empirically determine study-wide significance.

5.2.8 Global altered methylation analysis

For each cytosine context, the proportion of sites hypermethylated (defined as mean methylation in cases greater than zero) was calculated at a number of different p-value cutoffs (1, 0.5, 0.05, $5x1^{3}$, and $5x1^{4}$). To assign significance, this proportion was then compared to the proportion of sites hypermethylated in each of the bootstraps.

5.2.9 Lists of functional genomic categories

Lists for twenty-four different functional genomic categories to test for enrichment of hypermethylated cytosines within the CpH context were downloaded from three different sources: (**1**) the UCSC Genome Browser (mRNA, transcription factor binding sites (tfbs), DNase I hypersensitive sites (dnase), enhancers, CTCF binding sites (CTCF), segmental duplications (segdups), repetitive regions (repeats), and histone marks (H3K4m1 , H3K4m2 , H3K4m3 , H3K9Ac , H3K9m3 , H3K27Ac , H3K27m3 , H3K36m3 , H3K79m2 , and H4K20m1), (**2**) UCL Cancer Institute (beacons) , and (**3**) the ‘methylKit’ package[155] (promoters, exons, introns, transcription start sites (TSS), CpG Islands (CGI), and CGI shores).

5.2.10 Functional enrichment testing

To test for genomic enrichment of hypermethylated CpH sites in each genomic list and at each p-value cutoff from the differential methylation analysis (0.5, 0.1, 0.05, 0.01, 5x13, 1x1^{3}, and 5x1^{4}), a two-sided Fisher's exact 2x2 test was carried out. For each list and at each differential methylation p-value cutoff, odds ratios and p-values for enrichment were recorded.

5.2.11 Power calculation

Power calculations were carried out using the "pwr.t2n.test" function from the 'pwr' package in R. This two-sided t-test of means for samples of different sizes (N=34 controls and 29 cases) was carried out at the 0.05 significance level (Type I error probability).

5.3 Results

After the removal of sample outliers, 63 samples were included for analysis, comprising 29 autism cases and 34 controls (**Table 5.1**). Methylation was estimated at cytosines with greater than 10 reads across at least 20 cases and 20 controls, yielding methylation estimates at 1.0M CpG and 3.3M CpH sites (**Figure 5.3**). No individual CpG or CpH sites were significantly differentially methylated after correction for multiple testing (**Tables 5.2** and **5.3** and **Figure 5.4**).

In addition to testing for differential methylation at individual sites, where our power to detect differences is limited, we also measured global changes associated with hypo- or hypermethylation. Among sites demonstrating suggestive differential methylation ($p<0.05$), there is a consistent and statistically significant proportion of cytosines demonstrating increased methylation within the CpH (**Figure 5.5b**, $p=0.02$), but not the CpG (**Figure 5.5a**) context. Further, given the fact that more stringent p-value cut-offs for differentially methylated sites should enrich for true positives, we hypothesized that the global hypermethylation signal would increase in strength with increasingly stringent p-value cut-offs in the CpH methylations associated with global hypermethylation, but not with the CpG sites that do not show global differences. Indeed, as more stringent differential methylation p-value cutoffs were imposed, a greater skewing in the number of hypermethylated to hypomethylated sites were observed (**Figure 5.5b** and **Figure 5.6b**). This trend was not seen in the CpG sites as hypothesized (**Figure 5.5a** and **Figure 5.6a**). Taken together, these data suggest that small increases (CpH sites with a differentially methylated p-value < 0.001 demonstrate a median 1.8% increase in cases relative to controls) in methylation across many individual sites are found at cytosines outside of the classically studied CpG context in the autistic brain.

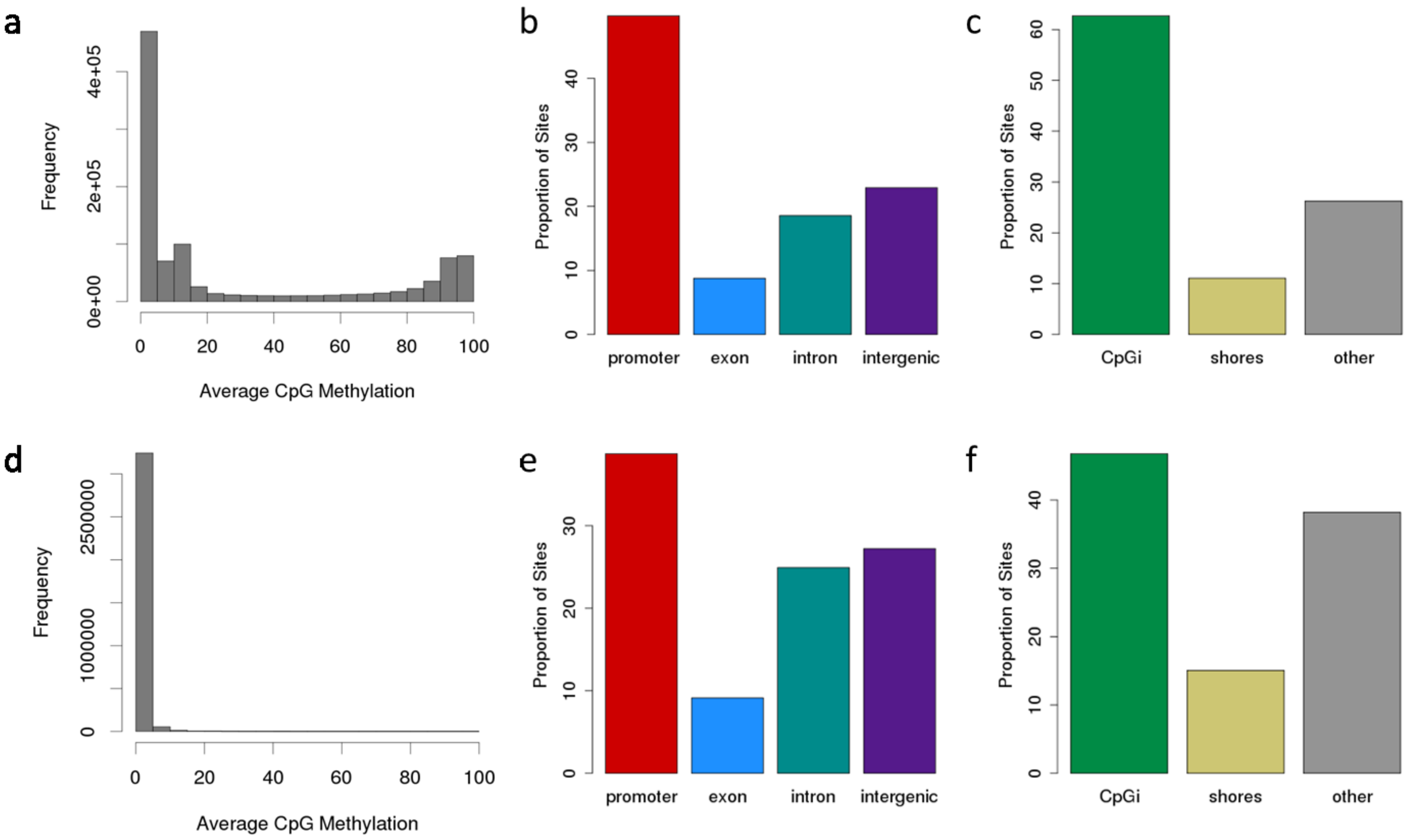


Figure 5.3 Summary of CpG and CpH sites.
Mean methylation distribution, genomic context proportions, and CpG Island proportions in CpG (**a-c**) and CpH (**d-f**) sites included for analysis.

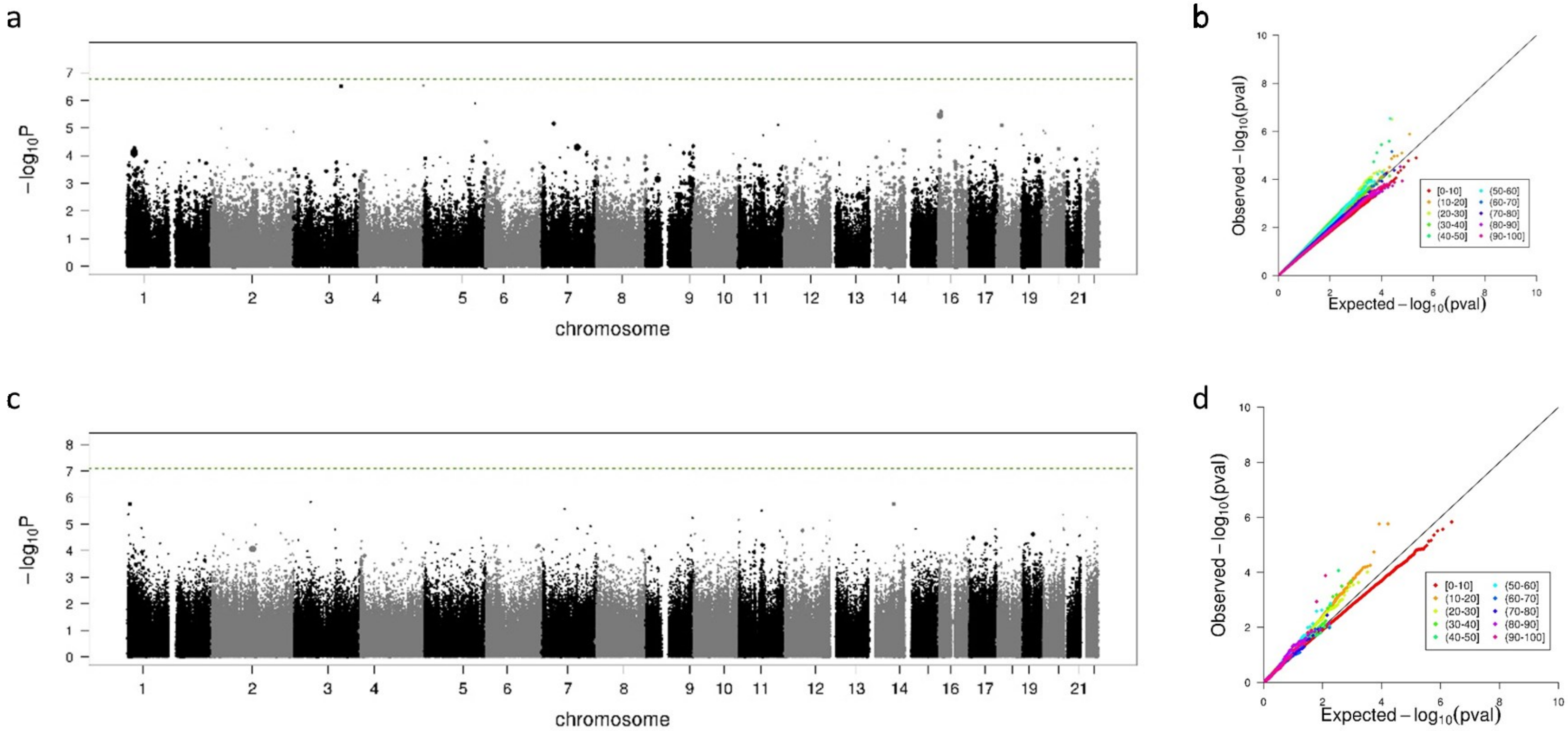
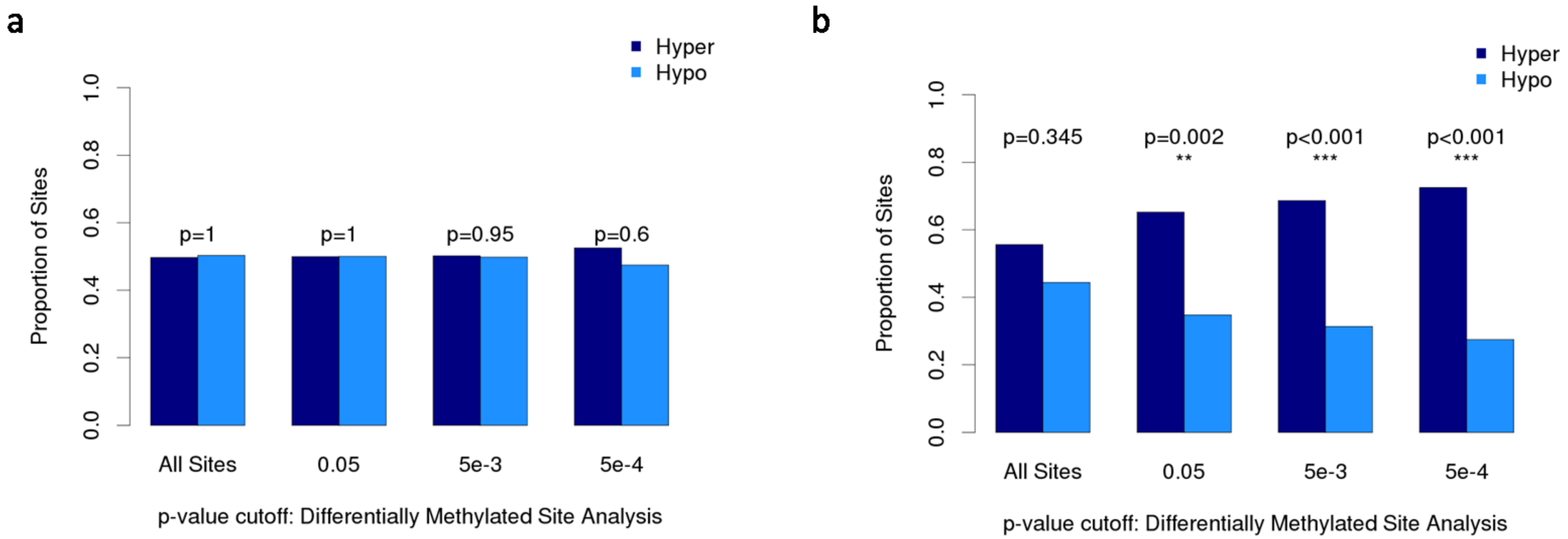


Figure 5.4 Single site differential methylation analysis.

Manhattan plots and QQ plots present results from the single site differential methylation analysis at CpG (**a-b**) and CpH sites (**c-d**) tested. Green dotted line on manhattan plots (**a,c**) indicate study-wide significance as determined by residual boostrapping. QQ plots (b,d) are colored by mean methylation value.



Number of Sites	769K	36K	3.3K	295
	382K	18K	1.7K	155
	387K	18K	1.7K	140

Number of Sites	2.5M	123K	10.8K	833
	1.4M	80K	7.4K	604
	1.1M	43K	3.4K	229

Figure 5.5 Proportion of hyper- and hypo- methylated sites in the CpG and CpH contexts.

Proportion of sites (y-axis) across increasingly stringent differentially methylated p-value cutoffs (x-axis). The number of cytosines at each differentially methylated p-value cutoff are displayed in the tables (below). (**a**) With approximately half of all sites demonstrating increased methylation (navy) and the other half decreased methylation (light blue), CpG sites behave as expected under the null. This pattern holds across increasingly stringent differential methylation p-value cut-offs demonstrating no global differences in methylation within the CpG context. (**b**) The proportion of cytosines demonstrating hypermethylation is not significantly different from the proportion demonstrating hypomethylation when looking at all CpH sites; however, with increasingly stringent differentially methylated p-value cutoffs, there is a significant proportion of hypermethylated CpH sites in the autistic brain.

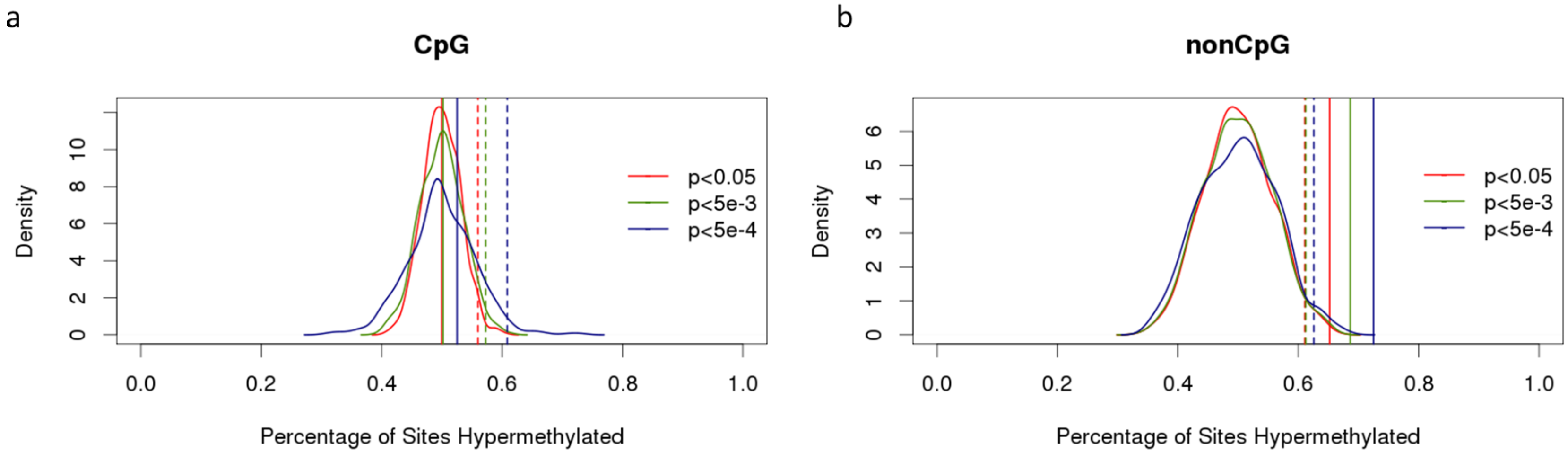


Figure 5.6 Null distributions of hypermethylation in CpG and CpH contexts.
Distributions plotted (solid curves) demonstrate expected proportion of hypermethylated sites under the null (as determined by residual boostrapping) at increasingly stringent p-value cutoffs from the differential methylation analysis ($p<0.05$ in red, $p< 0.005$ in green and $p<0.0005$ in blue) in the CpG (**a**) and CpH contexts (**b**). Dotted vertical lines indicate study-wide significance cutoffs. Solid vertical lines indicate signal in the data. Solid vertical lines to the right of its corresponding color dotted line indicates a study-wide significance (number of hypermethylated sites more extreme than expected under the null).

To better understand how altered mCH may be linked to the pathobiology of autism and aberrant neurodevelopment, we tested for enrichment of hypermethylated CpHs in various functional categories annotated across the genome. We used a Fisher's exact test to detect enrichment of hypermethylated cytosines in 24 different functional categories of the genome at several thresholds produced in the differential methylation analysis. This analysis highlights a role for increased methylation at CpH sites within repetitive regions of the genome and in regions that contain non-polymorphic human-specific CpGs, termed beacons23 (**Figure 5.7**). This finding implicates increased methylation within autism brain tissue at cytosines outside of the canonical CpG di-nucleotide It is not clear whether increased CpH methylation in autism is causal, protective, or benign in the etiology of disease. Given that mCH is specifically enriched in both the human and mouse brain[152], future studies can begin to probe the function of CpH methylation in successful and aberrant neurodevelopment.

To maximize the number of samples that could be sequenced, this work employed RRBS rather than whole genome bisulfite sequencing (WGBS). As RRBS enriches for CpG rich regions of the genome, we are unable to estimate methylation for cytosines outside of CpG rich regions. As sequencing costs continue to decline, WGBS of all the available brain tissue specimens will become more feasible and will undoubtedly add further insight into the role of methylation and other epigenetic phenomenon in autism. Additionally, given the extreme rarity of samples, sample size is always a cause for concern in post-mortem brain studies. Here, we report findings from the largest number of samples studied to date. As such, we are 80% powered to detect mean methylation differences greater than or equal 2.6% (**Figure 5.8**); however, cytosines of smaller effect could have been missed in these analyses.

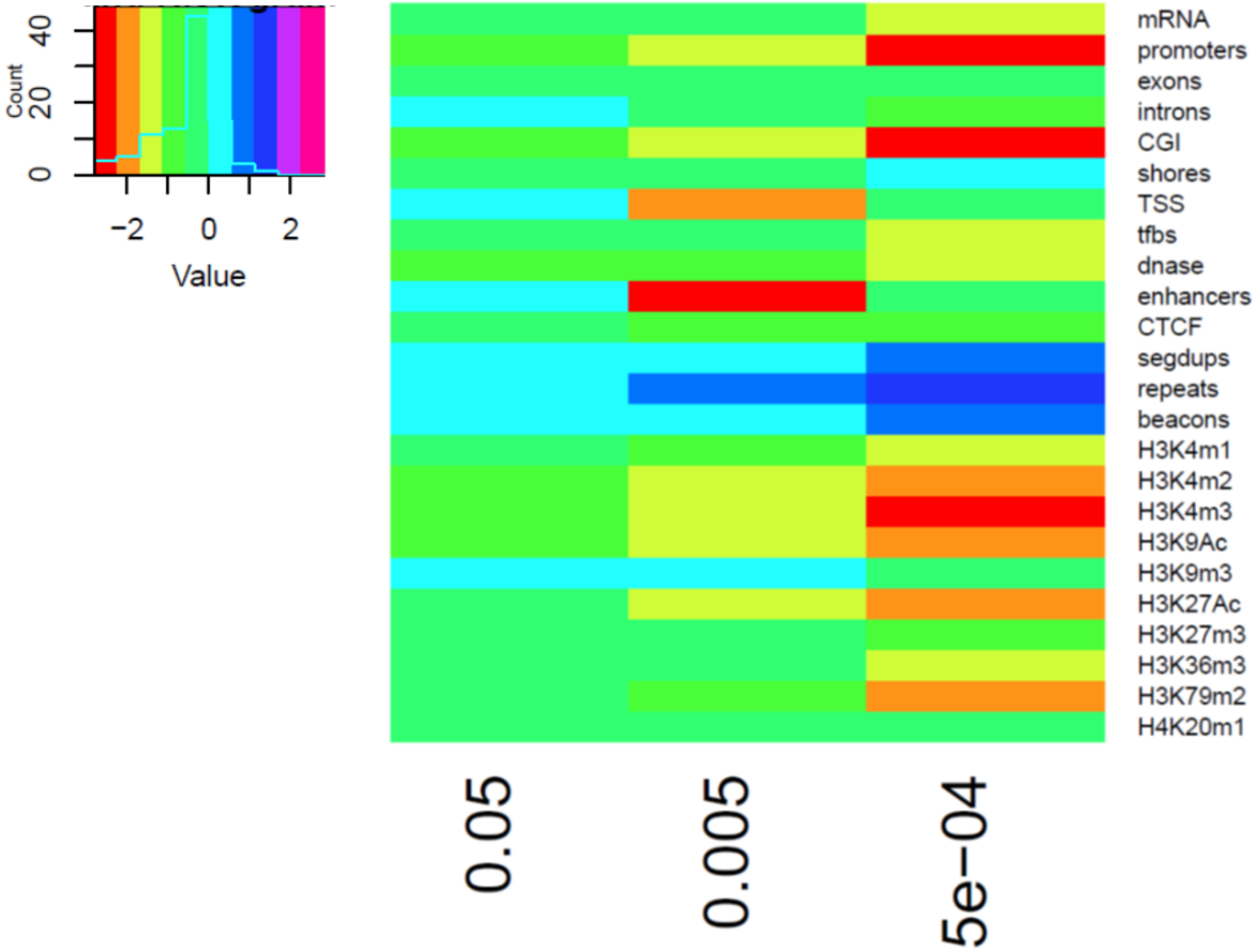


Figure 5.7 Functional genomic enrichment of hypermethylated CpH sites.
Enrichment of hypermethylated CpH sites at increasingly stringent differential methylation analysis p-value cutoffs (x-axis) was tested in 24 functional genomic categories. Odds ratios are plotted in the heatmap.

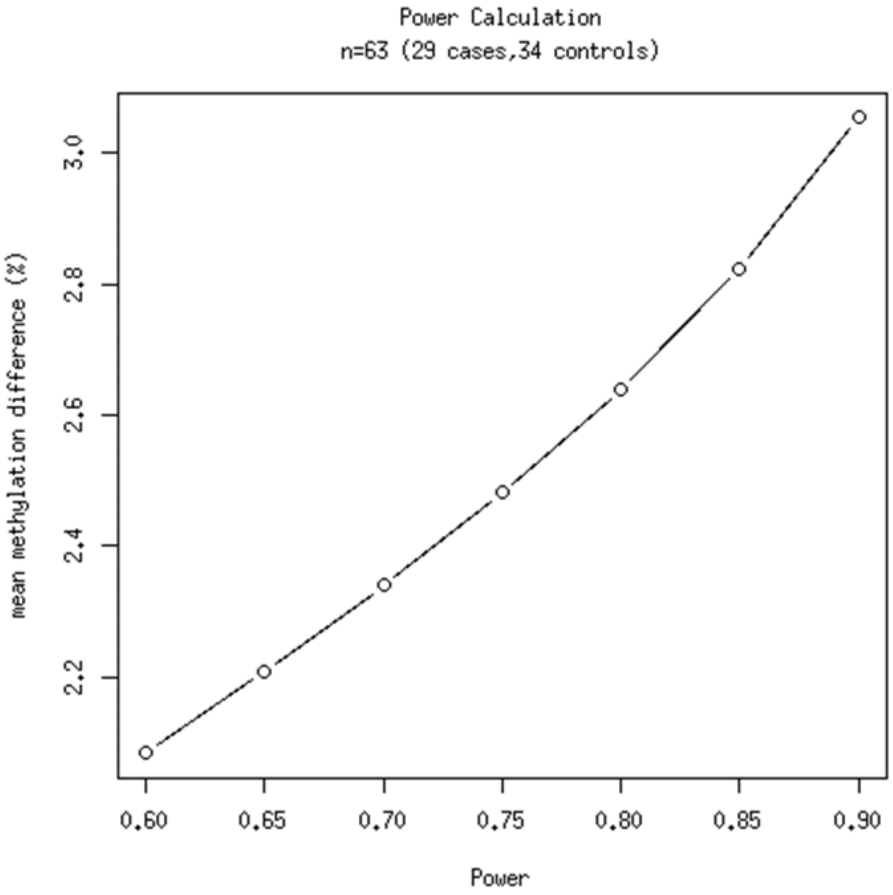


Figure 5.8 Power calculation.

Power calculation demonstrates that given our sample size, we are 80% powered to detect mean methylation differences between cases and controls greater than or equal to 2.6%.

5.4 Conclusions

In summary, this is the first genome-wide characterization of mCH methylation in autism affected brains, and while we do not detect any differences in individual sites in either the CpG or the CpH context after accounting for multiple testing, we report that increased CpH methylation occurs throughout the genome in DNA from autism affected brain. These CpH sites are strongly associated with repetitive regions and beacons, offering a first glimpse into how the epigenome may be affected in autism.

5.5 Tables

Table 5.1 Sample Information

	age	sex[†]	site[*]
Sample11	18	0	0
Sample14	2	1	0
Sample15	49	0	0
Sample16	29	0	0
Sample17	30	1	0
Sample22	27	1	0
Sample23	28	1	0
Sample25	5	0	0
Sample26	11	1	0
Sample27	15	1	0
Sample28	49	0	0
Sample2	60	1	0
Sample30	20	1	0
Sample31	56	1	0
Sample32	36	1	0
Sample33	42	1	0
Sample34	68	1	0
Sample35	41	1	0
Sample36	36	0	0
Sample37	38	1	0
Sample38	21	1	0
Sample3	16	1	0
Sample40	30	1	0
Sample41	32	0	0
Sample43	56	1	0
Sample44	9	0	1
Sample45	5	1	1
Sample46	20	0	1
Sample47	8	1	1
Sample48	8	1	1
Sample49	7	1	1
Sample4	14	1	1
Sample50	20	1	1
Sample51	4	0	1
Sample55	4	1	1
Sample56	5	0	1
Sample57	6	1	1
Sample60	4	1	1

Sample63	18	1	1
Sample64	18	1	1
Sample66	16	1	1
Sample67	20	0	1
Sample68	17	1	1
Sample70	39	1	0
Sample71	18	0	1
Sample72	20	1	1
Sample73	13	1	1
Sample74	16	1	1
Sample75	15	1	1
Sample76	19	1	1
Sample78	15	0	1
Sample79	13	1	1
Sample7	16	0	1
Sample80	5	1	0
Sample81	20	1	1
Sample82	16	0	1
Sample84	16	1	1
Sample85	20	1	1
Sample86	17	1	1
Sample88	56	1	0
Sample89	39	1	0
Sample8	11	0	0
Sample9	8	1	0

[†]Sex: 1=male, 0=female

[*]Site: 0=Harvard, 1=University of Maryland

Table 5.2 Top 10 differentially methylated CpG sites

chr	start	end	Mean Methylation	p-value	Methylation Difference
chr4	187079352	187079352	56.82	7.35E-07	-19.64
chr2	162279930	162279930	16.10	1.37E-06	-7.46
chr3	140823463	140823463	26.60	1.48E-06	-22.70
chr7	31557285	31557285	67.86	4.02E-06	-15.08
chr17	61781716	61781716	97.09	5.12E-06	6.30
chr22	50326315	50326315	95.10	6.11E-06	10.95
chr11	70851605	70851605	46.36	7.23E-06	-10.92
chr14	100197629	100197629	60.54	8.50E-06	-17.94
chr19	1395122	1395122	70.23	8.51E-06	18.20
chr22	36159533	36159533	98.10	1.04E-05	3.60

Table 5.3 Top 10 differentially methylated CpH sites

chr	start	end	Mean Methylation	p-value	Methylation Difference
chr4	184827961	184827961	2.37	1.60E-06	-4.93
chr17	11593366	11593366	8.32	2.52E-06	7.52
chr3	52099473	52099473	0.56	3.36E-06	2.13
chr1	12164559	12164559	2.42	3.40E-06	3.47
chr14	75430313	75430313	12.57	3.76E-06	8.14
chr8	144680079	144680079	0.87	5.33E-06	2.27
chr12	109251714	109251714	2.33	7.92E-06	-6.47
chr10	75532643	75532643	2.82	9.47E-06	-2.51
chrX	153979280	153979280	1.17	1.01E-05	2.38
chr15	75135664	75135664	1.70	1.04E-05	2.38

REFERENCES

1. American Psychiatric Association, American Psychiatric Association & DSM-5 Task Force. *Diagnostic and statistical manual of mental disorders: DSM-5.* (2013).
2. Matelski, L. & Van de Water, J. Risk factors in autism: Thinking outside the brain. *J. Autoimmun.* **67,** 1–7 (2016).
3. Lai, M.-C., Lombardo, M. V. & Baron-Cohen, S. Autism. *The Lancet* **383,** 896–910 (2014).
4. Developmental Disabilities Monitoring Network Surveillance Year 2010 Principal Investigators & Centers for Disease Control and Prevention (CDC). Prevalence of autism spectrum disorder among children aged 8 years - autism and developmental disabilities monitoring network, 11 sites, United States, 2010. *Morb. Mortal. Wkly. Rep. Surveill. Summ. Wash. DC 2002* **63,** 1–21 (2014).
5. Hallmayer J, Cleveland S, Torres A & et al. GEnetic heritability and shared environmental factors among twin pairs with autism. *Arch. Gen. Psychiatry* **68,** 1095–1102 (2011).
6. Sandin, S. *et al.* The familial risk of autism. *JAMA* **311,** 1770–1777 (2014).
7. Gaugler, T. *et al.* Most genetic risk for autism resides with common variation. *Nat. Genet.* **46,** 881–885 (2014).
8. Marshall, C. R. & Scherer, S. W. Detection and characterization of copy number variation in autism spectrum disorder. *Methods Mol. Biol. Clifton NJ* **838,** 115–135 (2012).
9. Sebat, J. *et al.* Strong Association of De Novo Copy Number Mutations with Autism. *Science* **316,** 445–449 (2007).

10. Levy, D. *et al.* Rare De Novo and Transmitted Copy-Number Variation in Autistic Spectrum Disorders. *Neuron* **70,** 886–897 (2011).

11. O'Roak, B. J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* (2012). doi:10.1038/nature10989

12. Sanders, S. J. *et al.* De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485,** 237–241 (2012).

13. Yu, T. W. *et al.* Using whole-exome sequencing to identify inherited causes of autism. *Neuron* **77,** 259–273 (2013).

14. Iossifov, I. *et al.* De Novo Gene Disruptions in Children on the Autistic Spectrum. *Neuron* **74,** 285–299 (2012).

15. Neale, B. M. *et al.* Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* (2012). doi:10.1038/nature11011

16. De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515,** 209–215 (2014).

17. Weiss, L. A., Arking, D. E., Daly, M. J. & Chakravarti, A. A genome-wide linkage and association scan reveals novel loci for autism. *Nature* **461,** 802–808 (2009).

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

19. Cross-Disorder Group of the Psychiatric Genomics Consortium *et al.* Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat. Genet.* **45,** 984–994 (2013).

20. Voineagu, I. *et al.* Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* **474,** 380–384 (2011).

21. Parikshak, N. N. *et al.* Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell* **155,** 1008–1021 (2013).

22. Willsey, A. J. *et al.* Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. *Cell* **155,** 997–1007 (2013).

23. Noh, H. J. *et al.* Network topologies and convergent aetiologies arising from deletions and duplications observed in individuals with autism. *PLoS Genet.* **9,** e1003523 (2013).

24. Geschwind, D. H. Autism: Many Genes, Common Pathways? *Cell* **135,** 391–395 (2008).

25. Geschwind, D. H. & State, M. W. Gene hunting in autism spectrum disorder: on the path to precision medicine. *Lancet Neurol.* **14,** 1109–1120 (2015).

26. Hu, V. W., Frank, B. C., Heine, S., Lee, N. H. & Quackenbush, J. Gene expression profiling of lymphoblastoid cell lines from monozygotic twins discordant in severity of autism reveals differential regulation of neurologically relevant genes. *BMC Genomics* **7,** 118 (2006).

27. Hu, V. W. *et al.* Gene expression profiling of lymphoblasts from autistic and nonaffected sib pairs: altered pathways in neuronal development and steroid biosynthesis. *PloS One* **4,** e5775 (2009).

28. Hu, V. W. *et al.* Gene expression profiling differentiates autism case-controls and phenotypic variants of autism spectrum disorders: evidence for circadian rhythm dysfunction in severe autism. *Autism Res. Off. J. Int. Soc. Autism Res.* **2,** 78–97 (2009).

29. Baron, C. A., Liu, S. Y., Hicks, C. & Gregg, J. P. Utilization of lymphoblastoid cell lines as a system for the molecular modeling of autism. *J. Autism Dev. Disord.* **36,** 973–982 (2006).

30. Nishimura, Y. *et al.* Genome-wide expression profiling of lymphoblastoid cell lines distinguishes different forms of autism and reveals shared pathways. *Hum. Mol. Genet.* **16,** 1682–1698 (2007).

31. LaSalle, J. M. Epigenomic strategies at the interface of genetic and environmental risk factors for autism. *J. Hum. Genet.* **58,** 396–401 (2013).

32. Horsthemke, B. & Wagstaff, J. Mechanisms of imprinting of the Prader–Willi/Angelman region. *Am. J. Med. Genet. A.* **146A,** 2041–2052 (2008).

33. Amir, R. E. *et al.* Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nat. Genet.* **23,** 185–188 (1999).

34. Oberlé, I. *et al.* Instability of a 550-base pair DNA segment and abnormal methylation in fragile X syndrome. *Science* **252,** 1097–1102 (1991).

35. Nardone, S. *et al.* DNA methylation analysis of the autistic brain reveals multiple dysregulated biological pathways. *Transl. Psychiatry* **4,** e433 (2014).

36. Ladd-Acosta, C. *et al.* Common DNA methylation alterations in multiple brain regions in autism. *Mol. Psychiatry* (2013). doi:10.1038/mp.2013.114

37. Bainbridge, M. N. *et al.* Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC Genomics* **7,** 246 (2006).

38. Allison, D. B., Cui, X., Page, G. P. & Sabripour, M. Microarray data analysis: from disarray to consolidation and consensus. *Nat. Rev. Genet.* **7,** 55–65 (2006).

39. Shendure, J. The beginning of the end for microarrays? *Nat. Methods* **5,** 585–587 (2008).

40. Nekrutenko, A. & Taylor, J. Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nat. Rev. Genet.* **13,** 667–672 (2012).

41. Moore, J. H., Asselbergs, F. W. & Williams, S. M. Bioinformatics challenges for genome-wide association studies. *Bioinforma. Oxf. Engl.* **26,** 445–455 (2010).

42. Weale, M. E. in *Genetic Variation* (eds. Barnes, M. R. & Breen, G.) **628,** 341–372 (Humana Press, 2010).

43. Kim, S., Cho, H., Lee, D. & Webster, M. J. Association between SNPs and gene expression in multiple regions of the human brain. *Transl. Psychiatry* **2,** 113 (2012).

44. Zou, F. *et al.* Brain Expression Genome-Wide Association Study (eGWAS) Identifies Human Disease-Associated Variants. *PLoS Genet* **8,** e1002707 (2012).

45. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45,** 580–585 (2013).

46. Leek, J. T. *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* **11,** 733–739 (2010).

47. Garber, M., Grabherr, M. G., Guttman, M. & Trapnell, C. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat. Methods* **8,** 469–477 (2011).

48. Mostafavi, S. *et al.* Normalizing RNA-sequencing data by modeling hidden covariates with prior knowledge. *PloS One* **8,** e68141 (2013).

49. Stegle, O., Parts, L., Durbin, R. & Winn, J. A Bayesian Framework to Account for Complex Non-Genetic Factors in Gene Expression Levels Greatly Increases Power in eQTL Studies. *PLoS Comput. Biol.* **6,** (2010).

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

51. Teschendorff, A. E., Zhuang, J. & Widschwendter, M. Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. *Bioinforma. Oxf. Engl.* **27,** 1496–1505 (2011).

52. Korn, J. M. *et al.* Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat. Genet.* **40,** 1253–1260 (2008).

53. Howie, B., Marchini, J. & Stephens, M. Genotype Imputation with Thousands of Genomes. *G3 Genes Genomes Genet.* **1,** 457–470 (2011).

54. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).

55. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–1111 (2009).

56. Risso, D., Schwartz, K., Sherlock, G. & Dudoit, S. GC-Content Normalization for RNA-Seq Data. *BMC Bioinformatics* **12,** 480 (2011).

57. 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 (2012).

58. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).

59. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).

60. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).

61. Xia, K. *et al.* seeQTL: a searchable database for human eQTLs. *Bioinformatics* **28,** 451–452 (2012).

62. Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28,** 1353–1358 (2012).

63. Aulchenko, Y. S., Ripke, S., Isaacs, A. & van Duijn, C. M. GenABEL: an R library for genome-wide association analysis. *Bioinforma. Oxf. Engl.* **23,** 1294–1296 (2007).

64. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55,** 997–1004 (1999).

65. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci.* **100,** 9440–9445 (2003).

66. Nica, A. C. *et al.* The Architecture of Gene Regulatory Variation across Multiple Human Tissues: The MuTHER Study. *PLoS Genet* **7,** e1002003 (2011).

67. Hansen, K. D., Irizarry, R. A. & Wu, Z. Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostat. Oxf. Engl.* **13,** 204–216 (2012).

68. ’t Hoen, P. A. C. *et al.* Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. *Nat. Biotechnol.* **31,** 1015–1022 (2013).

69. Nishida, N. *et al.* Evaluating the performance of Affymetrix SNP Array 6.0 platform with 400 Japanese individuals. *BMC Genomics* **9,** 431 (2008).

70. Yandell, M. & Ence, D. A beginner’s guide to eukaryotic genome annotation. *Nat. Rev. Genet.* **13,** 329–342 (2012).

71. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17,** 10–12 (2011).

72. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14,** R36 (2013).

73. Consortium, T. 1000 G. P. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).

74. Anders, S., Pyl, P. T. & Huber, W. HTSeq &#150; A Python framework to work with high-throughput sequencing data. *bioRxiv* (2014). doi:10.1101/002824

75. Ellis, S. E. *et al.* RNA-Seq optimization with eQTL gold standards. *BMC Genomics* **14,** 892 (2013).

76. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).

77. Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome Res.* **22,** 2008–2017 (2012).

78. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9,** 559 (2008).

79. Langfelder, P., Luo, R., Oldham, M. C. & Horvath, S. Is my network module preserved and reproducible? *PLoS Comput. Biol.* **7,** e1001057 (2011).

80. Cahoy, J. D. *et al.* A Transcriptome Database for Astrocytes, Neurons, and Oligodendrocytes: A New Resource for Understanding Brain Development and Function. *J. Neurosci.* **28,** 264–278 (2008).

81. Miller, J. A., Horvath, S. & Geschwind, D. H. Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 12698–12703 (2010).

82. Miller, J. A., Oldham, M. C. & Geschwind, D. H. A systems level analysis of transcriptional changes in Alzheimer's disease and normal aging. *J. Neurosci. Off. J. Soc. Neurosci.* **28,** 1410–1420 (2008).

83. Castillo-Davis, C. I. & Hartl, D. L. GeneMerge--post-genomic analysis, data mining, and hypothesis testing. *Bioinforma. Oxf. Engl.* **19,** 891–892 (2003).

84. Abrahams, B. S. *et al.* SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs). *Mol. Autism* **4,** 36 (2013).

85. Pinto, D. *et al.* Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. *Am. J. Hum. Genet.* **94,** 677–694 (2014).

86. Betancur, C. Etiological heterogeneity in autism spectrum disorders: more than 100 genetic and genomic disorders and still counting. *Brain Res.* **1380,** 42–77 (2011).

87. Steinberg, J. & Webber, C. The roles of FMRP-regulated genes in autism spectrum disorder: single- and multiple-hit genetic etiologies. *Am. J. Hum. Genet.* **93,** 825–839 (2013).

88. Talkowski, M. E. *et al.* Sequencing chromosomal abnormalities reveals neurodevelopmental loci that confer risk across diagnostic boundaries. *Cell* **149,** 525–537 (2012).

89. Darnell, J. C. *et al.* FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell* **146,** 247–261 (2011).

90. Nagata, T. *et al.* Profiling of genes associated with transcriptional responses in mouse hippocampus after transient forebrain ischemia using high-density oligonucleotide DNA array. *Brain Res. Mol. Brain Res.* **121,** 1–11 (2004).

91. Uddin, M. *et al.* Brain-expressed exons under purifying selection are enriched for de novo mutations in autism spectrum disorder. *Nat. Genet.* **46,** 742–747 (2014).

92. Esparza-Gordillo, J. *et al.* A common variant on chromosome 11q13 is associated with atopic dermatitis. *Nat. Genet.* **41,** 596–601 (2009).

93. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9,** 559 (2008).

94. Mantovani, A., Biswas, S. K., Galdiero, M. R., Sica, A. & Locati, M. Macrophage plasticity and polarization in tissue repair and remodelling. *J. Pathol.* **229,** 176–185 (2013).

95. Nicolae, D. L. *et al.* Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.* **6,** e1000888 (2010).

96. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337,** 1190–1195 (2012).

97. Delhaye, S. *et al.* Neurons produce type I interferon during viral encephalitis. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 7835–7840 (2006).

98. Cunningham, C. L., Martínez-Cerdeño, V. & Noctor, S. C. Microglia regulate the number of neural precursor cells in the developing cerebral cortex. *J. Neurosci. Off. J. Soc. Neurosci.* **33,** 4216–4233 (2013).

99. Miron, V. E. *et al.* M2 microglia and macrophages drive oligodendrocyte differentiation during CNS remyelination. *Nat. Neurosci.* **16,** 1211–1218 (2013).

100. Coull, J. A. M. *et al.* BDNF from microglia causes the shift in neuronal anion gradient underlying neuropathic pain. *Nature* **438,** 1017–1021 (2005).

101. Patterson, P. H. Maternal infection and immune involvement in autism. *Trends Mol. Med.* **17,** 389–394 (2011).

102. Courchesne, E. *et al.* Mapping early brain development in autism. *Neuron* **56,** 399–413 (2007).

103. Ben-David, E. & Shifman, S. Networks of Neuronal Genes Affected by Common and Rare Variants in Autism Spectrum Disorders. *PLoS Genet* **8,** e1002556 (2012).

104. Mäkinen, V.-P. *et al.* Integrative genomics reveals novel molecular pathways and gene networks for coronary artery disease. *PLoS Genet.* **10,** e1004502 (2014).

105. Derecki, N. C. *et al.* Wild-type microglia arrest pathology in a mouse model of Rett syndrome. *Nature* **484,** 105–109 (2012).

106. Morgan, J. T. *et al.* Abnormal microglial-neuronal spatial organization in the dorsolateral prefrontal cortex in autism. *Brain Res.* **1456,** 72–81 (2012).

107. Schafer, D. P. *et al.* Microglia sculpt postnatal neural circuits in an activity and complement-dependent manner. *Neuron* **74,** 691–705 (2012).

108. Rutter, M. Childhood schizophrenia reconsidered. *J. Autism Child. Schizophr.* **2,** 315–337 (1972).

109. World Health Organization. The ICD-10 Classification of Mental and Behavioural Disorders. (1993).

110. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders. 4th ed. (1994).

111. Smoller, J. W. & Finn, C. T. Family, twin, and adoption studies of bipolar disorder. *Am. J. Med. Genet. C Semin. Med. Genet.* **123C,** 48–58 (2003).

112. Carroll, L. S. & Owen, M. J. Genetic overlap between autism, schizophrenia and bipolar disorder. *Genome Med.* **1,** 102 (2009).

113. Ruderfer, D. M. *et al.* Polygenic dissection of diagnosis and clinical dimensions of bipolar disorder and schizophrenia. *Mol. Psychiatry* **19,** 1017–1024 (2014).

114. Cardno, A. G. & Owen, M. J. Genetic relationships between schizophrenia, bipolar disorder, and schizoaffective disorder. *Schizophr. Bull.* **40,** 504–515 (2014).

115. Crespi, B., Stead, P. & Elliot, M. Comparative genomics of autism and schizophrenia. *Proc. Natl. Acad. Sci.* **107,** 1736–1741 (2010).

116. Zhao, Z. *et al.* Transcriptome sequencing and genome-wide association analyses reveal lysosomal function and actin cytoskeleton remodeling in schizophrenia and bipolar disorder. *Mol. Psychiatry* **20,** 563–572 (2015).

117. Gupta, S. *et al.* Transcriptome analysis reveals dysregulation of innate immune response genes and neuronal activity-dependent genes in autism. *Nat. Commun.* **5,** (2014).

118. Ellis, S. E. *et al.* RNA-Seq optimization with eQTL gold standards. *BMC Genomics* **14,** 892 (2013).

119. The American Soldier: Vol. I: Adjustment During Army Life. by Samuel A. Stouffer; The American Soldier: Vol. II: Combat and Its Aftermath.

120. WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. Available at: http://nar.oxfordjournals.org/content/41/W1/W77.full. (Accessed: 25th August 2015)

121. Zhang, B., Kirov, S. & Snoddy, J. WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res.* **33,** W741-748 (2005).

122. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25,** 25–29 (2000).

123. Consortium, T. G. O. Gene Ontology Consortium: going forward. *Nucleic Acids Res.* **43,** D1049–D1056 (2015).

124. Sherman, B. T. *et al.* DAVID Knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis. *BMC Bioinformatics* **8,** 426 (2007).

125. YH Benjamini, Y. H. Controlling The False Discovery Rate - A Practical And Powerful Approach To Multiple Testing. *J R. Stat. Soc Ser. B* **57,** 289–300 (1995).

126. Chanda, P., Huang, H., Arking, D. E. & Bader, J. S. Fast association tests for genes with FAST. *PloS One* **8,** e68585 (2013).

127. Huang, H., Chanda, P., Alonso, A., Bader, J. S. & Arking, D. E. Gene-Based Tests of Association. *PLoS Genet* **7,** e1002177 (2011).

128. Talkowski, M. E. *et al.* Sequencing chromosomal abnormalities reveals neurodevelopmental loci that confer risk across diagnostic boundaries. *Cell* **149,** 525–537 (2012).

129. Kerr, M. K. & Churchill, G. A. Experimental design for gene expression microarrays. *Biostatistics* **2,** 183–201 (2001).

130. Tseng, G. C., Oh, M.-K., Rohlin, L., Liao, J. C. & Wong, W. H. Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res.* **29,** 2549–2557 (2001).

131. Yang, Y. H. *et al.* Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* **30,** e15–e15 (2002).

132. Kerr, M. K., Martin, M. & Churchill, G. A. Analysis of Variance for Gene Expression Microarray Data. *J. Comput. Biol.* **7,** 819–837 (2000).

133. Hattori, Y. *et al.* Identification of a neuron-specific human gene, KIAA1110, that is a guanine nucleotide exchange factor for ARF1. *Biochem. Biophys. Res. Commun.* **364,** 737–742 (2007).

134. Wei, N., Serino, G. & Deng, X.-W. The COP9 signalosome: more than a protease. *Trends Biochem. Sci.* **33,** 592–600 (2008).

135. Prevalence of Autism Spectrum Disorder Among Children Aged 8 Years — Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2010. Available at: http://www.cdc.gov/mmwr/preview/mmwrhtml/ss6302a1.htm. (Accessed: 23rd January 2016)

136. Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515,** 216–221 (2014).

137. Yu, T. W. *et al.* Using Whole-Exome Sequencing to Identify Inherited Causes of Autism. *Neuron* **77,** 259–273 (2013).

138. Sanders, S. J. *et al.* De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485,** 237–241 (2012).

139. O'Roak, B. J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **485,** 246–250 (2012).

140. Neale, B. M. *et al.* Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485,** 242–245 (2012).

141. Voineagu, I. *et al.* Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* **474,** 380–384 (2011).

142. Gupta, S. *et al.* Transcriptome analysis reveals dysregulation of innate immune response genes and neuronal activity-dependent genes in autism. *Nat. Commun.* **5,** 5748 (2014).

143. Amir, R. E. *et al.* Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nat. Genet.* **23,** 185–188 (1999).

144. Oberle, I. *et al.* Instability of a 550-base pair DNA segment and abnormal methylation in fragile X syndrome. *Science* **252,** 1097–1102 (1991).

145. Beuten, J. *et al.* Detection of imprinting mutations in Angelman syndrome using a probe for exon α of SNRPN. *Am. J. Med. Genet.* **63,** 414–415 (1996).

146. Nagarajan, R., Hogart, A., Gwye, Y., Martin, M. R. & LaSalle, J. M. Reduced MeCP2 Expression is Frequent in Autism Frontal Cortex and Correlates with Aberrant MECP2 Promoter Methylation. *Epigenetics* **1,** 172–182 (2006).

147. Gregory, S. G. *et al.* Genomic and epigenetic evidence for oxytocin receptor deficiency in autism. *BMC Med.* **7,** 62 (2009).

148. Nguyen, A., Rauch, T. A., Pfeifer, G. P. & Hu, V. W. Global methylation profiling of lymphoblastoid cell lines reveals epigenetic contributions to autism spectrum disorders and a novel autism candidate gene, RORA, whose protein product is reduced in autistic brain. *FASEB J.* **24,** 3036–3051 (2010).

149. Meissner, A. *et al.* Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res.* **33,** 5868–5877 (2005).

150. Smith, Z. D., Gu, H., Bock, C., Gnirke, A. & Meissner, A. High-throughput bisulfite sequencing in mammalian genomes. *Methods San Diego Calif* **48,** 226–232 (2009).

151. Lister, R. *et al.* Global Epigenomic Reconfiguration During Mammalian Brain Development. *Science* **341,** 1237905 (2013).

152. Guo, J. U. *et al.* Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain. *Nat. Neurosci.* **17,** 215–222 (2014).

153. Kinde, B., Gabel, H. W., Gilbert, C. S., Griffith, E. C. & Greenberg, M. E. Reading the unique DNA methylation landscape of the brain: Non-CpG methylation, hydroxymethylation, and MeCP2. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 6800–6806 (2015).

154. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27,** 1571–1572 (2011).

155. Akalin, A. *et al.* methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol.* **13,** R87 (2012).

156. Day, K. *et al.* Differential DNA methylation with age displays both common and dynamic features across human tissues that are influenced by CpG landscape. *Genome Biol.* **14,** R102 (2013).

157. Smith, A. K. *et al.* Methylation quantitative trait loci (meQTLs) are consistently detected across ancestry, developmental stage, and tissue type. *BMC Genomics* **15,** 145 (2014).

158. Zhang, D. *et al.* Genetic Control of Individual Differences in Gene-Specific Methylation in Human Brain. *Am. J. Hum. Genet.* **86,** 411–419 (2010).

159. Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28,** 1353–1358 (2012).

160. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* gkv007 (2015). doi:10.1093/nar/gkv007

161. Sanders, S. J. *et al.* De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* (2012). doi:10.1038/nature10945

162. Yu, T. W. *et al.* Using Whole-Exome Sequencing to Identify Inherited Causes of Autism. *Neuron* **77,** 259–273 (2013).

163. O'Roak, B. J. *et al.* Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nat. Genet.* **43,** 585–589 (2011).

APPENDIX 1: Using the human brain to understand “omics” in autism

Blog post for the Autism Science Foundation
May 19, 2016

Shannon E. Ellis

Autism is a complex neurodevelopmental disorder with a definitively established genetic basis[7]. However, complete understanding of the genetic etiology remains elusive. While the role for certain CNVs in autism has been definitively established[8,9] and exome sequencing studies have begun to uncover rare *de novo* mutations that play a role in the disorder[14–16,161–163], we are far from identifying the names of the hundreds of genes likely contributing to the disorder.

While naming the genes that play a role in autism is critical, it has become increasingly clear that changes in the DNA sequence are only a first step toward complete understanding. Additionally, scientists acknowledge that they must know what the genes do, not only what they are. Therefore, our lab has begun to get a handle on what is altered at the level of gene expression and DNA methylation within the primary affected tissue in autism – the brain. This is part of a larger field called “omics”: meaning “genomics” (the study of DNA sequence), “transcriptomics” (the study of gene expression), “epigenomics” (the study of how genes are expressed, “proteomics” (how proteins are regulated by DNA expression) and “metabolomics” (the chemical processes of making and breaking down compounds). Our study focused on transcriptomics and epigenomics.

To understand the processes in the brains of people with autism, we had to study the brains of people with autism. Therefore, we worked with the Autism BrainNet (formerly the Autism Tissue Program) to obtain brain samples and extract RNA. We were interested in RNA because it is produced from DNA. Using DNA as a template, RNA is formed as the first step in creating proteins. The proteins are what carry out the function of the cell. The number of copies of RNA is a reflection of the gene expression of the cell: the more copies, the more gene expression. While differences in an individual’s DNA are undoubtedly of interest to the study of autism, it is also important to look for differences in gene expression and protein levels to fully understand the disorder. We used the RNA to look at the gene expression of about 14,000 genes. Additionally, in these same individuals, DNA samples were used to estimate methylation levels at cytosines across cytosine-rich regions of the genome.

Last year, we published some results on the transcriptomic part. When we looked at the gene expression, we identified three groups of genes that showed different patterns of expression between autism cases and controls. The first two groups included genes that influence the way neurons work, how they interact with each other, and how they grow and communicate to form the human brain. Interestingly, the genes that showed differences in gene expression were different genes than those identified in previous studies that, rather than looking at gene expression differences, looked to identify DNA differences important to autism. This demonstrates that genes that have differences in their DNA are different from those genes showing downstream differences at the level of gene expression. The third group was made up primarily of M2-microglia genes, suggesting in increased immune response in the brains of autistic individuals. We want to stipulate that this does not mean that alterations in the immune system of the brain cause autism. It could be that abnormal gene expression in the brain triggers an M2 microglia response. Future work is required before we can determine if the increased immune response leads to or is a result of autism. Nevertheless, from a treatment standpoint, this work provides pathways that, despite variable genetic causes, can be targeted for treatment in affected individuals going forward.

We moved this further to compare the gene expression overlap in the brains of people with autism with other neuropsychiatric disorders. Previous studies have shown that there is an overlap in which genes have DNA differences across neuropsychiatric disorders. To establish if this overlap holds up at the level of the transcriptome, we compared gene expression differences across three disorders: autism, schizophrenia, and bipolar disorder. While there was little overlap between autism and bipolar disorder, there was significant overlap between the expression patterns in genes in autism and schizophrenia brains[11], with consistent decreased expression at neuronal and synaptic plasticity genes across these two disorders. This work extended the known genetic overlap between neuropsychiatric disorders by establishing a relationship between alterations in gene expression found in both autism and schizophrenia.

Finally, as gene expression is directly regulated by DNA methylation, we looked to determine if DNA methylation differences play a role in autism. Methylation is a process where a methyl group attaches to a part of a DNA sequence and turns down the expression of that gene. To look at DNA methylation, we looked at cytosines. Cytosines are places in our genome where these methyl groups normally attach. Thus far, most work has studied CpG methylation. This refers to when a methyl group attaches to a cytosine (C) that is directly next to a guanine (G) nucleotide. This CpG context is where methylation most frequently occurs in the genome. However, methylation can occur at cytosines next to other DNA nucleotides (C, T, or A), and this is referred to as CpH methylation. As CpH methylation occurs at higher levels in the brain relative to other tissues, we did not want to limit our study to CpG sites alone, but rather wanted to look for differences in CpH methylation as well. We found increased levels of methylation at CpH, but not CpG, sites globally within the autistic brain. We are currently working on why there is this difference and why it is seen in autism, but as CpH methylation is largely specific to the human brain (as compared to blood or other cells), it is a particularly compelling finding.

While we acknowledge we have not answered all the questions, we now have a better understanding of what is going on in the brain of individuals with autism. In particular, in addition to further establishing a relationship between schizophrenia and autism, this work not only highlights a role for increased immune activation within affected individuals, providing a particular pathway to target when considering future therapies, but also, for the first time, suggests a role for increased methylation at cytosines within the CpH context in the brains of autistic individuals.

References:

1. Gaugler, T. et al. Most genetic risk for autism resides with common variation. Nat. Genet. 46, 881–885 (2014).
2. Marshall, C. R. & Scherer, S. W. Detection and characterization of copy number variation in autism spectrum disorder. Methods Mol. Biol. Clifton NJ 838, 115–135 (2012).
3. Sebat, J. et al. Strong Association of De Novo Copy Number Mutations with Autism. Science 316, 445–449 (2007).
4. Sanders, S. J. et al. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature (2012).
5. Yu, T. W. et al. Using Whole-Exome Sequencing to Identify Inherited Causes of Autism. Neuron 77, 259–273 (2013).
6. Iossifov, I. et al. De Novo Gene Disruptions in Children on the Autistic Spectrum. Neuron 74, 285–299 (2012).
7. O'Roak, B. J. et al. Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. Nat. Genet. 43, 585–589 (2011).
8. Neale, B. M. et al. Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature (2012).
9. De Rubeis, S. et al. Synaptic, transcriptional and chromatin genes disrupted in autism. Nature 515, 209–215 (2014).
10. Gupta, S. et al. Transcriptome analysis reveals dysregulation of innate immune response genes and neuronal activity-dependent genes in autism. Nat. Commun. 5, 5748 (2014).
11. Ellis, S. E., Panitch, R., West, A. & Arking, D. E. Transcriptome Analysis of Cortical Tissue Reveals Shared Sets of Down-Regulated Genes in Autism and Schizophrenia. bioRxiv 29132 (2016).

APPENDIX 2: Analyses of gene activity can yield clues to roots of autism

Viewpoint piece for Spectrum
June 2016

Shannon E. Ellis and Dan E. Arking, Ph.D.

The number of genetic variants implicated in autism is large and growing, but it's increasingly clear that identifying these variants is only the beginning of the quest to understand the biology of autism.

Genes not only store information, they also serve as templates for RNA transcripts that then give rise to the proteins that function in a cell. So it is crucial that we understand which genes show differences in their RNA levels, too.

This is particularly important because although autism may affect large numbers of genes, their functions appear to converge on only a few biological pathways. These include the development of the brain's outer shell (cerebral cortex), the function of neuronal junctions (synapses), translation of the genetic code into protein, and the activation of brain immune cells called microglia[1].

Ultimately, these findings suggest that, despite multiple genetic causes, we may only need to target a few pathways to effectively treat people with autism.

Because a major source of biological regulation occurs at the level of gene expression, studying patterns of expression can highlight the key pathways in specific tissues such as the brain. This may then point us toward new therapies.

Deciphering data:

'Transcriptomic' studies refer to those in which researchers quantify gene expression across all genes present in a tissue sample and identify differences between people with autism and controls. This analysis generally reveals values for 10,000 to 20,000 genes, a number too large for the human brain to make sense of on its own.

Network analyses — software tools that identify biologically-relevant patterns from large datasets — are often the key to analyzing all these data.

In one type of analysis, called weighted gene co-expression network analysis (WGCNA), researchers group genes into modules based on the similarity of their expression patterns[2]. They then interpret the functional role of genes within each module, often intuiting these roles by looking at the genes in aggregate.

For example, in modules constructed from brain gene expression data, at least one module is likely to be enriched for genes expressed in neurons and another for genes in glia, cells that support neurons. These separate modules help to explain the cell types present in the tissue studied. From there, researchers can look at whether the expression pattern in a given module differs significantly between people with autism and those without.

Combining the functional information — for instance, what cell types are present and the function of genes in each module — with the expression patterns allows researchers to determine which pathways are altered in autism.

Imperfect analyses

Although network analyses help us make sense of large gene expression datasets, they have some limitations.

First, the number of individuals available for study could prevent a researcher from constructing biologically meaningful networks and prohibit useful interpretation of the data.

Second, because network analyses are designed to pick up subtle differences between groups, any systematic differences between cases and controls may lead researchers to incorrect conclusions.

For example, if samples from individuals with autism contain different cell types than those in controls (perhaps due to sampling of slightly different regions of the brain), WGCNA would pick up this difference, and a naïve researcher could incorrectly claim to have found an association with autism.

Accounting for such confounding factors is a key step for researchers to draw conclusions that can be trusted.

What's more, these analyses only decipher gene expression information. They are not designed to crunch data related to genetic variants, DNA modifications that affect gene expression, or information about the proteins themselves. Combining all of these datasets could provide an even more complete molecular understanding of autism. Groups are working on these multilevel analyses.

Finally, network analyses do not provide information on causality. In contrast to DNA, which generally does not change over a lifetime, gene expression levels differ over time. Mutations in DNA are almost certainly primary and may cause the condition, but differences in gene expression, as detected by network analyses, may or may not lead to autism. They might be compensatory — that is, they result from having autism.

Still, by pointing to pathways of interest, gene expression analyses can provide ideas for new treatments.

Promising pathways:

Given the potential for understanding gene expression using network analyses, it's important to consider what we can expect from this approach. In the past five years, researchers have published a number of studies looking at gene expression in postmortem brain samples[3,4]. This work has repeatedly highlighted differences in gene expression in the brains of individuals with autism in two main areas.

One of these highlights a role for altered expression of genes that are active in neurons. Intriguingly, however, the genes that harbor mutations are different from the ones showing altered expression. The neuronal pathways involving genes whose expression is altered in autism represent a promising set of drug targets separate from the genetic variants individuals may harbor.

The other reproducible pathway implicated by network analyses includes genes related to immune regulation. Many of the genes in this pathway tend to be active in a type of microglia that have an anti-inflammatory effect.

This result suggests that an exaggerated anti-inflammatory response occurs in the autism brain. Researchers will need to pin down whether this finding is causal or compensatory.

Network analyses with large sample sizes will undoubtedly improve our ability to identify which genes are driving these expression differences. They may also help us to detect genes whose expression differs only subtly in people with autism.

Taken together, network analyses of gene expression have identified two biologically interesting and testable pathways that can be targeted in therapeutic studies, emphasizing the utility of these approaches.

References:

1. Geschwind, D. H. & State, M. W. Gene hunting in autism spectrum disorder: on the path to precision medicine. Lancet Neurol. 14, 1109–1120 (2015).
2. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9, 559 (2008).
3. Voineagu, I. et al. Transcriptomic analysis of autistic brain reveals convergent molecular pathology. Nature 474, 380–384 (2011).
4. Gupta, S. et al. Transcriptome analysis reveals dysregulation of innate immune response genes and neuronal activity-dependent genes in autism. Nat. Commun. 5, 5748 (2014).

CURRICULUM VITAE

SHANNON E. ELLIS

733 N. Broadway Street • Miller Research Building, Room 420 • Baltimore, MD 21205
sellis18@jhmi.edu • (570) 793-7048

EDUCATION

Year	Degree	Institution	GPA	Discipline
2010—present	Ph.D.	Johns Hopkins, Baltimore, MD	--	Human Genetics
2006—2010	B.S.	King's College, Wilkes-Barre, PA	4.0	Biology & Spanish

RESEARCH EXPERIENCE

Graduate Student, 8/201present
Johns Hopkins University School of Medicine • Baltimore, MD
Laboratory of Dan E Arking, Ph.D.

- Utilized invaluable post-mortem cortical brain samples to better understand the largely elusive genetic basis of autism.
- Developed a method to guide RNA-Sequencing analysis using eQTLs as a gold standard.
- Analyzed RNA-Sequencing data to study alterations in gene expression in the brains of autistic individuals relative to controls. Identified an upregulation of activated M2 microglia genes in autism brains.
- Identified significant DNA hypermethylation at cytosines outside of the classically-studied CpG context in autism brains utilizing bisulfite sequencing.
- Wrote an R package ('methylarking') for one-step implementation of all methylation analyses.
- Analyzed data using R, Perl, and Python within a UNIX environment.

Rotation Student, 8/2010 - 12/2010
Johns Hopkins University School of Medicine • Baltimore, MD
Laboratory of Steven Brandt, M.D.

- Studied the association of three previously-identified Ulcerative Colitis susceptibility loci in African Americans, a population historically neglected in genetic studies.
- Utilized Taqman SNP genotyping or direct sequencing at the regions of interest to genotype seven loci in 293 Crohn's Disease patients and 268 healthy controls.
- Determined that one previously-identified locus (rs1801274) is associated with Ulcerative Colitis, but not Crohn's disease, identifying the first finding of association with UC in an African American population.

Undergraduate Researcher, 9/2006-5/2010
King's College • Wilkes-Barre, PA
Laboratory of Jeramia Ory, Ph.D.

- Studied copper's role on the pathogenesis of the opportunistic fungal pathogen, *Cryptococcus neoformans.*

- Identified genes that are differentially expressed at varying copper concentrations between a copper transporter knockout strain (*cuf1*-) and wild type strain (JEC21) of *C. neoformans* to both better understand which genes are involved in copper response and regulation and determine how these genes are altered in the avirulent *cuf1*- strain.
- Found that many genes in the *cuf1*- knockout strain are differentially expressed in low copper conditions relative to wild type and that these genes indicate general metabolic stress in the *cuf1*- strain, suggesting that altering oxidative phosphorylation in *C. neoformans* may help to minimize virulence in pathogenic strains.

Undergraduate Researcher, 6/2008-8/2008
Carnegie Mellon University • Pittsburgh, PA
Laboratory of Jonathan Minden, Ph.D.

- Studied the role of heat shock protein 27 (*Hsp27*) to better understand the activation of apoptosis.
- Determined that a post-translational modification to Hsp27 plays a pro-apoptotic role *in vivo* in *Drosophila* by a *reaper*-dependent mechanism.

PUBLICATIONS

1. **Ellis S.E.**, Panitch R., West A.B., Arking D.E. (2016). Transcriptome Analysis of Cortical Tissue Reveals Shared Sets of Down-Regulated Genes in Autism and Schizophrenia. ***Translational Psychiatry***.
2. Huang C, Haritunians T, Okou DT, Cutler DJ, Zwick ME, Taylor KD, Datta LW, Maranville JC, Liu Z, **Ellis S**, Chopra P, Alexander JS, Baldassano RN, Cross RK, Dassopoulos T, Dhere TA, Duerr RH, Hanson JS, Hou JK, Hussain SZ, Isaacs KL, Kachelries KE, Kader H, Kappelman MD, Katz J, Kellermayer R, Kirschner BS, Kuemmerle JF, Kumar A, Kwon JH, Lazarev M, Mannon P, Moulton DE, Osuntokun BO, Patel A, Rioux JD, Rotter JI, Saeed S, Scherl EJ, Silverberg MS, Silverman A, Targan SR, Valentine J, Wang MH, Simpson CL, Bridges SL, Kimberly RP, Rich SS, Cho JH, Di Rienzo A, Kao LW, McGovern DP, Brant SR, and Kugathasan S. (2015). Characterization of Genetic Loci That Affect Susceptibility to Inflammatory Bowel Diseases in African Americans. ***Gastroenterology*.**
3. Gupta, S., **Ellis, S.E.**, Ashar, F.N., Moes, A., Bader, J.S., West, A.B., and Arking, D.E. (2014). Transcriptome Analysis Reveals Deregulation of Innate Immune Response Genes and Neuronal Activity-Dependent Genes in Autism. ***Nature Communications***.
4. **Ellis, S.E.**, Gupta, S., Ashar, F.N., Bader, J.S., West, A.B., and Arking, D.E. (2013). RNA-Seq optimization with eQTL gold standards. BMC Genomics 14, 892.

CONFERENCES

Scientific Meetings Attended

2010, 2012—2015	American Society for Human Genetics.
2013—2014	Society for Neuroscience.
2009—2010	American Society for Microbiology.

Poster Presentations

1.**Ellis, S.E.**, Gupta S., Moes A, Absher D., West A.B. & Arking D.E. (Oct. 6-10, 2015). No Evidence That Differences In Cortical DNA Methylation Contribute to Autism. American Society for Human Genetics.

2. **Ellis, S.E.**, Gupta, S., Moes, A., West, A.B., and Arking, D.E. (Oct. 18-22, 2014). Assessing the role of methylation in autism brains. American Society for Human Genetics.

3. **Ellis, S.E.**, Gupta, S., Ashar, F.N., Bader, J.S., West, A.B., and Arking, D.E. (Oct. 22-26, 2013). RNA-Seq optimization with eQTL gold standards. American Society for Human Genetics.

4. **Ellis, S.E.**, Arking, D.E., Iacono, D., Pletnikova, O., Rudow, G., Talbot, C., O'Brien, R., Resnick, S. and Troncoso, J.C, (Nov. 9-13, 2013). Understanding the Transcriptome of Asymptomatic Alzheimer's Disease. Society for Neuroscience.

5. **Ellis, S.E.**, Doering, T.L., and Ory, J.J. (May 23-27, 2010). Microarray Analysis of a *cuf1* Strain of *Cryptococcus neoformans* Suggests Cuf1p is Involved in Both Repressor and Enhancer Activities. American Society for Microbiology.

TEACHING AND MENTORING EXPERIENCE

Teaching

Fall 2015 **Guest Lecturer**, Introdcution to Computational Genetics

- Instructed class of graduate students on data analysis techniques and pitfalls of RNA-Sequencing data analysis.
- Prepared two lectures and accompanying exercies for in-class instruction as well as take-home exercises to both assess comprehension and provide feedback to students.

2012—2014 **Tutor, Comprehensive Exam Preparation**

- Reviewed linkage and association studies for second year graduate students.
- Held mock exam practice sessions for students as they prepared for their oral comprehensive exams.

2013, Spring **Teaching Assistant, Advanced Topics in Human Genetics**

- Teaching assistant for 12 first year graduate students and three pediatric genetics fellows.
- Facilitated and guided discussion-based classes, and met with students to discuss the literature and help prepare in-class presentations.
- Wrote, administered, and graded the midterm exam.

2007—2010 **Tutor in Genetics, Biochemistry, and General Chemistry I & II**

- Individually tutored more than 45 undergraduate students.
- Reviewed lecture material, answered questions on assigned problem sets, and prepared and administered preparatory quizzes and exams.

Mentoring

2016, Summer Rebecca Panitch, Undergraduate Student

- Johns Hopkins University Center for Computational Biology Summer Program
- Project: Exploring the transcriptome of autistic individuals with a 15q duplication

2016, Winter Augusto Ramirez, Undergraduate Student

- Project: Utilizing gene expression profiles as a diagnostic tool in autism

2016, Winter Elizabeth Vincent, Graduate Student

- Rotation Project: The role of alternative splicing in the autistic brain transcriptome, Human Genetics graduate student rotation student

2015, Summer Rebecca Panitch, Undergraduate Student

- Johns Hopkins University Center for Computational Biology Summer Program
- Project: Reanalysis of schizophrenia and bipolar disorder gene expression data highlights the importance of incorporating surrogate variables in RNA-Seq studies

2014, Winter Heather Wick, Graduate Student

- Rotation Project: RNA expression in cingulate cortex of schizophrenia, bipolar disorder, and autism, Human Genetics graduate student rotation student

2014, Summer Edward Pang, Undergraduate Student

- Johns Hopkins University Center for Computational Biology Summer Program
- Project: Determining mtDNA copy number from sequencing data using GTeX data

2013, Summer James Miller, Undergraduate Student

- Johns Hopkins University Center for Computational Biology Summer Program
- Project: A new approach to visualizing DNA methylation data

ORGANIZATIONAL ACTIVITIES

Leadership Experience

2013—present Institute of Genetic Medicine Human Genetics Graduate Student Representative

2013—present Student Leader, Barton Childs Lecture Planning Committee

2011—2015 Committee Leader, Human Genetics Graduate Program New Student Recruitment

2014, Spring Assistant Women’s Lacrosse Coach, CCBC Essex

2014, Spring Student Leader, McKusick Lecture Planning Committee

Professional Societies

2010—present Member, The American Society of Human Genetics

HONORS AND AWARDS

2006—2010	Presidential Scholarship (a full academic scholarship to King’s College, Wilkes-Barre, PA)
2006—2010	Mendenhall-Tyson Scholarship
2010	Paul D. Laurence Best in Science Award
2010	Regina Award for Biology
2010	S. Idris Ley Memorial Award for the Highest Academic Achievement
2010	Josephine T. Moran Foreign Language Award
2009	Paul D. Laurence Best in Science Award
2009	American Society for Microbiology Student Travel Grant Award, 109th General Meeting
2009	American Society for Microbiology Undergraduate Research Fellowship
2008	National Science Foundation Undergraduate Research Fellowship