UNIVERSITY OF CALIFORNIA

Los Angeles

Analysis of cerebral cortical transcriptome dysregulation in autism and psychiatric disorders

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy in
Neuroscience

by

Jillian Roberta de Bree

2020

ABSTRACT OF THE DISSERTATION

Analysis of cerebral cortical transcriptome dysregulation in autism and psychiatric disorders

by

Jillian Roberta de Bree

Doctor of Philosophy in Neuroscience

University of California, Los Angeles, 2020

Professor Daniel H. Geschwind, Co-Chair

Professor Michael Jeffrey Gandal, Co-Chair

Psychiatric disorders are not well understood. Their diagnosis is based purely on behavioral symptoms and they lack a clearly defined pathology in brain, which challenges our ability to understand their biological roots. However, it is well established that psychiatric disorders are heritable, and large-scale genetic studies have begun to identify now thousands of psychiatric genetic risk variants.[1] Discovering how these genetic variants converge within discrete neurobiological pathways is a critical next step for understanding psychiatric disorder mechanisms and identifying new targets for therapeutic development. In search for these convergent pathways, transcriptomic studies have started to identify gene expression changes within human postmortem brain samples from psychiatric patients compared to neurotypical

controls. The transcriptome – the set of expressed RNA transcripts present in a given tissue or cellular samples – represents a snapshot of the cell-types and subcellular, molecular processes present and active in sequenced samples. As such, transcriptomic profiling of brain samples from psychiatric cases versus controls may provide increased resolution to identify a molecular pathology of disease not observed via traditional approaches. For example, in ASD, upregulation of microglial, astrocyte, and immune signaling genes, downregulation of specific synaptic genes, and attenuation of regional gene expression differences have been observed with transcriptomic analyses.[2,3] While transcriptomic studies have substantially improved our understanding of psychiatric neuropathology, they are limited in scope to single psychiatric disorders and few brain regions. Considering the growing evidence for genetic overlap between distinct psychiatric disorders,[4] it is a reasonable next step to determine if these disorders also share biological signatures in the brain. Comparing and contrasting gene expression changes across distinct psychiatric disorders – as well as across the entire cerebral cortex - will provide a fuller picture of the spatial landscape and specificity of molecular dysregulation in the psychiatric disease brain, pinpointing potential regions of particular vulnerability and biological pathways involved in psychiatric disease mechanisms.

To obtain this cross-disorder and multi-regional understanding of psychiatric gene expression changes, here I present a comprehensive set of transcriptomic investigations conducted by myself and others, spanning multiple psychiatric disorders and brain regions. In Chapter 2, I share our published mega-analysis of gene expression microarray datasets containing frontal cortex samples from subjects diagnosed with schizophrenia, bipolar disorder, ASD, and major depressive disorder subjects, compared with non-psychiatric controls.[5] We find that polygenic overlap parallels transcriptomic overlap, and that psychiatric genetic risk variants are associated with downregulated neuronal genes found in ASD, schizophrenia, and bipolar disorder. In Chapter 3, I present my contributions to our published collaborative work with the

PsychENCODE Consortium,[6] in which we compiled and uniformly processed genotype and RNA-sequencing data from more than 2,000 postmortem human brain samples to gain an understanding of how the entire transcriptome is impacted in frontal cortex samples from subjects diagnosed with schizophrenia, bipolar disorder, and ASD. Here, I detail my work integrating polygenic risk scores --measures of common genetic burden for psychiatric disease -- with transcriptomic changes to obtain a deeper understanding of how genetic variants directly regulate psychiatric gene expression changes. In Chapter 4, I present our work characterizing ASD transcriptomic across 11 distinct regions spanning the ASD cerebral cortex. We find widespread dysregulation across the cerebral cortex, with this dysregulation exhibiting the greatest magnitude of effect in the occipital region. ASD genetic risk variants are associated with genes downregulated cortex-wide that contribute to neuronal synaptic plasticity pathways, heavily implicating neuronal synaptic plasticity in ASD neuropathology. Together, these transcriptomic analyses expand our understanding of the molecular pathology of psychiatric disorders across distinct disorders and the cerebral cortex, implicating specific genes, cell-types, and biological pathways in psychiatric neuropathology.

Abstract Bibliography

1. Sullivan, P. F. & Geschwind, D. H. Defining the Genetic, Genomic, Cellular, and Diagnostic Architectures of Psychiatric Disorders. *Cell* **177**, 162–183 (2019).
2. I. Voineagu et al., Transcriptomic analysis of autistic brain reveals convergent molecular pathology. Nature. 474, 380–384 (2011).
3. Parikshak, N. N. *et al.* Genome-wide changes in lncRNA, splicing, and regional gene expression patterns in autism. *Nature* **540**, 423–427 (2016).
4. Bulik-Sullivan, B., Finucane, H., Anttila, V. et al. An atlas of genetic correlations across human diseases and traits. *Nat Genet* **47**, 1236–1241 (2015).
5. Gandal, M. J. *et al.* Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap. *Science* **359**, 693–697 (2018).
6. Gandal, M. J. *et al.* Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. *Science* **362**, (2018).

The dissertation of Jillian Roberta de Bree is approved.

Carrie E. Bearden

Roel A. Ophoff

Bogdan Pasanuic

Daniel H. Geschwind, Committee Co-Chair

Michael Jeffrey Gandal, Committee Co-Chair

University of California, Los Angeles

2020

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES..x

ACKNOWLEDGEMENTS xv

VITA...xx

CHAPTER 1: Leveraging RNA-seq in psychiatric disorder research 1

1.1: Introduction...2
1.2: Applications of RNA-seq in psychiatric disorder research 4
1.3: Effective experimental design for RNA-seq..9
1.4: Processing RNA-seq data 17
1.5: Gene prioritization with RNA-seq 24
1.6: Integration of RNA-seq results with orthogonal data 34
1.7: Functional interpretation and application of RNA-seq results.................................39
1.8: Discussion 45
1.9: Bibliography..47

CHAPTER 2: Shared molecular neuropathology across psychiatric disorders 55

2.1: Contributing authors 56
2.2: Experimental rationale and overlapping psychiatric cortical gene expression 56
2.3: Gene network analysis maps psychiatric gene expression to biological systems....61
2.4: Down-regulated neuronal modules are enriched for psychiatric genetic risk factors 64
2.5: Discussion 66
2.6: Materials and Methods 67
2.7: Bibliography..67

CHAPTER 3: Polygenic risk scores are associated with transcriptomic changes in ASD, schizophrenia, and bipolar disorder...74

3.1: Introduction and contributing authors..75
3.2: Project summary: psychiatric cross-disorder transcriptome-wide analysis 76
3.3: Identifying genetic drivers of transcriptomic dysregulation 77

3.4: Materials and Methods 81
3.5: Bibliography....................81

CHAPTER 4: Broad transcriptomic dysregulation across the cerebral cortex in ASD 90

4.1: Abstract 91
4.2: Transcriptomic changes across the cerebral cortex in ASD 91
4.3: Broad attenuation of transcriptomic regional identity....................95
4.4: Refining disrupted gene co-expression networks in ASD....................97
4.4.1: Cortex-wide dysregulation observed for ASD risk genes....................98
4.4.2: Magnitude of effect parallels anterior – posterior gradients 100
4.4.3: Cell-type changes mirror regional variation 103
4.5: Discussion....................106
4.6: Materials and Methods 108
4.7: Bibliography....................115

CHAPTER 5: Conclusions and future directions 119

5.1: Conclusions 120
5.2: Future directions 123
5.3: Bibliography 125

APPENDIX 127

A1: Supplementary Materials for Chapter 2....................128
A1.1: Extended Materials and Methods 128
A1.2: Extended Text....................143
A1.3: Extended Figures....................150
A1.4: Extended Tables 167
A1.5: Extended Bibliography 171

A2: Additional Results Accompanying Chapter 3 172
A2.1: Introduction 172
A2.2: Gene and isoform expression alterations in disease....................176
A2.3: Differential expression of the non-coding transcriptome 178

A2.4: Local splicing dysregulation in disease ... 179
A2.5: Transcriptome-wide association ... 182
A2.6: Co-expression networks refine shared cross-disorder signals ... 185
A2.7: Neuronal isoform networks capture disease specificity ... 189
A2.8: Distinct trajectories of neural-immune dysregulation ... 192
A2.9: Non-coding modules and lncRNA regulatory relationships ... 194
A2.10: Isoform network specificity and switching ... 197
A2.11: Discussion ... 198
A2.12: Materials and Methods ... 202
A2.14: Bibliography ... 202

A3: Supplementary Materials for Chapter 3 and Appendix Section A2 ... 203
A3.1: Extended Materials and Methods ... 203
A3.2: Extended Figures ... 233
A3.3: Extended Tables ... 256
A3.4: Extended Bibliography ... 258

A4: Supplementary Materials for Chapter 4 ... 259
A4.1: Extended Materials and Methods ... 259
A4.2: Extended Figures ... 273
A4.3: Extended Tables ... 288
A4.4: Extended Bibliography ... 289

LIST OF TABLES AND FIGURES

Chapter 1

Figure 1.1: The value of RNA-seq analysis in psychiatric research

Figure 1.2: Basic workflow for RNA-seq experiments

Table 1.1: RNA-seq basic terminology

Figure 1.3: Overview of RNA-seq data processing steps

Table 1.2: Overview of RNA-seq processing steps

Figure 1.4: Decision chart for GLM design

Table 1.3: Approaches for identifying DE gene sets

Table 1.4: Approaches for gene network analysis

Table 1.5: Approaches for orthogonal data integration with interesting gene sets

Figure 1.5: Direct applications of RNA-seq in psychiatric research

Chapter 2

Figure 2.1: Experimental rationale and design.

Figure 2.2: Cortical gene expression patterns overlap.

Figure 2.3: Network analysis identifies modules of coexpressed genes across disease.

Figure 2.4: Down-regulated neuronal modules are enriched for common and rare genetic risk factors.

Chapter 3

Figure 3.1: Overlaps and genetic enrichment among dysregulated transcriptomic features

Chapter 4

Figure 4.1: ASD transcriptomic differences across 11 cortical regions

Figure 4.2: Transcriptomic regional identity attenuation in ASD

Figure 4.3: Co-expression network analysis characterizes cortex-wide dysregulation of ASD risk genes

Figure 4.4: Functional characterization of regionally-variable transcriptomic dysregulation in ASD

Appendix Section A1: Supplementary Materials for Chapter 2

Figure A1.1: Microarray experiment quality control

Figure A1.2: Gene expression principal component loadings

Figure A1.3: Transcriptome overlap across disorders is robust to method differences

Figure A1.4: Top 50 most differentially expressed genes across disorders

Figure A1.5: Transcriptomic analysis of antipsychotic medications in primates

Figure A1.6: QC plots for RNAseq replication datasets

Figure A1.7: RNAseq replication plots

Figure A1.8: Robustness of the gene co-expression network

Figure A1.9: Linear regression adjusted R2 values for module eigengene associations

Figure A1.10: Module stability across individual studies

Figure A1.11: Co-expression Enrichment for Brain Enhancer Regulation

Figure A1.12: eQTL-GWAS enrichment

Figure A1.13: Leave-one-out analysis for polygenic ~ transcriptomic associations

Table A1.1: Gene expression microarray datasets

Table A1.2: RNAseq replication of differential gene expression

Table A1.3: GWAS summary statistics used in this study

Data Table A1.1: Differential gene expression summary statistics

Data Table A1.2: Gene co-expression module data

Data Table A1.3: Compilation of psychiatric risk variants

Data Table A1.4: LD score regression-based partitioned heritability analyses

Data Table A1.5: ASD GWAS summary statistics (from the iPSYCH Consortium)

Appendix Section A2: Additional Results Accompanying Chapter 3

Figure A2.1: Gene and isoform expression dysregulation in psychiatric brain

Figure A2.2: Aberrant local splicing and isoform usage in ASD, SCZ and BD

Figure A2.3: Transcriptome-wide association

Figure A2.4: Gene and isoform co-expression networks capture shared and disease-specific cellular processes and interactions

Figure A2.5: Two RBFOX1 isoform modules capture distinct biological and disease associations

Figure A2.6: Distinct neural-immune trajectories in disease

Figure A2.7: LncRNA annotation, ANK2 isoform switching & micro-exon enrichment

Appendix Section A3: Supplementary Materials for Chapter 3

Figure A3.1: Dataset composition, analysis and integration pipeline

Figure A3.2: Dataset demographics and quality control

Figure A3.3: Selection of Covariates

Figure A3.4: Validation of DGE Results

Figure A3.5: Validation of differential transcript expression and differential splicing

Figure A3.6: Annotation of individual ncRNAs

Figure A3.7: Covariate correction of DS

Figure A3.8: Annotation of DS events, Cross data DGE-DTE-DS overlaps

Figure A3.9: Additional differential local splicing examples

Figure A3.10: Age effects on differential gene expression

Figure A3.11: Assessment of psychiatric medication effects

Figure A3.12: Co-expression network cell type enrichments

Figure A3.13: Genetic enrichment analyses

Figure A3.14: Module-trait associations after SCZ downsampling

Figure A3.15: csuWGCNA identifies putative lncRNA negative regulatory relationships

Figure A3.16: LINC00643 and LINC01166 expression in human prefrontal cortex

Figure A3.17: Additional switch isoforms

Table A3.1: DE genes

Table A3.2: Annotation of psychiatric ncRNAs

Table A3.3: Differential splicing

Table A3.4: TWAS results

Table A3.5: WGCNA results

Table A3.6: csuWGCNA results

Table A3.7: Switch isoform and microexon characterization

Table A3.8: Splicing and isoform validation

Table A3.9: RNAscope

Appendix Section A4: Supplementary Materials for Chapter 4

Figure A4.1: Experiment workflow and sample overview

Figure A4.2: Quality control measures

Figure A4.3: Model covariates and previous studies across 11 cortical regions

Figure A4.4: Transcriptomic changes across 11 cortical regions

Figure A4.5: Transcriptomic regional identity attenuation in ASD

Figure A4.6: Additional ARI gene dysregulation

Figure A4.7: Gene-level co-expression network analysis module associations

Figure A4.8: Isoform-level co-expression network analysis module associations

Figure A4.9: Additional neuronal density associations and snRNA-seq

Figure A4.10: Results summary

Table A4.1: Metadata and sequencing quality metrics

Table A4.2: DE gene overlap analyses

Table A4.3: DE gene results

Table A4.4: Transcriptomic regional identity analysis results

Table A4.5: Genes matched to WGCNA modules

Table A4.6: Functional characterization of WGCNA modules

Table A4.7: Supporting data for analyses of regionally-variable ASD transcriptomic dysregulation

ACKNOWLEDGEMENTS

I want to offer my most sincere thanks to my doctoral committee co-chairs and mentors, Michael Gandal and Daniel Geschwind. When I began my neurogenetics journey in 2014, I was a young, eager, 4th year undergraduate student with no applied research experience. Dan and Mike, however, saw my potential, and took the risk to bring me on as a lab assistant. I am so incredibly grateful that they did, since the opportunity to work with them both has enabled me to become an independent, thoughtful, and insightful interdisciplinary scientist. They have both supported my career development in different ways. Dan's wealth of expertise, intuition for psychiatric neurogenetics, and general guidance have opened my eyes to the possibilities of integrative bioinformatic analyses in biology. Mike's weekly (and sometimes daily) support with tackling tough problems, his patience with me as I learned and grew, his willingness to discuss the full spectrum of challenges and considerations for psychiatric genetics with me, and his unwavering belief in my abilities have all helped me to become more confident, resilient, endlessly curious about biology and statistics, and (I hope) a researcher on the cutting edge of our field. In addition to both Mike and Dan's direct mentorship, I am also grateful for their financial support, and for the access that they have granted me to amazing datasets for my PhD research.

I also need to thank several other people who have contributed directly to my doctoral degree progress. Felix Schweizer and Jenny Lee from the NSIDP have provided endless guidance, and have kept me on track to complete my PhD research in a timely manner. With the NSIDP, they have created a program that is a foundation of support for all Neuroscience PhD students, and that foundation has surely enabled me to become a better scientist. My committee members, Bogdan Pasaniuc, Roel Ophoff, and Carrie Bearden, have all been willing to discuss my doctoral research with me and helped me to think critically about the results and implications of my research. I would also like to thank Ben Novitch and Bill Lowry, who gave me the opportunity to learn about *in vitro* techniques that enhanced my understanding of neuroscience. Finally, I

would like to sincerely thank Jenifer Sakai, Lauren Kawaguchi, and Luis Espana from the Geschwind lab, who have all ensured that I received my funding, got meeting time slots with Dan, and had access to the tools that I needed to conduct my research.

There are many people to thank in the Geschwind and Gandal labs, who have been my colleagues and friends for so many years. When I first began learning about neurogenetics as a lab assistant in the Geschwind lab, Neelroop Pariskshak, Vivek Swarup, Virpi Leppa, and Laura Perez-Cano all welcomed me, supported my efforts to learn, and engaged in conversations about our work that must have seemed very trivial to all of them. They were (and still are) my role models in many ways, and I am so grateful for the opportunities I had to learn from all of them. Later on, as I progressed and become a PhD student, I was fortunate to work with Gokul Ramaswami, Damon Polioudakis, Chris Hartl, Jennifer Lowe, and Jessica Rexach, who were all willing to mentor me and discuss current topics in psychiatric genomics with me, no matter how busy they were. Finally, as one of the founding members of the Gandal lab, I had the unique privilege to see how an academic lab grows and got to meet so many wonderful people in that process. Sepideh Parhami provided endless support and friendship for me, and together we began to establish the Gandal lab culture. I am also grateful for the opportunities I had to work with Minsoo Kim, Leanna Hernandez, and Gil Hoftman in the Gandal lab, who have all contributed substantially to my understanding of neuroscience and biostatistics. In general, I am so privileged to have worked with all of the smart, talented, and kind people who have passed through the Gandal and Geschwind labs over the years, and I am thankful to all of them for being my colleagues, mentors, and friends.

In chapter two, all of the work presented was published in *Science* in February of 2018, in a research article entitled “Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap” (volume 359, pages 693–697). Michael Gandal was the primary author of this work and completed all major analyses, such as the expression microarray mega-analysis. I, as the second author, conducted the RNA-seq validation analyses, calculated partitioned heritability scores for the identified gene co-expression modules, and generally supported all analyses and interpretation of results. Other co-authors included Neelroop N. Parikshak, Virpi Leppa, Gokul Ramaswami, Chris Hartl, Andrew J. Schork, Vivek Appadurai, Alfonso Buil, Thomas M. Werge, Chunyu Liu, Kevin P. White, and Steve Horvath. They all contributed to supporting analyses and interpretation of results, providing essential contributions for the publication. Daniel Geschwind was the senior author and project director. This work was associated with the PsychENCODE Consortium and iPSYCH-BROAD Working Group.

Chapter three contains work that was also published in *Science*, in December 2018, with the title "Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder" (volume 362, no page numbers). Michael Gandal was, again, the primary author of this work, and as a co-author I generated and analyzed polygenic risk scores for the psychiatric disorders we investigated (the focus of chapter three), conducted transcriptomic analyses that examined the effects of anti-psychotic drugs in primate neural tissue (also included in chapter three), and assisted with the interpretation and communication of results. Other co-authors included Pan Zhang, Evi Hadjimichael, Rebecca Walker, Chao Chen, Shuang Liu, Hyejung Won, Harm van Bakel, Merina Varghese, Yongjun Wang, Annie W. Shieh, Sepideh Parhami, Judson Belmont, Minsoo Kim, Patricia Moran Losada, Zenab Khan, Justyna Mleczko, Yan Xia, Rujia Dai, Daifeng Wang, Yucheng T. Yang, Min Xu, Kenneth Fish, Patrick R. Hof, Jonathan Warrell, Dominic Fitzgerald, Andrew E. Jaffe, Kevin White, Mette A. Peters, Mark Gerstein, Chunyu Liu, Lilia M. Iakoucheva, and Dalila Pinto. Daniel Geschwind was the senior author and main project director. All of these co-authors contributed to major and minor analyses for this project, and helped write, edit, and review the resulting manuscript. This work was associated with the PsychENCODE Consortium.

Chapter four consists of work that is currently in preparation, titled “Broad transcriptomic dysregulation across the cerebral cortex in ASD”. I am the primary author of this work and personally conducted (or substantially contributed to) every analysis. I also synthesized and interpreted all of the results, constructed all accompanying graphics and tables, and wrote the manuscript. Other co-authors assisted with drawing functional insights from transcriptomic data, contributed to and performed supporting analyses (such as the single nucleus RNA-seq analysis, performed by Brie Wamsley, and the cell-type deconvolution analysis, performed by Prashant Emani), and improved the quality and readability of the complete manuscript. These co-authors include Brie Wamsley, George T. Chen, Gil D. Hoftman, Sepideh Parhami, Diego de Alba, Gaurav Kale, Gokul Ramaswami, Christopher L. Hartl, Jing Ou, Ye Emily Wu, Neelroop N. Parikshak, Vivek Swarup, T. Grant Belgard, Prashant Emani, Nathan Chang, Daifeng Wang, and Bogdan Pasaniuc. Michael Gandal and Daniel Geschwind are joint senior authors and project leaders for this work.

I would also like to acknowledge my funding sources, for without this financial support none of this work would have been possible. My first year of support was provided by the UCLA NSIDP through the Brain Research Institute. For subsequent years, the following grants from Daniel Geschwind contributed to my work: NIMH R01MH110927, U01MH115746, P50-MH106438, and R01 MH-109912, R01 MH094714. The following grants from Michael Gandal also provided me with financial support: SFARI Bridge to Independence Award, NIMH R01-MH121521, NIMH R01-MH123922, NICHD-P50-HD103557. Finally, I would like to offer sincere thanks to the Achievement Rewards for College Scientists (ARCS) Foundation Los Angeles Founder Chapter for funding during the 2019 – 2020 academic year, which provided essential support as I neared the end of my doctoral studies.

To conclude, I offer my most humble appreciation and thanks for my family and friends who supported me and believed in me throughout my PhD journey. My parents, Rob and Robin

Haney, and my brother, Ed Haney, have created a foundation of kindness, love, and trust that has supported me since childhood, and that foundation has continued to carry me through my PhD years. Their willingness to listen to me explain and think through my work, over and over again, sometimes for hours, was so valuable for helping me to improve my science communication skills and think about my work with a fresh perspective. To my new family-in-law, Jim, Teresa, and Kevin de Bree, and Kristina de Bree and Brian Child, I thank you all for welcoming me into your family and supporting my endeavors to obtain a doctoral degree. To my friends, especially my fellow cohort members in the NSIDP, thank you for all of the laughter, the candid discussions about science and graduate student life, and the emotional support over these many years. And finally, to my husband James de Bree – your unwavering belief in me, your support in every aspect of our lives together, and your boundless love, have all empowered me to achieve this great feat and complete all of this pivotal, important research. I am so grateful to be your wife, and to share the rest of my scientific career and personal journey with you.

VITA

EDUCATION AND WORK EXPERIENCE

Ph.D. Candidate in Neuroscience — Fall 2016 – Present
University of California, Los Angeles
(UCLA) | GPA 4.0
Dr. Daniel H. Geschwind and Dr.
Michael J. Gandal Labs

Computational Genomics Internship — Summer 2020
Vertex Pharmaceuticals | Boston, MA

B.S. in Mathematics/Applied Science — Fall 2010 – Fall 2014
(Medical and Life Science Focus)
University of California, Los Angeles (UCLA) | GPA 3.5

HONORS AND LEADERSHIP

ARCS Graduate Scholar 2019 – 2020 — October 2019
Achievement Rewards for College Scientists Foundation Los Angeles Founder Chapter. $10,000 awarded for the 2019 - 2020 academic year.

UCLA AWiSE Administrative Coordinator — Fall 2017 - Present
UCLA Advancing Women in Science and Engineering (AWiSE). Responsible for managing and coordinating the efforts of AWiSE leadership team members to create a yearly calendar of events which supports the advancement of women in STEM-oriented careers.

UCLA Alumni Scholar (Undergraduate) — Fall 2010 – Fall 2014
Earned a $4,000 scholarship for undergraduate tuition and performed 120+ hours of community service.

Graduate of the UCLA Honors College (Undergraduate) — Fall 2010 – Fall 2014
Earned a 3.51 GPA and completed 44 honors units in both the sciences and humanities.

CONFERENCE PRESENTATIONS

Regional variation in transcriptional dysregulation and patterning in postmortem cerebral cortex in ASD.

- *Oral Presentation Finalist:*
  World Congress of Psychiatric Genetics 2019 in Los Angeles, CA — October 2019
- *Poster Presentation:*
  The 69th Annual Meeting of the American Society of Human Genetics in Houston, Texas — October 2019
- *Poster Presentation*
  11th Annual International Conference On Systems Biology Of Human Diseases at UCLA — June 2018

- *Travel Award and Poster Presentation*
  4th Annual Molecular Psychiatry Meeting in Maui, Hawaii — October 2016

PUBLICATIONS

- **Regional variation in transcriptional dysregulation and patterning in postmortem cerebral cortex in ASD.** *In Preparation.* — December 2020

  Jillian R. Haney, Brie Wamsley, … , Michael J. Gandal, Daniel H. Geschwind.

- **Integrative genomics identifies a convergent molecular subtype that links epigenomic with transcriptomic differences in autism.** *Nature Communications*. — September 2020

  Gokul Ramaswami, Hyejung Won, Michael J. Gandal, **Jillian Haney**, … , Daniel H. Geschwind.

- **Network signature of complement component 4 variation in the human brain identifies convergent molecular risk for schizophrenia.** *bioRxiv (In Submission)*. — March 2020

  Minsoo Kim, **Jillian R. Haney**, Pan Zhang, ... , Michael J. Gandal.

- **TGFβ superfamily signaling regulates the state of human stem cell pluripotency and competency to create telencephalic organoids.** *bioRxiv (In Submission)*. — December 2019

  Momoko Watanabe, **Jillian R. Haney**, … , Michael J. Gandal, Bennett G. Novitch.

- **Hepatic arginase deficiency fosters dysmyelination during postnatal CNS development.** *Journal of Clinical Insight*. — September 2019

  Xiao-Bo Liu, **Jillian R. Haney**, … , Stephen D. Cederbaum, and Gerald S. Lipshutz.

- **Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder.** *Science*. — December 2018

  Michael J. Gandal, Pan Zhang, …, **Jillian Haney**, … , Daniel H. Geschwind.

- **Banking on Polygenicity to Disentangle Psychiatric Comorbidity.** *Biological Psychiatry*. — July 2018

  Jillian R. Haney, Sepideh Parhami, Michael J. Gandal.

- **Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap.** *Science*. — February 2018

  Michael J. Gandal, **Jillian R. Haney**, …, Daniel H. Geschwind.

- **Autism and Turner's Syndrome: A Review.** *UCLA Undergraduate Science Journal*. — June 2014

  Jillian R. Haney and Andrew Shaner.

CHAPTER ONE

Leveraging RNA-seq in psychiatric disorder research

1.1: Introduction

Psychiatric disorders are broadly defined as mental health conditions influencing mood, cognition, emotion, and behavior (DSM-5).[1] Many distinct disorders fall under this definition, and each is considerably heterogeneous in presentation and severity across individuals. For many ‘adult onset disorders’, symptoms present during adulthood or late adolescence are adult mental illnesses, such as major depressive disorder, schizophrenia, and bipolar disorder. In contrast, neurodevelopmental disorders such as autism spectrum disorder (ASD) and attention-deficit hyperactivity disorder (ADHD) are present during childhood. In the United States, approximately 18.9% of adults meet criteria for a psychiatric disorder,[2] while approximately 7% of children are afflicted with developmental disorders.[3]

All psychiatric disorders are diagnosed and characterized by behavioral symptoms.[1] For example, an ASD diagnosis is made if an individual displays persistent deficits in social communication and interaction in addition to restricted and repetitive behaviors.[1] This reliance on descriptive behavioral symptoms creates challenges for both diagnosis and treatment. Differentiating and characterizing distinct disorders is inherently difficult, since behavioral symptoms often fluctuate, overlap across many distinct disorders, and vary widely across individuals.[1] Additionally, the lack of a clear biological understanding of psychiatric disease pathophysiology greatly hinders our ability to develop novel therapeutic interventions. Identifying such neurobiological mechanisms in psychiatric disorders has the potential to address both of these challenges, facilitating the identification of diagnostic biomarkers and the development of effective targeted therapeutics.

Hundreds of genetic risk variants have now been linked to different psychiatric disorders.[4] Environmental risk factors, such as maternal immune activation and stress, have also been associated with psychiatric disorders.[5] Discovering how, when, and where neural cell-types, biological processes, and systems are impacted by these risk variants and environmental influences is a critical next step towards identifying psychiatric neuropathological mechanisms. In other words, we must determine how these genetic variants and environmental exposures act across the ‘levels of neuroscience’ (**Figure 1.1a**) to contribute to the development of psychiatric disorders. While psychiatric behaviors and, more recently, genetic risk variants for psychiatric disorders have been extensively characterized,[4] whether there is an underlying, convergent brain-level molecular pathology has been – until recently – largely unknown. This is due to many reasons, with some of the key difficulties being the complexity of the human brain, the heterogeneity of psychiatric disorders, and challenges in developing psychiatric *in vivo* and *in vitro* models.

While there are many approaches for addressing these challenges, in this chapter I will discuss just one particularly effective and valuable methodology – transcriptomic analyses through RNA-sequencing (RNA-seq). The utility of high-throughput next generation sequencing methods such as RNA-seq is well known, and as such countless articles, websites, tutorials, blogs, and reviews exist that detail the intricacies of RNA-seq. But strategies and guidance for implementing this approach specifically in psychiatric disorder research are not as well established. This is problematic, since transcriptomics is capable of addressing many of the challenges that face psychiatric disorder research. In particular, RNA-seq analysis can reveal disrupted cell-types and biological processes in sequenced samples, making it a powerful tool for identifying molecular pathology in the brain. Therefore, in this chapter I will focus exclusively on the application of transcriptomic profiling in psychiatric disorder research. I will discuss different motivations for using RNA-seq, best practices for experimental design, main data processing and analytic steps, foundational guiding concepts, useful references and resources, and helpful workflows. This chapter will function as an approachable introductory overview of how RNA-seq can be applied effectively in psychiatric disorder research, where I will ultimately demonstrate how the strategic implementation of RNA-seq can substantially advance our understanding of psychiatric disorder mechanisms.

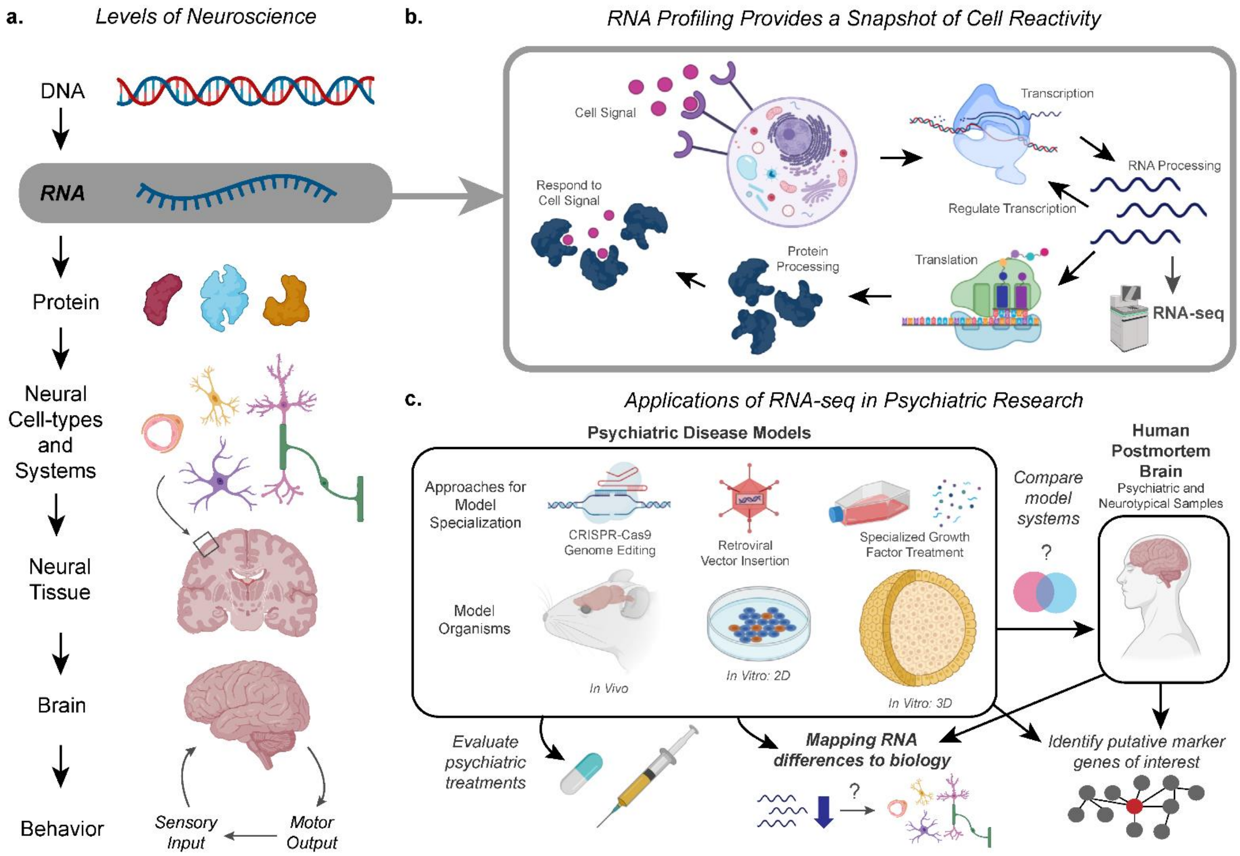

a.
Levels of Neuroscience
DNA
RNA
Protein
Neural Cell-types and Systems
Neural Tissue
Brain
Behavior
Sensory Input
Motor Output
b.
RNA Profiling Provides a Snapshot of Cell Reactivity
Cell Signal
Transcription
RNA Processing
Regulate Transcription
Respond to Cell Signal
Protein Processing
Translation
RNA-seq
c.
Applications of RNA-seq in Psychiatric Research
Psychiatric Disease Models
Approaches for Model Specialization
CRISPR-Cas9 Genome Editing
Retroviral Vector Insertion
Specialized Growth Factor Treatment
Model Organisms
In Vivo
In Vitro: 2D
In Vitro: 3D
Compare model systems
?
Human Postmortem Brain
Psychiatric and Neurotypical Samples
Evaluate psychiatric treatments
Mapping RNA differences to biology
?
Identify putative marker genes of interest

Figure 1.1: The value of RNA-seq analysis in psychiatric research. a. Hierarchical framework for understanding genomic relations in the context of neuroscience, highlighting where RNA falls in the progression from DNA to behavior. **b.** RNA profiling provides a snapshot of cell reactivity, enabling RNA-seq analysis to probe how neural systems are acting in sequenced samples. **c.** Summary of how RNA-seq can be applied in specific areas of psychiatric disorder research.

1.2: Applications of RNA-seq in psychiatric disorder research

One of the main functions of RNA in biological systems is to translate genetic information into proteins that participate in molecular pathways and create structures in cells. RNA also can play a regulatory role, influencing the activity of genes during neurodevelopment and after exposure to different extra-cellular and environmental stimuli. Following the central dogma of biology, genes are transcribed into RNA molecules, which in the case of protein-coding genes are then translated into proteins, or alternatively, non-coding genes are transcribed into a host of regulatory RNAs, including ribosomal RNAs, long-noncoding RNAs (lncRNAs), small nuceolar RNAs (snoRNAs), or microRNAs (miRNAs), among others. Each class of non-coding RNAs has distinct regulatory functions, such as the formation of ribosomal subunits for protein translation in the case of rRNA, the regulation of alternative mRNA splicing in the case of snoRNAs, or the modification of transcript abundance by miRNAs, especially for gene silencing purposes. In general, these regulatory RNAs either promote transcription that will ultimately create needed proteins, or inhibit transcription to reduce the amount of undesired proteins. Accordingly, RNA is constantly being synthesized (transcribed), modified (spliced, capped, polyadenylated), exported from the nucleus, trafficked to ribosomes, and translated. As single stranded mRNAs are inherently unstable, they are constantly undergoing degradation via ribonucleases (RNases) present within cells. RNA synthesis and degradation are dynamic processes, that can occur at either steady state or in transcriptional ‘bursts’ in response to different cellular stimuli.

Together, all of the classes of RNA comprise the complete transcriptome of a given cell or sample, making this RNA makeup a representation of all of the current reactive biological pathways in that sample. Therefore, through profiling the entire transcriptome with RNA-seq, we obtain a ‘snapshot’ - a single picture of a single moment in time - of how cells and biological processes are reacting in sequenced samples (**Figure 1.1b**). Importantly, this is distinct from surveying all proteins in a sample (the ‘proteome’) - while proteins represent cells as they are, RNA represents how cells seek to change (or remain at steady-state). In psychiatric disorder research, understanding how biological systems are being regulated (or dysregulated) across a set of samples is often a major aim, making RNA-seq an incredibly valuable tool in many distinct applications (**Figure 1.1c**).

The main application of RNA-seq in psychiatric disorder research, extending across all types of analyses and experiments, is the comparison of RNA differences across conditions of interest to biological functions (also referred to as differential expression). Tying gene expression differences in psychiatric patient samples or psychiatric experimental models to cell-types and biological processes provides concrete targets for follow-up experiments to validate and advance our understanding of psychiatric disorders.[6-8] Using RNA-seq to hone in on dysregulated biological pathways has already proven highly effective for work with human postmortem tissue donated from psychiatric patients. For example, analysis of postmortem tissue from subjects with ASD identified increased expression of astrocyte and microglial cell-type marker genes across the cerebral cortex, and that neuronal and synaptic genes are downregulated across the cerebral cortex in ASD.[6,7] In frontal cortex tissue from subjects with schizophrenia, neuronal synaptic vesicle cycle pathways and NFkB signaling pathways are both increased compared to neurotypical controls.[8] These results narrow in on distinct cell-types and signaling pathways that can be further investigated in psychiatric and neurodevelopmental model systems, bringing psychiatric disorder research one step closer to establishing concrete neuropathology in these disorders.

While the direct characterization of human postmortem brain tissue from psychiatric patients is one valuable application of RNA-seq, understanding RNA differences in model systems is equally as important. Human tissue can only be collected postmortem from heterogeneous patient populations, making the controlled study of prenatal timepoints and specific genetic or environmental risk factors for these disorders impossible with human tissue. This necessitates the use of model systems such as mice, cell cultures, and more recently neural spheroids to better understand psychiatric disorder mechanisms. However, developing these model systems has proven to be challenging - which is understandable considering the complexity of neural development - leading us to another effective application of RNA-seq in psychiatric disorder research: to validate psychiatric disorder model systems. Through comparing and contrasting the gene expression of new psychiatric disorder models with benchmark RNA-seq datasets - such as human postmortem tissue RNA-seq from the Allen Brain Atlas[9] and PsychENCODE,[8] and previously established *in vivo* and *in vitro* experimental models - these newer models can be improved and refined. RNA-seq can also be used to to compare different variations of a new model system (eg. different culture conditions, different mouse strains) in order to select a model which best achieves desired outcomes.

Understanding and evaluating transcriptomic dysregulation in model systems is a key approach for investigating psychiatric molecular pathology. Once a psychiatric behavior or biomarker is established in these model systems, RNA-seq can lend itself to a new application - evaluating potential therapeutic treatments. For high-throughput drug screening approaches, or for treatments precisely designed to target specific dysregulated biological pathways in psychiatric disorders, RNA-seq can reveal how treatments are impacting biological systems in models (for example, RNA-seq with a dementia iPSC *in vitro* model system revealed how suberanilohydroxamic acid can correct miR-203 induced dysregulated gene expression).[102] This knowledge can guide further development of treatments, indicating if a treatment approach is effective and/or if the treatment has any unintended effects. Additionally, for successful treatments, RNA-seq can reveal which disrupted biological systems are impacted by treatment, helping to pinpoint which aspects of molecular pathology may reflect pathogenic disease mechanisms. In summary, rather than running a vast array of screening experiments when evaluating a new treatment in a model system, performing a single RNA-seq experiment can provide an overview of how treatments are acting in these models, helping to inform decisions regarding how to proceed with putative psychiatric treatments.

Another advantageous application of RNA-seq is to identify specific marker genes that represent, and possibly regulate, the disrupted biological pathways implicated by transcriptomic analysis. Genes with the greatest differential gene expression signatures, genes with high connectivity in gene co-expression network modules dysregulated in psychiatric conditions ('hub' genes), and risk genes known *a priori* that exhibit transcriptomic dysregulation in a particular experiment could all be considered interesting marker genes in RNA-seq experiments. One of the best justifications for identifying and evaluating single marker genes is that they streamline experiments performed to validate the existence of dysregulated biological pathways implicated by RNA-seq. For example, performing an RNA FISH experiment with a hub gene from a highly dysregulated co-expression module in a psychiatric model system sample will confirm if this gene is differentially expressed compared to controls, support that the biological pathways associated with this gene's module may be altered in the psychiatric model, and localize this gene's RNA in cells - all of these results could further advance understanding of psychiatric molecular pathology. Investigating specific marker genes individually is also an effective approach, especially for hub genes from co-expression modules, as they may play a driving regulatory role in driving psychiatric molecular pathology.[106-107] For example, for a substantially dysregulated gene co-expression module in a model mouse system, it may be advantageous to create a new model mouse with that module's hub gene knocked out, and to evaluate psychiatric-linked behaviors in this new model mouse to predict if the dysregulation of this gene is causal or likely a consequence of psychiatric pathology - again, either of these results could advance our understanding of psychiatric disorders.

To conclude, RNA-seq is an effective tool with many distinct applications that can hone in on changes in biological pathways in psychiatric disorders. RNA-seq offers a distinct advantage over methods that can only investigate single facets of biological systems in psychiatric patient tissue and experimental models, considering that RNA-seq enables the implementation of transcriptomic analyses that empirically evaluate entire biological samples. Through taking this data-driven approach, RNA-seq analysis can pinpoint specific disrupted cell-types and biological processes as well as marker genes that can be used for validating the existence of implicated dysregulation. When used for this purpose, to prioritize specific cell-types and biological processes for investigation in psychiatric patient tissue samples and experimental models, RNA-seq - applied with effective experimental designs - is a powerful approach that can profoundly advance psychiatric disorder research.

1.3: Effective experimental design for RNA-seq

Some of the key advantages of an RNA-seq experiment are that it is high-throughput, fast, efficient, customizable, relatively easy to perform, and dependably capable of producing high-quality data. All of these factors help to ensure that resources spent on RNA-seq provide a significant return for laboratories. However, experimental design - one of the most critical aspects of any scientific experiment - usually influences outcomes far more than these other factors. If samples and RNA-seq parameters are not considered carefully, an RNA-seq experiment can lose a significant amount of value, even if the sequencing itself was of high-quality. To avoid this, it is advisable to spend a considerable amount of time and effort planning your RNA-seq experiment before executing it. Of the utmost importance, designing an experiment that clearly addresses your specific research goals is crucial. There are many factors to consider, primarily sample selection, type of RNA-seq (eg. bulk v. single-cell), sample library preparation for sequencing, depth and length of sequencing, and assignment of samples to sequencing batches. Explicitly defining how sample and sequencing choices will address your underlying questions and motivations for performing the experiment is helpful for deciding on these factors. For example, if interested in understanding how neurodevelopmental mechanisms contribute to psychiatric disorders, *in vivo* or *in vitro* model organisms carrying a psychiatric risk mutation may be suitable sample selections. And, while longer RNA reads are better for isoform-level quantifications, if studying a new model organism where very little is known, then shorter reads that can still reliably provide gene-level quantifications - which are still informative, albeit not as informative as isoform-level quantifications - may suffice for your purposes, especially if cost is an important deciding factor. These examples begin to illustrate how many distinct factors can guide experimental design, and how complicated it can get. Indeed, there are countless examples with a vast array of justifications for selecting sample types and customizing RNA-seq. To reduce this complexity, in this section I will review fundamental vocabulary (**Table 1.1**) and summarize some of the main considerations that guide experimental design for most psychiatric RNA-seq experiments, focusing on the core concepts (**Figure 1.2**).

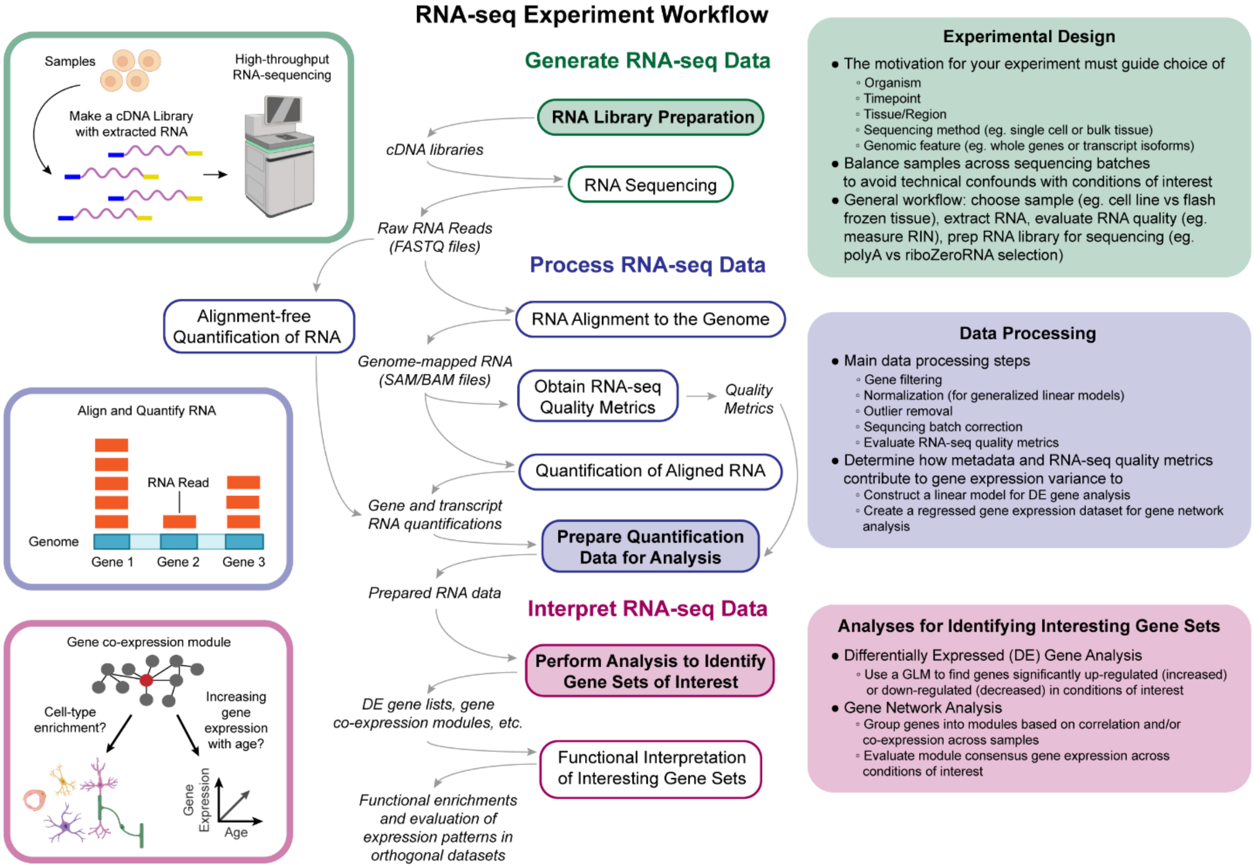

RNA-seq Experiment Workflow
Samples
High-throughput RNA-sequencing
Make a cDNA Library with extracted RNA
Generate RNA-seq Data
RNA Library Preparation
cDNA libraries
RNA Sequencing
Raw RNA Reads (FASTQ files)
Process RNA-seq Data
Alignment-free Quantification of RNA
RNA Alignment to the Genome
Genome-mapped RNA (SAM/BAM files)
Obtain RNA-seq Quality Metrics
Quality Metrics
Quantification of Aligned RNA
Gene and transcript RNA quantifications
Prepare Quantification Data for Analysis
Prepared RNA data
Interpret RNA-seq Data
Perform Analysis to Identify Gene Sets of Interest
DE gene lists, gene co-expression modules, etc.
Functional Interpretation of Interesting Gene Sets
Functional enrichments and evaluation of expression patterns in orthogonal datasets
Align and Quantify RNA
RNA Read
Genome
Gene 1
Gene 2
Gene 3
Gene co-expression module
Cell-type enrichment?
Increasing gene expression with age?
Gene Expression
Age
Experimental Design
• The motivation for your experiment must guide choice of
◦ Organism
◦ Timepoint
◦ Tissue/Region
◦ Sequencing method (eg. single cell or bulk tissue)
◦ Genomic feature (eg. whole genes or transcript isoforms)
• Balance samples across sequencing batches to avoid technical confounds with conditions of interest
• General workflow: choose sample (eg. cell line vs flash frozen tissue), extract RNA, evaluate RNA quality (eg. measure RIN), prep RNA library for sequencing (eg. polyA vs riboZeroRNA selection)
Data Processing
• Main data processing steps
◦ Gene filtering
◦ Normalization (for generalized linear models)
◦ Outlier removal
◦ Sepuncing batch correction
◦ Evaluate RNA-seq quality metrics
• Determine how metadata and RNA-seq quality metrics contribute to gene expression variance to
◦ Construct a linear model for DE gene analysis
◦ Create a regressed gene expression dataset for gene network analysis
Analyses for Identifying Interesting Gene Sets
• Differentially Expressed (DE) Gene Analysis
◦ Use a GLM to find genes significantly up-regulated (increased) or down-regulated (decreased) in conditions of interest
• Gene Network Analysis
◦ Group genes into modules based on correlation and/or co-expression across samples
◦ Evaluate module consensus gene expression across conditions of interest

Figure 1.2: Basic workflow for RNA-seq experiments. The main steps of an RNA-seq experiment are separated into three major phases, with extra considerations and illustrations for shaded steps.

Term	Definition	Relevance
Transcriptome	RNA (gene expression) across the entire genome	Profiling the transcriptome with RNA-seq provides a comprehensive overview of active biological processes and cell-types in biological samples of interest.
Read Depth	The number of reads obtained for each sample in an RNA-seq experiment.	Greater read depth will better capture lowly expressed genes. However, relatively smaller read depths are suitable for many purposes (eg. identifying distinct cell-types with scRNA-seq).
Whole Gene Quantification	Quantifying RNA across entire genes – all exons, and possibly all introns and UTRs	Whole gene quantification captures all reads aligning to a certain gene, but loses isoform resolution.
Isoform (transcript) Quantification	Quantifying RNA for specific gene isoforms – specific exons	Isoform quantification estimates how many reads align to distinct isoforms, but is limited to annotated isoforms.
Genome	All genes within an organism, encoded by DNA	While genetics is focused on the study of single genes, in an RNA-seq experiment we profile how the entire genome effects the transcriptome.
Epigenome	All epigenetic modifications within an organism, such as DNA methylation and histone acetylation	The epigenome influences the genome, which in turn effects the transcriptome that we measure with RNA-seq.
Proteome	All proteins that can be present within an organism	The transcriptome produces the proteome, in addition to regulatory noncoding RNAs.
Bulk Tissue RNA-seq	RNA-seq with whole tissue samples	Profiled RNA is a mixture from many distinct cell-types within tissue samples. However, compared to other approaches, more distinct samples can be profiled with greater read depth.
Single cell RNA-seq (scRNA-seq)/ Single nucleus RNA-seq (snRNA-seq)	RNA-seq with single cells or single cell nuclei	RNA is profiled within specific cell-types, adding cell-type specific resolution to downstream functional interpretation. However, obtaining many distinct samples and obtaining increased read depth for single cells remains challenging.
Spatial Transcriptomics	RNA-seq across tissue sections, enabling RNAs to be mapped to specific tissue locations	Profiled RNA is localized to specific locations on tissue sections, adding spatial knowledge to downstream functional interpretation. However, cell-type resolution is lost, obtaining many distinct samples may be cost prohibitive, and read depth is limited.
Biological Factors	Biological attributes (metadata) of a sample which may contribute to gene expression diagnosis/condition, age, treatment, sex, cell-type proportions, etc.)	These factors are of interest in RNA-seq experiments, and are important for making functional inference about interesting genes and gene sets.
Technical Factors	Technical attributes (metadata) of a sample which may contribute to gene expression (sequencing batch, RNA quality, etc.)	These factors are generally not of interest in RNA-seq experiments, but often do contribute to gene expression, making it essential to control for the contribution of these factors as much as possible.
Differentially Expressed (DE) Gene Analysis	Use of a generalized linear model to evaluate gene expression in biological factors	This type of approach will identify all of the genes that vary across conditions with detectable statistical significance.

Gene Network Analysis	Use of a gene clustering approach (eg. hierarchical clustering of gene connectivity, used in WGCNA) to identify groups of related genes (modules)	This approach identifies modules which capture the gene expression patterns of cell-types and biological processes in sequenced tissue samples. Therefore, identifying modules with variable expression across conditions of interest will implicate these specific cell-types and biological processes in those conditions.
Gene Set	Two or more genes which are similarly differentially expressed in one group compared to another group	Groups of differentially expressed genes with increased or decreased expression between conditions (eg. case v. control), and gene co-expression or correlation modules from a network analysis, are examples of gene sets.

Table 1.1: RNA-seq basic terminology. Overview of fundamental vocabulary for psychiatric transcriptomics.

The types of samples selected for an RNA-seq experiment will determine what questions can be addressed by the experiment. The first decision is the choice of organism. While human samples will directly inform questions regarding psychiatric disorders, the types of human samples available are limited to postmortem and *ex vivo* tissue, and studying specific genetic and environmental risk factors with human samples is immensely challenging if not impossible in many cases. These risk factors can be better investigated in model systems such as mice, rats, and *in vitro* 2D and 3D cell cultures, but it is important to consider how well these models recapitulate psychiatric disorder mechanisms. Model organisms offer the ability to study entire neural systems, but the applicability of findings with these models in humans always needs to be verified. With *in vitro* human cell culture models, while cells can be human, these models still do not recapitulate complete *in vivo* neural systems (in general, model systems are simpler and more experimentally tractable than true biological systems), making it important to validate results from these models as well. Using organisms where the transcriptome is well-defined (such as with human and mouse from GENCODE[10]) is another factor to consider - if an organism does not have an annotated transcriptome, experimenters will have to obtain it, adding additional complexity to the experiment.

After choosing a model system, it is important to consider which types of samples from these models can address project goals. Samples can be taken from a single area over a period of time, distributed across distinct regions, or span across different conditions. Sampling across specific timepoints and ages - also referred to as obtaining temporal or longitudinal resolution in an experiment - is advantageous for understanding how model systems change with age, or respond to treatments administered across a period of time. Alternatively, to determine how gene expression is changed across different organs or distinct areas in a model organism, it is necessary to have sufficient spatial resolution. This type of resolution is important for determining if changes seen in one area - such as a brain region - are present across multiple areas, or limited to just that single area. Finally, in an experiment geared towards understanding how distinct disorders, conditions, or exposures alter a model organism, it is important to obtain sufficient resolution across these groups of interest. In general, for every comparison and contrast that one seeks to make across groups - whether they be temporal, spatial, disorder, condition, or exposure groups - there must be a sufficient amount of samples within each distinct group to make those comparisons of interest. Conducting a preliminary power analysis[11] is often advisable to determine how many samples will most likely be needed. And since it is often challenging to obtain enough samples to have complete temporal, spatial, and disorder/condition/exposure resolution for an experiment, prioritizing the most important questions and planning experiments that obtain sufficient sample resolution for those questions is an effective strategy.

The type of RNA-seq that you will conduct - bulk, single cell/nucleus, or spatial - should also guide choice of samples. With bulk RNA-seq, while profiled RNA is a mixture from many distinct samples, it is more cost effective to sequence more samples with greater read depth (total number of RNA reads sequenced) than with other methods. Single cell or nucleus RNA-seq does provide single cell resolution, but what is gained with this cellular resolution often needs to be traded for resolution across distinct samples due to cost. Likewise, there are only a limited number of RNA molecules detectable within a given cell, with substantial dropout (genes with 0 counts). Substantial cDNA amplification is often required which can lead to PCR artifacts, although incorporation of a unique molecular index (UMI) can help to address this challenge. Read depth is also lower with single cell RNA-seq, and RNA extraction is still limited to polyA capture approaches, limiting the amount and diversity of RNA that can be collected with this method. Spatial RNA-seq provides spatial resolution on a micro-scale across sample slices, localizing RNA to specific cellular locations. However, spatial RNA-seq is subject to some of the same shortcomings as single cell/nucleus RNA-seq – cost, lower read depth, often a limited number of samples, and limited RNA diversity. All of the pros and cons of these different RNA-seq approaches should be considered during the experiment planning stage. In general, when interested in profiling the cellular diversity of specific samples and understanding how RNA is utilized in cells, single cell/nucleus and spatial RNA-seq are ideal choices. However, when seeking to profile diverse types of RNA, and/or many samples across time, regions, or conditions, bulk RNA-seq may be a more appropriate choice. Ideally, ultimately running multiple RNA-seq experiments to understand gene expression changes across time, space, conditions, cell-types, subcellular compartments, etc. is desirable. However, in all cases, it is important to think critically about how the type of RNA-seq selected will address key questions of interest. While newer and advanced techniques are often enticing, sometimes they may be unnecessary or, worse, detrimental towards achieving your experimental aims.

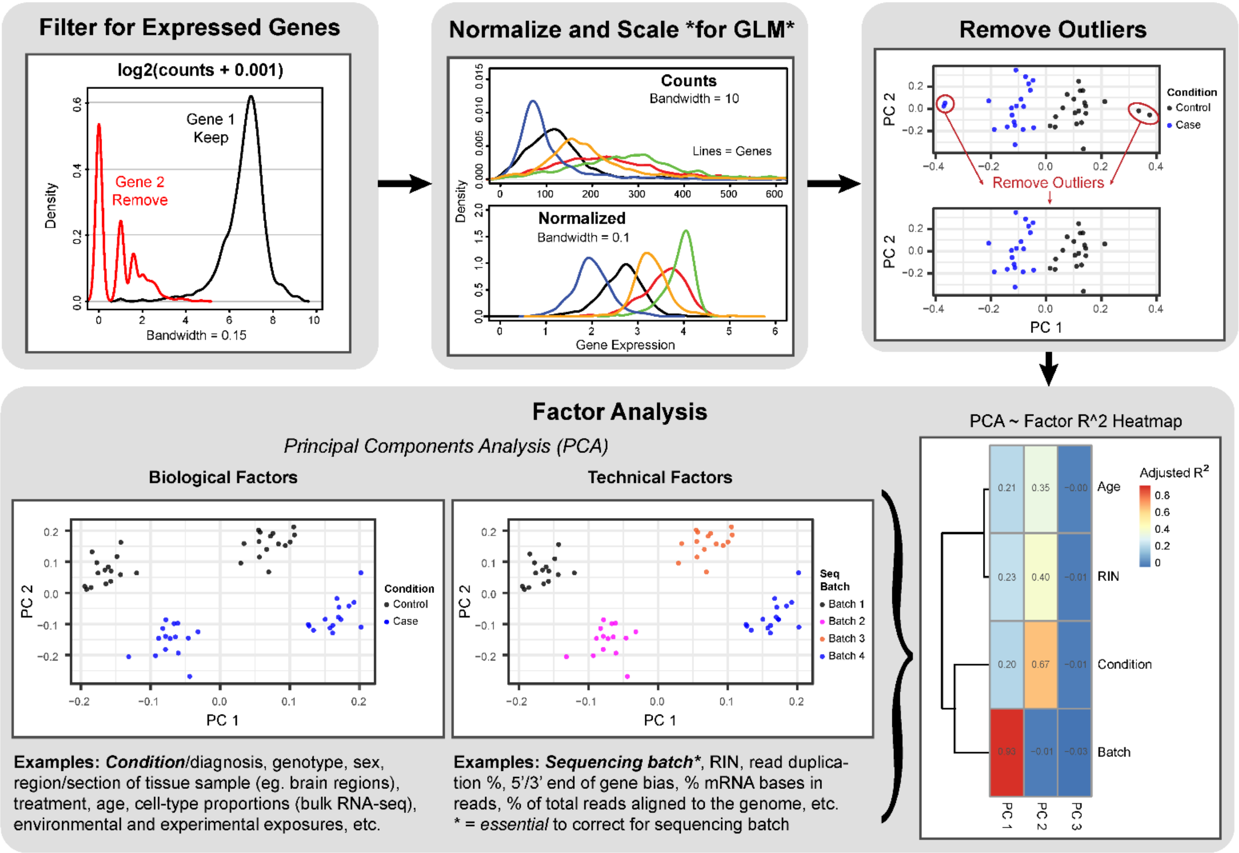
After samples are collected and the type of RNA-seq is chosen, the next task is to decide how RNA will be extracted, prepared for sequencing, and quantified in your experiment. Many resources exist to assist with making these choices,[12–16] and as this step is not necessarily specific to psychiatric disorder research I will not cover this topic comprehensively, but I will review some of the main considerations. There are many types of RNA that can be selected for - coding mRNA’s can be captured via polyA selection, noncoding RNA (such as lncRNAs) can be interrogated from total RNA libraries, which usually involve a set that depletes the otherwise very abundant levels of rRNA within a cell. Probe-based capture methods have also been developed which can enrich for specific genomic targets, as is done for whole exome sequencing, for examples.

Reads can be captured in experiments that range between 50 - 150 bp (or even longer) RNA reads. Paired and single end sequencing can both capture all of these types of RNA, however paired end reads are preferred since genome mapping rates (and consequently, read quantification) is improved when both RNA read ends are sequenced. However, when seeking to quantify distinct gene isoforms and identify differential RNA splicing and editing across samples, it is necessary to select for longer reads (at least 75 bp) and implement paired end RNA-seq. For microRNA and other experiments evaluating small noncoding RNAs, shorter read lengths - less than 50 bp - must be selected for. The type of RNA selection method also influences the type of RNA that will be sequenced. Poly-A selection methods use oligo-DT priming to generate cDNA libraries from mRNA and some pre-mRNA that have undergone poly-adenylation. rRNA depletion methods remove rRNA from extracted RNA, retaining all remaining types of RNA - beyond just protein coding RNA. In thinking about how to quantify RNA, there are two main choices - RNA can be quantified at the whole gene level (the sum of all exons), or the isoform level (estimate how reads aligned to the genome are distributed across isoforms). In general, for initial experiments which seek to obtain an initial understanding of how biological systems are acting in samples, approaches such as polyA RNA extraction with gene level quantifications are often sufficient. Alternatively, for experiments seeking to obtain finer, more detailed information about how distinct types of RNA are altered across samples where much is already known, it is advisable to perform rRNA depletion with longer RNA reads and isoform quantification. In general, if cost allows it is advisable to obtain rRNA depleted, longer RNA reads with greater read depth since these sample preparation choices offer more options for downstream analysis - but if funding is limited, for many types of experiments shorter RNA reads with relatively lower read depths can still be quite informative. It is necessary to reflect on experimental aims and choose sample preparation steps accordingly. Finally, it is important to evaluate RNA quality (RIN) and other RNA and sample quality metrics (such as postmortem interval with human samples, and pH) prior to sequencing, as these factors are known to effect sequencing quality.[17]

To end this discussion on effective experimental design, I will emphasize a critical planning step: the assignment of samples to RNA-seq batches. Implementing a randomized block design for sequencing batches, while seemingly trivial, is one of the most important aspects of RNA-seq experimental design. This is because sequencing batch contributes substantially to gene expression variance across samples in RNA-seq data, making it critical to account for it when conducting differential gene expression analysis. Therefore, it is extremely important to think about the comparisons and contrasts that will be made across groups of interest before sequencing, and distribute samples within these groups approximately equally across all sequencing batches. Ideally, all of your samples can be sequenced in one batch (this can be achieved through tagging RNAs from each sample with a specific barcode, a process referred to as sample multiplexing) - if this is the case, then there is no need to worry about randomizing samples across batches. However, for larger experiments that do require multiple batches, sorting approximately equal numbers of samples from each group of interest into each batch is absolutely necessary. If this is not done - for example, if all cases are in one sequencing batch and all controls are in another batch - then it will be impossible to tell sequencing batch effects and biological group effects apart in subsequent differential gene expression analyses, limiting the interpretability of the entire RNA-seq experiment. In this case, sequencing batch would be referred to as a confounding factor. To prevent this from happening, in addition to distributing equal numbers of samples from groups of interest across sequencing batches, it is advisable to run a simple chi-square test of independence to verify that groups of interest and sequencing batch are truly unassociated. Finally, while carefully planning sequencing batches is important, I note that randomizing samples across flow cell lanes is not required, as it has been well-documented that there is very little if any contribution of distinct flow cells to RNA-seq gene expression variance.[18]

1.4: Processing RNA-seq data

RNA-seq data processing is a rapidly evolving area of research where new tools and approaches are developed regularly. While these new tools generally enhance the accuracy and capability of RNA-seq analyses, with the vast array of RNA-seq processing pipelines available it is often difficult to determine which ones are appropriate for different experiments. To address this barrier, this section will focus on reviewing the main steps inherent to every RNA-seq processing pipeline (**Table 1.2**), as well as the justifications for performing these steps in different contexts (**Figure 1.2-1.3**). We will not go into detail comparing and contrasting specific tools, since this information is already generally well-documented and not unique to psychiatric disorder research. Instead, for help with choosing specific RNA-seq processing tools for your experiment, I recommend searching for recent reviews, blogs, and discussion boards that evaluate different tools of interest to help inform your decision. Websites such as Biostars[19] and SeqAnswers[20] are often good places to start. Here, I will provide essential knowledge for evaluating both the current and future RNA-seq processing tools that are comprehensively described in these other resources. Through discussing the concepts behind every main RNA-seq data processing step, this section will function as a universal and foundational guide for RNA-seq pipeline design.



Examples: ***Condition***/diagnosis, genotype, sex, region/section of tissue sample (eg. brain regions), treatment, age, cell-type proportions (bulk RNA-seq), environmental and experimental exposures, etc.

Examples: ***Sequencing batch****, RIN, read duplication %, 5'/3' end of gene bias, % mRNA bases in reads, % of total reads aligned to the genome, etc.
* = *essential* to correct for sequencing batch

Figure 1.3: Overview of RNA-seq data processing steps. Depictions of essential RNA-seq data processing steps: gene filtering, normalization, outlier removal. Then, processed RNA-seq data can be evaluated with a factor analysis to prepare for gene set prioritization analyses. The data shown here is randomly generated.

Processing Step	Description	Purpose	Considerations	Tools
RNA Read Alignment to the Genome	Match RNA reads to probable DNA/gene sources	This step is essential for variant calling and RNA splicing analyses, and is also necessary for some RNA read quantification tools	Not all RNA reads will align to the genome (eg. RNA derives from an unannotated genomic region), so some reads will be lost at this step	STAR,[21] HISAT2,[22]
RNA Read Quantification	Quantify the estimated number of RNA reads matched to each gene and/or isoform	Needed for downstream quantitative analyses	Estimated counts are the predicted counts of RNA reads, whereas TPM and FPKM are adjusted for differences in total RNA library size and gene length	Salmon,[25] Kallisto,[26] Rsubread,[27] RSEM,[28], featureCounts,[29] HTseq[30]
Genomic Variant Calling from RNA	Identify SNPs in RNA-seq data	Comparing SNPs from RNA-seq data with sample genotype data (separately collected) can be a useful quality control measure	SNP calling from RNA-seq data is not as accurate as whole genome sequencing or using a genotyping chip, and should not be used as the primary source of genotype information	GATK[31] (for sample swap detection, ancestry, and sex determination), https://github.com/brentp/somalier, verifyBAMID[23]
RNA Splicing Analysis	Identify splice sites present in RNA	To determine if there are any differentially spliced sites in the transcriptome across sample groups	Integration of differentially expressed isoforms with differentially spliced sites can enhance our understanding of these isoforms (we can predict if differential splicing is likely driving differential expression)	Leafcutter,[32] rMATS,[33] MAJIQ[34]
Gene Filtering	Remove genes with near zero expression in all samples	Only genes which are reasonably expressed can be assessed for differences across groups	Generally, genes with at least 0.1 CPM in 50% of samples is a reasonable filter; however, it is often helpful to evaluate multiple 'x' CPM in at least 'y%' of sample thresholds to select the optimal thresholds where the number of genes kept 'levels off' (increasing y does not greatly reduce the number of genes kept) and is reasonable (at least 15,000 genes for most experiments)	DESeq2,[35] or simply (in R): > filt = apply(cpm>0.1,1,sum) > keep = which(filt > 0.5*dim(cpm)[2]) > cpm_filt = cpm[keep,]
Normalization	Transform gene expression so that each gene is normally distributed	Normalizing enables samples to be fairly compared with a GLM	Taking log2 and adjusting for read depth variance across samples (TMM based size factors) are common approaches	Limma (voom),[36] DESeq2 (VST),[35] edgeR,[37] CQN (sqn)[38]
Outlier Removal	Remove sample outliers from gene expression data	Outliers can skew results and conclusions, so it is beneficial to remove them	Removing samples 3 standard deviations or more away from the mean of top PCs from PCA (covering 50% or more of expression variance) within groups contributing substantially to gene expression variance (usually condition/treatment and sequencing batch) is a reasonable approach	PCA, hierarchical clustering of samples, or with sample connectivity (in R with WGCNA[39]): > normadj = (0.5+0.5*bicor(expr))^2 > net = fundamentalNetworkConcepts(normadj) > ku = netsummary$Connectivity

Evaluate RNA-seq Quality Metrics	Determine how RNA-seq quality metrics contribute to gene expression data	RNA-seq metrics which contribute substantially to gene expression data should be corrected for in subsequent analyses, as this will help to identify biological effects	One approach for choosing RNA-seq quality metrics to correct for is to generate a heatmap of GLM associations (p-values, R^2, etc.) between expression data top PCs from PCA and RNA-seq quality metrics, and to correct for the RNA-seq quality metrics which are strongly associated with your top PCs	PicardTools,[40] RSeQC,[41] and STAR[21], QoRTs,[24] and multiQC[46] all collect RNA-seq quality metrics from RNA-seq data
Check for Associations Within Metadata	Evaluate associations across both biological and technical factors	Any confounds (associated factors) should be addressed, usually through keeping only one unique factor out of associated groups	Many RNA-seq quality metrics (and possibly biological factors too) will usually be associated with each other – it is advantageous to just select one of these metrics out of associated groups when evaluating how these metrics impact top expression PCs from PCA	Regression (similar to what is described for 'Evaluate Quality Metrics')
Sequencing Batch Correction	Account for sequencing batch in DE gene analysis and gene network analysis	Sequencing batch is often the largest factor contributing to gene expression, so it is important to account for it	Ideally, all samples can be sequenced in one batch, but if not make sure to balance sample groups of interest across batches	ComBat,[42] Regression (select sequencing batch as a model covariate)

Table 1.2: Overview of RNA-seq processing steps. Descriptions, justifications, considerations, and examples of tools for each main RNA-seq data processing step.

Raw RNA-seq data is returned in FASTQ files, which contain the sequences of RNA reads along with sequencing quality information. Several tools exist for checking the quality of raw RNA-seq data, with FastQC[43] being one of the most widely used. In general, it is advisable to evaluate quality metrics at this stage and take note of any irregularities (for example, large numbers of RNA read duplicates for some samples), since this information may guide quality control and differential gene expression analysis choices later on. After this preliminary quality check, for most RNA-seq experiments the objective is to quantify how the RNA reads in these files are distributed across genomic regions for downstream quantitative analysis. Genomic variants can be called from RNA-seq data, however the reliability of these variant calls is often lower than that of DNA sequencing or chip genotyping approaches. This makes RNA-seq variant calling most suitable for quality control purposes, such as for validating that RNA-seq and genotype chip data match when taken from the same sample (verifying that samples were not accidentally swapped during sample prep and/or sequencing). Optionally, it is sometimes advisable to trim sequencing adaptors from raw RNA reads, especially if RNA sequencing length was greater than the size of selected RNA reads (for example, if 75 bp RNA reads were selected, but sequenced reads are 100 bp, adaptors should be trimmed). However, it is better to plan your RNA-seq experiment so that sequencing length is approximately equivalent with selected read length.

RNA can be quantified in two main ways: genome alignment dependent, or alignment free. In alignment dependent pipelines, RNA must first be mapped to the genome, and then these alignments are counted. In general, we expect that approximately >80% or more of reads should align to the genome for good quality RNA-seq experiments. When the alignment rate is lower, it is advisable to go over the preceding sample preparation and RNA-sequencing steps to determine what might be contributing to the low alignment rate - it may be an important quality issue that could negatively impact downstream analyses. Aligned RNA (stored in binary BAM files, or uncompressed SAM files) is also needed for many quality control metric gathering tools. RNA-seq quality statistics are valuable for gene set prioritization approaches, as I will discuss in the subsequent section. In alignment free approaches, while a transcriptome reference is still needed, the alignment stage is skipped and quantification data is obtained directly from FASTQ files. RNA-seq data can be aligned and quantified with entire genes, from transcription start site to transcription end site, or with distinct gene isoforms. Importantly, for short read RNA-seq, isoform quantifications are estimates based on the distribution of RNA read alignments across the exons of a gene. The longer the RNA read, the more accurate these estimates are, and for long read RNA-seq (such as with PacBio or Oxford Nanopore approaches[44]) complete isoform alignments and quantifications can be made, although the read quality is lower. However, long read sequencing presents its own challenges, since the processing tools and pipelines for long read approaches are not as well established as those for short reads.[44] For all types of RNA quantification methods, output can be delivered as direct estimates of RNA read counts or values scaled for read depth and/or gene/isoform length (examples of scaled output are CPM, TPM, FPKM, and RPKM). While it is nearly always necessary to correct for the influences of variable read depth across samples for RNA quantification estimates, depending on downstream analysis steps, corrections for gene/isoform length may be unnecessary. Furthermore, some differential gene expression tools require the unscaled estimated counts as input, as these methods conduct corrections for the influences of read depth variance and/or gene/isoform length internally. In addition to the quantification of complete RNA reads, differential RNA splicing can also be quantified and compared across conditions of interest. Splicing analyses are often informative for interpreting differential expression results, particularly differentially expressed isoforms.

RNA-seq processing is generally conducted with a high-performance computing cluster, considering that RNA-seq data files are usually quite large (at least 10 gigabytes in size per sample, for compressed files). In addition to the advantage of being able to work with larger files, working on a computer cluster is advisable because it enables RNA-seq sample processing to be performed in parallel. Several strategies exist for how to organize and implement RNA-seq processing steps with multiple samples.[45] However, all processing pipelines generally end with aggregating results across all samples, to create consolidated matrices containing all genes/isoforms/features and all samples (for example, quantification data output could be a matrix where every row is a gene and every column is a sample, and entries are estimated counts of RNA from the genes). MultiQC[46] is one commonly used tool for assembling RNA-seq processing and quantification output into single files. Aggregating output in this way is always advisable for simplifying downstream processing steps that prepare data for approaches such as differential gene expression. In regards to selecting the preceding RNA-seq processing tools (aligner tools, quantification tools, etc.) for an RNA-seq pipeline, choices must depend on obtaining speed, a small memory footprint, usability, and accuracy. Fast, user friendly tools with a small memory footprint that are highly accurate are ideal choices. When forced to choose between tools that perform well in all of these areas, choosing the tool that is more established, or even randomly choosing between functionally equivalent tools if none are well established, are acceptable options. Additionally, when not much is known about any tools that are being considered, performing a quality analysis with your own data to compare and contrast tool performance is a good approach. Lastly, for all of the preceding processing steps discussed so far, it is important to always consider how each of these steps advances the RNA-seq experiment closer to addressing the motivation for the experiment. Quality control and read quantification are at the heart of RNA-seq processing, but additional steps - such as variant calling and RNA splicing analysis - may be advantageous for large, complex experiments that seek to characterize DNA regulatory relationships of RNA (for example, expression quantitative trait loci (eQTL) or splicing-QTL analyses). And, conversely, for smaller experiments which are geared towards gaining an initial functional understanding of biological systems in sequenced samples, obtaining simple gene-level RNA quantification data may be all that is required.

The following sections will focus on the analysis of RNA quantification data, as this is the most common application of RNA-seq, however many resources[47,48] exist which further describe the analysis of other types of processed RNA-seq data (such as genetic variant data and splicing data). With quantified RNA read data (estimated counts, CPM, TPM, etc.) there are three steps (**Figure 1.3**) that commonly take place before (or during) differential gene expression analysis and/or gene network analyses; from this point, I will refer to both of these approaches as 'gene prioritization analyses.' The first task after obtaining gene quantification data is a gene filtering step. Genes with zero expression across most samples are uninformative for gene prioritization analyses, requiring their removal from subsequent analysis. Next, for downstream approaches that utilize generalized linear models (GLMs) and/or gene network analyses, the quantification data must be normalized, typically by taking $\log_2$ after quantification data is corrected for read depth differences across samples. If gene network analysis is being performed, it is also advisable to correct for gene length influences on gene expression. Read depth differences can be corrected for by obtaining CPM (counts per million), whereas measures such as TPM (transcripts for million; generally preferred over RPKM/FPKM) can correct for both read depth and gene length. Another additional step that accounts for read depth influences on gene expression across samples is to correct for with trimmed mean of ‘m’ values (TMM).[49] Some differential gene expression tools, such DESeq2,[35] do not require users to normalize gene expression data prior to differential gene expression analysis - instead, DESeq2 integrates these corrections into differential gene expression calculations. The last step before gene prioritization steps can be implemented is outlier removal. Outliers can skew differential gene expression results and gene networks, so it is important to remove them before these steps. Evaluating top principal components as well as sample connectivity are both reasonable approaches to outlier removal. Additionally for this step, it is important to separately evaluate groups that contribute significantly to gene expression variance or are of special interest, such as sequencing batch and disorder/condition groups, since samples that appear to be outliers in the full dataset may just be a standard deviation away from its respective group mean. This can be especially true when group differences are large. Accordingly, when multiple groups contribute substantially to gene expression variance, outliers should be evaluated in combinations of groups (for example, all cases in sequencing batch 1, all controls in sequencing batch 1, etc.).

1.5: Gene prioritization with RNA-seq

To identify interesting genes with the greatest possible accuracy in RNA-seq experiments, it is important to understand how different technical and biological factors (also referred to as metadata) all contribute to gene expression variance across samples (**Table 1.2**). Biological factors are attributes of biological samples, such as age, bodily region, disorder status, genotype, treatment duration, etc. Technical factors are features of an experiment that are unrelated to the activity of biological systems in sequenced samples, such as sequencing batch, RIN, sample pH, etc. Other technical attributes of samples, such as percent of RNA read duplicates, 5’ end sequencing bias for reads, percent of coding RNA, etc., that can be obtained with QC metric acquisition tools also often explain gene expression changes, making it important to obtain these metrics during the preceding RNA-processing steps and to evaluate how they impact gene expression variance. While biological factor influences are usually what experimental aims are focused on, technical factors can have substantial impacts on gene expression variance as well, making it important to account for these factors. By including all of these types of factors in subsequent gene prioritization analyses, it is more likely that interesting genes can be found.

Once quantification data is cleaned and processed, then a factor analysis can be conducted to identify the factors (or ‘covariates’) that measurably contribute to gene expression variance (**Figure 1.3**). There are many valid approaches for this, and they are all based on evaluating covariates in the different types of generalized linear models (GLMs) used for gene prioritization analyses. Implementing the likelihood ratio (LR) test, Wald test, or the Lagrange multiplier test are all common choices for assessing and selecting GLMs, as they enable the comparison and optimization of different GLMs.[103] Here, I will describe how to conduct another simple approach, which is to visually inspect how covariates (which can include hidden covariates, when known covariates do not sufficiently explain gene expression variance)[104] contribute to gene expression variance as measured by principal component analysis (PCA). The first step is to measure the associations between the top principal components (PCs) of the gene expression data (the PCs that, together, explain greater than 50% of gene expression variance; or, at least the top 10 ranked PCs) and all known biological and technical factors. It is sufficient to conduct a simple linear regression for this, comparing every top PC with every known factor. The model adjusted R^2 is an advisable metric for comparison, since it is a single measure that applies to single numeric factors as well as multi-level categorical factors (eg. when three or more brain regions are included in an experiment). Importantly, I note that one should assess how correlated different factors are in an experiment - among correlated factors, only one should be evaluated for factor analyses. For example, cell line may be correlated with cell bank - in this case, just one of these factors (perhaps cell line) should be included for factor analysis. Additionally, factors should all be centered and scaled, or else the factors with larger numeric range will dominate PC associations. After the top PC associations with all factors are ascertained, it is then useful to visualize these associations on a heatmap where factors are hierarchically clustered based on top PC associations. With this heatmap, top factor associations with PCs will be clear. Selecting the top factor from each covariate cluster with a significant association with any of the topPCS is sufficient for choosing covariates for subsequent gene prioritization analyses. Tools such as variancePartition[50] and MARS[51] can be helpful for validating that selected covariates explain gene expression variance. Alternatively, these tools can be used independently to select GLM covariates that best explain gene expression variance, and then these covariates can be validated through measuring top PC associations.

Once factors that contribute substantially to gene expression changes are identified, a GLM must be designed that can account for these factors in gene prioritization analyses (**Figure 1.4; Table 1.3**). For simple experiments, GLM design is trivial, but for complex experiments great care must be taken to design GLMs that can best address experimental aims. Therefore, I will discuss how different types of covariates commonly present in psychiatric disorder experiments can be modeled with a GLM. Binary fixed effects are simple categorical covariates with only two choices, some examples being control v. case, frontal cortex v. occipital cortex, untreated v. treated, etc. For these types of factors, it is important to remember to set the appropriate reference group that will become the model intercept - in a case v. control experiment, for example, the reference group will likely be the control group if effects in cases are of interest. Continuous fixed effects are, as the name suggests, continuous quantitative measures that can be any value from a certain distribution, such as age, RIN, psychiatric diagnostic scores, number of observations of a behavior, cell-type proportions (often a major contributing factor to large variance components in bulk RNA-seq), etc. However, just because a covariate is numeric does not mean it is continuous. When discrete numeric values make up a covariate, such as three distinct treatment dosages, four possible test results, etc., this is referred to as an ordinal, multi-level covariate. Multi-level covariates are generally categorical, but in the case of ordinal factors then each category is represented by a number. In general, whenever a covariate has three or more distinct groups it is a multi-level factor, and it is especially important to be mindful of which group is set as the reference group for these covariates. To make comparisons between two groups within a multi-level covariate when neither is the reference group, it is necessary to perform linear contrasts. In more complicated cases, sometimes an experiment aims to evaluate more than one covariate. There are two main ways to evaluate more than one covariate with a GLM: the first is to investigate subsets of factors, such as comparing cases and controls within different brain regions, and the second is to explicitly determine how one factor interacts with another, such as measuring how sex effects case gene expression. When interested in subsets of factors, it is advantageous to combine factors together into these subsets and perform linear contrasts to directly address experimental aims. For example, for a schizophrenia postmortem human tissue sample from the temporal cortex, one may make a single combined disorder x region covariate that is equal to something such as SCZ_temporal for this sample. However, if interested in how the temporal cortex changes schizophrenic gene expression patterns in comparison to the frontal cortex, constructing an interaction term to capture this effect would be more appropriate. Finally, for experiments where multiple samples (at least three) are taken from a single subject - such as a longitudinal experiment where many samples are taken at different timepoints - it will likely be advantageous to utilize a mixed linear model with a random effect of subject. Since different individuals often vary in their baseline gene expression, for many different possible reasons, accounting for these differences will improve most GLMs. However, the likelihood ratio test or another model evaluation test can always be implemented to evaluate if a random effect is truly necessary.

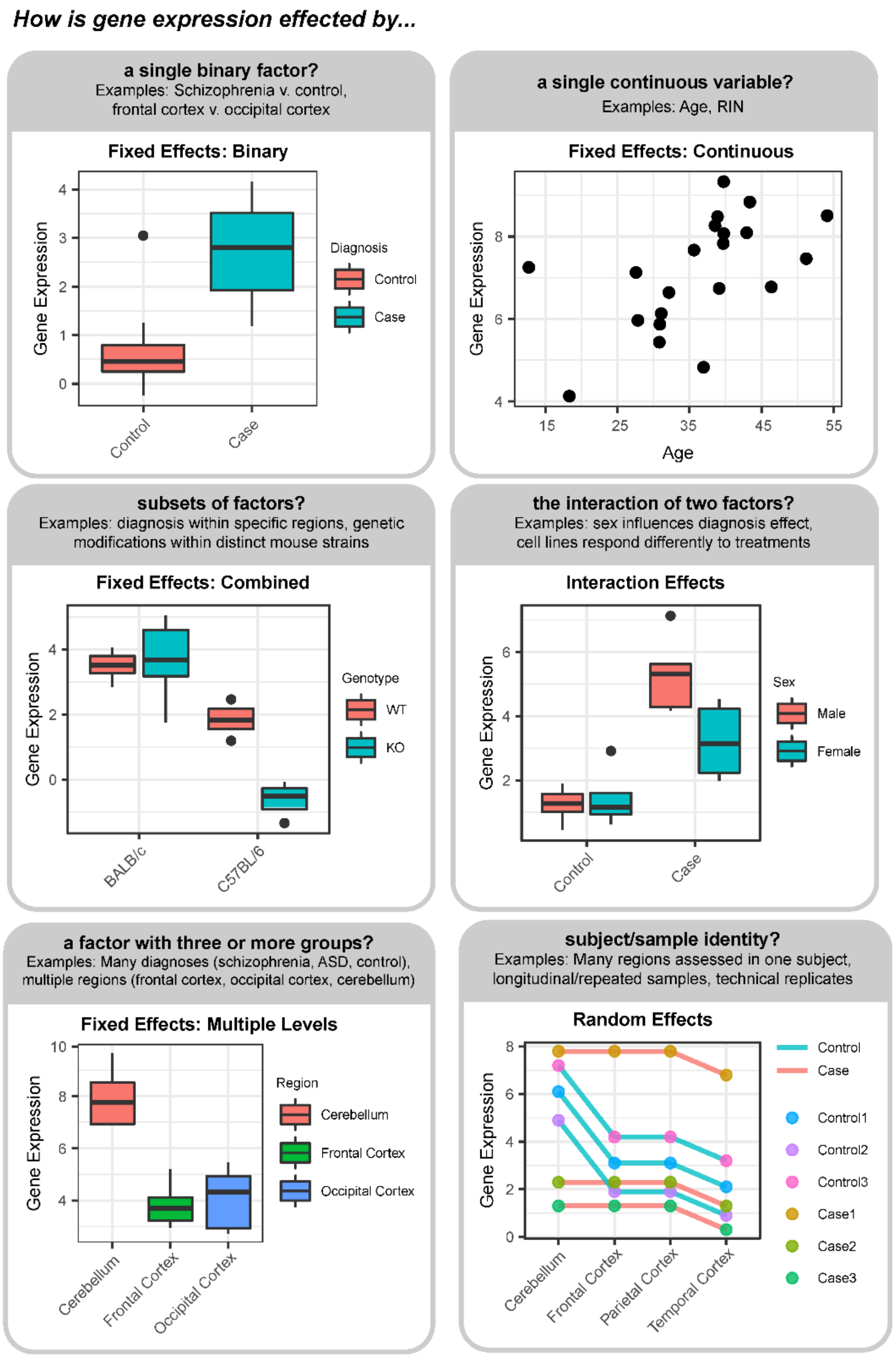


Figure 1.4: Decision chart for GLM design. Depictions of different factors that often influence gene expression variance in psychiatric transcriptomic experiments. Questions regarding how these factors effect gene expression guide GLM design.

Approach/ Term	Description	Strengths	Considerations	Examples
Generalized Linear Model (GLM)	A linear regression where response variables can have many types of statistical distributions beyond the normal distribution; DE gene effects are most commonly identified with GLMs	Many distinct types of data – particularly count data from RNA-seq – are readily analyzed with GLMs (such as with the negative binomial distribution)	Many types of GLMs can be implemented in RNA-seq with different tools, but for all approaches it is important to include covariates which contribute to gene expression variance	Limma/voom,[36] DESeq2,[35] edgeR,[37] Sleuth,[52] nlme,[53] lmer,[54] the base 'lm()' function in R
Choose Covariates for GLM	Include model covariates which explain gene expression variance	Biological effects of interest can be better detected when gene expression data is modeled as accurately as possible	Factor analysis is useful for determining which covariates to include in your GLM	MARS (such as with the 'earth' package in R),[51] variancePartition (in R),[50] see 'Evaluate Quality Metrics' in Box 2, caret[108]
Fixed Effects Model	Binary (such as case/control), continuous (such as age), and multivariate (such as three sequencing batches) covariates in a GLM	Easily interpretable covariates that can explain much of gene expression variance	Choice of reference group for categorical covariates is important for drawing inference from significant gene expression differences	Diagnosis, treatment, sequencing batch, genotype, age, treatment duration, sex, region (of a tissue or organ), tissue, organ, etc.
Mixed Effects Model	Often used for repeated samples, these models set different intercepts for 'random' effects	These models can better account for pervasive differences across sample groups	Sufficient numbers of samples (at least three) from each random effect group is required	Repeated samples from unique individuals, technical and biological replicates, longitudinal data, etc.
Linear Contrasts	Make many comparisons across and within distinct groups	Contrasts enable specific questions to be addressed without relying on a reference group, as with fixed effects	Combining factors together – such as diagnosis and brain region – is often advantageous for large and complex experiments (for example, compare region 1 in cases to region 2 in cases)	For comparing subgroups within multivariate factors, for comparing combinations of covariate groups
Interaction Effects	Determine if the interaction of two factors leads to an effect that differs from the effect of each factor alone	Interaction effects can reveal how covariates within an experiment influence each other to impact gene expression	Interaction effects work best for simple, pointed questions – such as 'how does sex effect case gene expression?'; for general surveys of differences across many multivariate factors, linear contrasts are usually easier to interpret	Should be implemented when interested in how one covariate effects another: how genotype influences case response, how age influences treatment response, etc.

Construct a Regressed Gene Expression Dataset	Remove the effects of undesired covariates (obtained from a GLM) from gene expression data	This is often advantageous for gene network analysis, to keep covariates such as sequencing batch from influencing gene network construction	Based on your experimental question, you may choose to remove biological covariate effects as well as technical covariate effects (for example, remove sex effects if you want a sex neutral gene network)	In R:
				y = gene expression
				y_reg = regressed y
				mod = model matrix
				effect = measured effects from GLM with model matrix
				> y_reg = y – mod %*% effect

Table 1.3: Approaches for identifying DE gene sets. Terms with descriptions, strengths, considerations, and examples that contribute to GLM design and DE gene set identification.

Once a GLM is constructed, a gene prioritization analysis can be performed. The most widely implemented of these, performed for nearly every RNA-seq experiment, is differential gene expression analysis (DE gene analysis). In this approach, a GLM is applied to every gene present in the processed gene expression data, and any gene with an adjusted p-value (corrected for multiple comparisons) less than 0.05 for a covariate of interest is deemed 'differentially expressed' across that covariate. Some of the most commonly used methods for adjusting p-values are the Bonferroni correction and the false discovery rate (Benjamini and Hochberg) correction.[105] There are many tools that implement DE gene analysis with GLMs that are specialized for RNA-seq data, most of which are accompanied by detailed instructions, illustrative tutorials, and comparative studies (**Table 1.3**). If time allows, it is advisable to implement at least two of these different tools and retain the high-confidence DE genes that are detected by all of the DE gene analysis methods implemented. Typically, DE genes are sorted into two groups to form gene sets: a group of upregulated genes and a group of downregulated genes. For example, for genes with an adjusted p-value less than 0.05 for a case/control covariate where the controls form the reference group, genes with a positive effect are increased/upregulated in cases compared to controls, whereas genes with a negative effect are decreased/downregulated in cases compared to controls. In the next section, I will describe different functional analyses that can reveal how biological systems are impacted by these groups of up- and downregulated genes.

Other than DE gene analysis, the other major type of gene prioritization analysis is gene network analysis (**Table 1.4**). For this approach, genes are clustered together into groups (also referred to as modules) based on patterns of gene correlation, co-expression, and/or connectivity across samples.[39] Gene network formation methods differ in how gene connections are precisely defined, but all methods endeavor to measure gene relatedness across samples in an RNA-seq dataset and use this information to build a network from which modules can be identified. To perform gene network analysis, it is usually advantageous to use a regressed gene expression dataset (**Table 1.3**). A regressed dataset has the effects of technical covariates removed so that only biological covariates influence gene network formation. Here, I will point out three different types of gene network analysis: regulatory network formation, seeded co-expression analysis, and unseeded co-expression analysis. Regulatory gene networks (such as those formed with ARACNE[55]) are designed to extract modules containing direct regulatory relationships, whereas co-expression clustering approaches such as WGCNA[39] are better for identifying more expansive modules that can represent the activity of entire biological processes and cell-types. However, since regulatory and co-expression gene relationships are often correlated, ARACNE and WGCNA are both capable of identifying regulatory and co-expression modules to some extent. A seeded analysis only captures correlation patterns with a single gene, whereas unseeded approaches evaluate all gene correlation patterns to form modules. The type of gene network approach chosen should depend on experimental aims. If interested in identifying regulatory relationships, a regulatory network should be used. Or, if interested in generally characterizing a disrupted biological system, an unseeded co-expression network analysis may be more appropriate. However, for all types of gene networks, it is important to understand how the underlying assumptions behind network formation support experimental aims, and to systematically evaluate clustering parameters so as to optimize them for your experiment.

Approach/Term	Description	Strengths	Considerations	Examples
Co-expression Network Analysis (Unseeded)	Empirical, data-driven clustering of genes based on co-expression patterns across samples (connectivity is the co-expression measure used for WGCNA)	Reduces data dimensionality (1,000s of genes reduced to modules) and matches gene expression to distinct cell-types and biological processes	Not all genes are grouped into modules, so some DE genes may not be included in modules	WGCNA[39], MEGENA[109]
Co-expression Network Analysis (Seeded)	Networks are formed based on association with a single 'seed' gene, producing one module of correlated genes, and one module of anti-correlated genes	Determine how the entire transcriptome is effected by the expression of a single gene	Seeded module formation is ideal for experiments where only a single gene is of interest – however, since seeded modules are designed to capture direct effects of the seed gene alone, any indirect and/or related transcriptomic effects will be missed	Created through correlation with a seed gene that passes a certain threshold
Regulatory Network	Identify genes that are in specific, direct regulatory networks (eg. transcription factors that directly influence the expression of other genes)	Modules represent direct regulatory relationships, so any group differences in these modules will represent dysregulation in specific regulatory pathways	Unlike co-expression networks, these networks will not capture transcriptome-wide patterns that represent activity in biological systems – these networks are only focused on identifying regulatory relationships	ARACNE[55]
Module Eigengene	The first principal component (PC 1) of a gene network module (can be thought of as 'average module expression')	Summarizes module gene expression for evaluation of the module as a whole	Module eigengenes can replace single genes for DE gene analysis approaches, so that entire modules with clear functional associations can be evaluated across sample groups of interest	All network approaches
Hub Genes	Genes to prioritize in a module, typically which are highly significantly associated with the module eigengene	These genes help to understand the function of a module better, as they are more likely to 'drive' module function than other genes in the module – they may be transcription factors in unseeded co-expression approaches	Hub genes can be defined in many ways, but they are always genes of key interest in a module based on some statistical measure – correlation with the seed gene (for seeded approaches) and increased regulatory activity in the module (for regulatory networks) may also determine which module genes are hubs	All network approaches

Table 1.4: Approaches for gene network analysis. Terms with descriptions, strengths, considerations, and examples that summarize different types of gene network analysis and core concepts/vocabulary.

Once modules are obtained, a key advantage of all types of gene network analysis is that modules can be analyzed much like the groups of DE genes previously discussed - but with much reduced dimensionality (less than 100 modules typically, compared to thousands of genes). And while groups of DE genes are usually quite broad, as they are the summation of all up- and downregulated genes, modules are constructed to capture specific biological patterns, making their functional interpretation typically more informative. Modules also contain 'hub' genes, which are those genes that are highly correlated with all other module genes. Hub genes are of particular interest, since they are likely to play driving roles in regulating module gene expression, possibly as transcription factors or through another biological mechanism.To determine how modules are impacted by covariates of interest, the first principal component (or 'module eigengene') of modules can be treated much like the expression of a single gene and analyzed with a GLM. However, while many genes with covariate effects that do not quite pass GLM statistical significance thresholds can be rescued by gene network analysis, it is also common for many DE genes to not be sorted into modules, making it advantageous to perform a DE gene analysis along with a gene network analysis to ensure that all dysregulated genes are identified and evaluated. In addition to these considerations, it is important to note that gene network analysis is generally harder to implement than DE gene analysis, even though many helpful tutorials[56,57] and resources[58,59] exist. Because of this, experiments with limited time and resources may only be able to perform DE gene analysis, which is entirely sufficient for many projects. Additionally, gene network analysis still proves to be a challenging and quickly evolving field for single cell/nucleus RNA-seq experiments,[60] creating another barrier to implementation for these projects. However, it is still advisable to use gene network analysis - especially for bulk RNA-seq experiments - if time and resources allow, considering that this approach is more specific and oftentimes more functionally informative than DE gene analysis.

1.6: Integration of RNA-seq results with orthogonal data

Once an interesting gene set has been identified with DE gene analysis or gene network analysis, the next step is to predict what the gene set represents in biological samples. This generally involves linking gene sets to the active cell-types and biological processes that likely drive gene set expression. To obtain this understanding, it is necessary to integrate gene sets with orthogonal biological data sources that can offer functional insights. Here, I will focus on two main types of orthogonal data integration approaches: functional enrichment analysis, and the evaluation of gene sets in external gene expression datasets (**Table 1.5**). Choosing which orthogonal data integrations to perform for an RNA-seq experiment is at the heart of gene set functional characterization, since all of these orthogonal data associations inform each other and, ultimately, must all be interpreted jointly to predict how biological systems are effected in conditions of interest. I will discuss this topic - the functional interpretation of interesting gene sets - in detail in the next section. Here, I will prepare for this discussion by providing an overview of orthogonal data integration approaches that are useful in psychiatric disorder research.

Does my gene set contribute to…	Approach	Tools and Resources	Orthogonal Data Integration Category	Strengths	Limitations
Biological processes?	Gene ontology enrichment	*Tools:* Metascape,[61] gProfileR,[62] GSEA,[63] PantherDB,[64] Enrichr,[65] topGO,[66] ReactomePA,[67] PathFinder,[68] WebGestalt,[69] DAVID[70] *Resources:* GO Consortium,[7171,72] KEGG,[73–75] Reactome	**Functional Enrichment** Determine if your gene set significantly overlaps with a known functional category	▪ Any list of interesting and relevant genes (transcription factors, previously identified co-expression modules, etc.) can be compared to your gene set using methods such as the hypergeometric test. ▪ Large curated lists and databases exist for functional enrichment analyses. · These approaches can be quickly implemented to gain functional insights from your gene set. ▪ For genetic variant enrichments: with human samples, these gene sets are more likely to contribute to causal pathology. This knowledge helps to prioritize and guide future research.	▪ Many of the functional enrichment terms present in large databases may not be relevant for your sample type (eg. a ‘kidney nephron development’ enrichment is largely uninformative for neural tissue derived gene sets). ▪ A lack of functional enrichments for your gene set does not mean there are none – especially if the gene set and functional enrichment terms are small. ▪ These analyses are limited to known lists of functionally related genes. ▪ For genetic variant enrichments: Some genetic variant databases are incomplete or underpowered, especially for highly heterogeneous psychiatric disorders, possibly leading to false negatives.
Cell-types?	Cell-type marker enrichment	*Tools:* Hyper-geometric test (Fisher’s Exact test), pSI,[76] EWCE[77] *Resources:* scRNA-seq,cell-type specific datasets			
Protein-protein interactions (PPIs)?	PPI enrichment	*Tools:* Dapple,[78] STRING[79] *Resources:* BioGRID,[80,81] InWeb[82]			
Causal pathology?	Common genetic variant enrichment	*Tools:* LDSC,[83] MAGMA[84] *Resource*: GWAS			
	Rare genetic variant enrichment	*Tool:* Logistic regression *Resources:* Databases such as SFARI,[85] SCHEMA[86]			

Neuro-developmental stages?	Evaluate gene expression across developmental periods	*Resource:* Allen Developing Brain Atlas[87]	**Evaluate Expression Patterns in External Datasets** Examine how your gene set – using the first principal component, or top ranked genes – is expressed in a relevant external gene expression dataset obtained from a publicly available database	· Determine if your gene set exists, and/or exhibits similar patterns of expression, in other related studies (is this gene set robust?). · Obtain greater resolution (eg. longitudinal) in your functional understanding of your gene set (does my gene set have consistent expression over time?).	· If samples are not well-matched or technically flawed, observed patterns may not be informative for your identified gene set. · Orthogonal publicly available datasets often only contain neurotypical controls, limiting interpretations to control samples.
Distinct brain regions?	Evaluate gene expression across brain regions	*Resource:* GTEX,[88] Allen Brain Atlas[9]			
Human-specific processes?	Evaluate gene expression in humans compared to other species	*Resource:* Allen Brain Atlas (mouse, non-human primate, human)[89]			

Table 1.5: Approaches for orthogonal data integration with interesting gene sets. Terms with descriptions, strengths, considerations, and examples that summarize different types of orthogonal data integration techniques.

Functional enrichment analysis is best suited for categorical types of orthogonal data, including gene ontologies, cell-type markers, protein-protein interaction lists, and groups of genes linked to genetic risk variants for psychiatric disorders. Many functional enrichment approaches employ some variation of the hypergeometric test to determine if an RNA-seq derived gene set overlaps with another gene set category more than expected by chance. Significant overlaps indicate that the orthogonal gene set category is over-represented in the RNA-seq derived gene set. While each approach has its own considerations and accounts for different influencing factors (eg. linkage disequilibrium for genetic enrichment analysis), this idea of 'identifying significant overlaps between gene lists' is behind most enrichment analyses. The implication of a significant overlap varies depending on the orthogonal gene set category. Gene ontology term enrichments indicate that the RNA-seq derived gene set is likely involved in the biological processes that correspond with the gene ontology term. However, it is important to consider how different gene ontology terms overlap with each other, and if positive gene ontology enrichments make sense for the experimental sample type. For example, a kidney nephron development enrichment with a hippocampal sample may indicate that the gene set is simply involved in developmental genes that overlap with the kidney nephron development gene ontology list.

Another type of functional enrichment approach is to examine cell-type marker overlap with interesting gene lists. This type of analysis is most applicable to bulk RNA-seq experiments, which will contain gene expression from the complete mix of cell-types present in samples. Positive enrichments indicate that the overlapping cell-type may be responsible for driving gene set expression. However, again it is important to consider how cell-type markers overlap across all of the cell-types examined - when this is the case, it is likely that a gene set that is significantly enriched for one cell-type will likely be significantly enriched for all other overlapping cell-types too. To address this, some approaches constrain cell-type markers to cell-type 'specificity' markers, which distinguish single cell-types from all other cell-types examined for the enrichment analysis. One such method, EWCE,[77] even employs a permutation approach - rather than the hypergeometric test - to establish cell-type enrichments with cell-type specificity markers. In addition to examining cell-type marker overlaps, another informative functional enrichment approach is to determine if direct protein-protein interactions are present in a gene list more than expected by chance. Positive enrichments imply that a gene list likely serves to produce proteins that participate in a biological process, rather than regulatory RNAs.

The last type of functional enrichment analysis that I will discuss here, which is specifically relevant for human postmortem and *ex vivo* tissue experiments, is genetic variant enrichment with gene sets. Both common and rare genetic risk variants can be compared with gene sets to determine if genes proximal to these risk variants are significantly over-represented. Positive enrichments indicate that this gene set may be causal, or at least contribute to causal processes in some way, in disorder pathology. Importantly, for all of the types of functional enrichment approaches I have touched on here, functional inference should not be made based on negative enrichments (the lack of an enrichment). In this case, it is impossible to know if a lack of statistical power, or other statistical or experimental influences, prevent a true positive enrichment from being discovered. For example, when comparing a small gene set with a small list of cell-type markers, even if the gene set is derived from that cell-type a hypergeometric test may not be significant due to the small input list sizes.

The second major type of orthogonal data integration approach is to evaluate gene set expression patterns in external gene expression datasets. These datasets include RNA-seq, and possibly gene expression microarray, experiments that capture neurodevelopmental trajectories (**Table 1.5**; Allen Developing Brain Atlas[87]), span distinct brain regions (**Table 1.5**; GTEX,[88] Allen Brain Atlas[9]), or contain multiple distinct species (**Table 1.5**; Allen Brain Atlas (mouse, non-human primate, human)[89]). Any external datasets used for this type of analysis should offer insights into longitudinal, spatial, and/or species trajectories that are absent from the original RNA-seq experiment. To evaluate gene sets in these external datasets, the first principal component of the gene set can be used, or select genes of particular interest within the gene set (eg. module hub genes, top 10 most significantly differentially expressed genes). I will refer to both of these as 'representative' genes from an interesting gene set. Representative genes can be evaluated with GLMs in external datasets. Of course, care should be taken to process these external datasets and account for impactful covariates, just as before with the original RNA-seq dataset.

The objective for analyzing representative genes in orthogonal gene expression data is to understand how a gene set may function in contexts beyond that of the RNA-seq experiment the gene set was identified in. Examining representative genes in datasets with good longitudinal and/or neurodevelopmental resolution can indicate how gene set expression changes over time. For example, if gene set expression is low in fetal developmental periods but increases continuously after birth, this indicates that the gene set is only active postnatally and becomes more active with maturity. Another valuable type of external gene expression dataset, multi-region datasets, offer increased spatial resolution that can reveal how a gene set is expressed across many distinct regions. For example, a multi-region RNA-seq dataset containing a span of cortical areas may show that a gene set identified in just one of these areas displays increased gene expression in anterior regions relative to posterior regions, indicating that this gene is more active in the frontal areas. Multi-species gene expression datasets can also help to interpret gene sets, especially those identified in a non-human model species. If representative genes for a gene set are highly expressed in a certain model system of a disorder, but not differentially expressed in samples from the human form of the disorder (where regions and ages are matched as well as possible), then the gene set may not be informative for understanding human disorder mechanisms. Finally, for modules identified with gene network analysis, module preservation analysis can be conducted to see if the correlation patterns that created a module in one RNA-seq experiment are recapitulated in an external dataset. This type of analysis can be especially helpful when comparing species, to determine if a module found in a model system (such as mouse) is also present in humans.

1.7: Functional interpretation of RNA-seq results to advance research

Three distinct factors contribute to functionally interpreting a gene set: integrations with orthogonal biological datasets (as discussed in the preceding section), biological attributes of sequenced samples, and the directionality of gene expression changes for biological covariates. The descriptive qualities of sequenced samples will determine how gene set integration with orthogonal datasets should be carried out and interpreted. For example, for RNA-seq performed with human samples, cell-type markers should be defined with human samples that match the sequenced brain areas and ages as closely as possible. Additionally, for this example any significant enrichments should be interpreted based on the external dataset in which cell-type markers were defined. If cell-type markers were defined in parietal cortex samples, then cell-type enrichments for gene sets identified in the frontal cortex should be interpreted as enrichments for frontal cell-types resembling parietal cortex cell-types. This example illustrates that the type of orthogonal data selected for integrative analyses, and the conclusions made based on significant results, both heavily depend on the biological attributes of sequenced samples. In general, functional interpretations of RNA-seq gene sets depend heavily on the types of samples sequenced - interpretations can only be made based on the temporal, spatial, and species contexts in which samples were acquired. For example, if RNA-seq was conducted with juvenile mice, then findings cannot be extrapolated to adult mice. The investigation of interesting gene sets in a new dataset containing adult mice (possibly an orthogonal gene expression dataset) would be the only way to predict how a gene set identified in the original RNA-seq experiment may behave in adult mice.

The other factor that contributes substantially to gene set functional interpretation is the directionality of gene expression changes for biological covariates. For example, if a bulk RNA-seq gene set is upregulated in a disorder group, and astrocyte cell-type markers are enriched in that gene set, then we can infer that astrocyte activity and/or cell-type proportion is increased in that disorder in the samples that were sequenced; the converse is true for downregulated gene sets. This same type of reasoning can be applied for continuous biological covariate associations. For example, if this same gene set that is enriched for astrocyte cell-type markers is also positively associated with age, this indicates that this gene set is most highly expressed in mature astrocytes. Lastly, gene expression effects in all biological covariates should be examined together - along with orthogonal data integrations and sample attributes - to ensure that functional interpretation is complete, accurate, and informative. For example, consider a spheroid model system with multiple different genetic knock-outs, where samples were taken at different timepoints. A gene set is identified that is enriched for excitatory neuron cell-type markers and neuronal projection gene ontology terms. If this gene set is highly expressed in a single genotype at an early developmental time point, the interpretation would be that only this genotype has early increased neuronal projection activity. However, if instead this gene set is equally expressed in all genotypes and is highly expressed at the final time point, this would indicate that neuronal projection processes are equally active across the genotypes and expressed later in development. Evidently, different gene expression effects in distinct covariates can heavily influence how a gene set is interpreted, even when orthogonal data integration results are the same. And, as this example illustrates, if individual covariates are examined alone then important characteristics - such as longitudinal trajectories, specific genotype effects, etc. - that further describe a gene set will be missed, leading to incomplete and minimally informative functional interpretations. As demonstrated by all of these examples, when functionally interpreting interesting gene sets it is advantageous to examine all relevant orthogonal data integrations, sample qualities, and covariate effects all together. This comprehensive approach enhances and optimizes gene set functional interpretation, leading to stronger and more informative biological insights that advance experimental aims.

Once gene expression changes in interesting gene sets are identified and linked to putative biological functions, the final step in an RNA-seq experiment is to use these functional insights to measurably advance project goals (**Figure 1.5a**). I describe this process with a workflow that can be broken down into four major stages. The first is the identification of interesting gene sets, which I have described in the preceding sections. The second step is to use orthogonal data integration techniques, also described earlier, to determine which features of biological systems are functionally associated with interesting gene sets. Third, sample characteristics and the directionality of interesting gene set expression changes should be integrated with functional associations to predict how biological systems may be altered in conditions of interest (such as psychiatric disorder patient groups, model mice for a psychiatric disorder, etc.). The last step is to prioritize which predictions are the most interesting and relevant in addressing experimental aims, and to then validate these predictions with follow-up experiments. To place this workflow into the context of psychiatric disorder research, consider the following example (also depicted in **Figure 1.5b**, column 1). An RNA-seq experiment is performed to understand how biological systems in human postmortem cortical tissue are effected in a psychiatric disorder where underlying neuropathology is largely unknown. RNA-seq analysis reveals a gene set that is downregulated (step 1) in psychiatric disorder patient cortical samples that are enriched for excitatory neuron cell-type markers and presynaptic vesicle release gene ontology terms (step 2). One important prediction from these results is that presynaptic vesicle release activity may be decreased in these patients (step 3). A validation experiment should then be planned to validate this result, such as examining presynaptic vesicle release in a human *in vitro* spheroid model of the psychiatric disorder (step 4). A particularly effective spheroid model would have a representative gene from the interesting gene set knocked out so that gene set downregulation can be directly compared to any decreases in presynaptic vesicle release.

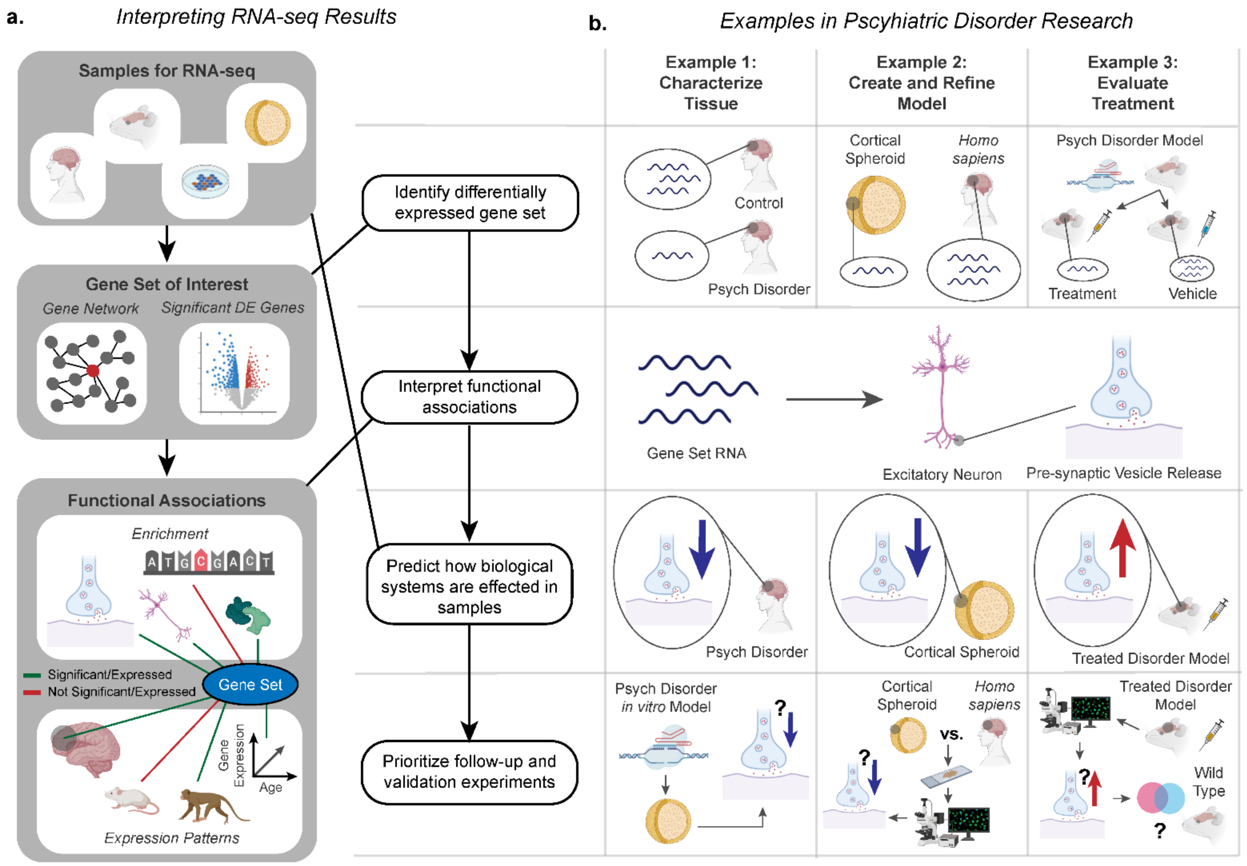


Figure 1.5: Direct applications of RNA-seq in psychiatric research. a. Fundamental workflow for interpreting interesting gene sets identified with RNA-seq analysis. **b.** Hypothetical examples of how this workflow can be applied in different areas of psychiatric disorder research to advance our understanding of psychiatric disorders. This workflow is explicitly demonstrated in Refs. 6, 8, 90-97.

As this example illustrates, this workflow can accept as input a sample type relevant to psychiatric disorder research where little is known and produce precise, data-driven predictions regarding how biological systems are altered in conditions of interest. These predictions can be tested experimentally, and if predictions are validated our knowledge of psychiatric disorders will be measurably expanded and research can move forward. In the case where predictions are unsupported, it is important to understand why a prediction turned out to be untrue. Returning to each previous RNA-seq analysis step and validating that gene expression dysregulation findings and functional interpretations were accurate, as well as evaluating how well the chosen validation experiment can test predictions, are just a few approaches that can help to reveal why a prediction from RNA-seq was inaccurate. The specific factors that contribute to a prediction being unvalidated will dictate how experimental results should be interpreted, and will also serve to develop and improve RNA-seq analysis methodology for future psychiatric disorder research projects. Advancing technical knowledge is just as important as directly expanding biological knowledge, since improving technical knowledge will surely enhance our ability to effectively apply RNA-seq across many distinct areas of psychiatric disorder research.

To further demonstrate how the RNA-seq analysis workflow described in this section can advance psychiatric disorder research efforts, I will discuss several examples that illustrate how this workflow can be successfully applied in different contexts (**Figure 1.5b**). One purpose for performing an RNA-seq experiment in psychiatric disorder research is to enhance our understanding of how biological systems are disrupted in a psychiatric disorder. Through sequencing human postmortem samples from psychiatric disorder patients, or samples from organisms that model a psychiatric disorder, RNA-seq analysis can identify and implicate different cell-types and biological processes in disorder pathology. Validation experiments can then confirm if the dysregulation identified with RNA-seq does likely contribute to psychiatric disorder mechanisms. To share just one example, RNA-seq analysis with human cortical samples from ASD patients has found that neuronal energetic function and synaptic genes are downregulated in ASD.[6,8] Since these findings, experiments with model mice have validated this result: neuronal energetic pathway downregulation was also observed in mouse models harboring human CNVs associated with ASD,[90] and another ASD model mouse (with the ASD-linked CNTNAP2 mutation knocked out) also exhibited decreased prefrontal synaptic activity.[91] Both of these experiments find that the neuronal and synaptic dysfunction observed with RNA-seq in ASD patient cortical samples is recapitulated in ASD model organisms, supporting that neurons play a role in ASD pathological mechanisms. Together, these results serve to expand and strengthen our understanding of ASD.

Another application of this RNA-seq analysis workflow in psychiatric disorder research is to validate and strengthen model systems. Using transcriptomic profiles to develop and validate model systems is important for ensuring that functional inference can be drawn from these models. This approach has proven especially useful for human cortical spheroid models, through comparing spheroid transcriptomic profiles with human fetal gene expression.[92,93] To share one last example, this RNA-seq analysis workflow can also be used to develop psychiatric disorder treatments. RNA-seq experiments can already reveal how current treatments effect patients,[94–97] and in the future RNA-seq can be used to evaluate the efficacy of new therapeutics targeting underlying causal mechanisms (once they are established). In this case, RNA-seq can reveal if transcriptomic pathology is mitigated by a treatment, returning closer to typical expression levels. Finally, beyond the examples that I have described here, there are certainly other contexts in which this RNA-seq analysis workflow is effective. When the fundamental steps that I have discussed in this chapter are followed, an RNA-seq analysis has the potential to substantially advance the aims of any psychiatric research project.

1.8: Discussion

In this chapter, I have demonstrated how RNA-seq analysis is a valuable and effective methodology for psychiatric disorder research. I provided an overview of how RNA-seq can be applied in different areas of psychiatric disorder research and discussed how to design experiments that directly address specific aims. Main processing steps for RNA-seq data analysis were justified and described, and different types of gene prioritization techniques were compared. Finally, I demonstrated how to functionally interpret interesting gene sets and established how these interpretations can lead to biological insights that measurably advance psychiatric disorder research. Together, all of these sections form an approachable overview of the foundational concepts that guide RNA-seq analysis in psychiatric disorder applications. A central theme of this chapter is that RNA-seq analysis can streamline and optimize research projects, serving to simplify experiments and highlight relevant cell-types and biological processes for continued investigation. These qualities are especially desirable for psychiatric disorder research, considering that little is often known about underlying neuropathology. Predetermined biological targets are unnecessary for RNA-seq experiments, since they profile the entire transcriptome. Functional interpretation analyses can connect gene expression changes to cell-types and biological processes, enabling RNA-seq experiments to provide initial characterizations of entire biological systems. However, I note that RNA-seq analysis is not entirely data-driven - while all gene expression changes are captured by RNA-seq, hypotheses must guide GLM design and functional association analyses. Prior knowledge is what enables gene expression changes to be identified and interpreted. The spatiotemporal context in which samples were sequenced, biological attributes of samples, and technical influences on gene expression also contribute critically to RNA-seq analysis. To ensure that an RNA-seq experiment can address experimental aims, it is essential to carefully design experiments with all of these contributing factors in mind.

In the following chapters, I will show how the concepts and workflows established here enabled me to refine our understanding of psychiatric disorders. In chapter two, I will demonstrate how the integration of gene expression microarray data and RNA-seq data across different psychiatric disorder experiments enhanced our understanding of how distinct psychiatric disorders compare to each other. In chapter three, I will build off of these findings and implement more advanced transcriptomic analyses such as isoform analysis, splicing analysis, polygenic risk score analysis, and TWAS to obtain a deeper understanding of transcriptomic similarities and differences across psychiatric disorders. Finally, in chapter four, I will show how a multi-regional transcriptomic study of the ASD cerebral cortex revealed widespread gene expression changes that heavily implicate neuronal synaptic plasticity processes in ASD disorder mechanisms. Together, all of these chapters will continuously demonstrate how the guidelines and strategies emphasized in this initial chapter can lead to robust and impactful advances in psychiatric disorder research.

1.9: Bibliography

1. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)*. (American Psychiatric Pub, 2013).

2. Jonaki Bose Sarra L. Hedden Rachel N. Lipari Eunice Park-Lee. Key Substance Use and Mental Health Indicators in the United States: Results from the 2017 National Survey on Drug Use and Health (HHS Publication No. SMA 18-5068, NSDUH Series H-53). *SAMHSA* (2018).

3. Zablotsky, B., Black, L. I. & Blumberg, S. J. Estimated Prevalence of Children With Diagnosed Developmental Disabilities in the United States, 2014-2016. *NCHS Data Brief* 1–8 (2017).

4. Sullivan, P. F. & Geschwind, D. H. Defining the Genetic, Genomic, Cellular, and Diagnostic Architectures of Psychiatric Disorders. *Cell* **177**, 162–183 (2019).

5. Schmitt, A., Malchow, B., Hasan, A. & Falkai, P. The impact of environmental factors in severe psychiatric disorders. *Frontiers in Neuroscience* vol. 8 (2014).

6. Parikshak, N. N. *et al.* Genome-wide changes in lncRNA, splicing, and regional gene expression patterns in autism. *Nature* **540**, 423–427 (2016).

7. Haney, J. R. *et al.* Broad transcriptomic dysregulation across the cerebral cortex in ASD. *bioRxiv* (2020).

8. Gandal, M. J. *et al.* Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. *Science* **362**, (2018).

9. Hawrylycz, M. J. *et al.* An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* **489**, 391–399 (2012).

10. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research* vol. 47 D766–D773 (2019).

11. Fitzner, K. & Heckinger, E. Sample size calculation and power analysis: a quick review. *Diabetes Educ.* **36**, 701–707 (2010).

12. Van Den Berge, K. *et al.* RNA sequencing data: hitchhiker’s guide to expression analysis. doi:10.7287/peerj.preprints.27283.

13. RNA-Seq Blog. https://rna-seqblog.com.

14. Sahraeian, S. M. E. *et al.* Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis. *Nat. Commun.* **8**, 59 (2017).

15. Robles, J. A. *et al.* Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. *BMC Genomics* **13**, 484 (2012).

16. Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biol.* **17**, 13 (2016).

17. Zhu, Y., Wang, L., Yin, Y. & Yang, E. Systematic analysis of gene expression patterns associated with postmortem interval in human tissues. *Sci. Rep.* **7**, 5435 (2017).

18. Knight, V. B. & Serrano, E. E. Expression analysis of RNA sequencing data from human neural and glial cell lines depends on technical replication and normalization methods. *BMC Bioinformatics* **19**, 412 (2018).

19. Parnell, L. D. *et al.* BioStar: an online question & answer resource for the bioinformatics community. *PLoS Comput. Biol.* **7**, e1002216 (2011).

20. Li, J.-W. *et al.* SEQanswers: an open access community for collaboratively decoding genomes. *Bioinformatics* **28**, 1272–1273 (2012).

21. Dobin, A. & Gingeras, T. R. Mapping RNA-seq Reads with STAR. *Current Protocols in Bioinformatics* 11.14.1–11.14.19 (2015) doi:10.1002/0471250953.bi1114s51.

22. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).

23. G. Jun, M. Flickinger, K. N. Hetrick, Kurt, J. M. Romm, K. F. Doheny, G. Abecasis, M. Boehnke,and H. M. Kang, *Detecting and Estimating Contamination of Human DNA Samples in Sequencing and Array-Based Genotype Data*, American journal of human genetics doi:10.1016/j.ajhg.2012.09.004 (volume 91 issue 5 pp.839 - 848).

24. Hartley SW, Mullikin JC. QoRTs: a comprehensive toolset for quality control and data processing of RNA-Seq experiments. *BMC Bioinformatics*. doi: 10.1186/s12859-015-0670-5.

25. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).

26. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).

27. Liao, Y., Smyth, G. K. & Shi, W. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Res.* **47**, e47 (2019).

28. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).

29. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).

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

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

32. Li, Y. I. *et al.* Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.* **50**, 151–158 (2018).

33. Shen, S. *et al.* rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E5593–601 (2014).

34. Vaquero-Garcia, J. *et al.* A new view of transcriptome complexity and regulation through the lens of local splicing variations. *Elife* **5**, e11752 (2016).

35. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

36. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).

37. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).

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

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

40. Picard Toolkit. *Broad Institute, GitHub repository*.

41. Wang, L., Wang, S. & Li, W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* **28**, 2184–2185 (2012).

42. Zhang, Y., Parmigiani, G. & Johnson, W. E. : batch effect adjustment for RNA-seq count data. *NAR Genom Bioinform* **2**, lqaa078 (2020).

43. Andrews, S. *et al.* FastQC. (2010).

44. Amarasinghe, S. L. *et al.* Opportunities and challenges in long-read sequencing data

analysis. *Genome Biol.* **21**, 30 (2020).

45. Wilson, G. *et al.* Good enough practices in scientific computing. *PLoS Comput. Biol.* **13**, e1005510 (2017).

46. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).

47. Mehmood, A. *et al.* Systematic evaluation of differential splicing tools for RNA-seq studies. *Brief. Bioinform.* (2019) doi:10.1093/bib/bbz126.

48. Liu, F. *et al.* Systematic comparative analysis of single-nucleotide variant detection methods from single-cell RNA sequencing data. *Genome Biology* vol. 20 (2019).

49. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).

50. Hoffman, G. E. & Schadt, E. E. variancePartition: interpreting drivers of variation in complex gene expression studies. *BMC Bioinformatics* **17**, 483 (2016).

51. Milborrow. Derived from mda:mars by T. Hastie and R. Tibshirani., S. *earth: Multivariate Adaptive Regression Splines*. http://CRAN.R-project.org/package=earth (2011).

52. Pimentel, H., Bray, N. L., Puente, S., Melsted, P. & Pachter, L. Differential analysis of RNA-seq incorporating quantification uncertainty. *Nat. Methods* **14**, 687–690 (2017).

53. Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. & R Core Team. nlme: Linear and Nonlinear Mixed Effects Models. (2020).

54. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Usinglme4. *Journal of Statistical Software* vol. 67 (2015).

55. Margolin, A. A. *et al.* ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinformatics* vol. 7 (2006).

56. Langfelder, P. & Horvath, S. WGCNA Tutorials. https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/Tutorials.

57. Califano, A. ARACNe. *Califano Laboratory of Systems Biology* http://califano.c2b2.columbia.edu/aracne.

58. Parikshak, N. N., Gandal, M. J. & Geschwind, D. H. Systems biology and gene networks in neurodevelopmental and neurodegenerative disorders. *Nat. Rev. Genet.* **16**, 441–458 (2015).

59. Sedaghat, N., Saegusa, T., Randolph, T. & Shojaie, A. Comparative study of computational methods for reconstructing genetic networks of cancer-related pathways. *Cancer Inform.* **13**, 55–66 (2014).

60. Chen, G., Ning, B. & Shi, T. Single-Cell RNA-Seq Technologies and Related Computational Data Analysis. *Front. Genet.* **10**, 317 (2019).

61. Zhou, Y. *et al.* Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat. Commun.* **10**, 1523 (2019).

62. Raudvere, U. *et al.* g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* **47**, W191–W198 (2019).

63. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* vol. 102 15545–15550 (2005).

64. Thomas, P. D. *et al.* PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res.* **13**, 2129–2141 (2003).

65. Chen, E. Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* vol. 14 128 (2013).

66. Alexa, A. & Rahnenfuhrer, J. topGO: Enrichment Analysis for Gene Ontology. doi:2020.

67. Yu, G. & He, Q.-Y. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Molecular BioSystems* vol. 12 477–479 (2016).

68. Goesmann, A., Haubrock, M., Meyer, F., Kalinowski, J. & Giegerich, R. PathFinder: reconstruction and dynamic visualization of metabolic pathways. *Bioinformatics* **18**, 124–129 (2002).

69. Wang, J., Vasaikar, S., Shi, Z., Greer, M. & Zhang, B. WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. *Nucleic Acids Res.* **45**, W130–W137 (2017).

70. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57 (2009).

71. The Gene Ontology Consortium & The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research* vol. 47 D330–D338 (2019).

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

73. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).

74. Kanehisa, M. Toward understanding the origin and evolution of cellular organisms. *Protein Sci.* **28**, 1947–1951 (2019).

75. Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M. & Tanabe, M. KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res.* (2020) doi:10.1093/nar/gkaa970.

76. Xu, X., Wells, A. B., O'Brien, D. R., Nehorai, A. & Dougherty, J. D. Cell type-specific expression analysis to identify putative cellular mechanisms for neurogenetic disorders. *J. Neurosci.* **34**, 1420–1431 (2014).

77. Skene, N. G. & Grant, S. G. N. Identification of Vulnerable Cell Types in Major Brain Disorders Using Single Cell Transcriptomes and Expression Weighted Cell Type Enrichment. *Front. Neurosci.* **10**, 16 (2016).

78. Rossin, E. J. *et al.* Proteins Encoded in Genomic Regions Associated with Immune-Mediated Disease Physically Interact and Suggest Underlying Biology. *PLoS Genetics* vol. 7 e1001273 (2011).

79. Szklarczyk, D. *et al.* STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).

80. Stark, C. *et al.* BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* **34**, D535–9 (2006).

81. Oughtred, R. *et al.* The BioGRID interaction database: 2019 update. *Nucleic Acids Res.* **47**, D529–D541 (2019).

82. Li, T. *et al.* A scored human protein-protein interaction network to catalyze genomic interpretation. *Nat. Methods* **14**, 61–64 (2017).

83. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).

84. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* **11**, e1004219 (2015).

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

86. Singh, T. *et al.* Exome sequencing identifies rare coding variants in 10 genes which confer substantial risk for schizophrenia. *medRxiv* (2020) doi:10.1101/2020.09.18.20192815.

87. Miller, J. A. *et al.* Transcriptional landscape of the prenatal human brain. *Nature* **508**, 199–206 (2014).

88. Melé, M., Ferreira, P. G., Reverter, F. & DeLuca, D. S. The human transcriptome across tissues and individuals. (2015).

89. Sunkin, S. M. *et al.* Allen Brain Atlas: an integrated spatio-temporal portal for exploring the

central nervous system. *Nucleic Acids Res.* **41**, D996–D1008 (2013).

90. Gordon, A. *et al.* Transcriptomic networks implicate neuronal energetic abnormalities in three mouse models harboring autism and schizophrenia-associated mutations. *Mol. Psychiatry* (2019) doi:10.1038/s41380-019-0576-0.

91. Lazaro, M. T. *et al.* Reduced Prefrontal Synaptic Connectivity and Disturbed Oscillatory Population Dynamics in the CNTNAP2 Model of Autism. *Cell Rep.* **27**, 2567–2578.e6 (2019).

92. Yoon, S.-J. *et al.* Reliability of human cortical organoid generation. *Nature Methods* vol. 16 75–78 (2019).

93. Camp, J. G. *et al.* Human cerebral organoids recapitulate gene expression programs of fetal neocortex development. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 15672–15677 (2015).

94. Ramaker, R. C. *et al.* Post-mortem molecular profiling of three psychiatric disorders. *Genome Med.* **9**, 72 (2017).

95. Foley, D. L. & Mackinnon, A. A systematic review of antipsychotic drug effects on human gene expression related to risk factors for cardiovascular disease. *Pharmacogenomics J.* **14**, 446–451 (2014).

96. de Bartolomeis, A. *et al.* Immediate-Early Genes Modulation by Antipsychotics: Translational Implications for a Putative Gateway to Drug-Induced Long-Term Brain Changes. *Front. Behav. Neurosci.* **11**, 240 (2017).

97. Troudet, R. *et al.* Gene expression and response prediction to amisulpride in the OPTiMiSE first episode psychoses. *Neuropsychopharmacology* **45**, 1637–1644 (2020).

98. Patrick, E. *et al.* Deconvolving the contributions of cell-type heterogeneity on cortical gene expression. *PLoS Comput. Biol.* **16**, e1008120 (2020).

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

100. Hernandez, L. M. *et al.* Transcriptomic Insight Into the Polygenic Mechanisms Underlying Psychiatric Disorders. *Biol. Psychiatry* (2020) doi:10.1016/j.biopsych.2020.06.005.

101. Choi, S. W., Mak, T. S.-H. & O'Reilly, P. F. Tutorial: a guide to performing polygenic risk score analyses. *Nat. Protoc.* **15**, 2759–2772 (2020).

102. Swarup, V., Hinz, F.I., Rexach, J.E. *et al*. Identification of evolutionarily conserved gene networks mediating neurodegenerative dementia. *Nat Med* **25,** 152–164 (2019).

103. Buse, A. The Likelihood Ratio, Wald, and Lagrange Multiplier Tests: An Expository Note. *The American Statistician* **36(3)**, 153-157 (1982).

104. Mostafavi S, Battle A, Zhu X, et al. Normalizing RNA-sequencing data by modeling hidden covariates with prior knowledge. *PLoS One* **8(7)**, (2013).

105. Noble WS. How does multiple testing correction work?. *Nat Biotechnol* **27(12)**, 1135-1137 (2009).

106. Horvath S, Zhang B, Carlson M, Lu KV, Zhu S, Felciano RM, Laurance MF, Zhao W, Qi S, Chen Z, Lee Y, Scheck AC, Liau LM, Wu H, Geschwind DH, Febbo PG, Kornblum HI, Cloughesy TF, Nelson SF, Mischel PS. Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target. *Proc Natl Acad Sci U S A* **103(46)**, 17402 - 17407 (2006).

107. Zhang B, Gaiteri C, Bodea LG, Wang Z, McElwee J, Podtelezhnikov AA, Zhang C, Xie T, Tran L, Dobrin R, Fluder E, Clurman B, Melquist S, Narayanan M, Suver C, Shah H, Mahajan M, Gillis T, Mysore J, MacDonald ME, Lamb JR, Bennett DA, Molony C, Stone DJ, Gudnason V, Myers AJ, Schadt EE, Neumann H, Zhu J, Emilsson V. Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. *Cell* **153(3)**, 707-20 (2013).

108. Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang and Can Candan. (2016). caret: Classification and Regression Training. R package version 6.0-71. https://CRAN.R-project.org/package=caret.

109. Song WM, Zhang B. Multiscale Embedded Gene Co-expression Network Analysis. *PLOS Computational Biology* **11(11)**, (2015).

CHAPTER TWO

Shared molecular neuropathology across psychiatric disorders

2.1: Contributing authors

All of the work presented in this chapter was published in *Science* in February of 2018, in a research article entitled “Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap” (volume 359, pages 693–697). Michael Gandal was the primary author of this work and completed all major analyses, such as the expression microarray mega-analysis. I, as the second author, conducted the RNA-seq validation analyses, calculated partitioned heritability scores for the identified gene co-expression modules, and generally supported all analyses and interpretation of results. Other co-authors included Neelroop N. Parikshak, Virpi Leppa, Gokul Ramaswami, Chris Hartl, Andrew J. Schork, Vivek Appadurai, Alfonso Buil, Thomas M. Werge, Chunyu Liu, Kevin P. White, and Steve Horvath. They all contributed to supporting analyses and interpretation of results, providing essential contributions for the publication. Daniel Geschwind was the senior author and project director. This work was associated with the PsychENCODE Consortium and iPSYCH-BROAD Working Group.

2.2: Experimental rationale and overlapping psychiatric cortical gene expression

Despite remarkable success identifying genetic risk factors for major psychiatric disorders, it remains unknown how genetic variants interact with environmental and epigenetic risk factors in the brain to impart risk for clinically distinct disorders (*1*, *2*). We reasoned that brain transcriptomes, a quantitative, genome-wide molecular phenotype (*3*), would allow us to determine whether disease-related signatures are shared across major neuropsychiatric disorders with distinct symptoms and whether these patterns reflect genetic risk.

We first analyzed published gene-expression microarray studies of cerebral cortex across five major neuropsychiatric disorders (*3-11*) using 700 cerebral cortical samples from subjects with ASD (n=50 samples), SCZ (n=159), BD (n= 94), MDD (n=87), AAD (n=17), and matched controls (n=293) (*12*). These disorders are prevalent and disabling, contributing substantially to global disease burden. Inflammatory bowel disease (IBD, n=197) was included as a non-neural comparison.

Individual datasets underwent stringent quality control and normalization (**Fig. 2.1**; (*12*)), including re-balancing to alleviate confounding between diagnosis and biological (e.g., age, sex) or technical (e.g., post-mortem interval, pH, RIN, batch, 3' bias) covariates (**Figs. A1.1, A1.2**). Transcriptome summary statistics for each disorder were computed with a linear mixed-effects model to account for any sample overlap across studies (*12*). Comparison of differential gene expression (DGE) $\log_2$ fold change ($\log_2$FC) signatures revealed a significant overlap among ASD, SCZ, and BD and SCZ, BD, and MDD (**Fig. 2.2A**; all Spearman's $r \geq 0.23$, $P < 0.05$, 40,000 permutations). The regression slopes between ASD, BD, and MDD $\log_2$-FC effect sizes compared to SCZ (5.08, 0.99, and 0.37) indicate a gradient of transcriptomic severity with ASD > SCZ ≈ BD > MDD (**Fig. 2.2B**). To ensure robustness, we compared multiple methods for batch correction, probe summarization, and feature selection, including use of integrative correlations, none of which changed the qualitative observations (**Fig. A1.3;** (*12*)). Results were also unaltered after first regressing gene-level RNA degradation metrics, suggesting that systematic sample quality issues were unlikely to drive these correlations (**Fig. A1.3**). Further, the lack of (or negative) overlap between AAD and other disorders suggests that similarities are less likely due to comorbid substance abuse, poor overall general health, or general brain-related post-mortem artefacts.

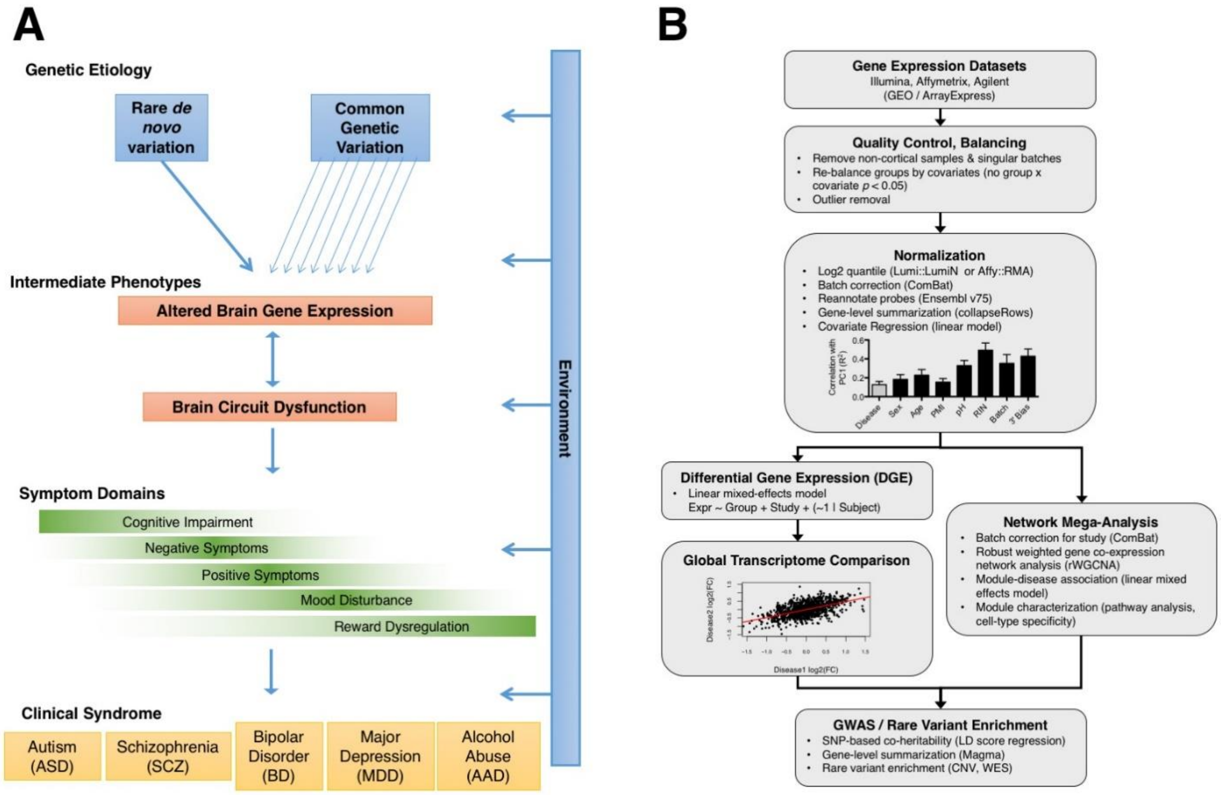

A
Genetic Etiology
Rare de novo variation
Common Genetic Variation
Intermediate Phenotypes
Altered Brain Gene Expression
Brain Circuit Dysfunction
Symptom Domains
Cognitive Impairment
Negative Symptoms
Positive Symptoms
Mood Disturbance
Reward Dysregulation
Clinical Syndrome
Autism (ASD)
Schizophrenia (SCZ)
Bipolar Disorder (BD)
Major Depression (MDD)
Alcohol Abuse (AAD)
Environment
B
Gene Expression Datasets
Illumina, Affymetrix, Agilent
(GEO / ArrayExpress)
Quality Control, Balancing
Remove non-cortical samples & singular batches
Re-balance groups by covariates (no group x covariate p < 0.05)
Outlier removal
Normalization
Log2 quantile (Lumi::LumiN or Affy::RMA)
Batch correction (ComBat)
Reannotate probes (Ensembl v75)
Gene-level summarization (collapseRows)
Covariate Regression (linear model)
Correlation with PC1 (R²)
0.0
0.2
0.4
0.6
Disease
Sex
Age
PMI
pH
RIN
Batch
3' Bias
Differential Gene Expression (DGE)
Linear mixed-effects model
Expr ~ Group + Study + (~1 | Subject)
Global Transcriptome Comparison
Disease2 log2(FC)
Disease1 log2(FC)
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
Network Mega-Analysis
Batch correction for study (ComBat)
Robust weighted gene co-expression network analysis (rWGCNA)
Module-disease association (linear mixed effects model)
Module characterization (pathway analysis, cell-type specificity)
GWAS / Rare Variant Enrichment
SNP-based co-heritability (LD score regression)
Gene-level summarization (Magma)
Rare variant enrichment (CNV, WES)

Fig. 2.1: Experimental rationale and design. (**A**) Model of psychiatric disease pathogenesis. (**B**) Flowchart of the cross-disorder transcriptome analysis pipeline (*12*). Cortical gene expression datasets were compiled from cases of ASD (n=50 samples), SCZ (n=159), BD (n=94), MDD (n=87), AAD (n=17), and matched non-psychiatric controls (n=293) (*12*) (see **Table A1.1**).

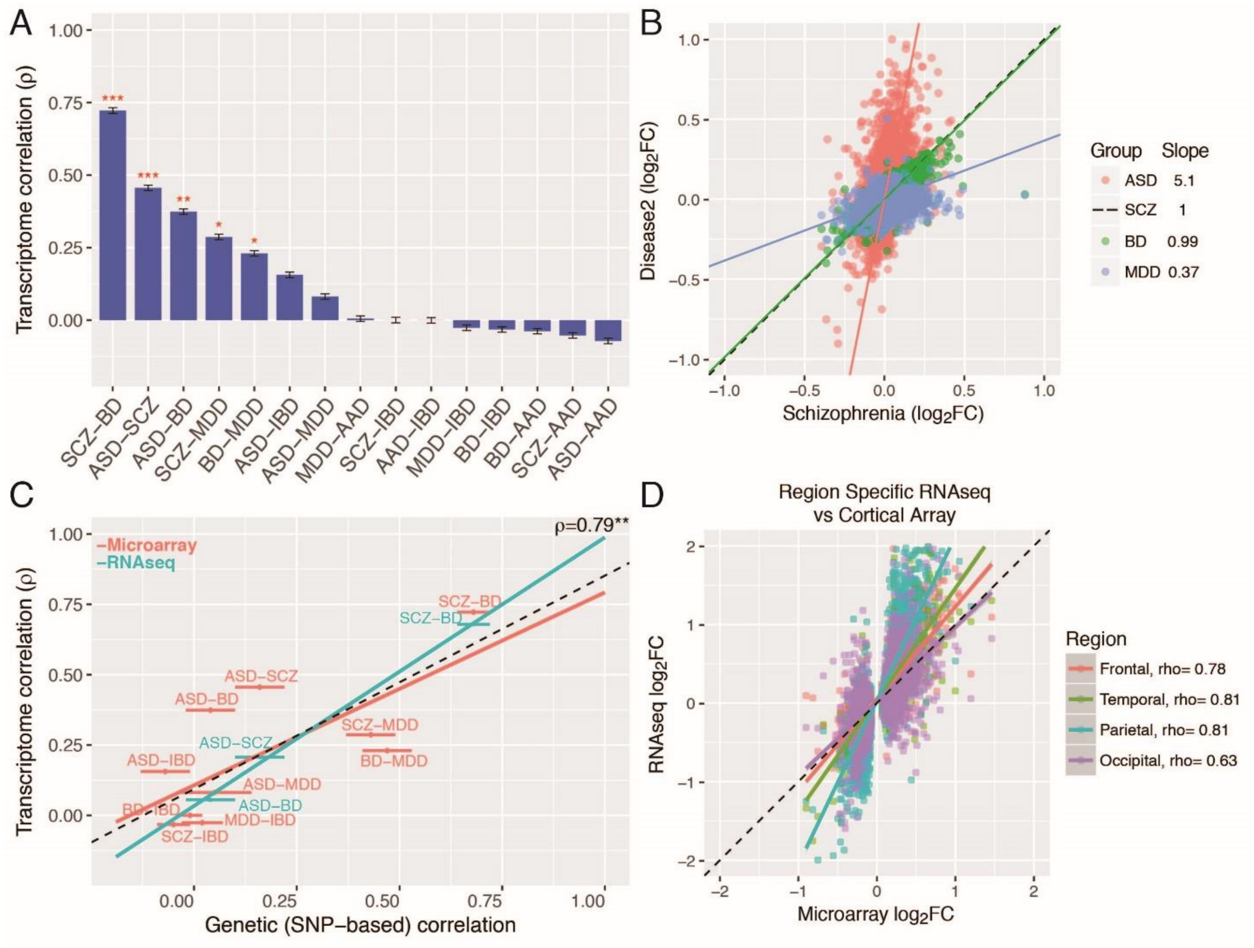


Fig. 2.2: Cortical gene expression patterns overlap. Cortical gene expression patterns overlap. (**A**) Rank order of microarray transcriptome similarity for all disease pairs, as measured by Spearman's correlation of differential expression $\log_2$FC values. (**B**) Comparison of the slopes among significantly associated disease pairs indicates a gradient of transcriptomic severity, with ASD > SCZ ~ BD > MDD. (**C**) Overlapping gene expression patterns across diseases are correlated with shared common genetic variation, as measured by SNP co-heritability (*22*). The Y-axis shows transcriptome correlations using microarray-based (discovery, red) and RNAseq (replication, blue) datasets. (**D**) RNAseq across all cortical lobes in ASD replicates microarray results and demonstrates a consistent transcriptomic pattern. Spearman's □ is shown for comparison between microarray and region-specific RNAseq replication datasets (all P's < 10^{-14}). Plots show mean +/- SEM. *P < 0.05, **P < 0.01, ****P* < 0.001.

Disease-specific DGE summary statistics (**Data Table A1.1**) provide human *in vivo* benchmarks for determining the relevance of model organisms, *in vitro* systems, or drug effects (*13*, *14*). We identified a set of concordantly down- and upregulated genes across disorders (**Fig. A1.4**), as well as those with more specific effects. Complement component 4A (*C4A*), the top GWAS-implicated SCZ disease gene (*15*), was significantly upregulated in SCZ ($\log_2$FC=0.23, P=6.9x10^{-6}) and in ASD (RNAseq; $\log_2$FC=0.91, P=0.014; **Data Table A1.1**) but not BD, MDD, or AAD. To investigate potential confounding by psychiatric medications, we compared disease signatures with those from non-human primates treated with acute or chronic dosing of antipsychotic medications. Significant negative overlap (**Fig. A1.5;** (*12*)) was observed, indicating that antipsychotics are unlikely to drive, but rather may partially normalize, these transcriptomic alterations, whereas the psychotomimetic PCP partially recapitulates disease signatures.

To validate that these transcriptomic relationships are generalizable, we generated independent RNAseq datasets for replication for 3 out of the 5 disorders (**Fig. A1.6;** (*12*)). We identify 1099 genes whose DGE is replicated in ASD (OR 6.4, P=3.3x10^{-236}, Fisher's exact test; **Table A1.2**), 890 genes for SCZ (BrainGVEX; OR 4.5, P=7.6x10^{-155}), and 112 genes for BD (BrainGVEX; OR 3.9, P=4.6x10^{-26}), which is likely due to the relatively smaller RNAseq sample size for BD (*12*). We observed similarly high levels of transcriptomic overlap among ASD, SCZ, and BD, and a similar gradient of transcriptomic severity (**Figs. 2.2C; A1.7**). The SCZ and BD patterns were further replicated in the CommonMind dataset, although gene-level overlap was lower (*12*, *16*) (**Fig. A1.7**). The ASD signature was largely consistent across the four major cortical lobules, indicating that this pattern is not caused by regional differences (**Fig. 2.2D**).

2.3: Gene network analysis maps psychiatric gene expression to biological systems

To more specifically characterize the biological pathways involved, we performed robust weighted gene co-expression network analysis (rWGCNA; (*12*, *17*)), identifying several shared and disorder-specific co-expression modules (**Fig. 2.3**). Modules were stable (**Fig. A1.8)**, showed greater association with disease than other biological or technical covariates (**Fig. A1.9**), and were not dependent on corrections for covariates or batch effects (**Fig. A1.10**). Moreover, each module was enriched for protein-protein interactions (**Fig. A1.8**) and brain enhancer-RNA co-regulation (**Fig. A1.11**) derived from independent data, which provides anchors for dissecting protein complexes and regulatory relationships.

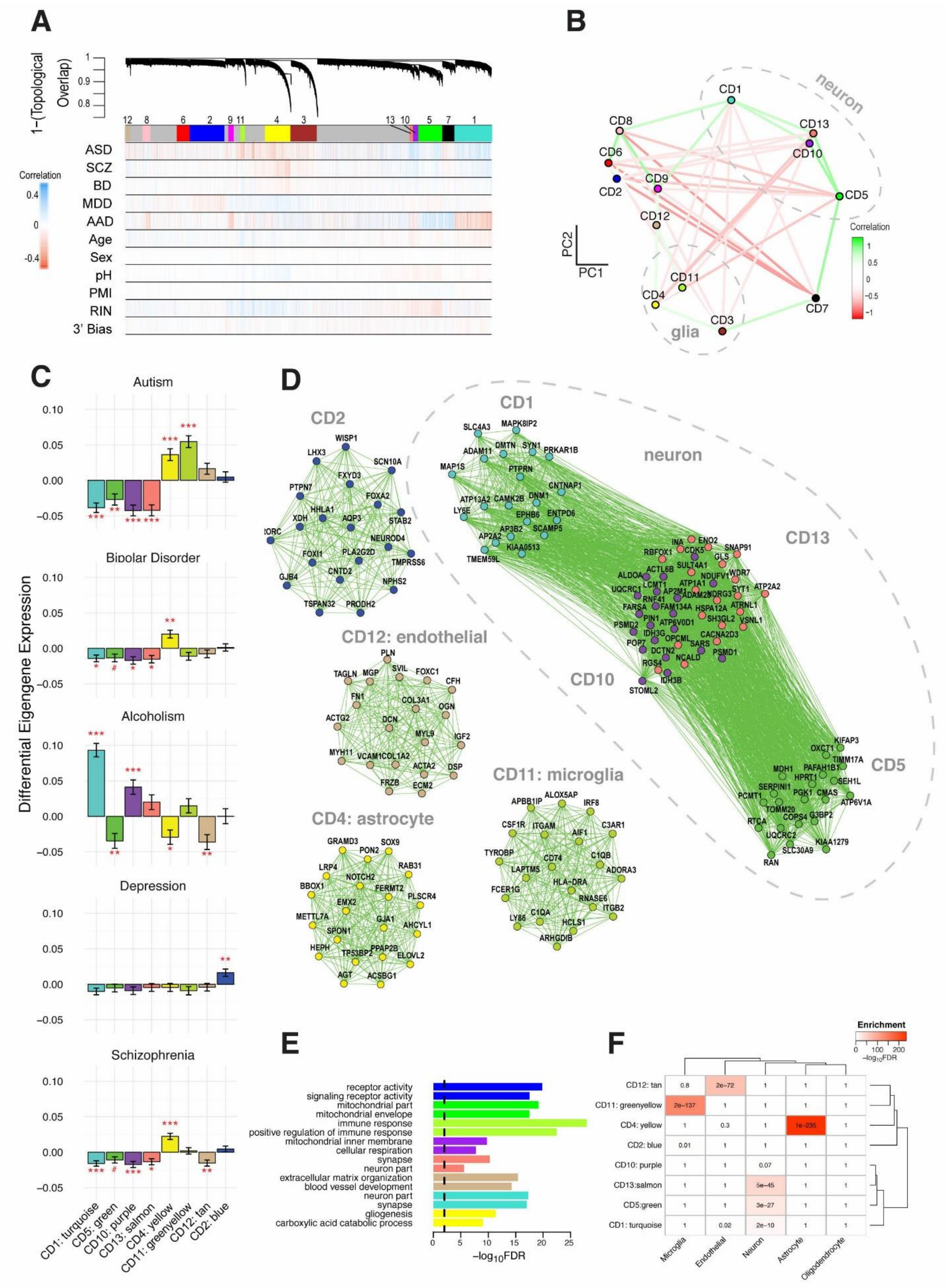

A
1-(Topological Overlap)
1
0.9
0.8
12 8 6 2 9 11 4 3 13 10 5 7 1
ASD
SCZ
BD
MDD
AAD
Age
Sex
pH
PMI
RIN
3' Bias
Correlation
0.4
0
-0.4
B
neuron
glia
CD1
CD8
CD13
CD10
CD6
CD9
CD2
CD5
CD12
CD11
CD4
CD3
CD7
PC2
PC1
Correlation
1
0.5
0
−0.5
−1
C
Differential Eigengene Expression
Autism
Bipolar Disorder
Alcoholism
Depression
Schizophrenia
0.10
0.05
0.00
−0.05
***
**
#
*
CD1: turquoise
CD5: green
CD10: purple
CD13: salmon
CD4: yellow
CD11: greenyellow
CD12: tan
CD2: blue
D
CD2
WISP1
LHX3
SCN10A
FXYD3
PTPN7
FOXA2
HHLA1
XDH
AQP3
STAB2
RORC
NEUROD4
FOXI1
PLA2G2D
TMPRSS6
GJB4
CNTD2
NPHS2
TSPAN32
PRODH2
CD1
neuron
SLC4A3
MAPK8IP2
ADAM11
DMTN
SYN1
PRKAR1B
MAP1S
PTPRN
CNTNAP1
ATP13A2
CAMK2B
DNM1
LY6E
EPHB6
ENTPD6
AP3B2
SCAMP5
AP2A2
KIAA0513
TMEM59L
CD13
INA
ENO2
SNAP91
RBFOX1
CDK5
GLS
SULT4A1
WDR7
ALDOA
ACTL6B
NDUFV1
ATP2A2
UQCRC1
LCMT1
ATP1A1
SYT1
RNF41
AP2M1
ADAM23
NDRG3
ATRNL1
FARSA
FAM134A
HSPA12A
PSMD2
PIN1
ATP6V0D1
SH3GL2
VSNL1
IDH3G
OPCML
CACNA2D3
POP7
DCTN2
SARS
RGS4
NCALD
PSMD1
STOML2
IDH3B
CD10
CD5
KIFAP3
OXCT1
TIMM17A
MDH1
PAFAH1B1
HPRT1
SEH1L
SERPINI1
PGK1
CMAS
PCMT1
ATP6V1A
TOMM20
COPS4
G3BP2
RTCA
UQCRC2
KIAA1279
SLC30A9
RAN
CD12: endothelial
PLN
TAGLN
MGP
SVIL
FOXC1
CFH
FN1
COL3A1
OGN
ACTG2
DCN
MYL9
IGF2
MYH11
VCAM1
COL1A2
ACTA2
DSP
FRZB
ECM2
CD11: microglia
APBB1IP
ALOX5AP
IRF8
CSF1R
ITGAM
C3AR1
AIF1
TYROBP
C1QB
CD74
ADORA3
LAPTM5
HLA-DRA
FCER1G
RNASE6
ITGB2
C1QA
LY86
HCLS1
ARHGDIB
CD4: astrocyte
GRAMD3
SOX9
PON2
LRP4
RAB31
NOTCH2
BBOX1
FERMT2
PLSCR4
EMX2
METTL7A
GJA1
AHCYL1
SPON1
HEPH
TP53BP2
PPAP2B
ELOVL2
AGT
ACSBG1
E
receptor activity
signaling receptor activity
mitochondrial part
mitochondrial envelope
immune response
positive regulation of immune response
mitochondrial inner membrane
cellular respiration
synapse
neuron part
extracellular matrix organization
blood vessel development
neuron part
synapse
gliogenesis
carboxylic acid catabolic process
0
5
10
15
20
25
−log10FDR
F
Enrichment
0
100
200
−log10FDR
CD12: tan
CD11: greenyellow
CD4: yellow
CD2: blue
CD10: purple
CD13:salmon
CD5:green
CD1: turquoise
0.8
2e−72
2e−137
1e−235
0.3
0.01
0.07
5e−45
3e−27
0.02
2e−10
1
Microglia
Endothelial
Neuron
Astrocyte
Oligodendrocyte

Fig. 2.3. Network analysis identifies modules of co-expressed genes across disease. (**A**) Network dendrogram from co-expression topological overlap of genes across disorders. Color bars show correlation of gene expression with disease status, biological, and technical covariates. (**B**) Multidimensional scaling plot demonstrates relationship between modules and clustering by cell-type - relationship. (**C**) Module-level differential expression is perturbed across disease states. Plots show beta values from linear mixed-effect model of module eigengene association with disease status (FDR-corrected #P<0.1, *P<0.05, **P<0.01, ***P<0.001). **D)** The top twenty hub genes are plotted for modules most disrupted in disease. See **Data Table A1.2** for a complete list of genes' module membership (kME). Edges are weighted by the strength of correlation between genes. Modules are characterized by (**E)** Gene Ontology enrichment (top two pathways shown for each module) and (**F)** cell-type specificity, on the basis of RNAseq of purified cell populations from healthy human brain samples (*25*).

An astrocyte-specific module (CD4, hubs *GJA1*, *SOX9*) was broadly upregulated in ASD, BD, and SCZ (FDR-corrected P's < 0.05, **Fig. 2.3C**, **Data Table A1.2;** (*12*)) and enriched for glial cell differentiation and fatty-acid metabolism pathways. In contrast, a module strongly enriched for microglial markers (CD11) was upregulated specifically in ASD (two-sided t-test, FDR-corrected P=4×10^{-9}). Hubs include canonical microglial markers (*HLA-DRA, AIF1),* major components of the complement system (*C1QA, C1QB*) and *TYROBP*, a microglial signalling adapter protein (*18*). Results fit with convergent evidence for microglial upregulation in ASD and an emerging understanding that microglia play a critical role regulating synaptic function during neurodevelopment (*19*).

One module was upregulated specifically in MDD (CD2, FDR-corrected P=0.009; **Data Table A1.2**) and was enriched for G-protein coupled receptors, cytokine-cytokine interactions, and hormone activity pathways, suggesting a link between inflammation and dysregulation of the hypothalamic-pituitary (HPA) axis, consistent with current models of MDD pathophysiology (*20*). Several modules annotated as neuronal/mitochondrial were downregulated across ASD, SCZ, and BD (CD1, CD10, CD13; **Fig. 2.3C**, **Data Table A1.2;** (*12*)). The overlap of CD10 with a mitochondrial gene-enriched module previously associated with neuronal firing rate (*21*) links energetic balance, synaptic transmission, and psychiatric disease (**Data Table A1.2)**.

2.4: Down-regulated neuronal modules are enriched for psychiatric genetic risk factors

The transcriptome may reflect the cause or the consequence of a disorder. To refine potential causal links, we compared SNP-based genetic correlations between disease pairs (*22*) with their corresponding transcriptome overlap. SNP co-heritability was significantly correlated with transcriptome overlap across the same disease pairs (**Fig. 2.2C,** Spearman's r=0.79, 95% confidence interval [0.43–0.93], P=0.0013), suggesting that a major component of these gene-expression patterns reflects biological processes coupled to underlying genetic variation.

To determine how disease-associated variants may influence specific biological processes, we investigated whether any modules harbor genetic susceptibility for specific disorders or for relevant cognitive or behavioral traits (*12*). We identified significant enrichment among several of the downregulated, neuronal co-expression modules (CD1, CD10, CD13) for GWAS signal from SCZ and BD, as well as for educational attainment and neuroticism (**Fig. 2.4A;** FDR-corrected P's < 0.05, Spearman; (*12*)). We also observe enrichment for the three downregulated neuronal co-expression modules in the iPSYCH Consortium (*23*) ASD GWAS cohort (**Fig 2.4A**; **Table A1.3**; (*12*)). In contrast, these modules showed no enrichment for MDD, AAD, or IBD. Further, none of the microglial- or astrocyte-specific modules showed psychiatric GWAS enrichment. Extending this analysis to disease-associated rare variants (**Data Table A1.3;** (*2*, *12*)), we found that the CD1 neuronal module was enriched for genes harbouring rare, non-synonymous *de novo* mutations identified in ASD (OR 1.36, FDR-corrected P=0.03, logistic regression) and SCZ cases (OR 1.82, FDR-corrected P=0.014) but not unaffected controls (**Fig. 2.4B**). A similar CD1-enrichment was observed for genes affected by rare, recurrent copy-number variation (CNV) in ASD (OR 2.52, FDR-corrected P=0.008) and SCZ (OR 2.46, FDR-corrected P=0.014). These results suggest convergence of common and rare genetic variation acting to downregulate synaptic function in ASD and SCZ.

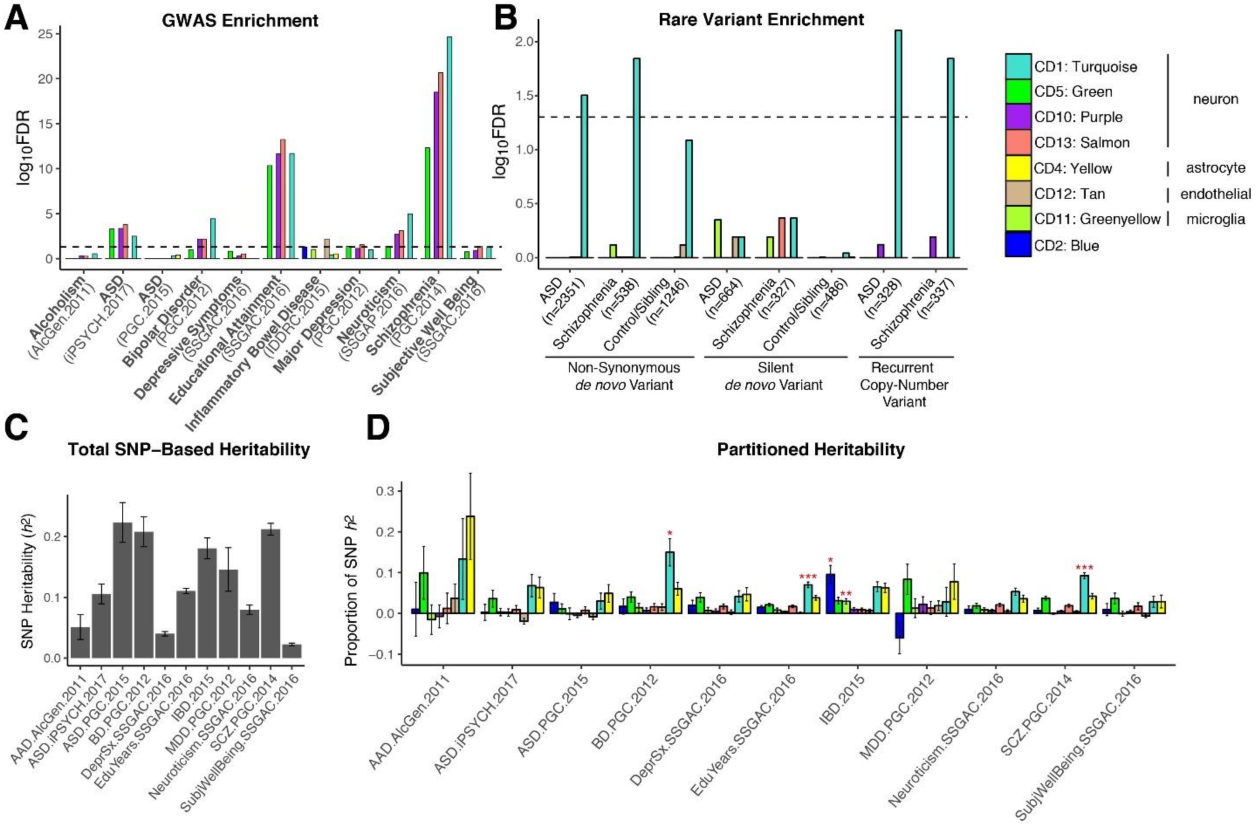


Fig. 2.4. Downregulated neuronal modules are enriched for common and rare genetic risk factors. (**A**) Significant enrichment is observed for SCZ-, ASD-, and BD-associated common variants from GWAS among neuron/synapse & mitochondrial modules (*12*). GWAS datasets are listed in **Table A1.3**. (**B**) The CD1 neuronal module shows significant enrichment for ASD- and SCZ-associated non-synonymous *de novo* variants from whole exome sequencing. The number of genes affected by different classes of rare variants is shown in parentheses. Significance was calculated using logistic regression, correcting for gene length. P-values are FDR corrected. (**C**) Total SNP-based heritability (liability scale for psychiatric diagnoses) calculated from GWAS using LD-score regression. (**D**) Proportion of heritability for each disorder or trait that can be attributed to individual co-expression modules. Significance (FDR-corrected *P<0.05, **P<0.01, ***P<0.001) is from enrichment statistics comparing the proportion of SNP heritability within the module divided by the proportion of total SNPs represented. The CD1 module shows significant enrichment in SCZ, BD, and educational attainment.

We next used LD score regression (*24*) to partition GWAS heritability (**Fig. 2.4C**; **Data Table A1.4**) into the contribution from SNPs located within genes from each module ((*12*); **Fig. 2.4D**). CD1 again showed significant enrichment for SCZ (2.5 fold, FDR-corrected P=8.9x10^{-11}), BD (3.9 fold, FDR-corrected P<0.014), and educational attainment (1.9 fold, FDR-corrected P<0.0008; χ^2) GWAS, accounting for ~10% of SNP-based heritability within each dataset, despite containing only 3% of the SNPs. This illustrates how gene network analysis can begin to parse complex patterns of common variants, each of small effect size, to implicate specific biological roles for common variant risk across neuropsychiatric disorders.

2.5: Discussion

These data provide a quantitative, genome-wide characterization of the cortical pathology across five major neuropsychiatric disorders, providing a framework for identifying the responsible molecular signalling pathways and interpreting genetic variants implicated in neuropsychiatric disease risk. We observe a gradient of synaptic gene down-regulation, with ASD > SZ ≈ BD. BD and SCZ appear most similar in terms of synaptic dysfunction and astroglial gene up-regulation, which may represent astrocytosis, activation, or both. ASD, an early-onset disorder, shows a distinct upregulated microglial signature, which may reflect the role for microglia in regulation of synaptic connectivity during neurodevelopment (*19*). MDD shows neither the synaptic nor astroglial pathology, but does exhibit dysregulation of HPA-axis and hormonal signalling not observed in the other disorders.

Our data suggest that shared genetic factors underlie a substantial proportion of cross disorder expression overlap. Given that a minority of these relationships represent eQTL (**Fig. A1.12**), most of the genetic effects are likely acting indirectly, through a cascade of developmental and cell-cell signalling events rooted in genetic risk. Genetic variation is also not the only driver of expression variation; there is undoubtedly a contribution from environmental effects. Hidden confounders could introduce a correlation structure that matches SNP-level genetic correlations, but parsimony and hidden covariate correction suggests that this is unlikely. Diagnostic misclassification could artificially elevate shared signals, but the results are robust to disorder removal (**Fig. A1.13)**, and misclassification would not account for the substantial overlap we observe with ASD, which has a highly distinct phenotypic trajectory from later onset disorders. Finally, we have replicated broad transcriptomic and cell-type specific patterns independently for ASD, SCZ and BD, providing an organizing pathological framework for future investigation of the mechanisms underlying specific gene and isoform-level transcriptomic alterations in psychiatric disease.

2.6: Materials and Methods

Please see the Appendix (section A1.1) for all materials and methods.

2.7: Bibliography

1. D. H. Geschwind, J. Flint, Genetics and genomics of psychiatric disease. *Science*. **349**, 1489–1494 (2015).

2. M. J. Gandal, V. Leppä, H. Won, N. N. Parikshak, D. H. Geschwind, The road to precision psychiatry: translating genetics into disease mechanisms. *Nat. Neurosci*. **19**, 1397–1407 (2016).

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

4. K. Garbett *et al.*, Immune transcriptome alterations in the temporal cortex of subjects with autism. *Neurobiol. Dis.* **30**, 303–311 (2008).

5. M. L. Chow *et al.*, Age-Dependent Brain Gene Expression and Copy Number Anomalies in Autism Suggest Distinct Pathological Processes at Young Versus Mature Ages. *PLoS Genet*. **8** (2012).

6. C. Chen *et al.*, Two gene co-expression modules differentiate psychotics and controls. *Mol. Psychiatry*. **18**, 1308–1314 (2012).

7. P. R. Maycox *et al.*, Analysis of gene expression in two large schizophrenia cohorts identifies multiple changes associated with nerve terminal function. *Mol. Psychiatry*. **14**, 1083–1094 (2009).

8. K. Iwamoto, M. Bundo, T. Kato, Altered expression of mitochondria-related genes in postmortem brains of patients with bipolar disorder or schizophrenia, as revealed by large-scale DNA microarray analysis. *Hum. Mol. Genet.* **14**, 241–253 (2005).

9. S. Narayan *et al.*, Molecular profiles of schizophrenia in the CNS at different stages of illness. *Brain Res.* **1239**, 235–248 (2008).

10. L.-C. Chang *et al.*, A Conserved BDNF, Glutamate- and GABA-Enriched Gene Module Related to Human Depression Identified by Coexpression Meta-Analysis and DNA Variant Genome-Wide Association Studies. *PLoS ONE*. **9**, (2014).

11. I. Ponomarev, S. Wang, L. Zhang, R. A. Harris, R. D. Mayfield, Gene Coexpression Networks in Human Brain Identify Epigenetic Modifications in Alcohol Dependence. *J. Neurosci*. **32**, 1884–1897 (2012).

12. See supplemental materials.

13. D. Prilutsky *et al.*, iPSC-derived neurons as a higher-throughput readout for autism: promises and pitfalls. *Trends Mol. Med*. **20**, 91–104 (2014).

14. J. Lamb *et al.*, The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*. **313**, 1929–1935 (2006).

15. A. Sekar *et al.*, Schizophrenia risk from complex variation of complement component 4. *Nature*. **530**, 177–183 (2016).

16. M. Fromer *et al.*, Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat. Neurosci*. **19**, 1442–1453 (2016).

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

18. B. Zhang *et al.*, Integrated Systems Approach Identifies Genetic Nodes and Networks in Late-Onset Alzheimer's Disease. *Cell*. **153**, 707–720 (2013).

19. M. W. Salter, B. Stevens, Microglia emerge as central players in brain disease. *Nat. Med*. **23**, 1018–1027 (2017).

20. P. W. Gold, The organization of the stress system and its dysregulation in depressive illness. *Mol. Psychiatry*. **20**, 32–47 (2014).

21. K. D. Winden *et al.*, The organization of the transcriptional network in specific neuronal classes. *Mol. Syst. Biol.* **5**, 291 (2009).

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

23. Pedersen, C., Bybjerg-Grauholm, J., Pedersen, M. *et al.* The iPSYCH2012 case–cohort sample: new directions for unravelling genetic and environmental architectures of severe mental disorders. *Mol Psychiatry* **23,** 6–14 (2018).

24. H. K. Finucane *et al.*, Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet*. **47**, 1228–1235 (2015).

25. Y. Zhang *et al.*, Purification and Characterization of Progenitor and Mature Human Astrocytes Reveals Transcriptional and Functional Differences with Mouse. *Neuron*. **89**, 37–53 (2016).

26. P. Du, W. A. Kibbe, S. M. Lin, lumi: a pipeline for processing Illumina microarray. *Bioinformatics*. **24**, 1547–1548 (2008).

27. L. Gautier, L. Cope, B. M. Bolstad, R. A. Irizarry, affy--analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*. **20**, 307–315 (2004).

28. G. K. Smyth, Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol*. **3**, Article3 (2004).

29. M. C. Oldham, P. Langfelder, S. Horvath, Network methods for describing sample relationships in genomic datasets: application to Huntington's disease. *BMC Syst. Biol*. **6**, 63 (2012).

30. W. E. Johnson, C. Li, A. Rabinovic, Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. **8**, 118–127 (2006).

31. S. Durinck, P. T. Spellman, E. Birney, W. Huber, Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc*. **4**, 1184–1191 (2009).

32. J. A. Miller *et al.*, Strategies for aggregating gene expression data: The collapseRows R function. *BMC Bioinformatics*. **12**, 322 (2011).

33. P. Langfelder, B. Zhang, S. Horvath, Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics*. **24**, 719–720 (2008).

34. N. N. Parikshak *et al.*, Integrative Functional Genomic Analyses Implicate Specific Molecular Pathways and Circuits in Autism. *Cell*. **155**, 1008–1021 (2013).

35. M. J. Mason, G. Fan, K. Plath, Q. Zhou, S. Horvath, Signed weighted gene co-expression network analysis of transcriptional regulation in murine embryonic stem cells. *BMC Genomics*. **10**, 327 (2009).

36. A. C. Zambon *et al.*, GO-Elite: a flexible solution for pathway and ontology over-representation. *Bioinformatics*. **28**, 2209–2210 (2012).

37. J. Reimand, T. Arak, J. Vilo, g:Profiler--a web server for functional interpretation of gene lists (2011 update). *Nucleic Acids Res.* **39**, W307–W315 (2011).

38. J. D. Dougherty, E. F. Schmidt, M. Nakajima, N. Heintz, Analytical approaches to RNA profiling data for the identification of genes enriched in specific cells. *Nucleic Acids Res.* **38**, 4218–4230 (2010).

39. X. Xu, A. B. Wells, D. R. O'Brien, A. Nehorai, J. D. Dougherty, Cell type-specific expression analysis to identify putative cellular mechanisms for neurogenetic disorders. *J. Neurosci.* **34**, 1420–1431 (2014).

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

41. Psychiatric GWAS Consortium Bipolar Disorder Working Group, Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nat. Genet*. **43**, 977–983 (2011).

42. S. Ripke *et al.*, A mega-analysis of genome-wide association studies for major depressive disorder. *Mol. Psychiatry*. **18**, 497–511 (2012).

43. Autism Spectrum Disorders Working Group of The Psychiatric Genomics Consortium, Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24.32 and a significant overlap with schizophrenia. *Mol. Autism* **8**, 21 (2017).

44. G. Schumann *et al.*, Genome-wide association and genetic functional studies identify autism susceptibility candidate 2 gene (AUTS2) in the regulation of alcohol consumption. *Proc. Natl. Acad. Sci. U.S.A*. **108**, 7119–7124 (2011).

45. J. Z. Liu *et al.*, Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet*. **47**, 979–986 (2015).

46. A. Okbay *et al.*, Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nat. Genet*. **48**, 624–633 (2016).

47. A. Okbay *et al.*, Genome-wide association study identifies 74 loci associated with educational attainment. *Nature*. **533**, 539–542 (2016).

48. C. A. de Leeuw, J. M. Mooij, T. Heskes, D. Posthuma, MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLoS Comput. Biol*. **11**, (2015).

49. P. Munk-Jørgensen, P. B. Mortensen, The Danish Psychiatric Central Register. *Dan. Med. Bull*. **44**, 82–84 (1997).

50. C. B. Pedersen *et al.*, A comprehensive nationwide study of the incidence rate and lifetime risk for treated mental disorders. *JAMA Psychiatry*. **71**, 573–581 (2014).

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

52. B. Howie, J. Marchini, M. Stephens, Genotype imputation with thousands of genomes. *G3 (Bethesda)*. **1**, 457–470 (2011).

53. N. Patterson, A. L. Price, D. Reich, Population structure and eigenanalysis. *PLoS Genet*. **2**, (2006).

54. A. Manichaikul *et al.*, Robust relationship inference in genome-wide association studies. *Bioinformatics*. **26**, 2867–2873 (2010).

55. C. C. Chang *et al.*, Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*. **4**, 7 (2015).

56. J. de Ligt *et al.*, Diagnostic Exome Sequencing in Persons with Severe Intellectual Disability. *N. Engl. J. Med.* **367**, 1921–1929 (2012).

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

58. M. Fromer *et al.*, De novo mutations in schizophrenia implicate synaptic networks. *Nature*. **506**, 179–184 (2014).

59. S. L. Girard *et al.*, Increased exonic de novo mutation rate in individuals with schizophrenia. *Nat Genet*. **43**, 860–863 (2011).

60. S. Gulsuner *et al.*, Spatial and Temporal Mapping of De Novo Mutations in Schizophrenia

to a Fetal Prefrontal Cortical Network. *Cell*. **154**, 518–529 (2013).

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

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

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

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

65. A. Rauch *et al.*, Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet*. **380**, 1674–1682 (2012).

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

67. B. Xu *et al.*, De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia. *Nat. Genet*. **44**, 1365–1369 (2012).

68. G. Mi, Y. Di, S. Emerson, J. S. Cumbie, J. H. Chang, Length Bias Correction in Gene Ontology Enrichment Analysis Using Logistic Regression. *PLoS ONE*. **7**, (2012).

69. D. Moreno-De-Luca *et al.*, Using large clinical data sets to infer pathogenicity for rare copy number variants in autism cohorts. *Mol. Psychiatry*. **18**, 1090–1095 (2012).

70. S. J. Sanders *et al.*, Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron*. **87**, 1215–1233 (2015).

71. B. K. Bulik-Sullivan *et al.*, LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet*. **47**, 291–295 (2015).

72. S. Akbarian et al. *et al.*, The PsychENCODE project. *Nat. Neurosci*. **18**, 1707–1712 (2015).

73. K. D. Hansen, R. A. Irizarry, Z. WU, Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics*. **13**, 204–216 (2012).

74. A. E. Jaffe *et al.*, qSVA framework for RNA quality correction in differential expression analysis. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 7130–7135 (2017).

75. V. Reinhart *et al.*, Evaluation of TrkB and BDNF transcripts in prefrontal cortex, hippocampus, and striatum from subjects with schizophrenia, bipolar disorder, and major depressive disorder. *Neurobiol. Dis.* **77**, 220–227 (2015).

76. L. Cope, X. Zhong, E. Garrett, G. Parmigiani, MergeMaid: R tools for merging and cross-study validation of gene expression data. *Stat. Appl. Genet. Mol. Biol*. **3**, Article29–13 (2004).

77. G. Parmigiani, E. S. Garrett-Mayer, R. Anbazhagan, E. Gabrielson, A cross-study comparison of gene expression studies for the molecular classification of lung cancer. *Clin. Cancer Res.* **10**, 2922–2927 (2004).

78. M. V. Martin, K. Mirnics, L. K. Nisenbaum, M. P. Vawter, Olanzapine Reversed Brain Gene Expression Changes Induced by Phencyclidine Treatment in Non-Human Primates. *Mol. Neuropsychiatry*. **1**, 82–93 (2015).

79. P. Yao *et al.*, Coexpression networks identify brain region-specific enhancer RNAs in the human brain. *Nat. Neurosci.*. **18**, 1168–1174 (2015).

80. The GTEx Consortium, The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*. **348**, 648–660 (2015).

81. A. van B. Granlund *et al.*, Whole genome gene expression meta-analysis of inflammatory bowel disease colon mucosa demonstrates lack of major differences between Crohn’s disease and ulcerative colitis. *PLoS One*. **8**, e56818 (2013).

82. C. L. Noble *et al.*, Regional variation in gene expression in the healthy colon is dysregulated in ulcerative colitis. *Gut*. **57**, 1398–1405 (2008).

83. S. Girirajan *et al.*, Refinement and Discovery of New Hotspots of Copy-Number Variation Associated with Autism Spectrum Disorder. *Am J Hum Genet*. **92**, 221–237 (2013).

84. D. Malhotra, J. Sebat, CNVs: Harbingers of a Rare Variant Revolution in Psychiatric Genetics. *Cell*. **148**, 1223–1241 (2012).

85. S. J. Sanders *et al.*, Multiple Recurrent De Novo CNVs, Including Duplications of the 7q11.23 Williams Syndrome Region, Are Strongly Associated with Autism. *Neuron*. **70**, 863–885 (2011).

86. L. A. Weiss *et al.*, Association between Microdeletion and Microduplication at 16p11.2 and Autism. *N. Engl. J. Med.* **358**, 667–675 (2008).

87. C. R. Marshall *et al.*, Structural Variation of Chromosomes in Autism Spectrum Disorder. *Am J Hum Genet*. **82**, 477–488 (2008).

88. D. Moreno-De-Luca *et al.*, Deletion 17q12 Is a Recurrent Copy Number Variant that Confers High Risk of Autism and Schizophrenia. *Am J Hum Genet*. **87**, 618–630 (2010).

89. J. T. Glessner *et al.*, Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature*. **459**, 569–573 (2009).

90. E. K. Green *et al.*, Copy number variation in bipolar disorder. *Mol Psychiatry*. **21**, 89–93 (2015).

91. J. T. Glessner *et al.*, Duplication of the SLIT3 Locus on 5q35.1 Predisposes to Major Depressive Disorder. *PLoS One*. **5**, (2010).

92. H. Stefansson *et al.*, Large recurrent microdeletions associated with schizophrenia. *Nature*. **455**, 232–236 (2008).

93. A. Corvin *et al.*, Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature*. **455**, 237–241 (2008).

94. D. F. Levinson *et al.*, Copy Number Variants in Schizophrenia: Confirmation of Five Previous Findings and New Evidence for 3q29 Microdeletions and VIPR2 Duplications. *Am J Psychiatry.* **168**, 302–316 (2011).

95. V. Vacic *et al.*, Duplications of the neuropeptide receptor gene VIPR2 confer significant risk for schizophrenia. *Nature*. **471**, 499–503 (2011).

96. D. Rujescu et al., Disruption of the neurexin 1 gene is associated with schizophrenia. *Hum. Mol. Genet*. **18,** 988–996 (2009).

97. S. E. McCarthy *et al.*, Microduplications of 16p11.2 are associated with schizophrenia. *Nat Genet*. **41**, 1223–1227 (2009).

98. J. G. Mulle *et al.*, Microdeletions of 3q29 Confer High Risk for Schizophrenia. *Am J Hum Genet*. **87**, 229–236 (2010).

CHAPTER THREE

Polygenic risk scores are associated with transcriptomic changes in ASD, schizophrenia, and bipolar disorder

3.1: Introduction and contributing authors

Understanding of the pathophysiology of psychiatric disorders, including autism spectrum disorder (ASD), schizophrenia (SCZ), and bipolar disorder (BD), lags behind most other fields of medicine. In the absence of clearly defined pathology, the diagnosis and study of psychiatric disorders are dependent on behavioral, symptomatic characterization. Defining genetic contributions to disease risk provides a substantial foothold for biological understanding. But, leveraging genetic risk to infer disease mechanisms is challenged by substantial genetic complexity and polygenicity, and the lack of a cohesive neurobiological model through which to interpret genetic findings. Recent work demonstrates that the transcriptome represents a quantitative phenotype that provides biological context for understanding the molecular pathways disrupted in major psychiatric disorders (*1*, *2*). RNA sequencing (RNA-Seq) in a large cohort of cases and controls could substantially advance knowledge of the biology disrupted in each disorder and provide a foundational resource for integration with other genomic and genetic data. Therefore, other researchers and I sought to integrate genotype and RNA-sequencing in brain samples from 1,695 subjects with autism (ASD), schizophrenia (SCZ), bipolar disorder (BD) and controls, to identify putative genetic drivers of psychiatric transcriptomic dysregulation. This massive project characterized molecular pathology across three major psychiatric disorders and provides a comprehensive resource for mechanistic insight and therapeutic development.

In this chapter, I present one of my contributions to this project: the generation of psychiatric polygenic risk scores, and the comparison of these polygenic risk scores with psychiatric gene expression. Michael Gandal was the primary author of this work. As a co-author, in addition to generating and analyzing polygenic risk scores for the psychiatric disorders we investigated, I conducted transcriptomic analyses that examined the effects of anti-psychotic drugs in primate neural tissue (also mentioned in this chapter), and assisted with the interpretation and communication of results. Other co-authors included Pan Zhang, Evi Hadjimichael, Rebecca

Walker, Chao Chen, Shuang Liu, Hyejung Won, Harm van Bakel, Merina Varghese, Yongjun Wang, Annie W. Shieh, Sepideh Parhami, Judson Belmont, Minsoo Kim, Patricia Moran Losada, Zenab Khan, Justyna Mleczko, Yan Xia, Rujia Dai, Daifeng Wang, Yucheng T. Yang, Min Xu, Kenneth Fish, Patrick R. Hof, Jonathan Warrell, Dominic Fitzgerald, Andrew E. Jaffe, Kevin White, Mette A. Peters, Mark Gerstein, Chunyu Liu, Lilia M. Iakoucheva, and Dalila Pinto. Daniel Geschwind was the senior author and main project director. All of these co-authors contributed to major and minor analyses for this project, and helped write, edit, and review the resulting manuscript. This work was associated with the PsychENCODE Consortium. Additional results for this extensive project are included in the appendix (section A2). Supporting materials for both this chapter and for appendix A2 are also included in the appendix (section A3).

3.2: Project summary: psychiatric cross-disorder transcriptome-wide analysis

In this project, we integrated genotype and RNA-sequencing in brain samples from 1695 subjects with autism, schizophrenia, bipolar disorder and controls. Analysis of multiple levels of transcriptomic organization – gene expression, local splicing, transcript isoform expression, and co-expression networks for both protein-coding and non-coding genes – provided an in-depth view of ASD, BD, and SCZ molecular pathology. Over 25% of the transcriptome exhibits differential splicing (DS) or expression (DE) in at least one disorder, including 916 non-coding RNAs (ncRNAs), most of which have unexplored functions, and as a group are under increased constraint in humans relative to ncRNA genome-wide. Local splicing analysis permitted identification of genes exhibiting isoform switching across disorders and cell types. Changes at the isoform-level, rather than gene-level, showed the largest effect sizes, genetic enrichment, and greatest disease specificity. We identified 61 co-expression modules associated with at least one disorder, the majority of which show enrichment for cell type-specific marker genes, with 5 modules significantly dysregulated across all three disorders. These modules allow parsing of previously shared downregulated neuronal and synaptic components into a variety of cell type-

and disease-specific signals, including multiple excitatory neuron and distinct interneuron modules with differential patterns of disease association. We also refined the glial-immune signal, demonstrating shared disruption of the blood-brain-barrier and upregulation of NFkB-associated genes, as well as disease-specific alterations in microglial, astrocyte and interferon-response modules. To identify candidate causal drivers, we integrated polygenic risk scores (PRS) and performed a transcriptome-wide association study (TWAS). Dozens of genes are significantly associated with PRS, and TWAS prioritizes novel candidate risk genes likely mediated by cis-effects on brain expression, including 12 in BD, 5 in ASD, and 107 in SCZ.

By integrating RNA-sequencing and genetic data in an unprecedented cohort to refine the shared and distinct molecular pathology of ASD, BD, and SCZ, we provided a quantitative, genome-wide resource for mechanistic insight and therapeutic development. These data inform the molecular pathways and cell types involved in these psychiatric disorders, emphasizing the importance of local splicing and isoform-level gene regulatory mechanisms in defining cell type and disease specificity and - when integrated with GWAS - permit the discovery of new candidate risk genes.

3.3: Identifying genetic drivers of transcriptomic dysregulation

We sought to determine whether changes observed across multiple levels of transcriptomic organization in post mortem brain from patients who were diagnosed with three major psychiatric disorders (ASD, SCZ, and BD) are reflective of the same, or distinct, underlying biological processes. Furthermore, such transcriptomic changes may represent a causal pathophysiology or may be a consequence of disease. To begin to address this, we assessed the relationships among transcriptomic features with polygenic risk scores (PRS) for disease, which provide a causal genetic anchor (**Fig 3.1A**). Across all three disorders, there was strong concordance among differential gene, isoform, and ncRNA signals, as summarized by their first principal component (**Fig 3.1A**). Notably, differential splicing shows greatest overlap with the ncRNA signal, suggesting a role for non-coding genes in regulating local splicing events.

Significant associations with PRS were observed for DGE and DTE signal in SCZ, with greater polygenic association at the isoform level in accordance with the larger transcript isoform effect sizes observed. Concordantly, transcript-level differential expression also showed the most significant enrichment for SCZ SNP-heritability, as measured by stratified LD score regression (*21*, *39*) (**Fig 3.1B**). The overall magnitude of genetic enrichment was modest, however, suggesting that most observed transcriptomic alterations are less a proximal effect of genetic variation and more likely the consequence of a downstream cascade of biological events following earlier acting genetic risk factors.

We were also interested to determine the degree to which genes showed increases in the magnitude of DE over the duration of illness, as a positive relationship would be expected if age-related cumulative exposures (e.g. drugs, smoking) were driving these changes. To assess this, we fit local regression models to case and control sample-level expression measurements as a function of age and computed age-specific DE effect-sizes (**Fig A3.10**). Of 4821 DE genes in SCZ, only 143 showed even nominal association ($p < 0.05$, uncorrected for multiple comparisons) between effect size magnitude and age. Similar associations were seen in 29 of 1119 DE genes in BD and 85 of 1611 DE genes in ASD. Consequently, this would not support substantial age-related environmental exposures as the mechanism for the vast majority of differentially expressed genes.

Using gene expression data from animal models, we investigated whether exposure to commonly used psychiatric medications could recapitulate observed gene expression changes in disease (**Fig A3.11**). Overall, with the exception of lithium, chronic exposure to medications including antipsychotics (clozapine, haloperidol), mood stabilizers (lamotrigine), and antidepressants had a minimal effect on the transcriptome, in most cases with no differentially expressed genes at traditional FDR thresholds (*21*). Even at more liberal thresholds, the overlap between medication-driven and disease signal remains sparse. One notable exception was a module that reflects major components of a well-described (*40*) neural activity-dependent gene expression program, whose disease relationships are refined in the network analysis section below. Finally, we do note that there are other factors that were not measured that can potentially contribute to gene expression variation in post-mortem tissue, including agonal events and pH (*22*, *41*, *42*) in addition to those measured and used as covariates, such as RNA integrity and post mortem interval (PMI). We used surrogate variable correction in our analyses to account for such unmeasured confounders (*43*), which is a standard approach (*44*).

\

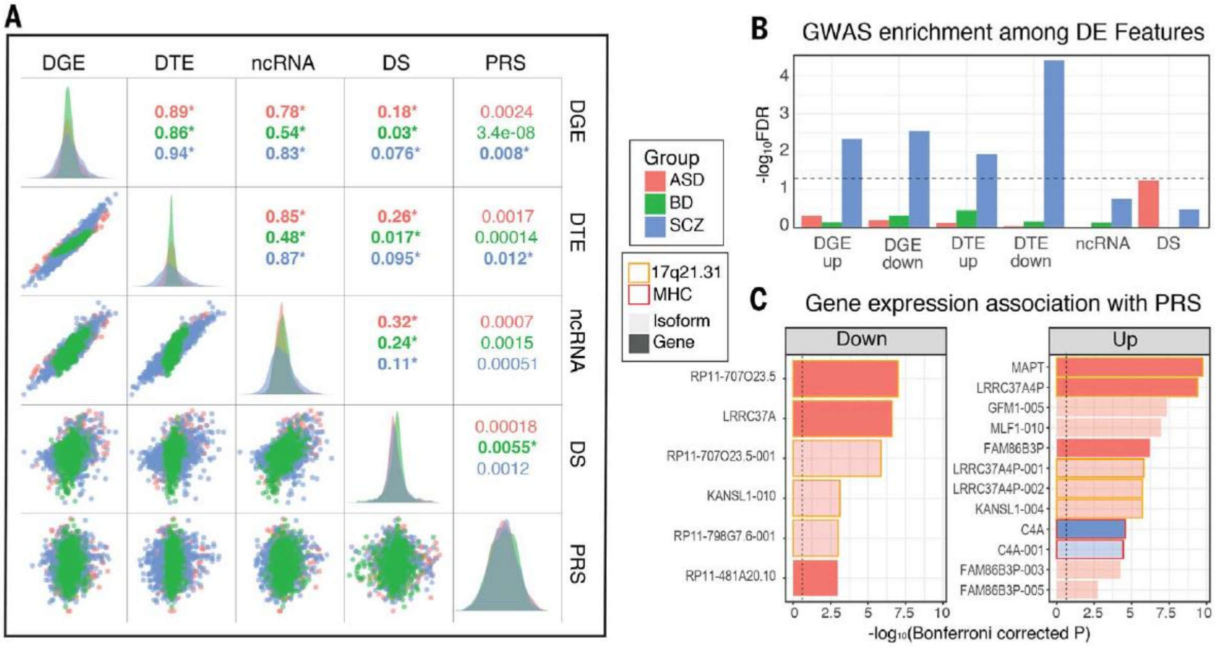


Figure 3.1. Overlaps and genetic enrichment among dysregulated transcriptomic features. A) Scatterplots demonstrate overlap among dysregulated transcriptomic features, summarized by their first principle component across case and control subjects (R^2 values; *P<0.05). Polygenic risk shows greatest association with differential transcript signal in SCZ, although the magnitude is small . **B)** Stratified LD-score regression identifies enrichment of GWAS SNP-heritability in SCZ among multiple differentially expressed transcriptomic features, with downregulated isoforms showing must substantial association. **C)** Polygenic risk scores created separately for each disorder are significantly (FDR<0.05) associated with expression for multiple individual genes and isoforms. Plots are split between upregulated and downregulated associations with increasing PRS. Several associations with SCZ PRS are located within the MHC region of the genome, which harbors the largest GWAS association signal but also highly complex LD structure.

We next sought to leverage this transcriptome-wide dataset to prioritize potential risk genes whose brain expression may be altered by disease-associated genetic variation. We first assessed whether gene or isoform level expression measures were significantly associated with PRS for each disorder (*21*), identifying 45 genes or isoforms whose expression was significantly associated with PRS (FDR < 0.05), including 32 in ASD, 2 in BD, and 11 in SCZ (**Fig 3.1C**; **Table A3.4**). In ASD, the majority of associations map to 17q21.31, which harbors a known common inversion polymorphism and rare deleterious structural variants associated with intellectual disability (*45*). PRS for BD was associated with isoforms of the neuronal calcium sensor *NCALD* and *SNF8*, an endosomal sorting protein. In SCZ, we identify upregulation of the established risk gene *C4A* as the most significant PRS association (*5*). Concordantly, we find a strong positive correlation between *C4A* expression and genetically imputed *C4A* copy number (R=0.25, P=7x10^{-14}), as well as with imputed number of *C4-HERV* elements (R=0.24, P=1.2x10^{-12}), but not *C4B* copy number (R=0.006, P=0.85) (*21*). Additional associations with PRS were observed in the MHC region in SCZ, which harbors the largest GWAS association comprised of multiple independent signals (*5*), but is difficult to parse due to complex patterns of LD. These included two lncRNAs, *HCG17* and *HCG23,* as well as the MHC class I heavy chain receptor *HLA-C*. However, expression of all three of these genes were also significantly (P<0.05) correlated with *C4A* copy number, indicating a need for further validation given the complex LD structure in this region. Together, these results point to specific genes that exhibit increasing gene expression dysregulation with increasing psychiatric polygenic risk, indicating that these genes may be convergent targets of psychiatric genetic risk variants. Determining how the transcriptomic dysregulation of these genes may contribute to psychiatric disorder mechanisms will continue to enhance our understanding of how psychiatric genetic risk variants induce psychiatric neuropathology.

3.4: Materials and Methods

Please see the appendix (section A3.1) for all materials and methods.

3.5: Bibliography

This bibliography is comprehensive for this chapter, appendix section A2, and appendix section A3.

1. M. J. Gandal *et al.*, Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap. *Science*. **359**, 693–697 (2018).
2. M. Fromer *et al.*, Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat. Neurosci.* **19**, 1442–1453 (2016).
3. H. A. Whiteford, A. J. Ferrari, L. Degenhardt, V. Feigin, T. Vos, The global burden of mental, neurological and substance use disorders: an analysis from the Global Burden of Disease Study 2010. *PLoS One*. **10** (2015).
4. M. J. Gandal, V. Leppa, H. Won, N. N. Parikshak, D. H. Geschwind, The road to precision psychiatry: translating genetics into disease mechanisms. *Nat. Neurosci.* **19**, 1397–1407 (2016).
5. A. Sekar *et al.*, Schizophrenia risk from complex variation of complement component 4. *Nature*. **530**, 177–183 (2016).
6. S. J. Sanders, First glimpses of the neurobiology of autism spectrum disorder. *Curr. Opin. Genet. Dev.* **33**, 80–92 (2015).
7. M. T. Maurano *et al.*, Systematic localization of common disease-associated variation in regulatory DNA. *Science*. **337**, 1190–1195 (2012).
8. L. D. Ward, M. Kellis, Interpreting noncoding genetic variation in complex traits and human disease. *Nat. Biotechnol.* **30**, 1095–1106 (2012).
9. A. Visel, E. M. Rubin, L. A. Pennacchio, Genomic views of distant-acting enhancers. *Nature*. **461**, 199–205 (2009).
10. GTEx Consortium, Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*. **348**, 648–660 (2015).
11. S. K. Reilly *et al.*, Evolutionary genomics. Evolutionary changes in promoter and enhancer activity during human corticogenesis. *Science*. **347**, 1155–1159 (2015).
12. The ENCODE Project Consortium, Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*. **447**, 799 (2007).
13. Roadmap Epigenomics Consortium *et al.*, Integrative analysis of 111 reference human epigenomes. *Nature*. **518**, 317–330 (2015).

14. R. Andersson *et al.*, An atlas of active enhancers across human cell types and tissues. *Nature*. **507**, 455–461 (2014).

15. N. N. Parikshak, M. J. Gandal, D. H. Geschwind, Systems biology and gene networks in neurodevelopmental and neurodegenerative disorders. *Nat. Rev. Genet.* **16**, 441–458 (2015).

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

17. PsychENCODE Consortium *et al.*, The PsychENCODE project. *Nat. Neurosci.* **18**, 1707–1712 (2015).

18. Daifeng Wang, Shuang Liu, Jonathan Warrell, Hyejung Won, Xu Shi, Fabio Navarro, Declan Clarke, Mengting Gu, Prashant Emani, Min Xu, Yucheng T. Yang, Jonathan J. Park, Suhn Kyong Rhie, Kasidet Manakongtreecheep, Holly Zhou, Aparna Nathan, Jing Zhang, Mette Peters, Eugenio Mattei, Dominic Fitzgerald, Tonya Brunetti, Jill Moore, PsychENCODE Consortium, Nenad Sestan, Andrew E. Jaffe, Kevin White, Zhiping Weng, Daniel H. Geschwind, James Knowles, Mark Gerstein, Comprehensive functional genomic resource and integrative model for the adult brain. *Science*. **362** (6420), (2018).

19. N. N. Parikshak *et al.*, Genome-wide changes in lncRNA, splicing, and regional gene expression patterns in autism. *Nature*. **540**, 423–427 (2016).

20. H. J. Kang *et al.*, Spatio-temporal transcriptome of the human brain. *Nature*. **478**, 483–489 (2011).

21. See supplemental methods.

22. A. E. Jaffe *et al.*, qSVA framework for RNA quality correction in differential expression analysis. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 7130–7135 (2017).

23. B. Li, C. N. Dewey, RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*. **12**, 323 (2011).

24. A. E. Jaffe *et al.*, Developmental and genetic regulation of the human cortex transcriptome illuminate schizophrenia pathogenesis. *Nat. Neurosci.* **21**, 1117–1125 (2018).

25. Y. I. Li *et al.*, RNA splicing is a primary link between genetic variation and disease. *Science*. **352**, 600–604 (2016).

26. P. J. Batista, H. Y. Chang, Long noncoding RNAs: cellular address codes in development and disease. *Cell*. **152**, 1298–1307 (2013).

27. J. di Iulio *et al.*, The human noncoding genome defined by genetic diversity. *Nat. Genet.* (2018).

28. C. K. Vuong, D. L. Black, S. Zheng, The neurogenetics of alternative splicing. *Nat. Rev. Neurosci.* **17**, 265–281 (2016).

29. Y. I. Li *et al.*, Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.* **50**, 151–158 (2018).

30. M. Irimia *et al.*, A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell*. **159**, 1511–1523 (2014).

31. N. Akula *et al.*, RNA-sequencing of the brain transcriptome implicates dysregulation of neuroplasticity, circadian rhythms and GTPase binding in bipolar disorder. *Mol. Psychiatry*. **19**, 1179–1185 (2014).

32. J.-A. Lee *et al.*, Cytoplasmic Rbfox1 Regulates the Expression of Synaptic and Autism-Related Genes. *Neuron*. **89**, 113–128 (2016).

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

34. E. Sebestyén *et al.*, Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks. *Genome Res.* **26**, 732–744 (2016).

35. J. Gauthier *et al.*, Truncating mutations in NRXN2 and NRXN1 in autism spectrum disorders and schizophrenia. *Hum. Genet.* **130**, 563–573 (2011).

36. F. F. Hamdan *et al.*, Excess of de novo deleterious mutations in genes associated with glutamatergic systems in nonsyndromic intellectual disability. *Am. J. Hum. Genet.* **88**, 306–316 (2011).

37. B. Treutlein, O. Gokce, S. R. Quake, T. C. Südhof, Cartography of neurexin alternative splicing mapped by single-molecule long-read mRNA sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E1291–E1299 (2014).

38. Grove, J., Ripke, S., Als, T.D. et al. Identification of common genetic risk variants for autism spectrum disorder. *Nat Genet* **51**, 431–444 (2019).

39. H. K. Finucane *et al.*, Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).

40. A. E. West, M. E. Greenberg, Neuronal Activity–Regulated Gene Transcription in Synapse Development and Cognitive Function. *Cold Spring Harb. Perspect. Biol.* **3** (2011).

41. Y. Zhu, L. Wang, Y. Yin, E. Yang, Systematic analysis of gene expression patterns associated with postmortem interval in human tissues. *Sci. Rep.* **7**, 5435 (2017).

42. J. Z. Li *et al.*, Systematic changes in gene expression in postmortem human brains associated with tissue pH and terminal medical conditions. *Hum. Mol. Genet.* **13**, 609–616 (2004).

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

44. GTEx Consortium, Genetic effects on gene expression across human tissues. *Nature*. **550**, 204 (2017).

45. M. C. Zody *et al.*, Evolutionary toggling of the MAPT 17q21.31 inversion region. *Nat. Genet.* **40**, 1076–1083 (2008).

46. A. Gusev *et al.*, Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).

47. E. R. Gamazon *et al.*, A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**, 1091–1098 (2015).

48. M. C. Oldham *et al.*, Functional organization of the transcriptome in human brain. *Nat. Neurosci.* **11**, 1271–1282 (2008).

49. B. Zhang, S. Horvath, A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4**, Article17 (2005).

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

51. C. Chen *et al.*, Two gene co-expression modules differentiate psychotics and controls. *Mol. Psychiatry*. **18**, 1308–1314 (2013).

52. R. Daneman *et al.*, The mouse blood-brain barrier transcriptome: a new resource for understanding the development and function of brain endothelial cells. *PLoS One*. **5**, e13741 (2010).

53. B. Obermeier, R. Daneman, R. M. Ransohoff, Development, maintenance and disruption of the blood-brain barrier. *Nat. Med.* **19**, 1584–1596 (2013).

54. G. Genovese *et al.*, Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia. *Nat. Neurosci.* **19**, 1433–1441 (2016).

55. S. E. Ellis, R. Panitch, A. B. West, D. E. Arking, Transcriptome analysis of cortical tissue reveals shared sets of downregulated genes in autism and schizophrenia. *Transl. Psychiatry*. **6**, e817 (2016).

56. R. C. Ramaker *et al.*, Post-mortem molecular profiling of three psychiatric disorders. *Genome Med.* **9**, 72 (2017).

57. C. K. Vuong *et al.*, Rbfox1 Regulates Synaptic Transmission through the Inhibitory Neuron-Specific vSNARE Vamp1. *Neuron*. **98**, 127–141.e7 (2018).

58. A. F. Pardiñas *et al.*, Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nat. Genet.* (2018).

59. L. T. Gehman *et al.*, The splicing regulator Rbfox1 (A2BP1) controls neuronal excitation in the mammalian brain. *Nat. Genet.* **43**, 706–711 (2011).

60. B. L. Fogel *et al.*, RBFOX1 regulates both splicing and transcriptional networks in human neuronal development. *Hum. Mol. Genet.* **21**, 4171–4186 (2012).

61. A. M. Bond *et al.*, Balanced gene regulation by an embryonic brain ncRNA is critical for adult hippocampal GABA circuitry. *Nat. Neurosci.* **12**, 1020–1027 (2009).

62. N. G. Skene *et al.*, Genetic identification of brain cell types underlying schizophrenia. *Nat. Genet.* **50**, 825–833 (2018).

63. B. Labonté *et al.*, Sex-specific transcriptional signatures in human depression. *Nat. Med.* **23**, 1102–1111 (2017).

64. A. de Bartolomeis *et al.*, Immediate-Early Genes Modulation by Antipsychotics: Translational Implications for a Putative Gateway to Drug-Induced Long-Term Brain Changes. *Front. Behav. Neurosci.* **11**, 240 (2017).

65. R. Birnbaum *et al.*, Investigating the neuroimmunogenic architecture of schizophrenia. *Mol. Psychiatry* (2017).

66. S. G. Fillman *et al.*, Increased inflammatory markers identified in the dorsolateral prefrontal cortex of individuals with schizophrenia. *Mol. Psychiatry*. **18**, 206–214 (2013).

67. R. Pacifico, R. L. Davis, Transcriptome sequencing implicates dorsal striatum-specific gene network, immune response and energy metabolism pathways in bipolar disorder. *Mol. Psychiatry*. **22**, 441–449 (2017).

68. J. S. Rao, G. J. Harry, S. I. Rapoport, H. W. Kim, Increased excitotoxicity and neuroinflammatory markers in postmortem frontal cortex from bipolar disorder patients. *Mol. Psychiatry*. **15**, 384–392 (2010).

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

70. I. Rusinova *et al.*, Interferome v2.0: an updated database of annotated interferon-regulated genes. *Nucleic Acids Res.* **41**, D1040–6 (2013).

71. A. Necsulea *et al.*, The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature*. **505**, 635–640 (2014).

72. R. Dai, Y. Xia, C. Liu, C. Chen, csuWGCNA: a combination of signed and unsigned WGCNA to capture negative correlations. *bioRxiv* (2019).

73. P. Pruunsild, C. P. Bengtson, H. Bading, Networks of Cultured iPSC-Derived Neurons Reveal the Human Synaptic Activity-Regulated Adaptive Gene Program. *Cell Rep.* **18**, 122–135 (2017).

74. P. P. Amaral *et al.*, Complex architecture and regulated expression of the Sox2ot locus during vertebrate development. *RNA*. **15**, 2013–2027 (2009).

75. T. R. Mercer, M. E. Dinger, S. M. Sunkin, M. F. Mehler, J. S. Mattick, Specific expression of long noncoding RNAs in the mouse brain. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 716–721 (2008).

76. K. Aberg, P. Saetre, N. Jareborg, E. Jazin, Human QKI, a potential regulator of mRNA expression of human oligodendrocyte-related genes involved in schizophrenia. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 7482–7487 (2006).

77. G. Barry *et al.*, The long non-coding RNA Gomafu is acutely regulated in response to neuronal activation and involved in schizophrenia-associated alternative splicing. *Mol. Psychiatry*. **19**, 486–494 (2014).

78. M. Koga *et al.*, Involvement of SMARCA2/BRM in the SWI/SNF chromatin-remodeling complex in schizophrenia. *Hum. Mol. Genet.* **18**, 2483–2494 (2009).

79. I. D. Krantz *et al.*, Cornelia de Lange syndrome is caused by mutations in NIPBL, the human homolog of Drosophila melanogaster Nipped-B. *Nat. Genet.* **36**, 631–635 (2004).

80. S. J. Sanders *et al.*, Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron*. **87**, 1215–1233 (2015).

81. M. Melé *et al.*, Human genomics. The human transcriptome across tissues and individuals. *Science*. **348**, 660–665 (2015).

82. M. Quesnel-Vallières *et al.*, Misregulation of an Activity-Dependent Splicing Network as a Common Mechanism Underlying Autism Spectrum Disorders. *Mol. Cell*. **64**, 1023–1034

(2016).

83. M. L. Estes, A. K. McAllister, Immune mediators in the brain and peripheral tissues in autism spectrum disorder. *Nat. Rev. Neurosci.* **16**, 469–486 (2015).

84. U. Meyer, J. Feldon, O. Dammann, Schizophrenia and autism: both shared and disorder-specific pathogenesis via perinatal inflammation? *Pediatr. Res.* **69**, 26R–33R (2011).

85. J. D. Rosenblat *et al.*, Inflammation as a neurobiological substrate of cognitive impairment in bipolar disorder: Evidence, pathophysiology and treatment implications. *J. Affect. Disord.* **188**, 149–159 (2015).

86. T. Steijger *et al.*, Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods*. **10**, 1177–1184 (2013).

87. M. I. Love, J. B. Hogenesch, R. A. Irizarry, Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation. *Nat. Biotechnol.* **34**, 1287–1291 (2016).

88. T. Lappalainen, Functional genomics bridges the gap between quantitative genetics and molecular biology. *Genome Res.* **25**, 1427–1431 (2015).

89. M. Brandt, T. Lappalainen, SnapShot: Discovering Genetic Regulatory Variants by QTL Analysis. *Cell*. **171**, 980–980. (2017).

90. D. M. Ruderfer *et al.*, Genomic Dissection of Bipolar Disorder and Schizophrenia, Including 28 Subphenotypes. *Cell*. **173**, 1705–1715. (2018).

91. B. B. Lake *et al.*, Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat. Biotechnol.*(2017).

92. T. Goldmann *et al.*, Origin, fate and dynamics of macrophages at central nervous system interfaces. *Nat. Immunol.* **17**, 797–805 (2016).

93. H. Keren-Shaul *et al.*, A Unique Microglia Type Associated with Restricting Development of Alzheimer’s Disease. *Cell*. **169**, 1276–1290. (2017).

94. A. Zeisel *et al.*, Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*. **347**, 1138–1142 (2015).

95. Y. Zhang *et al.*, Purification and Characterization of Progenitor and Mature Human Astrocytes Reveals Transcriptional and Functional Differences with Mouse. *Neuron*. **89**, 37–53 (2016).

96. B. Wilkinson *et al.*, The autism-associated gene chromodomain helicase DNA-binding protein 8 (CHD8) regulates noncoding RNAs and autism-related genes. *Transl. Psychiatry*. **5**, e568 (2015).

97. K. E. Samocha *et al.*, A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* **46**, 944–950 (2014).

98. I. Iossifov *et al.*, Low load for disruptive mutations in autism genes and their biased transmission. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E5600–7 (2015).

99. K. J. Karczewski *et al.*, The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic Acids Res.* **45**, D840–D845 (2017).

100. D. Zhang *et al.*, Genetic control of individual differences in gene-specific methylation in human brain. *Am. J. Hum. Genet.* **86**, 411–419 (2010).

101. A. E. Jaffe *et al.*, Developmental And Genetic Regulation Of The Human Cortex Transcriptome In Schizophrenia. *Nat Neurosci* **21**, 1117–1125 (2018).

102. M. C. Oldham, P. Langfelder, S. Horvath, Network methods for describing sample relationships in genomic datasets: application to Huntington’s disease. *BMC Syst. Biol.* **6**, 63 (2012).

103. J. Reimand, T. Arak, J. Vilo, g:Profiler—a web server for functional interpretation of gene lists (2011 update). *Nucleic Acids Res.* **39**, W307–W315 (2011).

104. M. Lek *et al.*, Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. **536**, 285–291 (2016).

105. B. B. Lake *et al.*, Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science*. **352**, 1586–1590 (2016).

106. S. Darmanis *et al.*, A survey of human brain transcriptome diversity at the single cell level. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 7285–7290 (2015).

107. J. P. Doyle *et al.*, Application of a translational profiling approach for the comparative analysis of CNS cell types. *Cell*. **135**, 749–762 (2008).

108. A. Dobin *et al.*, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. **29**, 15–21 (2013).

109. F. Hahne, R. Ivanek, Visualizing Genomic Data Using Gviz and Bioconductor. *Methods Mol. Biol.* **1418**, 335–351 (2016).

110. I. Letunic, P. Bork, 20 years of the SMART protein domain annotation resource. *Nucleic Acids Res.* **46**, D493–D496 (2018).

111. R. D. Finn *et al.*, The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44**, D279–85 (2016).

112. A. G. Baltz *et al.*, The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Mol. Cell*. **46**, 674–690 (2012).

113. A. Castello *et al.*, Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell*. **149**, 1393–1406 (2012).

114. S. C. Kwon *et al.*, The RNA-binding protein repertoire of embryonic stem cells. *Nat. Struct. Mol. Biol.* **20**, 1122–1130 (2013).

115. E. L. Huttlin *et al.*, The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell*. **162**, 425–440 (2015).

116. T. S. Keshava Prasad *et al.*, Human Protein Reference Database--2009 update. *Nucleic Acids Res.* **37**, D767–72 (2009).

117. K. Lage *et al.*, A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.* **25**, 309–316 (2007).

118. J. Das, H. Yu, HINT: High-quality protein interactomes and their applications in understanding human disease. *BMC Syst. Biol.* **6**, 92 (2012).

119. A. Chatr-Aryamontri *et al.*, The BioGRID interaction database: 2017 update. *Nucleic Acids Res.* **45**, D369–D379 (2017).

120. K. Zuberi *et al.*, GeneMANIA prediction server 2013 update. *Nucleic Acids Res.* **41**, W115–22 (2013).

121. D. Szklarczyk *et al.*, The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Res.* **45**, D362–D368 (2017).

122. A. Ruepp *et al.*, CORUM: the comprehensive resource of mammalian protein complexes—2009. *Nucleic Acids Res.* **38**, D497–D501 (2010).

123. K. S. Pollard, M. J. Hubisz, K. R. Rosenbloom, A. Siepel, Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).

124. A. Siepel *et al.*, Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).

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

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

127. M. J. Mason, G. Fan, K. Plath, Q. Zhou, S. Horvath, Signed weighted gene co-expression network analysis of transcriptional regulation in murine embryonic stem cells. *BMC Genomics*. **10**, 327 (2009).

128. C. S. Benton *et al.*, Evaluating genetic markers and neurobiochemical analytes for fluoxetine response using a panel of mouse inbred strains. *Psychopharmacology* . **221**, 297–315 (2012).

129. N. R. Wray *et al.*, Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat. Genet.* (2018).

130. A. Okbay *et al.*, Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nat. Genet.* **48**, 624–633 (2016).

131. A. Okbay *et al.*, Genome-wide association study identifies 74 loci associated with educational attainment. *Nature*. **533**, 539–542 (2016).

132. A. P. Morris *et al.*, Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat. Genet.* **44**, 981–990 (2012).

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

134. Autism Spectrum Disorders Working Group of The Psychiatric Genomics Consortium, Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24.32 and a significant overlap with schizophrenia. *Mol. Autism*. **8**, 21 (2017).

135. Psychiatric GWAS Consortium Bipolar Disorder Working Group, Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4.

Nat. Genet. **43**, 977–983 (2011).

136. International HapMap 3 Consortium *et al.*, Integrating common and rare genetic variation in diverse human populations. *Nature*. **467**, 52–58 (2010).

137. B. J. Vilhjálmsson *et al.*, Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am. J. Hum. Genet.* **97**, 576–592 (2015).

138. J. Yang, S. H. Lee, M. E. Goddard, P. M. Visscher, GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).

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

140. H. T. Nguyen *et al.*, Integrated Bayesian analysis of rare exonic variants to identify risk genes for schizophrenia and neurodevelopmental disorders. *Genome Med.* **9**, 114 (2017).

141. D. Polioudakis *et al.*, A single cell transcriptomic analysis of human neocortical development. *Neuron* **103**, 785–801 (2019).

142. A. Untergasser *et al.*, Primer3--new capabilities and interfaces. *Nucleic Acids Res.* **40**, e115 (2012).

143. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

144. Mingfeng Li, Gabriel Santpere, Yuka Imamura Kawasawa, Oleg V. Evgrafov, Forrest O. Gulden, Sirisha Pochareddy, Susan M. Sunkin, Zhen Li, Yurae Shin, Robert R. Kitchen, Ying Zhu, Donna M. Werling, Andre M.M. Sousa, Hyojung Kang, Mihovil Pletikos, Jinmyung Choi, Sydney Muchnik, Xuming Xu, Daifeng Wang, Shuang Liu, Paola Giusti-Rodríguez, Christiaan A de Leeuw, Antonio Pardinas, BrainSpan Consortium, PsychENCODE Consortium: Developmental Subgroup, Ming Hu, Fulai Jin, Yun Li, Michael Owen, Michael O'Donovan, James Walters, Danielle Posthuma, Patrick Sullivan, Patt Levitt, Daniel R. Weinberger, Joel E. Kleinman, Daniel H. Geschwind, Stephan Sanders, Michael J. Hawrylycz, Matthew State, Mark B. Gerstein, Ed S. Lein, James A. Knowles, Nenad Sestan, Integrative Functional Genomic Analysis of Human Brain Development and Neuropsychiatric Risk Convergence. *Science* **362** (6420), (2018).

CHAPTER FOUR

Broad transcriptomic dysregulation across the cerebral cortex in ASD

4.1: Abstract

Classically, psychiatric disorders have been considered to lack defining pathology, but recent work has demonstrated consistent disruption at the molecular level, characterized by transcriptomic and epigenetic alterations.[1–3] In ASD, upregulation of microglial, astrocyte, and immune signaling genes, downregulation of specific synaptic genes, and attenuation of regional gene expression differences are observed.[1,2,4–6] However, whether these changes are limited to the few cortical regions profiled is unknown. Here, we perform RNA-sequencing (RNA-seq) on 725 brain samples spanning 11 distinct cortical areas in 112 ASD cases and neurotypical controls. We identify substantially more genes and isoforms that differentiate ASD from controls than previously observed. These alterations are pervasive and cortex-wide, but vary in magnitude across regions, roughly showing an anterior to posterior gradient, with the strongest signal in visual cortex, followed by parietal cortex and the temporal lobe. We find a notable enrichment of ASD genetic risk variants among cortex-wide downregulated synaptic plasticity genes and upregulated protein folding gene isoforms. Finally, using snRNA-seq we determine that regional variation in the magnitude of transcriptomic dysregulation reflects changes in cellular proportion and cell-type-specific gene expression, particularly impacting L3/4 excitatory neurons. These results highlight widespread, genetically-driven neuronal dysfunction as a major component of ASD pathology in the cerebral cortex, extending beyond association cortices to involve primary sensory regions.

4.2: Transcriptomic changes across the cerebral cortex in ASD

Similar to other neuropsychiatric disorders, the risk for autism spectrum disorder (ASD) involves substantial genetic liability, which is profoundly complex and heterogeneous.[7,8] Despite this causal heterogeneity, molecular profiling studies consistently show common patterns of shared transcriptomic and epigenetic dysregulation in the majority of ASD cases.[1–3,5] But, whether this represents focal, regional, or more generalized dysfunction is not known. To address this question cortex-wide, we conducted strand-specific RNA-sequencing (RNA-seq) to identify gene and isoform (transcriptomic) changes in 725 samples across 11 brain regions spanning all four cortical lobules (frontal, parietal, temporal, and occipital), from 49 subjects with idiopathic ASD and 54 matched neurotypical controls (**Fig. 4.1a**, **Methods**, **Table A4.1**, and **Fig. A4.1-3**). Previous work using gene expression microarrays and RNA-seq identified gene co-expression modules representing specific pathways differentially expressed in ASD frontal and temporal cortices.[4,5] The number of samples profiled here is more than five times greater than these prior studies, so we first used this multi-region RNA-seq resource to replicate and extend these previous findings. We observed widespread dysregulation across all 11 cortical regions that replicated the previously identified patterns of dysregulation in temporal and frontal cortices (**Fig. 4.1b**, **Methods**, **Table A4.2**, **Fig. A4.3**). However, the magnitude of effect varied across regions, with the primary visual cortex (V1; Brodmann Area (BA) 17) exhibiting the greatest degree of dysregulation, followed by parietal cortex (BA7) and posterior superior temporal gyrus (BA 41/42/22) in terms of fold changes and the number of genes differentially expressed (**Fig. 4.1b**, **Fig. A4.3**). To show that this was not due to regional variation in sample sizes, we performed permutation testing, which indicated that this increased signal was not biased by regional sample size differences (**Methods**, **Table A4.2**).

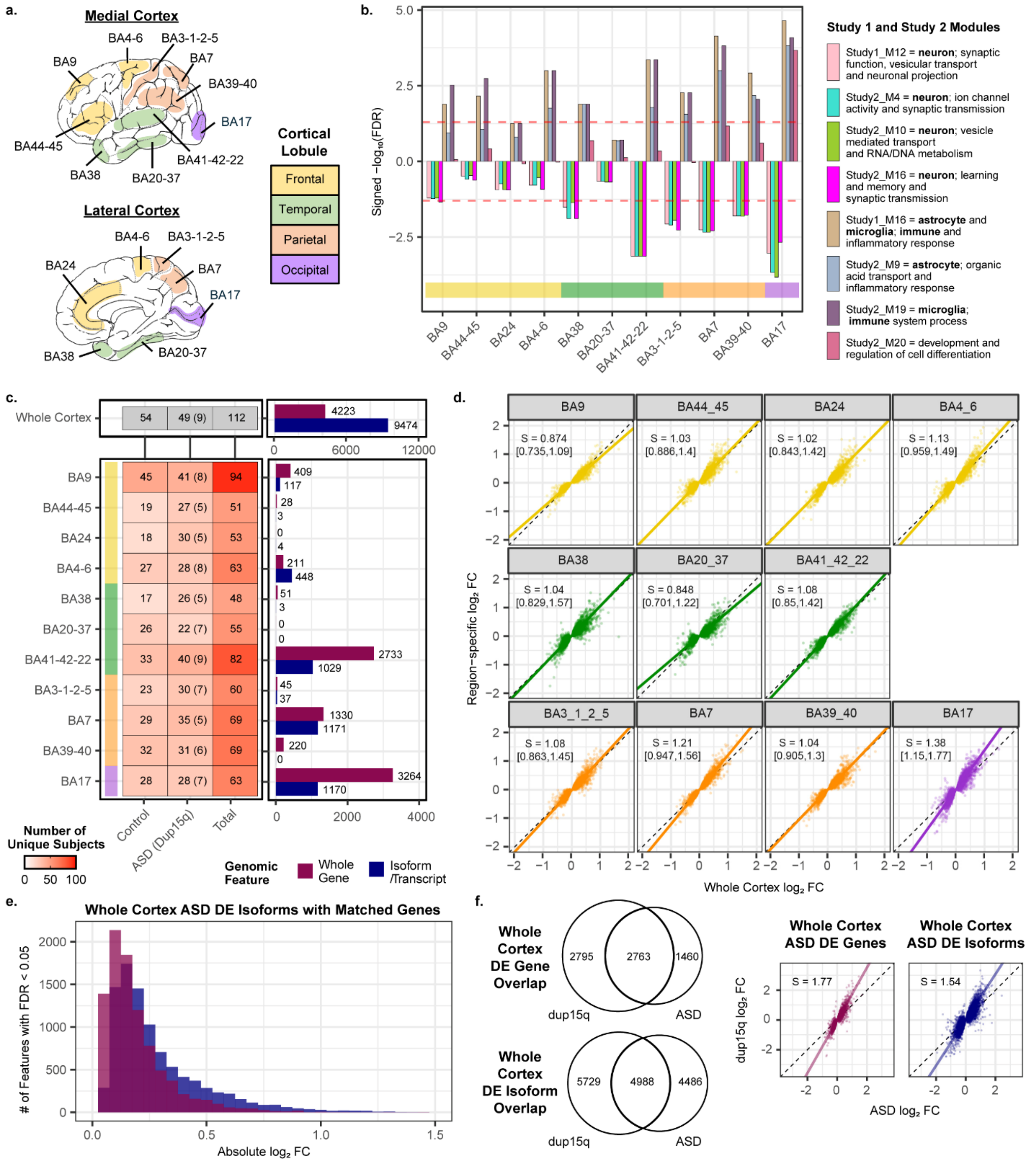


Figure 4.1 | ASD transcriptomic differences across 11 cortical regions. a. Human cortical Brodmann Areas (BA) with cortical lobules indicated. Cortical lobule colors are consistent throughout the figure. **b.** Dysregulation of previously identified co-expressed gene modules (Study 1: Voineagu et al., Nature 2011; Study 2: Parikshak et al., Nature 2016)[4-5] across cortical regions. Red dashed line marks FDR < 0.05. **c.** Unique subjects by region and diagnosis (left), with number of differentially expressed (DE; linear mixed model FDR < 0.05) genes and isoforms (right) across the whole cortex (top) or within individual regions (bottom). Gene and isoform colors are consistent throughout this figure. **d.** log2 fold change (FC) of individual regions compared to the whole cortex log2 FC for the 4,223 whole cortex DE genes. Slope (S,

with 95% confidence interval in brackets) is calculated with principal components regression. **e.** For the whole cortex DE isoforms, a histogram of all isoforms DE across the whole cortex in ASD with their matched genes. **f.** Left: venn diagrams depicting the number of genes and isoforms DE across the whole cortex in dup15q samples compared to idoipathic ASD samples. Right: for the ASD whole cortex DE genes (left) and isoforms (right), the idiopathic ASD whole cortex log2 FC compared to the dup15q whole cortex log2 FC. Slope (S) is calculated with principal components regression.

Given that qualitatively similar transcriptomic changes were observed across regions (**Fig. 4.1b**), we next combined all regions to increase our statistical power to detect previously unrecognized differentially expressed (DE) genes and isoforms. We used a linear mixed model framework to control for individual effects and identify changes across all 11 regions examined as well as within individual regions, separately (**Fig. 4.1c**, **Methods**, **Table A4.3**, **Fig A4.4**). We found 4,223 genes and 9,474 isoforms (FDR < 0.05) DE across all cortical regions, a notable increase compared to previous analyses (**Fig. 4.1c**, **Fig. A4.3**). We again observed the greatest signal in BA17, and 59% of DE genes in BA17 alone overlapped with what was observed globally (**A4.2**, **Fig. A4.4**). Additionally, DE gene effect sizes in BA17 and BA7 were the highest in magnitude, more than other regions assessed (**Fig. 4.1d**, **Methods**). In comparing DE genes and isoforms across all regions, we found both conserved and distinct dysregulation (**Fig. A4.4**, **Table A4.2**). Notably, as previously observed in frontal and temporal cortex[1] we observed that DE isoforms exhibited greater effect size changes in ASD than their matched genes (**Fig. 4.1e**, **Table A4.2**, **Fig. A4.4**).

We next evaluated differential gene and isoform expression in an additional 83 pan-cortical samples from 9 subjects with dup15q syndrome, a rare genetic disorder with high penetrance for ASD, which previously was shown to strongly parallel changes in idiopathic ASD in frontal and temporal cortex, but with greater magnitude of effect.[5] We replicated these previous results broadly across the cortical regions examined, finding substantial overlap in transcriptomic changes between dup15q and idiopathic ASD and with dup15q exhibiting a greater magnitude of dysregulation overall (**Fig. 4.1f**, **Table A4.2**, **Fig. A4.4**). BA17 also exhibited the greatest number of DE genes in dup15q (**Fig. A4.4**). These results demonstrate that the molecular pathology shared by this genetic form of ASD and idiopathic ASD is widespread across distinct regions of the cortex, and that some commonalities in regional variance of effect exist.

4.3: Broad attenuation of transcriptomic regional identity

We previously observed an attenuation of typical gene expression differences between two regions, frontal and temporal lobe in ASD,[4,5] which we refer to here as an "Attenuation of Transcriptomic Regional Identity" (ARI). To assess whether this was a broader phenomenon, we systematically contrasted all unique pairs of 11 cortical regions (55 comparisons in all) using a conservative statistical approach to account for differences in sample size across regions, while correcting stringently for multiple comparisons (**Fig. 4.2a**, **Methods**). We further validated the identified transcriptomic regional identity patterns in our control samples with those from an external data source, the Allen Brain Atlas[9] (**Table A4.4**, **Fig. A4.5**, **Methods**). Ten pairs of regions exhibited significantly greater ARI patterns in ASD compared to controls, with an additional 31 out of the 55 pairs of regions exhibiting a trend towards attenuation in ASD (**Fig. 4.2b**, **Table A4.4**, **Fig. A4.5**, **Methods**). These results provide evidence in support of widespread ARI across the cerebral cortex in ASD for the first time, across both gene and isoform levels (**Fig. A4.5**). Additionally, we observed a regional anterior - posterior gradient, with nine of the ten region pairs exhibiting significant ARI in ASD containing either BA17 or BA39-40 (**Fig. 4.2c-d**). Notably, BA17 was also one of the regions with the largest case-control differences in gene expression. To determine how gene expression changes were dispersed across regions in these pairs, we used a conservative filtering process to identify individual genes exhibiting ARI (**Methods**, **Table A4.4**). Although these genes were widely dysregulated, the posterior regions BA17 and BA39-40 exhibited the greatest changes (**Fig. 4.2c-d**, **Fig. A4.6**). ARI genes were also comparably disrupted in the dup15q samples (**Fig. A4.6**), suggesting that transcriptomic regional identity attenuation in the cerebral cortex is shared across heterogenous forms of ASD.

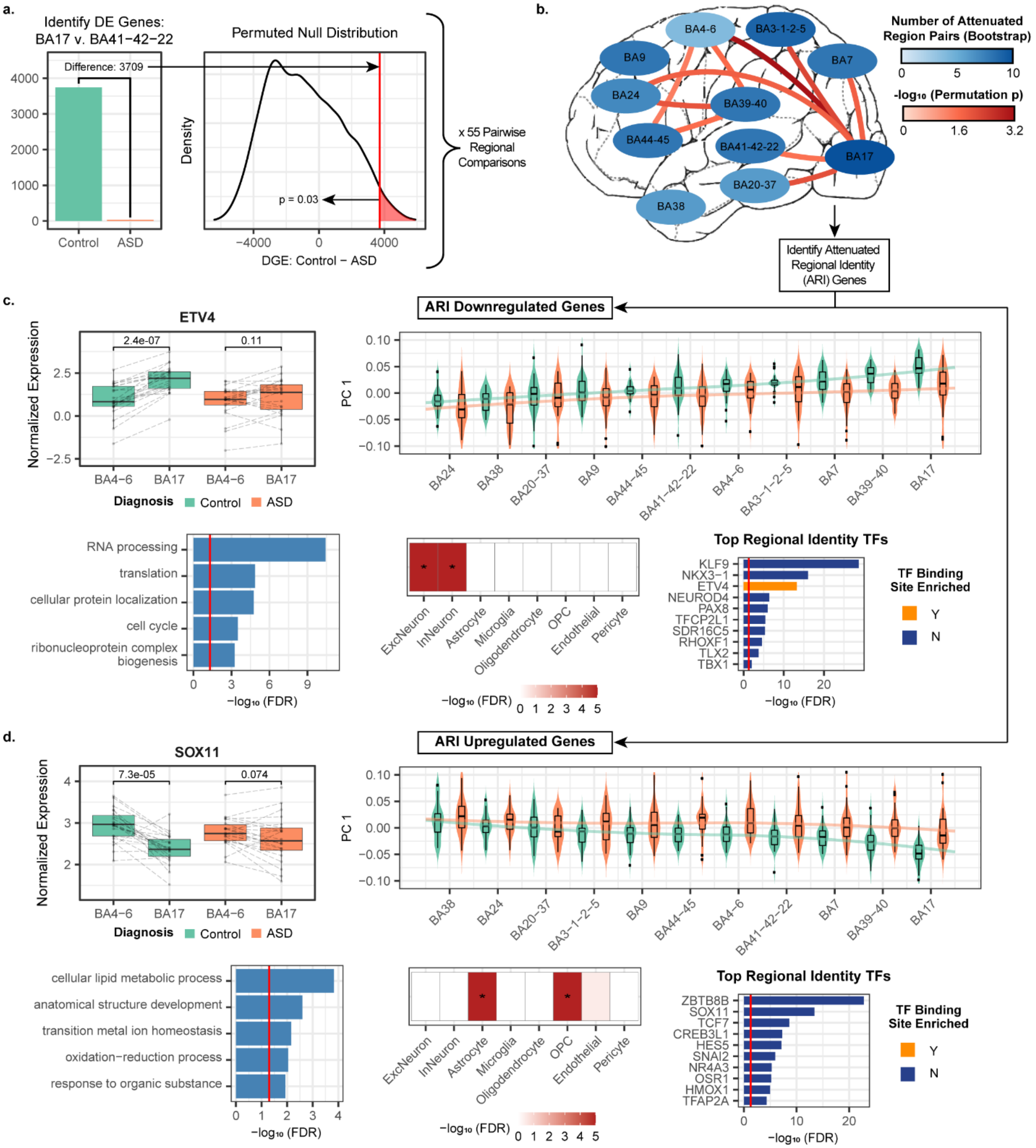


Figure 4.2 | Transcriptomic regional identity attenuation in ASD. a. Methods overview for identifying differences in transcriptomic regional identity in ASD. The regional comparison of BA17 v. BA41-42-22 is used here as an example. Number of DE genes between regions is calculated in controls and ASD samples (left). A permuted null distribution is then used to determine the significance of the difference in DE genes between controls and ASD samples (right). **b.** Regional comparisons with attenuation of transcriptomic regional identity in ASD with $p < 0.05$ are connected with a bar. Attenuated regional identity (ARI) genes are extracted from these regional comparisons (Methods). Cortical lobules are also depicted. **c-d.** Overview of ARI downregulated (c.) and upregulated (d.) genes. Top left, select attenuated transcription factors in BA17 and BA4-6. Lines link paired samples from the same subject, and the paired Wilcoxon signed-rank

test p-value is plotted above boxplots. Top right, PC 1 of ARI genes across all regions. Bottom left and bottom center, gene ontology and cell-type enrichment, respectively. Bottom right, top 10 attenuated transcription factors (TFs), where FDR is representative of how well these TFs distinguish BA17 and BA39-40 from the other nine cortical regions assessed here in controls (Methods). Enrichment for transcription factor binding sites is also depicted (Bonferroni-corrected p-value < 0.05 required for enrichment).

To identify the biological processes contributing to ARI gene dysregulation in ASD, we grouped together all of the ARI genes that were either downregulated (1,881 genes) or upregulated (1,695 genes) with a pronounced posterior effect in ASD (**Methods**). The downregulated set of ARI genes showed broad enrichment for neuronal cell-type-specific markers and RNA processing pathways, and contained many transcription factors (**Fig. 4.2c**, **Table A4.4**). The upregulated ARI genes also contained many transcription factors, and were enriched for oligodendrocyte progenitor cell (OPC) and astrocyte cell-type markers along with metabolic and development pathways. ARI gene dysregulation was further characterized by subsequent co-expression network analysis, which further refined the topology and pathways involved.

4.4: Refining disrupted gene co-expression networks in ASD

We next used weighted gene correlation network analysis (WGCNA)[10] across all samples to partition genes into co-expression modules capturing potentially shared biological functions or regulation (**Methods**). We identified a total of 35 gene modules, of which 9 were downregulated and 15 were upregulated in ASD (**Table A4.5-6**, **Fig. A4.7**). We further generated networks using isoform-level quantifications, identifying 61 isoform modules. Of these, 39 were distinct from the gene modules, with 5 downregulated and 9 upregulated in ASD (**Table A4.5-6**, **Fig. A4.8**). In total, 38 gene and isoform modules were dysregulated in at least one region in ASD. These fell into two broad groups - either dysregulated (1) cortex-wide with comparable magnitude across regions, or (2) with significantly variable magnitude across regions. Again, dup15q effects were similar to ASD effects, but were greater in magnitude (**Fig. A4.7-8**, **Table A4.6**).

4.4.1: Cortex-wide dysregulation observed for ASD risk genes

Thirty-five gene and isoform modules exhibited a consistent pattern of dysregulation in ASD across all cortical regions assessed (linear mixed model, FDR < 0.05; **Fig. 4.3a**, **Fig. A4.7-8**, **Table A4.6**). These include GeneM9, an upregulated neuronal module with a significant enrichment for non-coding genes; GeneM32, a strongly upregulated reactive astrocyte module with the greatest overall magnitude of dysregulation; and GeneM24, a downregulated module enriched for endothelial and pericyte marker genes which are involved in blood-brain-barrier functions (**Fig. 4.3b, Fig. A4.7, Table A4.6**). These modules replicate previous findings of neuronal upregulation, astrocyte reactivity, and BBB disruption in ASD,[1,4–6] but extend these findings by demonstrating that these processes are widespread across the cerebral cortex.

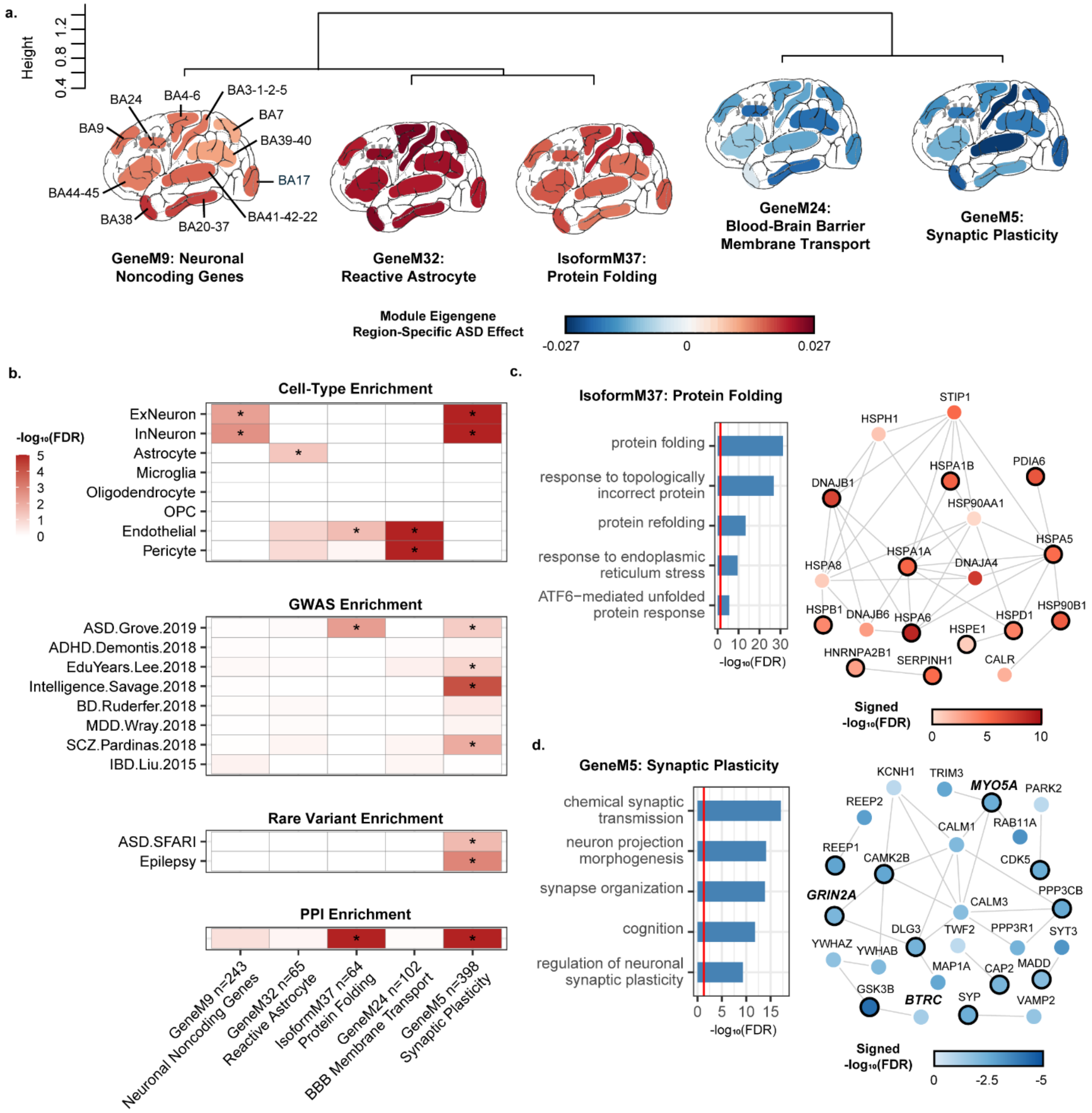


Figure 4.3 | Co-expression network analysis characterizes cortex-wide dysregulation of ASD risk genes. a. Average linkage hierarchical clustering of the biweight midcorrelation of the top 5 most dysregulated gene and isoform co-expression module eigengenes (first principal component of the module) with regionally-consistent patterns of ASD dysregulation. The module eigengene ASD effect is indicated for each cortical region examined (Methods). **b.** -log10(FDR) for cell-type, GWAS, rare variant, and protein-protein interaction (PPI) enrichment for the modules depicted in a. '*' indicates a significant enrichment (FDR < 0.05 for cell-type, rare variant, and PPI enrichment, and FDR < 0.1 for GWAS enrichment). 'n' indicates the number of genes/isoforms in each module. **c-d.** For ASD GWAS enriched modules Isoform M37 (**c**) and GeneM5 (**d**), top gene ontology terms (left) and hub genes (module genes within the top 20 genes with the highest correlation with the module's eigengene) that participate in a protein-protein interaction (PPI) with any other module gene are depicted along with their PPI partners (right). Node color

is the signed -log10(FDR) of the whole cortex ASD effect, edges denote direct PPIs, and hub genes are indicated with a black outline. SFARI database14 gene names are in bold italic.

Two other modules - GeneM5 and IsoM37 - demonstrated cortex-wide dysregulation along with significant enrichment for ASD-associated common genetic variation (**Fig. 4.3b-d**).[11] GeneM5 is down-regulated in ASD, contains many neuronal genes involved in synaptic plasticity, and significantly overlaps with the synaptic module CTX.M16 previously identified by Parikshak et al.[5] (**Fig. 4.3a**, **Fig. 4.3d**, **Table A4.5-6**). GeneM5 is also significantly enriched for genes containing rare *de novo* protein altering mutations associated with ASD, including the high-confidence risk genes *GRIN2A*, *MYO5A*, and *BTRC*[12] (**Table A4.5-6**, **Methods**). This demonstrates convergence of rare and common risk variants on shared biological processes in ASD. GeneM5 is enriched in cortical lower layer 4-6 excitatory neuron cell-type markers (**Fig. A4.7**),[13] identifying them as a point of convergence for rare and common genetic risk in ASD. Finally, IsoM37 is enriched for ASD common genetic risk variants (but not rare mutations), is upregulated in ASD, and contains genes involved in protein folding (**Fig. 4.3a**, **Fig. 4.3c**, **Table A4.6**). To our knowledge, this is the first report of an upregulated ASD transcriptomic signature that is associated with known ASD risk variants.

4.4.2: Magnitude of effect parallels anterior-posterior gradients

In addition to observing profound cortex-wide dysregulation in ASD, we found 13 modules that exhibited their most pronounced ASD effect in BA17, as measured against a permuted distribution containing all regions (**Fig. A4.7**, **Table A4.6**, **Methods**). Of these, 12 showed significant enrichment for ARI genes (half up-regulated and half down-regulated in ASD) and all 13 had anterior - posterior gradients of expression in neurotypical samples, indicating that these modules contribute to transcriptomic regional identities that are observed in neurotypical controls, but attenuated in ASD (**Fig. A4.7**, **Table A4.6**). Six of these modules were more highly expressed in posterior regions in neurotypical subjects and were observed to be downregulated in ASD across the cortex (**Fig. 4.4a**, **Fig. A4.7**, **Table A4.6**). These include GeneM23, an oligodendrocyte-specific module consisting of genes important for organelle regulation and intracellular restructuring; GeneM14, a neuronal module that contains genes involved in neurite morphogenesis and is also strongly downregulated in BA41-42-22; and GeneM3, a neuronal module enriched for energy generation and neuronal processes that are highly energy dependent, such as vesicle transport (**Fig. 4.4b-c**, **Table A4.5-6**). GeneM3 is also significantly enriched for cell-type markers specific to layer 4-5 excitatory neurons (**Fig. A4.7**).[13] The other six modules that were more highly expressed in anterior regions in neurotypical subjects exhibited cortex-wide upregulation in ASD **(Fig. 4.4a**, **Fig. A4.7**, **Table A4.6**). These include GeneM8, a microglial module containing genes involved in immune signaling and phagocytosis; and GeneM7, an immune response module containing genes such as NF-kB and interferon response pathways (**Fig. 4.4b**, **Table A4.5-6**). Although neuronal and oligodendrocyte downregulation along with immune and microglia upregulation have been previously reported in ASD,[1,4–6] these findings indicate that this dysregulation is widespread across the cerebral cortex, with increased magnitude in posterior regions, a pattern most pronounced in BA17.

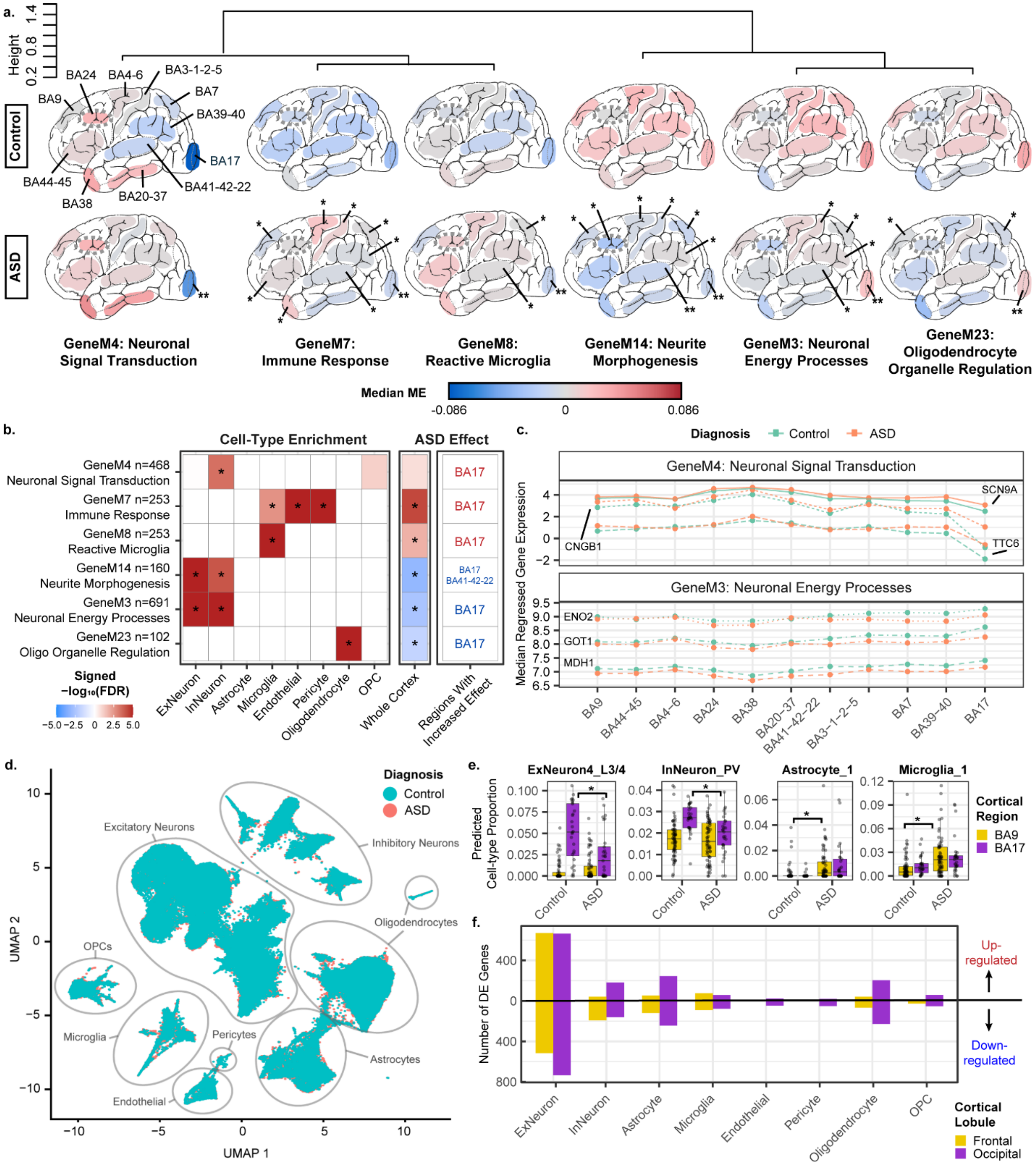


Figure 4 | Functional Characterization of Regionally-variable Transcriptomic Dysregulation in ASD. a. Average linkage hierarchical clustering of the biweight midcorrelation of the top 6 most dysregulated gene co-expression module eigengenes (ME; first principal component of the module) with regionally-variable patterns of dysregulation. The median of the ME, stratified by diagnosis, is depicted for each cortical region examined. Significant region-specific dysregulation in ASD is marked with '*', and regions with a significantly increased magnitude of effect compared to the whole cortex effect is marked with '**' (Methods). **b.** Left and center, signed -log10(FDR) for cell-type enrichment (left) and whole cortex ASD effect for the modules depicted in a (center). '*' indicates a significant enrichment (FDR < 0.05). Right, regions marked with '**' in a are listed. Red text indicates upregulation and blue text indicates downregulation in ASD. 'n' is the number of genes in each module. **c.** Median regressed gene expression for the top three hub genes for GeneM4 (top) and GeneM3 (bottom). **d.** UMAP plot for snRNA-seq, containing matched ASD and neurotypical control samples from both the frontal (BA9 and BA4_6; ASD: 4, Control: 2) and occipital (BA17; ASD:2, Control: 2) cortical lobules. **e.** Predicted neural cell-type proportions obtained from cell-type deconvolution with bulk RNA-seq samples, using snRNA-seq cell-type markers. Cell-types with significant proportion differences in ASD are shown, with significant ASD changes marked with '*'. **f.** DE genes identified with snRNA-seq. Total number of down-regulated (bottom) and up-regulated (top) genes in ASD samples compared to controls is shown (the sum of all cell subtype DE genes within each broad cell-type).

One neuronal module, GeneM4, was observed to be significantly upregulated in ASD only in BA17. GeneM4 contains many genes important for various intracellular signaling and maturation processes, such as *SCN9A* (**Fig. 4.4a-c**, **Table A4.5-6**). Additionally, GeneM4 is significantly enriched for lincRNAs and for previously reported gene modules associated with upregulated pathways related to development[5] and signaling[1,6] in ASD, although we observe this effect in BA17 for the first time (**Fig. A4.7**). We also identified two modules exhibiting strong region-specific dysregulation in regions other than BA17 (**Fig. A4.7**). For example, the module GeneM34, which contains genes involved in cellular stress response regulatory processes, is upregulated with the greatest magnitude in BA4-6 and shows no significant effect in BA17 (**Fig. A4.7**, **Table A4.5-6**). None of the gene modules with regionally-variable magnitudes of ASD effect were significantly enriched for known ASD genetic risk variants.

4.4.3: Cell-type changes mirror regional variation

We finally sought to determine what might be driving the observed changes in magnitude of ASD effect across regions. It is well established that BA17 is the most neuronally dense region in the human brain, with a notable expansion in the thickness of L3/4, compared with other cortical regions.[18] Likewise, there is somewhat of an anterior-posterior gradient in neuronal density observed in mice and primates.[14–17] As such, we posited that regional variation in cell density could be contributing to regional differences in magnitude of ASD effect. Regional neuronal density across multiple brain regions has not been quantitatively studied in the human brain, but such gradients have been established across some regions in non-human primates.[15,16] Therefore, we compared the region-specific ASD effect size changes in our gene modules to regional neuronal nuclei density measured in primates[15] for 6 matched regions across species. We observed a significant association between neuronal density and the effect sizes for several modules dysregulated in ASD (seven with FDR < 0.05, and an additional eight with FDR < 0.1, **Fig. A4.9**, **Table A4.7**). Further, L3/4 thickness was also associated with the region-specific ASD effect sizes in dysregulated modules (**Table A4.7**).

These observations motivated us to perform single-nucleus RNA sequencing (snRNA-seq) in a small cohort of individuals to help evaluate how distinct neural cell-types could be contributing to the regional variance in ASD transcriptomic dysregulation identified with bulk RNA-seq (**Fig. 4.4d**, **Fig. A4.9**, **Table A4.7, Methods**). We sequenced over 150,000 nuclei from ASD and control samples across frontal and occipital cortices with matching bulk RNA-seq. From these data, we identified 35 distinct cell clusters and 4,953 cell-type-specific DE genes in ASD subjects in the frontal and occipital cortex. The vast majority of these were DE in excitatory neurons in both regions, and exhibited larger effects overall in the occipital lobe (**Fig. 4.4f**). While statistical power limited our ability to detect significant cell-type proportion differences between regions or diagnoses (**Methods**), we do observe that excitatory neurons are increased in proportion by ~5% in BA17 across both control and ASD subjects compared to frontal regions (**Fig. A4.9**), corresponding with the primate neuronal density measurements. To predict how cell-type proportions may vary across our entire bulk RNA-seq dataset, next we utilized cell-type markers from our snRNA-seq to perform cell-type deconvolution in all samples (**Methods**). We identified 11 significant cell subtype proportion changes present across six different regions in ASD, characterized by neuronal decreases and astrocyte and microglia increases (**Fig. 4.4e**, **Fig. A4.9**, **Table A4.7**). We also found many anterior-posterior cell-type proportion gradients in control subjects that are attenuated in ASD (**Fig. A4.9**, **Table A4.7**), mirroring patterns observed with our bulk RNA-seq transcriptomic regional identity analysis.

When directly comparing cell-type-specific DE and deconvolved proportional changes in ASD with regional variability in the larger bulk transcriptome sample, we observed a convergent signal within excitatory neurons – in particular, those in L3/4 (ExNeuron4; **Fig. 4.4e-f**, **Fig. A4.9**). Recapitulating the known increase in thickness of L3/4 in BA17 compared with other cortical regions[18], we observed a significant increase in the estimated proportion of the ExNeuron4_L3/4 cluster in posterior regions, peaking in BA17 (**Fig. 4.4e; Fig. A4.9g).** This regional pattern was significantly attenuated in ASD, with a ~2-fold median reduction in estimated Ex4 neuronal proportion in BA17 compared with controls. This cell cluster, with marker genes *RORB*, *PCP4*, *CUX2*, *PHACTR2*, and *EYA4*, also exhibited substantially greater cell-type-specific DE in snRNA-seq profiling from BA17 compared with frontal cortex (90 vs 0 DE genes, respectively; **Fig. A4.9d**). Similarly, BA17 shows a substantially greater upregulation of inhibitory neuron genes in the single cell data, consistent with the observed greater up-regulation of GeneM4 (interneuron) in BA17 (**Fig. 4.4f**). Substantial changes in gene expression are also evident in other cell subtypes (**Fig. A4.9, Table A4.7**), such as microglia_2, which shows a strong and specific increase in DE genes in ASD BA17 compared to frontal regions. These observed intracellular/cell-type changes in neuronal and microglial gene expression are further supported by another snRNA-seq dataset containing a small ASD cohort, which assessed a single region.[19] Here, through performing multi-region snRNA-seq and cell-type deconvolution, we show that predicted cell-type proportions as well as cell-type-specific gene expression profiles are impacted across the ASD cerebral cortex. Importantly, we see increased cell-type-specific transcriptomic dysregulation and lowered neuronal proportions with a notable convergence within L3/4 excitatory neurons in ASD BA17, a region where neuronal proportions are neurotypically abundant. These changes likely contribute to the pronounced ASD effect we observe with bulk RNA-seq in this region.

4.5: Discussion

Overall, the findings presented here substantially expand our understanding of ASD pathology beyond the previously established 'downregulated neuron' and 'upregulated glia/immune' functional categories observed in frontal and temporal lobe. We identify gene and isoform expression changes in ASD that extend across the cerebral cortex, many neural cell-types, and specific biological processes (**Fig. A4.10**).[1,4–6] We find that the recently observed reactive astrocyte upregulation and blood-brain barrier membrane transport downregulation[1] is extended cortex-wide in ASD. Furthermore, we find that other dysregulated pathways observed before in ASD - particularly upregulated immune response and reactive microglia genes, along with downregulated neurite morphogenesis and neuronal energy pathway genes - are not only impacted cortex-wide in ASD, but impacted in a regional gradient that reflects fundamental elements of cortical cytoarchitecture, such as neuronal density. It is also notable that the magnitude of region-level differences in ASD parallels regional variance in attenuation of transcriptomic identity, suggesting that they reflect related processes. That the gradient of region-specific changes between ASD and controls coincides with both neuronal proportion differences and cell-type-specific transcriptomic dysregulation further suggests that the interplay of cytoarchitecture and cell-type gene expression, rather than a single one of these features, influences our ability to observe transcriptomic changes in bulk tissue. Given the connection between regional cytoarchitecture, local circuits and long-range brain connectivity,[20,21] parsimony suggests that in addition to developmental patterning contributions,[5,22] the diminution of transcriptomic regional identity reflects changes in local neuronal circuit dysfunction and deficits in synaptic plasticity and homeostasis that are widely propagated.[20] This is supported by our observation that the gene co-expression module representing synaptic plasticity genes is downregulated cortex-wide and is significantly enriched with common and rare ASD genetic risk variants, further emphasizing that synaptic plasticity is a convergent pathway in ASD. Given this result, along with our observations of profound neuronal dysregulation present throughout the ASD cortex, future work should determine which specific aspects of synaptic plasticity may contribute to causal mechanisms in the disorder across specific brain regions and developmental timepoints.

Several additional factors should guide the interpretation of these results. The samples utilized in this work were obtained from heterogeneous postmortem cortical tissue, meaning that the results reported here are broadly applicable to the postnatal ASD cortex across both sexes and a span of ages from two to 68 years old, and they should be interpreted in this context. Rigorous methodology was utilized at every step to account for biological and technical variability, ensuring that the results reported here are conservative and widely applicable. Additionally, bulk tissue RNA-seq, in contrast to single cell and nucleus RNA-seq, does not have the cellular resolution to assess dissection variability across cortical regions and cell-type specificity of transcriptomic changes. We addressed this by performing snRNA-seq, which significantly enhanced our understanding of regional variation in ASD transcriptomic dysregulation. However, snRNA-seq also has its own limitations. While snRNA-seq can profile tens of thousands of cells, snRNA-seq experiments typically have fewer unique samples than bulk RNA-seq experiments, and the comparability of snRNA-seq cell-type proportions to true sample cell-type proportions is currently unclear.[23] It is also challenging to estimate isoform quantifications with single cell RNA-seq approaches, whereas this remains a strength of bulk tissue RNA-seq.[24] Leveraging this, we subsequently identified an upregulated isoform-specific co-expression module enriched with ASD GWAS variants, implicating increased protein folding dysfunction for the first time as a putative pathway contributing to ASD causal mechanisms. Interestingly, upregulated proteostasis is also implicated in Down's Syndrome,[25,26] supporting that protein folding machinery may be an affected biological process in multiple neurodevelopmental disorders. The utilization of methods that have greater cellular resolution is necessary for the improved and continued mapping of the results presented here to specific cortical cell-types. As we seek to gain a complete understanding of ASD neural pathology, future approaches which integrate different sources of biological data - including this cortex-wide transcriptomic resource - to determine how ASD risk genes are acting in the brain will be essential.

4.6: Materials and Methods

Sample Acquisition and Preparation for RNA-seq

Postmortem cortical brain samples were acquired from the Harvard Brain Bank as part of the Autism BrainNet project (formerly the Autism Tissue Project, ATP) and the University of Maryland Brain Banks (UMDB). A total of 842 samples from subjects with ASD, dup15q syndrome, and non-psychiatric controls (112 unique subjects) across 11 cortical regions encompassing all major cortical lobes – frontal: BA4/6, BA9, BA44/45, BA24; temporal: BA38, BA41/42/22, BA20/37; parietal: BA3/1/2/5, BA7, BA39/40; and occipital, BA17 - were acquired. These included 253 samples previously published in Parikshak et al., Nature 2016[5] from BA9 and BA41/42/22 and/or Gandal et al., Science 2018b[1,5] from BA9, BA4/6, and BA41/42/22. An ASD diagnosis was confirmed by the Autism Diagnostic Interview-Revised (ADIR) in 30 of the subjects. In the remaining 19 subjects, diagnosis was supported by clinical history. Frozen brain samples were stored at -80 deg C. To extract RNA from these samples, first approximately 50-100mg of tissue were dissected from the cortical regions of interest on dry ice in a dehydrated dissection chamber to reduce degradation effects from sample thawing or humidity. Then, RNA was isolated from each sample using the miRNeasy kit with no modifications (Qiagen). For each RNA sample, RNA quality was quantified using the RNA Integrity Number (RIN) on an Agilent Bioanalyzer.

RNA-seq and RNA Data Processing

Initial sequencing in BA9 and BA41/42/22 was performed in three batches as published by Parikshak et al., Nature 2016.[5] The remaining regions, along with additional BA9 and BA41/42/22 samples, were sequenced across three new batches. For all of these batches, strand-specific RNA-seq libraries were prepared. For the first two batches, the TruSeq Stranded Total RNA sample prep kit with RiboZero Gold (Illumina) was used to obtain rRNA-depleted libraries. The remaining batch was prepared with the TruSeq RNA Exome sample prep kit (formerly the TruSeq RNA Access sample prep kit; Illumina). All libraries were randomly pooled to multiplex 24 samples per lane using Illumina TruSeq barcodes. Each lane was sequenced five times on an Illumina HiSeq 2500 or 4000 instrument using high output mode with standard chemistry and protocols for 50, 69, or 100 bp paired-end reads (read length varied by batch) to achieve a target depth of 70 million reads.

After sequencing, the resulting sample FASTQ files from all batches (including the Parikshak et al.[5] samples) were subjected to the same processing pipeline. First, FASTQ files were assessed with FastQC[27] (v0.11.2) to verify that quality was sufficient for further processing. FASTQ files were then aligned to the human reference genome (GRCh37[28] Ensembl v75) with STAR[29] (v2.5.2b). Picard tools[30] (v2.5.0) was used with the resulting BAM files to collect various read quality measures, in addition to the quality measures collected by STAR. verifyBAMID[31] was also used with these BAM files along with known sample genotypes from Parikshak et al.[5] to validate that sample identity was correct for all BAM files. Additionally, the expression of XIST (a female-specific gene) was assessed to contribute to sample identity verification. Finally, RSEM[32] (v1.3.0) was used for quantification (Gencode[33] release 25lift37) to obtain expected read counts at the gene and isoform levels.

Expected gene and isoform read counts were then subjected to several processing steps in preparation for downstream analysis, mainly using R.[34] First, Counts Per Million (CPM) were obtained from counts for gene and isoform filtering purposes. Genes and isoforms were filtered such that genes/isoforms with a CPM > 0.1 in at least 30% of samples were retained. Genes/isoforms were also removed which had an effective length (measured by RSEM) of less than 15 bp. Isoforms were additionally filtered such that all isoforms corresponded with genes in the gene-level analysis. The counts for the remaining genes (24,836) and isoforms (99,819) passing these filters were normalized using the limma-trend approach in the limma[35] R package. Briefly, the limma-trend approach obtains normalized expression data through taking the $\log_2$(CPM) of read counts with an adjustment for sample read depth variance. An offset value calculated with CQN[36] accounting for GC content bias and gene/isoform effective length bias in read quantification was also incorporated during the normalization process. With this normalized expression data, sample outliers were identified in each sequencing batch by cortical lobe (frontal, parietal, temporal, and occipital) group that had both (1) an absolute z-score greater than 3 for any of the top 10 expression principal components (PCs) and (2) a sample connectivity score less than -2. Sample connectivity was calculated using the fundamentalNetworkConcepts function in the WGCNA[10] R package, with the signed adjacency matrix (soft power of 2) of the sample biweight midcorrelation as input. This process identified 34 outliers, resulting in a final total of 808 samples (341=Control, 384=ASD, 83=dup15q) which were carried forward for analysis.

Evaluating Previous Co-Expression Modules and ASD DE Genes/Isoforms Cortex-wide

Linear models for all subsequent analyses are described in the Appendix (Extended Methods, section A4.1).

To determine how gene co-expression modules previously identified in Parikshak et al.[5,35] and Voineagu et al.[4] were effected across distinct cortical regions, we first created a regressed gene expression dataset that only contained the effects of biological covariates (subject, diagnosis, region, sequencing batch, sex, ancestry, age, and age^2). This regressed dataset was created with the 'lmerTest'[37] package in R through subtracting the effects of technical covariates from each gene, leaving only the random intercept, biological covariate effects, and the residual. ASD-associated module eigengene region-specific ASD effects were identified using contrasts (eg. Control_BA17 - ASD_BA17) with the limma[34] R package with this regressed expression dataset, accounting for all biological covariates. Region-specific contrasts with a p-value < 0.05 were considered significant (FDR-correction was unwarranted since only eight module eigengenes were examined).

To identify genes and isoforms dysregulated in ASD both within specific regions and cortex-wide, the limma[35] R package was applied with the gene and isoform expression data using our full gene and isoform models (both biological and technical covariates). The standard limma[35] workflow was implemented as recommended for linear mixed models. Region-specific dysregulation was identified as described above for the Parikshak et al.[5] and Voineagu et al.[4] modules. Whole cortex dysregulation was established through subtracting the sum of the ASD region-specific effects from the sum of the Control region-specific effects. For both region-specific and whole cortex effects, genes and isoforms with an FDR-corrected p-value < 0.05 were considered significantly dysregulated. dup15q region-specific and whole cortex dysregulation was also established in this manner. The fixed effects of sex, age, and age^2 were also acquired (shared in **Table A4.3**) using the full gene and isoform models.

The methodology used to evaluate region-specific ASD effects compared to whole cortex ASD effects is described in the Appendix (Extended Methods, section A4.1) Methods.

Transcriptomic Regional Identity Analysis

To identify differentially expressed genes and isoforms between all 55 pairs of cortical regions, a regressed gene expression dataset containing only the random effect of subject and the fixed effects of diagnosis and region (along with the model residual) was used. Regression was performed as described for evaluation of previously identified co-expression modules. Significant attenuation of DE genes between each pair of regions (a reduction in transcriptomic regional identity differences) in ASD was established through the following process. (1) ASD and

Control subjects containing each region in the regional pair were extracted for use in the analysis. (2) Separately in ASD and Control subjects, the number of DE genes between regions was calculated using the paired Wilcoxon signed-rank test. Genes with an FDR-corrected p-value < 0.05 were considered DE. (3) The difference in the number of DE genes between regions for ASD v Control subjects was calculated (the 'true' difference). (4) A permuted distribution of the difference in DE genes between regions for ASD v Control subjects was generated to test the 'true' difference. Each permutation (10,000 in total) randomly assigned 'ASD' and 'Control' status to subjects, but kept the number of ASD and Control subjects consistent with the true number of ASD and Control subjects. (5) A two-tailed p-value was obtained from testing the 'true' difference against the permuted distribution. If the regional comparison p-value < 0.05, with the number of DE genes between regions in ASD less than that in Controls, then the regional comparison was considered significantly attenuated in ASD. Otherwise, the regional comparison was considered over-patterned in ASD. This procedure was repeated with isoform level regressed gene expression data (similarly, only containing the random effect of subject and the fixed effects of diagnosis and region, along with the model residual) to identify altered transcriptomic identities in ASD at the isoform-level.

The previously described permutation approach was designed to identify differences in transcriptomic regional identity in ASD. Importantly, this method is not appropriate for assessing variance in expected numbers of DE genes between regions across regional pairs and diagnoses, since the number of ASD and Control subjects varied across regional pairs. To examine this, for each regional comparison we subset to 10 pairs of ASD and Control subjects (10 was selected since every regional comparison had at least this many subjects). When subsetting, subjects were removed such that the remaining subjects were closest in age to the median age of the available samples for that regional comparison. A bootstrap approach was then used to calculate the number of DE genes (p-value < 0.05) between regions separately in Control and ASD subjects through sampling subjects with replacement (mean taken across 10,000 bootstraps). The same regressed expression dataset used for the permutation approach was utilized for this bootstrap analysis. Any regional comparison in which the number of DE genes between regions was less in ASD than in Control subjects was considered trending towards attenuation in ASD.

To validate our bootstrapped estimates for the number of DE genes between pairs of regions in Controls, we compared these estimates to those of the Allen Brain Atlas[9], which is the best publicly available work for comparison. Allen Brain Atlas regions were matched to Brodmann regions (**Table A4.4**) and matching regional pairs were extracted for comparison with this work. When the Allen Brain Atlas had two or more regional pairs matching one regional pair in this work, the mean was taken across the Allen Brain Atlas regional pairs. A p-value for the association of the number of DE genes between regions in Controls obtained in this work compared to the Allen Brain Atlas was calculated from a linear model (cortex-wide bootstrap mean ~ allen brain atlas mean).

We applied a stringent filtering process to identify high-confidence attenuated regional identity (ARI) genes from each significantly attenuated regional comparison identified with the permutation procedure described above. First, for each of the attenuated regional comparisons, we extracted the genes which were identified as DE between regions in subjects labeled as Controls in each of the 10,000 permutations. Then, we calculated how many times each of the genes truly DE between pairs of regions in the Control subjects were present in their respective permuted groups (ranging from a possible 0 to 10,000 occurrences). Those ‘true’ DE genes which were present in less than 95% of their respective permutations were retained as ARI genes for each attenuated regional comparison. For each set of ARI genes (ten total), each gene was matched to the region in which it had higher expression in Control subjects. The paired

Wilcoxon signed-rank p-values identified for these genes in Controls (those subjects used for the permutation analysis) were also extracted and are shared in **Table A4.4**.

ARI gene groups (ARI downregulated genes, those highly expressed in BA17 and BA39-40 relative to other regions in Controls; ARI upregulated genes, those lowly expressed in BA17 and BA39-40 relative to other regions in Controls) were created through taking the union (without duplicates) across all ten identified ASD-attenuated regional comparisons, and sorting genes into the two groups based on gene expression profiles across regions. The details of this process are described in the Appendix (Extended Methods, section A4.1), along with functional annotation procedures.

Network-Based Functional Characterization

Standard workflows, as previously described in Parikshak et al.[5] and Gandal et al.,[1] were followed (with minor modifications) to identify gene and isoform co-expression modules using Weighted Gene Co-Expression Network Analysis (WGCNA).[10] Details regarding network formation, module identification, and module functional characterization are described in the Appendix (Extended Methods, section A4.1).

snRNA-seq and Cell-type Deconvolution

Cell types were annotated based on expression of known marker genes visualized on the UMAP plot, violin plots, and by performing unbiased gene marker analysis. To gain insight into the regional enrichment or diagnostic enrichment of cell types, the relative proportion of the number of nuclei in each cell type was normalized to the total number of nuclei captured from each library. Average cell-type proportions and standard errors (across libraries) were scaled such that each Lobule x Diagnosis group sums to 100%, so that cell-type proportions in these groups could be fairly compared across all cell-types. To determine if any changes in cell-type proportion were statistically significant, we implemented scDC[38] to bootstrap proportion estimates for our samples (**Table A4.7**). We employed a linear mixed model (random effect of subject) to determine if any changes in cell-type proportion were present across regions and diagnoses. None of the model covariates were statistically significant ($p > 0.05$ for all model covariates). However, we did find several significantly different predicted cell-type proportions in ASD with cell-type deconvolution analysis. We describe methods for cell-type deconvolution in detail in the Supplementary Methods (Extended Methods, section A4.1). To identify genes differentially expressed in ASD compared to control in each cell type, the non-parametric Wilcoxon rank sum test was applied including gene detection rate and sequencing depth within the model. We compared frontal cortex ASD cells to frontal cortex control cells within each cluster and likewise for the occipital cortical cells. The bars in **Figure 4.4e** are the summation of all differentially expressed genes identified in each cell subtype for the broader cell-type (eg. all excitatory neuron subtype DE genes are summed to obtain the number of DE genes in the broad excitatory neuron cell class). Further details regarding the snRNA-seq analysis are included in the Supplementary Methods (Extended Methods, section A4.1).

4.7: Bibliography

1. Gandal, M. J. *et al.* Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. *Science* **362**, (2018).
2. Wu, Y. E., Parikshak, N. N., Belgard, T. G. & Geschwind, D. H. Genome-wide, integrative analysis implicates microRNA dysregulation in autism spectrum disorder. *Nat. Neurosci.* **19**, 1463–1476 (2016).
3. Sun, W. *et al.* Histone Acetylome-wide Association Study of Autism Spectrum Disorder. *Cell* **167**, 1385–1397.e11 (2016).
4. Voineagu, I. *et al.* Transcriptomic analysis of autistic brain reveals convergent molecular

pathology. *Nature* **474**, 380–384 (2011).

5. Parikshak, N. N. *et al.* Genome-wide changes in lncRNA, splicing, and regional gene expression patterns in autism. *Nature* **540**, 423–427 (2016).

6. Gandal, M. J. *et al.* Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap. *Science* **359**, 693–697 (2018).

7. Sullivan, P. F. & Geschwind, D. H. Defining the Genetic, Genomic, Cellular, and Diagnostic Architectures of Psychiatric Disorders. *Cell* **177**, 162–183 (2019).

8. de la Torre-Ubieta, L., Won, H., Stein, J. L. & Geschwind, D. H. Advancing the understanding of autism disease mechanisms through genetics. *Nat. Med.* **22**, 345–361 (2016).

9. Hawrylycz, M. J. *et al.* An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* **489**, 391–399 (2012).

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

11. Grove, J. *et al.* Identification of common genetic risk variants for autism spectrum disorder. *Nat. Genet.* **51**, 431–444 (2019).

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

13. Hodge, R. D. *et al.* Conserved cell types with divergent features in human versus mouse cortex. *Nature* vol. 573 61–68 (2019).

14. Keller, D., Erö, C. & Markram, H. Cell Densities in the Mouse Brain: A Systematic Review. *Frontiers in Neuroanatomy* vol. 12 (2018).

15. Collins, C. E., Airey, D. C., Young, N. A., Leitch, D. B. & Kaas, J. H. Neuron densities vary across and within cortical areas in primates. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 15927–15932 (2010).

16. Collins, C. E. *et al.* Cortical cell and neuron density estimates in one chimpanzee hemisphere. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 740–745 (2016).

17. Cadwell, C. R., Bhaduri, A., Mostajo-Radji, M. A., Keefe, M. G. & Nowakowski, T. J. Development and Arealization of the Cerebral Cortex. *Neuron* **103**, 980–1004 (2019).

18. Wagstyl, K. *et al.* BigBrain 3D atlas of cortical layers: Cortical and laminar thickness gradients diverge in sensory and motor cortices. *PLoS Biol.* **18**, e3000678 (2020).

19. Velmeshev, D. et al. Single-cell genomics identifies cell type-specific molecular changes in autism. Science 364, 685–689 (2019).

20. Ecker, C. *et al.* Intrinsic gray-matter connectivity of the brain in adults with autism spectrum disorder. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 13222–13227 (2013).

21. Hilgetag, C. C., Beul, S. F., van Albada, S. J. & Goulas, A. An architectonic type principle integrates macroscopic cortico-cortical connections with intrinsic cortical circuits of the primate brain. *Netw Neurosci* **3**, 905–923 (2019).

22. Parikshak, N. N. *et al.* Integrative Functional Genomic Analyses Implicate Specific Molecular Pathways and Circuits in Autism. *Cell* vol. 155 1008–1021 (2013).

23. Patrick, E. *et al.* Deconvolving the contributions of cell-type heterogeneity on cortical gene expression. *PLoS Comput. Biol.* **16**, e1008120 (2020).

24. Arzalluz-Luque, Á. & Conesa, A. Single-cell RNAseq for the study of isoforms—how is that possible? *Genome Biology* vol. 19 (2018).

25. Zhu, P. J. *et al.* Activation of the ISR mediates the behavioral and neurophysiological abnormalities in Down syndrome. *Science* **366**, 843–849 (2019).

26. Halliday, M. & Mallucci, G. R. Translating translation in Down syndrome. *Science* vol. 366 797–798 (2019).

27. Andrews, S. *et al.* FastQC. (2010).

28. Yates, A. D. *et al.* Ensembl 2020. *Nucleic Acids Res.* (2019) doi:10.1093/nar/gkz966.

29. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

30. Picard Toolkit. *Broad Institute, GitHub repository*.

31. Jun, G. *et al.* Detecting and Estimating Contamination of Human DNA Samples in Sequencing and Array-Based Genotype Data. *The American Journal of Human Genetics* vol. 91 839–848 (2012).

32. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).

33. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research* vol. 47 D766–D773 (2019).

34. Website. R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

35. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).

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

37. Kuznetsova, A., Brockhoff, P. B. & Christensen, R. H. B. lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software* vol. 82 (2017).

38. Cao, Y. *et al.* scDC: single cell differential composition analysis. *BMC Bioinformatics* **20**, 721 (2019).

CHAPTER FIVE

Conclusions and future directions

5.1: Conclusions

In this work, I demonstrated how transcriptomic analyses can measurably advance psychiatric disorder research. In both cross-disorder studies (Chapters 2 and 3), shared and distinct gene expression changes across disorders are connected to psychiatric genetic risk variants and signatures of specific cell-types and biological processes. We find a broad disruption of neuronal and synaptic processes as well as upregulation of astrocyte-specific transcriptional programs. In contrast, microglia cell-type markers are distinctly upregulated in ASD, differentiating it from the other psychiatric disorders surveyed. Integrating genomic and transcriptomic data further refines our understanding of multiple psychiatric disorders, as this approach narrows in on genetic variants contributing to observed molecular pathology in the brain. Finally, I show how multi-region transcriptomic analyses in ASD uncovered shared molecular disruptions that extend across the cerebral cortex, and that these changes strongly implicate disrupted neuronal synaptic plasticity mechanisms in ASD. All of these transcriptomic experiments exemplify how core concepts and approaches in psychiatric neurogenetics – strategic experimental design with robust statistical control, dimensionality reduction and feature selection through integration of orthogonal biological datasets, and unbiased interrogation of gene expression changes within a cellular and molecular neuroscience framework – can substantially enhance our understanding of psychiatric disorders.

Strategic experimental design enables transcriptomic experiments to profile different axes of biological variation relevant to psychiatric disorder research. In the work presented here, samples were selected to expand our understanding of psychiatric disorders across two major axes of biological variation: a cross-disorder axis, and a spatial/regional axis. Interrogating how the transcriptome varies across these axes both enhanced our ability to interpret previous findings and revealed new psychiatric molecular pathology, leading to valuable biological insights. For example, in chapter four we saw that previously observed decreased neuronal gene expression in ASD was expanded across the cerebral cortex, and also found that this effect was greatest in the occipital cortical lobule. The occipital cortical lobule is the most neuronally dense region in the brain, with a prominent expansion of Layer 3/4 neurons compared with other regions of the cortex.[1-5] These results emphasized that neuronal dysregulation is a core component of ASD pathology. The cross-disorder experiments in chapters two and three also enhanced and expanded our knowledge of multiple psychiatric disorders through comparing gene expression changes across these disorders. All of these works show that characterizing psychiatric transcriptomic pathology across orthogonal axes of biological variation – such as the neurodevelopmental/temporal, spatial/regional, cross-disorder, and cellular axes – substantially improves our understanding of psychiatric disorders. Through profiling every facet of the psychiatric transcriptome, we enhance our ability to draw meaningful functional insights from gene expression changes and narrow in on relevant psychiatric disorder mechanisms. Data-driven, empirical analyses that continue to survey the transcriptome across spatiotemporal, multi-disorder, and cellular axes will surely continue to empower psychiatric disorder research.

The work presented here also relied on the integration of different bioinformatic techniques, datasets, and approaches to obtain a fuller picture of how cell-types and biological processes are altered in psychiatric disorders. While transcriptomic analyses are informative, they are restricted to quantifying a single molecular feature, RNA, and therefore do not capture the entire landscape of molecular changes occurring in samples, limiting how much we can learn about biological systems with this approach. Combining transcriptomic profiles with genomic, epigenomic, proteomic, and other empirical biological assays refines our understanding of how entire biological systems are impacted by psychiatric disorders. For example, in chapter three genomic data is combined with transcriptomic data to predict how genetic variants may contribute to psychiatric gene expression dysregulation. Polygenic risk score comparisons connect common genetic variants with genes exhibiting psychiatric transcriptomic dysregulation, highlighting specific regulatory mechanisms that may contribute to causal pathology in psychiatric disorders. In addition to these findings, we see that other integrative ‘omics analyses also advance experimental aims in other chapters. In chapter two, primary findings made with a gene expression microarray metanalysis were validated with RNA-seq analysis, supporting the robustness of those findings. Additionally, in chapter four, the integration of single nucleus RNA-seq with bulk RNA-seq showed that increased gene expression dysregulation in the occipital cortical lobule was associated with differential gene expression increases in occipital cell-types relative to frontal region cell-types. Across all of these experiments, we see that integrating different types of data and methodologies provides a deeper understanding of biological systems, enabling these analyses to hone in on major components of psychiatric disorder mechanisms.

Reflecting on the main results of the work presented here, it is evident that combining data-driven analyses with hypotheses rooted in prior knowledge was essential for elucidating major findings. To emphasize a key finding, in chapters two and four we see that neuronal downregulation and astrocyte upregulation across ASD, schizophrenia, and bipolar disorder is apparent, along with a distinct microglial upregulation signature in ASD. In ASD, we find that these effects are present cortex-wide, with a general increase in magnitude of effect in the occipital cortical lobule relative to all other regions examined.

In further analyzing patterns of neuronal gene downregulation across the cerebral cortex in ASD, we find that some of these genes are involved in synaptic plasticity processes, and that common and rare genetic risk variants for ASD (such as *GRIN2A*, *MYO5A*, and *BTRC*) converge on these specific genes. Additionally, we find that many downregulated neuronal genes, including several involved in neuronal energetic pathways, exhibit the greatest ASD gene expression effects in the occipital region. With the knowledge that this region is one of the most neuronally dense in the brain, we find that the magnitude of region-specific ASD neuronal gene expression is positively associated with regional neuronal density, and additionally with single nucleus RNA-seq we find that occipital cell-types are particularly severely dysregulated compared to frontal regions. Through integrating differential gene expression, ASD GWAS and rare variants, and measures of neuronal density and cell-type-specific transcriptomic dysregulation, we gain a fuller picture of the ASD brain, and hone in on neurons (and synaptic plasticity regulation in particular, as ASD risk variants converge on genes associated with this biological process) as a critical cell-type in ASD disorder mechanisms.

Finally, to review upregulated gene expression signatures, we find that NF-kB pathways, particularly in astrocytes, along with interferon pathways are strongly upregulated in ASD and schizophrenia, with ASD showing the greatest dysregulation signature. This shows that immune reactivity pathways and astrocytes contribute strongly to psychiatric disorder mechanisms. Additionally, the robust observation of microglial activation in ASD alone implicates this cell-type in pathological processes specific to ASD. Together, all of these major findings demonstrate how empirical transcriptomic methodologies guided by prior knowledge of neurodevelopment, psychiatric genetics, and neural cytoarchitecture have significantly advanced our understanding of psychiatric disorders.

5.2: Future directions

As we seek to continually improve our understanding of psychiatric disorders and elucidate the core components of psychiatric neuropathology, the refinement and improvement of methodologies for assaying different types of biological data will be essential. The development of long read RNA-sequencing[8] methods will enhance isoform and differential splicing quantification methods, improving our ability to identify transcriptomic dysregulation in psychiatric disorders, as evidenced extensively in chapter three. For single cell and nucleus RNA-seq, as well as the newer spatial RNA-seq[9] technology that can localize RNA in histological slices, increasing read depth and expanding RNA capture beyond poly-A reads will allow for single cell, nucleus, and spatial RNA-seq methods to capture the diversity of the transcriptome and provide increasingly specific insights into psychiatric transcriptomic pathology. For psychiatric GWAS, and for ASD GWAS in particular, greater sample sizes are needed to identify missing heritability and better assign psychiatric risk to genetic variants. This will not only improve our basic understanding of psychiatric genomics, but will also improve the accuracy of bioinformatic integrative analyses that utilize GWAS results. Improved phenotyping for GWAS, and generally for all biological data assays, will enable us to investigate how the genome may contribute to distinct phenotypes within psychiatric disorders, improving our ability to understand how molecular pathology and resultant behaviors may vary across these distinct psychiatric phenotypes. Expanding the types of empirical biological assays at our disposal, while challenging, will also enhance our ability to perform integrative bioinformatic analyses. As I have demonstrated here, these analyses can greatly improve our ability to understand psychiatric disorder mechanisms. Thus, expanding access to different types of biological assays – such as empirical proteomic assays – will support these efforts. Finally, improving methodologies for integrating genomic, transcriptomic, epigenomic, etc. data together will also serve to refine psychiatric bioinformatics experiments and provide fuller pictures of psychiatric biological systems.

In addition to bioinformatics method development, improved experimental and model systems are needed to investigate the genes, cell-types, and biological pathways implicated in psychiatric disorder mechanisms with analyses such as those presented in this work. To support the development of all psychiatric model organisms and systems, we must enhance our understanding of human neurodevelopment. Only with a thorough understanding of human neurodevelopment will we be capable of comprehensively evaluating how psychiatric genetic risk variants contribute to psychiatric disorders. Continued experiments and bioinformatic analyses utilizing human fetal tissue and *in vitro* models utilizing hESCs and iPSCs, such as 3D neural spheroids,[10] will ultimately achieve this objective. Refining *in vitro* models to replicate typical neurodevelopment as closely as possible will not only directly improve our understanding of neurodevelopment, but it will also provide a suitable model system for evaluating genes, cell-types, and biological processes implicated in psychiatric molecular pathology. Work incorporating glial cell-types and vasculature into 3D neural spheroid models is particularly promising,[6,7] as these models will better replicate true human neural systems. Mice and other model organisms that can be evaluated *in vivo* will also contribute to improving our understanding of psychiatric disorder mechanisms, especially for efforts to understand complete neural systems and altered brain connectivity. Continuing efforts to compare and contrast different aspects of the human brain with that of these model organisms – through transcriptomic experiments and other means – will be essential for planning experiments in these organisms that can truly recapitulate relevant characteristics of human psychiatric disorders. Efforts to find human brain imaging correlates of underlying psychiatric disorder mechanisms, such as with fMRI techniques,[11] will also advance our understanding of psychiatric disorders at the level of whole brain connectivity. Overall, refining experimental model systems, improving technologies for investigating those models, and evaluating how well these technologies capture true underlying biology will all support the advancement of psychiatric disorder research.

5.3: Bibliography

1. Keller, D., Erö, C. & Markram, H. Cell Densities in the Mouse Brain: A Systematic Review. *Frontiers in Neuroanatomy* vol. 12 (2018).
2. Collins, C. E., Airey, D. C., Young, N. A., Leitch, D. B. & Kaas, J. H. Neuron densities vary across and within cortical areas in primates. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 15927–15932 (2010).
3. Collins, C. E. *et al.* Cortical cell and neuron density estimates in one chimpanzee hemisphere. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 740–745 (2016).

4. Cadwell, C. R., Bhaduri, A., Mostajo-Radji, M. A., Keefe, M. G. & Nowakowski, T. J. Development and Arealization of the Cerebral Cortex. *Neuron* **103**, 980–1004 (2019).

5. Wagstyl, K. *et al.* BigBrain 3D atlas of cortical layers: Cortical and laminar thickness gradients diverge in sensory and motor cortices. *PLoS Biol.* **18**, e3000678 (2020).

6. Cakir, B., Xiang, Y., Tanaka, Y. et al. Engineering of human brain organoids with a functional vascular-like system. *Nat Methods* **16**, 1169–1175 (2019).

7. Song L, Yuan X, Jones Z, Vied C, Miao Y, Marzano M, Hua T, Sang QA, Guan J, Ma T, Zhou Y, Li Y. Functionalization of Brain Region-specific Spheroids with Isogenic Microglia-like Cells. *Sci Rep* **9:11055**, (2019).

8. Oikonomopoulos S, Bayega A, Fahiminiya S, Djambazian H, Berube P, Ragoussis J. Methodologies for Transcript Profiling Using Long-Read Technologies. *Front Genet* **11***,* 606 (2020).

9. Moncada, R., Barkley, D., Wagner, F. *et al.* Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nat Biotechnol* **38,** 333–342 (2020).

10. Sloan SA, Andersen J, Pașca AM, Birey F, Pașca SP. Generation and assembly of human brain region-specific three-dimensional cultures. *Nat Protoc* **13(9)**, 2062-2085 (2018).

11. Christopher deCharms, R. Applications of real-time fMRI. *Nat Rev Neurosci* **9,** 720–729 (2008).

APPENDIX

A1: Supplementary Materials for Chapter 2

A1.1: Extended Materials and Methods

Raw Data

Raw microarray gene expression data from 700 post-mortem cortical brain samples across 14 studies of multiple neuropsychiatric disorders was obtained from the Gene Expression Omnibus (GEO), ArrayExpress, or from the study authors directly (see **Table A1.1;** (*3-11, 75, 81-82*)). Each study was processed separately and analyzed according to the general workflow as described below. Only data from cortical samples was used (except in Inflammatory Bowel Disease datasets).

Quality Control and Normalization

Illumina microarrays were $\log_2$ transformed and quantile normalized using the *lumi* package in R (*26*). Affymetrix microarrays were RMA normalized (background correction, $\log_2$ transformation, quantile normalization, and probe summarization) using the *affy* package in R (*27*). Agilent microarrays were normalized using the *limma* package in R (read.maimages, backgroundCorrect, normalizeWithinArrays, normalizeBetweenArrays, getEAWP functions) (*28*). Strict care was taken to ensure data integrity. All efforts were made to integrate available biological (e.g., sex, age, brain region) and technical covariates (e.g., experimental batch, RIN, post-mortem interval, pH) for each study from GEO, the study supplement, or directly from the authors (see **Fig. A1.1**). For Affymetrix microarrays, chip scan date was used as a surrogate for experimental batch, extracted from the metadata. Normalized 5'/3' bias, a measure strongly influenced by RNA degradation, was calculated for Affymetrix arrays using the AffyRNAdeg function. This was not available to calculate for Illumina arrays due to probe design and location. Correlation panel plots were created to assess the influence of each covariate on gene expression, as summarized by its first 5 principal components (see **Fig. 2.1**; **Fig. A1.2**). We balanced case/control status across available biological and technical covariates such that for each study, case/control status was not significantly associated with any measured covariate ($p > 0.05$). This included removing all singular batches and experimental batches confounded with case/control status (see **Fig. 2.1**).

Outliers were defined as samples with standardized sample network connectivity Z scores < -2, as described (*29*), and were removed (see Supplemental Text "Description of Datasets"). Batch effects were corrected with the ComBat function of the *sva* package in R (*30*). Similar results were achieved using alternative batch correction methods, such as linear regression or including batch in the final mixed effect model (**Fig. A1.3**).

To provide a systematic nomenclature for assessment of gene expression across platforms, microarray probes were re-mapped to Ensembl gene IDs (v75; Feb 2014 data freeze) using the *biomaRt* package in R (*31*), taking the maximum mean signal across all probes available for each gene, using the collapseRows function. The collapseRows MaxMean function was explicitly developed to perform cross-platform microarray meta-analysis and has been extensively validated in its ability to increase between-study consistency and enhance reproducibility (*32*). We note that choosing the 3'-most probe for each gene leads to similar results (**Fig. A1.3**).

Finally, all available biological and technical covariates except for diagnostic group were regressed from each individual expression dataset prior to differential gene expression (DGE) meta-analysis.

Differential Gene Expression (DGE)

DGE was calculated using a linear mixed-effects model using the *nlme* package in R, with fixed effects of diagnostic group and study and a random effect for unique subject. This statistical framework enabled calculation of meta-analytic $\log_2$ fold-change values ($\log_2$FC) for each gene and disease, collapsing results from multiple studies while accounting for any subjects overlapping between studies with a random effect term. Genes were then filtered to include only those that were present across all studies (11,245 Ensembl gene IDs; listed in **Data Table A1.1**). Spearman's r was used to compare DGE meta-analysis $\log_2$FC signatures across all disease pairs, as shown in **Fig. 2.2A**.

Significance thresholds were determined using permutation testing to account for any study-specific factors that could potentially bias results. Within each individual study, we randomly permuted case/control status 40,000 times and repeated the linear mixed-effect model meta-analysis as described in the previous section. This generated $\log_2$FC summary statistics for each of the six "disease" groups. We then assessed transcriptome overlap between each "disease" pair using Spearman's correlation and recorded the resulting test statistics (r values). This process was repeated to generate a null distribution of 40,000 r values.

Gene Co-Expression Network Mega-Analysis

To place results from individual genes within their systems-level network architecture, we performed Weighted Gene Co-Expression Network Analysis (WGCNA). Individual (covariate-regressed) expression datasets were combined together using the 11,245 genes present across all studies. ComBat was used to mitigate batch effects (*30*), as shown in **Fig. A1.2**. This normalized mega-analysis expression set was then used for all downstream network analyses.

Network analysis was performed with the *WGCNA* package (*17*) using signed networks. A soft-threshold power of 9 was used for all studies to achieve approximate scale-free topology (R^2>0.8). Networks were constructed using the blockwiseModules function. The network dendrogram was created using average linkage hierarchical clustering of the topological overlap dissimilarity matrix (1-TOM). Modules were defined as branches of the dendrogram using the hybrid dynamic tree-cutting method (*33*). Modules were summarized by their first principal component (ME, module eigengene) and modules with eigengene correlations of >0.9 were merged together. A robust version of WGNCA (rWGCNA) was run to reduce the influence of potential outlier samples on network architecture (*34*). Module robustness was ensured by randomly resampling (2/3 of the total) from the initial set of samples 100 times followed by consensus network analysis, a meta-analytic approach, to define modules using a consensus quantile threshold of 0.2. Modules were defined using biweight midcorrelation (bicor), with a minimum module size of 50, deepsplit of 4, merge threshold of 0.1, and negative pamStage. Modules are labelled by a number CD# and color for illustration purposes. Genes that did not fall within a specific module are assigned the color grey (CD0).

Dynamic tree cut methods and signed networks were used as they have been shown to be more biologically meaningful, compared to static tree cut methods and unsigned networks (*33*, *35*). Soft threshold power was chosen to be the smallest value such that approximate scale-free topology was achieved, defined as $R^2 > 0.8$ for the frequency distribution of network connectivity on a log scale as described (*13*). We chose a minimum module size of 50, as modules with smaller sizes are more likely to capture noise. A deep split parameter of 4 creates more specific modules, which is enabled by the large sample size employed in this study. In general, we have found that with large sample sizes, WGCNA is robust to changes in module parameters (**Fig. A1.8**).

Module (eigengene)-disease associations were evaluated using a linear mixed-effects model, using a random effect of subject, to account for any subject overlap across initial datasets. We also used linear regression to test for association between module eigengenes and several covariates or confounders (sex, age, PMI, pH, RIN, normalized 5’/3’ bias). Significance values were FDR-corrected to account for multiple comparisons. Results from module-eigengene association tests are reported in **Data Table A1.2** and shown in **Fig. A1.9**.

Genes within each module were prioritized based on their module membership (kME), defined as correlation to the module eigengene. The top 20 hub genes for seven of the modules are shown in **Fig. 2.3D**, with top connections plotted.

Gene Set Enrichment

Functional enrichment of Gene Ontology pathways was assessed with GO-Elite v1.2.5 (*36*) as well as using the *gProfiler* (*37*) R package, using GO and KEGG databases. Only pathways containing between 10 and 2000 genes were used. For gProfiler, "moderate" hierarchical filtering was used. A custom background set consisted of the (11,245) genes present across all studies and microarray platforms. The top pathways reaching significance with FDR-adjusted $P < 0.05$ are shown in **Fig. 2.3E** and **Data Table A1.2**. The two methods had highly concordant results. Enrichment for putative transcription factor binding sites (TFBSs) within the promoters of genes from each co-expression module was performed using the gProfileR package, which integrates annotations from the TRANSFAC database. Default parameters were used. Results are compiled in **Data Table A1.2**.

Cell-type specific expression analysis of genes within each module was performed using the *pSI* package (specificity index; http://genetics.wustl.edu/jdlab/psi_package/) in R (*38*, *39*). Cell-type specific gene expression data was obtained from an RNAseq study of purified populations of neurons, astrocytes, oligodendrocytes, microglia, and endothelial cells derived from adult human cerebral cortex (*25*). Raw data (FPKM) was downloaded from GEO (GSE73721). Gene symbols were mapped to Ensembl gene identifiers using the *biomaRt* R package. Expression values were $\log_2$ normalized and averaged across cell-type replicates. Specificity for the five CNS cell types was calculated with the *specificity.index* function. Significance was assessed using Fisher's exact test with a pSI threshold set to 0.05, followed by FDR-correction of p values.

Transcriptome "Severity" Measures

The transcriptome overlap between disease pairs was assessed using Spearman's correlation of $\log_2$FC values, as described above. As a global measure of the severity of the transcriptomic phenotype, we sought to compute the slope of the linear regression of $\log_2$FC values between disease pairs (**Fig 2.2B**). However, the linear regression slope is dependent on the (arbitrary) ordering of the response (Y) and predictor (X) variables, and we found that in practice the slope can change considerably according to this order. To circumvent this issue, we used principal component regression (a linear case of "orthogonal regression" or "total least squares") which provides an estimate of slope that is invariant to the choice of predictor and response variables.

GWAS Enrichment

We compiled a set of GWAS summary statistics for several neuropsychiatric disorders, cognitive, and behavioral traits (**Table A1.3**; (*40-47*)). Summary statistics from GWAS meta-analyses of ASD, schizophrenia, bipolar disorder, and major depression were downloaded from the PGC website (https://www.med.unc.edu/pgc/downloads). Results from GWAS studies of alcoholism, inflammatory bowel disease, educational attainment, depressive symptoms, and neuroticism were obtained from the respective studies (**Table A1.3**). Given the relatively small sample size for the PGC ASD GWAS, we performed a new ASD GWAS using samples from the iPSYCH Consortium, described in the next section. Gene-level analysis of GWAS results was performed by MAGMA v1.04, a gene-set annotation framework that accounts for linkage disequilibrium (LD) between SNPs (*48*). LD was calculated using the 1000 Genomes European ancestry reference dataset. An annotation step was performed first in which SNPs were mapped to genes (either hg18 or hg19 genome build, depending on the study) based on the presence of a SNP in the region between a gene's start and stop sites. Gene-level analysis was then performed to create aggregate statistics for each gene.

To quantify enrichment of GWAS signal within each gene co-expression module, we calculated Spearman's correlation between the module membership (kME) of each gene and the $-\log_{10}$ p-value for that gene for each GWAS study. kME is a measure between 0 and 1 of the centrality of a gene within a module; "hub genes" have kME values approaching 1, whereas genes that are not present in a module generally have kME less than 0.5. This process was performed for all module x GWAS combinations, and p values were FDR-corrected.

ASD iPSYCH GWAS

The iPSYCH Autism sample is part of the larger iPSYCH Danish Case-Cohort Study (iDCCS2012; http://ipsych.au.dk/; (*23*)) consisting of 86,189 individuals (57,377 psychiatric cases) born in Denmark between 1981 and 2005. The 86,189 represent all individuals in the population birth cohort (N=1,472,762) with an ICD record of ADHD, affective disorder, anorexia, autism, bipolar disorder, or schizophrenia in the Danish Psychiatric Research Register (*49*) as of 2012 plus 30,000 randomly sampled individuals as representative controls. More details on the register-based phenotyping can be found in (*50*) which defines the aggregation used here. DNA was extracted and amplified from dried neonatal bloodspots stored in the Danish Neonatal Screening Biobank and then genotyped in 23 waves using the Illumina Infinium PsychArray v1.0. SNP genotypes underwent extensive quality control both by individual (>99% call rate, concordant genotype and recorded sex, typical levels of heterozygosity, non-duplicated samples) and by SNP (good clustering, >97.5% call rate, Hardy-Weinberg equilibrium, no association with genotyping wave). Data from all genotyping waves were merged together, phased using SHAPEIT3 (https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html), imputed to the 1000

Genomes phase 3 reference haplotypes using Impute2 (*51*, *52*) and cleaned again (info score > 0.2, MAF > 0.001, good and comparable quality in cases and controls) resulting in 11,600,723 estimated SNP dosages for analysis. We used principal component analysis implemented in Eigensoft (smartPCA; (*53*)) to select individuals with homogenous genetic ancestry and estimated pairwise relatedness with KING (*54*), removing individuals to ensure no pair was related closer than 2nd degree relatives, leaving 65,534 individuals. We defined cases as the subset of the remaining cohort with autism as the only ascertained diagnosis (N=8,605 of 12,371 possible autism cases; ICD codes F84.0, F84.1, F84.5, F84.8 and/or F84.9), while our controls (N=19,526) were the subset of the random cohort with no diagnoses in the Danish Psychiatry Research Register, ascertained or otherwise (ICD F00-F99). GWAS summary statistics were generated using logistic regression in plink v1.9 (*55*) including age, sex and ten genetic ancestry PCs as covariates. Summary statistics are available at https://github.com/mgandal/Shared-molecular-neuropathology-across-major-psychiatric-disorders-parallels-polygenic-overlap/tree/master/raw_data/GWAS

Rare Variant Enrichment

A composite list of rare *de novo* variants (RDNVs) was compiled from several recent whole-exome sequencing (WES) studies of trios (parents & proband) or quads (parents, proband, and unaffected sibling) affected by ASD, schizophrenia, or intellectual disability (ID), summarized in the **Data Table A1.3** (*56-67*). Unaffected siblings were used as controls. RDNVs were categorized as non-synonymous (amino acid change) or silent. Enrichment of RDNVs among genes within each co-expression module was assessed using logistic regression controlling for gene length (*68*), an important potential confounding variable, as gene length is known to be strongly correlated with mutation rate (*61*). Odds ratios (OR) were calculated as ln(beta) from the regression model and P-values were FDR-corrected for multiple comparisons. Only significant enrichments (FDR corrected P<0.05) with an odds-ratios >1 (e.g., over-representation) are shown in **Fig. 2.4.**

To investigate the potential contribution of CNV-affected genes on transcriptome modules, we searched for syndromic CNVs for neuropsychiatric disorders using PubMed and Google Scholar. Search terms were: CNV, deletion, duplication, and disorder name. CNVs that were identified in at least two studies with a p<0.01 or passed p-value threshold of 10^{-6} in one study were included. We filtered to only include studies with adequate power and sample size (>500 subjects per group). Only one high-confidence CNV was identified in each of depression and bipolar disorder studies. As such, these disorders were excluded in downstream analysis. Genes for the regions were retrieved from RefGene (UCSC download, hg19) using either maximal regions provided by the publications in question, or from recent autism publications (*69*, *70*). If the borders were reported in hg18, we used the UCSC LiftOver tool to convert them to hg19. Enrichment of CNV affected genes was computing using logistic regression, as above, including gene length as a covariate. See **Data Table A1.3** for a compilation of CNVs.

LD Score Regression

To further dissect the relationship between transcriptome alterations and disease-associated genetic variants, we used stratified LD score regression to partition disease heritability within functional categories represented by gene co-expression modules (*24*, *71*). Using GWAS summary statistics and LD explicitly modelled from ancestry-matched 1000 Genomes reference panel, this method calculates the proportion of genome-wide SNP-based heritability that can be attributed to SNPs within explicitly defined functional categories (see https://github.com/bulik/ldsc/). Functional categories for each module were defined by taking all SNPs within gene-body annotation (transcription start to stop sites) for all genes within the module. To improve model accuracy, these categories were added to the ‘full baseline model’ which included 53 functional categories capturing a broad set of functional elements, as defined in (*24*).

Enrichment is defined as the proportion of SNP-heritability accounted for by each module divided by the proportion of total SNPs within the module. Significance is assessed using a block jacknife procedure, as described (*24*), followed by FDR-correction of p-values. Modules with FDR-corrected enrichment p-values of less than 0.05 were considered significant heritability contributors. See **Data Table A1.4** for proportion of heritability and enrichment along with other relevant statistics for each module and GWAS.

Psychencode BrainGVEX RNAseq Replication

RNAseq data was newly generated from 153 postmortem frontal cortex brain samples from subjects with schizophrenia (n=53), bipolar disorder (n=47), and non-psychiatric controls (n=53), as part of the BrainGVEX study (Synapse accession doi:10.7303/syn4590909) within the PsychEncode Consortium (https://www.synapse.org/pec) (*72*). Brain samples were collected as part of the "Array Collection" and the "New Collection" from the Stanley Medical Research Institute (SMRI). Protocols for RNA extraction and sequencing can be found on the Sage Synapse website (https://www.synapse.org/#!Synapse:syn4616686; https://www.synapse.org/#!Synapse:syn4640744). Briefly, 50-60mg of fresh-frozen brain tissue was dissected on dry ice. Total RNA was extracted using trizol/chloroform and purified by isopropanol precipitation. RNA quality (RIN) was measured using an Agilent Bioanalyzer. Strand-specific, rRNA-depleted RNAseq libraries were prepared using TruSeq Stranded Total RNA sample prep kit with RiboZero Gold HMR (Illumina) kits. Libraries were multiplexed (3 per lane) and sequenced with 100 bp paired end reads on Illumina HiSeq2000 with read depth >70 million reads on average.

FastQC was used for initial quality control. Reads were mapped to the hg19 human genome build with Ensembl v75 annotations using STAR RNAseq aligner v2.5.0a. Aligned reads were coordinate-sorted and read pairs mapping to different chromosomes were removed from the BAM file using the samtools view –f 0x0002 command. BAM files for the same sample were merged across sequencing runs using samtools. Quality control after read alignment was performed using Picard Tools v1.131 (CollectAlignmentSummaryMetrics, CollectRNA-seqMetrics, CollectGcBiasMetrics, MarkDuplicates). To control for differences in RNA quality, read depth and other sequencing-related technical artifacts across subjects, we created a matrix of “sequencing statistics” corresponding to the aggregate of above Picard Tools metrics. Two sequencing statistics, seqPC1 and seqPC2, were calculated as the first and second principal components of this matrix. These sequencing statistics were used as covariates in downstream analyses.

Aligned reads were quantified using HTSeq Counts (v0.6.0) in union exon mode. Counts were normalized for read-depth, GC content, and gene length and $\log_2$-transformed using the *cqn* package in R (*73*). Genes were filtered to include only those with at least $\log_2$(FPKM) of 1 in 50% of samples.

Outliers were detected by calculating standardized sample network connectivity Z scores, and samples with Z < -2 were removed from downstream analysis as described above for microarray studies. Library preparation date was used to denote experimental batches. Singular batches were removed and groups were then balanced such that case/control status was not significantly associated with any measured covariate ($p > 0.05$).

Correction for library batch (which was collinear with brain bank) was performed using the ComBat function from the *sva* package in R. Finally, differential gene expression of log2(normalized FPKM) expression values was calculated using *limma* with empiric Bayes moderated t-statistics, including the following covariates: diagnosis, age, sex, RIN, RIN^2, ethnicity, PMI, pH, seqPC1, and seqPC2. Regression coefficients ($\log_2$FC beta values) for each gene were calculated for each group (BD and SCZ) and Spearman's correlation used to assess global transcriptome overlap, as above. Data QC are shown in **Fig. A1.6**, results are shown in **Fig. A1.7**, and summary statistics are included in **Data Table A1.1**.

ASD Pan-Cortical RNAseq Replication

RNAseq data for replication was newly generated from 88 postmortem cortex brain samples from subjects with ASD (53 samples from 24 subjects) and non-psychiatric controls (35 samples from 17 subjects), across four cortical regions encompassing all major cortical lobes – frontal, BA4/6; temporal, BA38; parietal, BA7; and occipital, BA17, as part of the UCLA-ASD study (Synapse accession doi:10.7303/syn4587609) within the PsychEncode Consortium (https://www.synapse.org/pec) (*72*). Brain samples were obtained from the Harvard Brain Bank as part of the Autism Tissue Project (ATP). An ASD diagnosis was confirmed by the Autism Diagnostic Interview-Revised (ADIR) in 22 of the subjects. In the remaining two subjects, diagnosis was supported by clinical history. Frozen brain regions were dissected on dry ice in a dehydrated dissection chamber to reduce degradation effects from sample thawing or humidity. Approximately 50-100mg of tissue across the cortical region of interest was isolated from each sample using the miRNeasy kit with no modifications (Qiagen). For each RNA sample, RNA quality was quantified using the RNA Integrity Number (RIN) on an Agilent Bioanalyzer. Strand-specific, rRNA-depleted RNAseq libraries were prepared using TruSeq Stranded Total RNA sample prep kit with RiboZero Gold (Illumnia) kits. Libraries were randomly pooled to multiplex 24 samples per lane using Illumina TruSeq barcodes. Each lane was sequenced five times on an Illumina HiSeq 2500 instrument using high output mode with standard chemistry and protocols for 50 bp paired-end reads to achieve a target depth of 70 million reads.

FastQC was used for initial quality control. Aligned reads were coordinate sorted and read pairs mapping to different chromosomes were removed from the BAM file using the samtools view –f 0x0002 command. BAM files for the same sample were merged across sequencing runs using samtools. Quality control after read alignment was performed using Picard Tools v1.131 (CollectAlignmentSummaryMetrics, CollectRNA-seqMetrics, CollectGcBiasMetrics, MarkDuplicates). To control for differences in RNA quality, read depth and other sequencing-related technical artifacts across subjects, we created a matrix of "sequencing statistics" corresponding to the aggregate of above Picard Tools metrics. Two sequencing statistics, seqPC1 and seqPC2, were calculated as the first and second principal components of this matrix. These sequencing statistics were used as covariates in downstream analyses.

Aligned reads were quantified using HTSeq Counts (v0.6.0) in union exon mode. Counts were normalized for read-depth, GC content, and gene length and $\log_2$-transformed using the *cqn* package in R (*73*). Genes were filtered to include only those with at least $\log_2$(FPKM) of 1 in 50% of samples. Outliers were detected by calculating standardized sample network connectivity Z scores, and samples with $Z < -2$ were removed from downstream analysis as described above for microarray studies. Library preparation date was used to denote experimental batches, which was the same for all samples. Groups were balanced such that case/control status was not significantly associated with any measured covariate ($p > 0.05$).

Finally, differential gene expression of $\log_2$(normalized FPKM) expression values was calculated using *limma* with empiric Bayes moderated t-statistics, including the following covariates: diagnosis, age, sex, RIN, RIN^2, PMI, BrainRegion, seqPC1, and seqPC2. The limma::duplicateCorrelation function was used to account for the non-independence of samples derived from the same subject across multiple brain regions. Regression coefficients ($\log_2$FC beta values) for each gene were assessed for global transcriptome overall via Spearman's correlation as above. Data QC are shown in **Fig. A1.6**, results are shown in **Fig. A1.7**, and summary statistics are included in **Data Table A1.1**.

CommonMind RNAseq Analysis

RNAseq data from 604 total human postmortem dorsolateral prefrontal cortex (DLPFC) brain samples were obtained from subjects with schizophrenia (n=262), bipolar disorder (n=47), major depression, and neurotypical controls (n=295), as part of the CommonMind Consortium available through dbGap and Sage Synapse (https://www.synapse.org/cmc; doi:10.7303/syn2759792) as recently published (*16*). Details of sample collection and processing are described here (https://www.synapse.org/#!Synapse:syn2759792/wiki/194729). Briefly, samples were acquired through brain banks at three institutions, Mount Sinai, University of Pennsylvania, and University of Pittsburgh. Total RNA was extracted from 50 mg of homogenized DLPFC brain tissue using RNeasy kit. Samples with RIN < 5.5 (n=51) were excluded. RNAseq library preparation was performed using ribosomal RNA depletion, with the Ribozero Magnetic Gold Kit. Samples were barcoded, multiplexed (n=10/lane), and 100 bp paired end sequencing was performed on Illumina HiSeq 2500 with an average of 85 million reads.

Reads were mapped to human genome build hg19 with Ensembl v70 annotations using TopHat version 2.0.9. Quantification was performed using HTSeq-Counts v0.6.0 in intersection-strict mode. The resulting count level data was made available for downstream analysis through Sage Synapse. Available metadata included demographics as well as ancestry PCs, experimental batches, and sequencing statistics calculated with RNA-SeQC (mapped reads, exonic rate, intronic rate, intergenic rate, genes detected, transcripts detected, expression profiling efficiency, rRNA rate, total reads, percent aligned). Quality control after read alignment was also performed using Picard Tools v1.131 (CollectAlignmentSummaryMetrics, CollectRNA-seqMetrics, CollectGcBiasMetrics, MarkDuplicates). To control for differences in RNA quality, read depth and other sequencing-related technical artifacts across subjects, we created a matrix of “sequencing statistics” corresponding to the aggregate of above Picard Tools and RNA-SeQC metrics. Two sequencing statistics, seqPC1 and seqPC2, were calculated as the first and second principal components of this matrix. These sequencing statistics were used as covariates in downstream analyses.

Read counts were normalized for read-depth, GC content, and gene length and $\log_2$-transformed using the *cqn* package in R (*73*). Genes were filtered to include only those with at least $\log_2$(FPKM) of 1 in 50% of samples. Batch correction was performed for sequencing library batches using the ComBat function from the *sva* package in R. Outliers were detected by calculating standardized sample network connectivity Z scores, and samples with $Z < -2$ were removed from downstream analysis as described above for microarray studies. Groups were then balanced such that case/control status was not significantly associated with any measured covariate ($p > 0.05$). Given a substantial age difference between BD and SCZ samples, we split the dataset into two case/control subsets matched for demographics as shown in **Fig. A1.6**. Finally, differential gene expression of log2(normalized FPKM) expression values was calculated using *limma* with empiric Bayes moderated t-statistics, including the following covariates: diagnosis, age, sex, institution, RIN, RIN^2, PMI, seqPC1, seqPC2 and top 5 ancestry PCs. Regression coefficients ($\log_2$FC beta values) for each gene were calculated for each group (BD and SCZ) and Spearman’s correlation used to assess global transcriptome overlap, as above. Data QC are shown in **Fig. A1.6**, results are shown in **Fig. A1.7**, and summary statistics are included in **Data Table A1.1**.

qSVA Assessment

We performed qSVA analyses to mitigate the impact of potential differences in RNA quality across case and control groups on cross disorder transcriptomic similarity measures (*74*). qSVA is based on experimentally defined genomic regions susceptible to RNA degradation in brain, specific to the method of RNAseq library preparation. As all RNAseq datasets in this study used RiboZero library preparation, we counted paired end reads uniquely mapped to RiboZero-based RNA degradation regions using featureCounts. Counts were normalized by read depth and interval length to yield FPK80M values (fragments per kilobase 80 million reads mapped) and were $\log_2$ transformed with an offset of 1. Principal component analysis was then performed on this normalized degradation matrix and qSVs were defined as the top n principal components, where n is calculated using the method of Buja and Eyuboglu with 100 permutations. These n qSVs were then included as covariates in downstream differential expression analyses, shown in **Fig. A1.7**. The number of qSVs for each dataset were: 5 for ASD-pancortical, 11 for BrainGVEX, 6 for CommonMind SCZ-matched subset, and 4 for CommonMind BD-matched subset.

A1.2: Extended Text

<u>Description</u> <u>of</u> <u>Datasets</u>

A. Voineagu et al, [GSE28521]

This dataset initially included 58 brain samples from frontal and temporal cortex from subjects with ASD (n=29) and controls (n=29) (*3*). Samples were run on Illumina HumanRef-8 v3.0 expression beadchip arrays. The data was $\log_2$-transformed and quantile normalized. Three outlier samples were identified and removed. Two singular or imbalanced batches were removed. Two samples were removed due to covariate confounding, for a final cohort of 48 samples. Batch effects were corrected with ComBat. Probes were reannotated and collapsed to

Ensembl v75 gene definitions. The following covariates were then regressed: Brain Region, Sex, Age, RIN, and PMI. See **Fig. A1.1A**.

B. Chen et al., [GSE35978]

The initial dataset consisted of 160 parietal cortex brain samples from subjects with MDD (n=14), BD (n=45), SCZ (n=51), and controls (N=50) (*6*). Samples were run on Affymetrix Human Gene 1.0 ST microarrays. Due to significant covariate confounding, MDD samples were removed from the analysis. Data were RMA normalized and log2 transformed using the *Affy* package in R. RNA 3' bias was extracted from the slope of the *AffyRNAdeg* function. Chip scan date was extracted as a proxy for microarray batch. One singular batch was removed. There were 7 outlier samples identified and removed. 14 samples were removed due to covariate confounding, for a final cohort of 124 samples. Batch effects were corrected with ComBat. Probes were reannotated and collapsed to Ensembl v75 gene definitions. The following covariates were then regressed: Sex, Age, PMI, pH, RNA 5'/3' Bias. See **Fig. A1.1B**.

C. Garbett et al, 2008

The initial dataset consisted of 12 samples from temporal cortex from subjects with ASD and controls, which were all used in this analysis (*4*). Samples were run on Affymetrix Human Genome 133 plus 2 microarrays. Data were RMA normalized and log2 transformed using the *Affy* package in R. RNA 3' bias was extracted from the slope of the *AffyRNAdeg* function. Chip scan date was extracted as a proxy for microarray batch. One singular batch was removed, for a final cohort of 11 samples. Batch effects were corrected with ComBat. Probes were reannotated and collapsed to Ensembl v75 gene definitions. The following covariates were then regressed: Sex, Age, PMI, RIN, Normalized 5'/3' Bias. See **Fig. A1.1C**.

D. Lanz et al., [GSE53987]

The initial dataset consisted of 68 PFC brain samples from subjects with SCZ (n=15), BD (n=17), MDD (n=17), and controls (n=19) (*75*). Two outlier samples were identified and removed. One sample was removed due to covariate confounding, for a final count of 65 samples. Data were RMA normalized and log2 tranformed using the *Affy* package in R. RNA 3' bias was extracted from the slope of the *AffyRNAdeg* function. Chip scan date was extracted as a proxy for microarray batch. See **Fig. A1.1D**.

E. Chow et al., [GSE28475]

The initial dataset consisted of 33 samples from frontal cortex from male subjects with autism and controls (*5*). Samples were run on Illumina HumanRef-8 v3.0 expression beadchip. Data were log2 transformed and quantile normalized using the R lumi package and downloaded from GEO. Two samples had missing RIN values, which were imputed from the group mean. Three outlier samples were removed for a final cohort of 30 samples. Batch effects were corrected with ComBat. Probes were reannotated and collapsed to Ensembl v75 gene definitions. The following covariates were then regressed: Age, PMI, RIN. See **Fig. A1.1E**.

F. Iwamoto et al., [GSE12649]

The initial dataset consisted of 101 samples from frontal cortex (BA46) from subjects with SCZ (n=35), BD (n=32), and controls (n=34) (*8*). Samples were run on Affymetrix Human Genome U133A Arrays. Data were RMA normalized and log2 transformed using the *Affy* package in R. RNA 3' bias was extracted from the slope of the *AffyRNAdeg* function and used as a measure of RNA quality. Chip scan date was extracted as a proxy for microarray batch. One singular batch was removed as were 4 outlier samples. Ten samples were removed due to covariate confounding, leaving 86 arrays in the final analysis. Batch effects were corrected with ComBat. Probes were reannotated and collapsed to Ensembl v75 gene definitions. The following covariates were then regressed: Sex, Age, PMI, pH, RNA quality. See **Fig. A1.1F**.

G. Maycox et al., [GSE17612]

The initial dataset consisted of 51 prefrontal cortex (BA10) brain samples from subjects with SCZ (n=28) and controls (n=23) (*7*). Samples were run on Affymetrix Human Genome U133 Plus 2.0 Arrays. Data were RMA normalized and log2 transformed using the *Affy* package in R. RNA 3' bias was extracted from the slope of the *AffyRNAdeg* function. Chip scan date was extracted as a proxy for microarray batch. There were no singular or confounded batches. Individual sample RIN values were not available although all samples were noted to have RIN >6. RNA 5'/3' bias did not differ between groups and was included as a measure of RNA quality. There were 5 outlier samples identified and removed, leading to 46 in the final cohort. Batch effects were corrected with ComBat. Probes were reannotated and collapsed to Ensembl v75 gene definitions. The following covariates were then regressed: Sex, Age, PMI, RNA 5'/3' Bias. Of note, pH was not included as a covariate as there was concern regarding the rigorousness of the pH measurements. See **Fig. A1.1G**.

H. Sibille et al., [GSE54567, GSE54568, GSE54571, GSE54572].

The initial dataset consisted of 140 samples across five matched cohorts from anterior cingulate cortex (BA25) and prefrontal cortex (BA9) from subjects with MDD (n=70) and controls (n=70) (*10*). All samples were run on Affymetrix Human Genome U133 Plus 2.0 Arrays. Data were normalized and log2 transformed using the *Affy::RMA* function in R. RNA 3' bias was extracted from the slope of the *AffyRNAdeg* function. Chip scan date was extracted as a proxy for microarray batch. One singular batch was removed as were 6 outlier samples, for a final count of 133. Batch effects were corrected with ComBat. Probes were reannotated and collapsed to Ensembl v75 gene definitions. The following covariates were then regressed: Brain Region, Sex, Age, Race, PMI, pH, RIN, RNA quality (5'/3' bias). See **Fig. A1.1H**.

I. Mayfield et al., [GSE29555].

The initial dataset initially consisted of 32 superior frontal cortex from subjects with alcoholism and match controls (*11*). All samples were run on Illumina HumanHT-12 V3.0 expression beadchip arrays. Data was quantile normalized and log2 transformed using the Lumi::LumiN package in R. There were no singular batches and 2 samples were marked as outliers, leading to a final cohort of 30. Probes were reannotated and collapsed to Ensembl v75 gene definitions. The following covariates were then regressed: Age, Sex, PMI, pH, and RIN. See **Fig. A1.1I**.

J. Narayan et al., [GSE21138]

The initial dataset consisted of 59 frontal cortex samples (BA46) from subjects with SCZ and controls (*9*). Samples were run on Affymetrix Human Genome U133 Plus 2.0 Arrays. Data were RMA normalized and log2 transformed using the *Affy* package in R. RNA 3' bias was extracted from the slope of the *AffyRNAdeg* function. Individual sample RIN values were not available although RNA 5'/3' bias was used a measure of RNA quality. Chip scan date was extracted as a proxy for microarray batch. One singular batch was removed as were two outlier samples. Five samples were removed due to covariate confounding, leading to 51 final remaining samples. Batch effects were corrected with ComBat. Probes were reannotated and collapsed to Ensembl v75 gene definitions. The following covariates were then regressed: Sex, Age, PMI, pH, RNA quality. See **Fig. A1.1J**.

Assessment of Statistical Robustness

We performed extensive exploratory data analysis to assess the factors contributing most substantially to the variance in gene expression across studies and disorders in our final mega-analytic dataset. **Fig. A1.2** shows the correlation between the top 5 principal components of gene expression and several biological and technical factors known to influence gene expression. As demonstrated, while many factors are associated with these top 5 expression PCs, diagnostic group has the highest loading across all 5 PC's compared with other covariates. Additionally, we show that gene co-expression modules are largely stable across individual datasets (**Fig. A1.10**). Finally, we plot the association between gene co-expression module eigengenes and biological and technical covariates (**Fig. A1.9**). Similar to the top 5 expression PCs, diagnostic group shows the largest association with module eigengene values compared with other biological and technical covariates.

We have performed extensive assessment of the methods employed to validate that results are robust regardless of specific methodological details. We compared multiple batch correction methods, including ComBat (original), linear regression, and including batch in the final mixed effect model (**Fig. A1.3, A-C**). As demonstrated, the choice of batch correction method does not appreciably alter results. This applies to the co-expression modules which show a stable pattern in individual datasets, with or without any batch correction procedure (**Fig. A1.10**). We assessed two methods for summarization of gene expression values from microarray probes, including the collapseRows function (maxMean summarization) as well as selection of the 3' most probe for each gene. Again, there were no qualitative differences between these methods (**Fig. A1.3D**). When measuring cross disorder transcriptome overlap, we used all genes measured across all the different datasets. However, some gene features may be more influenced by platform-specific technical factors than others and therefore be less reliable when aggregating across studies. To address this, we performed an analysis using integrative correlations, selecting gene features based on cross-study replicability. Using the *MergeMaid* and *metaArray* packages in R (*76*, *77*), we selected only those genes whose average integrative correlation across studies is >0.95 quantile of the null distribution.

Repeating our transcriptome overlap analyses using these genes does not alter the results, and in fact, increases the magnitude of cross disorder correlations (**Fig. A1.3E**).

To ensure that results were not being driven by one specific disease population, we repeated our comparison between transcriptome and genetic similarity after systematically removing one disorder (**Fig. A1.13**). Some of the datasets investigating SCZ and BD had an overlapping set of controls, which might increase the biological signal we detect for SCZ-BD relative to other disease pairs. To ensure that the strong SCZ-BD transcriptome overlap was not being driven by these shared controls, we recomputed the relationship after randomly splitting the controls. There was a slight reduction in the SCZ-BD transcriptome overlap, but the relationship remains highly significant and qualitatively the same (r=0.60, P<0.001).

Jaffe and colleagues have recently generated a dataset of postmortem human control brain samples allowed to degrade for set intervals to define the susceptibility of each genomic region to RNA degradation (*74*). We compared the RNA degradation T-statistics for each gene with disease differential gene expression effect sizes in this study and, with the exception of the alcoholism disease signature, find negligible overlap between these measures. Furthermore, the cross disorder transcriptome overlaps remain robust even after first regressing out these degradation statistics (**Fig. A1.3F**). Finally, we employed full qSVA correction using RiboZero-based degradation features on RNAseq data as shown in **Fig. A1.7**. These different methods for normalization and correction of potential confounders did not lead to qualitatively distinct results from the primary analysis.

Gene-Level RNAseq Replication

This global view of transcriptome-wide replication does not focus on individual genes. Here, we look at the degree to which genes identified as differentially expressed in the discovery (microarray) datasets (FDR < 0.05) are replicated in the RNAseq data (P < 0.05 with concordant direction of effects), as shown in **Table A1.2**. For comparability, we restricted our background to the 8874 genes present across all microarray and RNAseq datasets (listed in **Data Table A1.1**). Among these 8874 genes, **Table A1.2** shows the number of genes identified as differentially expressed in the discovery (microarray) dataset at FDR < 0.05. Among these discovery DGE genes, we identify the number that replicate in RNAseq datasets, at P<0.05 and with a concordant direction of effect. We used Fisher's exact test to calculate odds ratios and the statistical significance of overlap between microarray DGE genes and the RNAseq replication set (all genes with P<0.05 and concordant direction). Finally, we define the replication rate as the percentage of discovery DGE genes replicated in a given RNAseq dataset. BD shows less overlap than ASD or SCZ (although still highly significant) likely due to the relatively small RNAseq sample size.

A1.3: Extended Figures

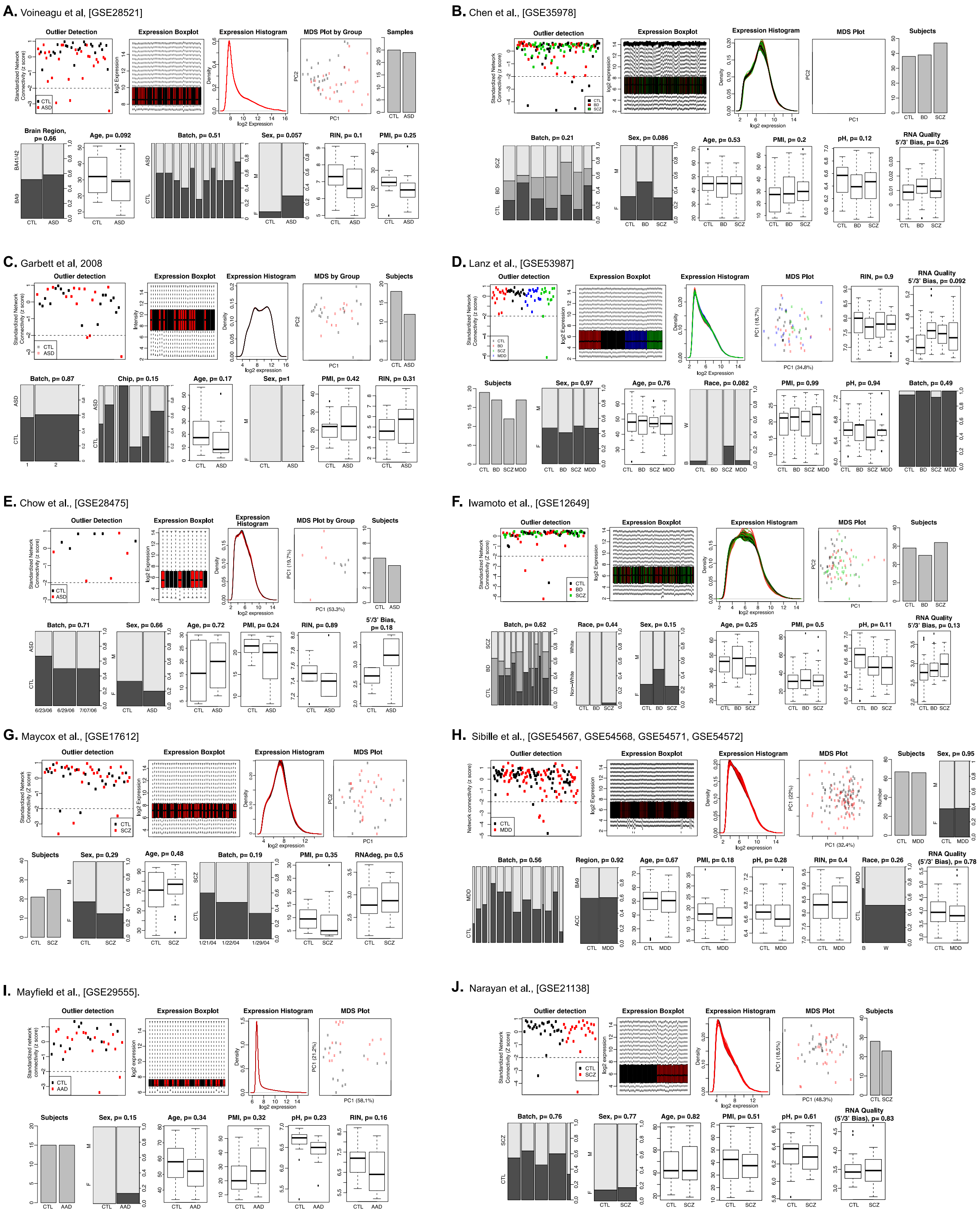


Fig. A1.1.

For each microarray dataset, we show several quality control plots including expression boxplot and histograms. Outlier detection was determined based on standardized network connectivity z-scores.

Multidimensional scaling (MDS) plots show sample clustering by the first two expression principal components. Groups were balanced by available covariates and potential confounding factors.

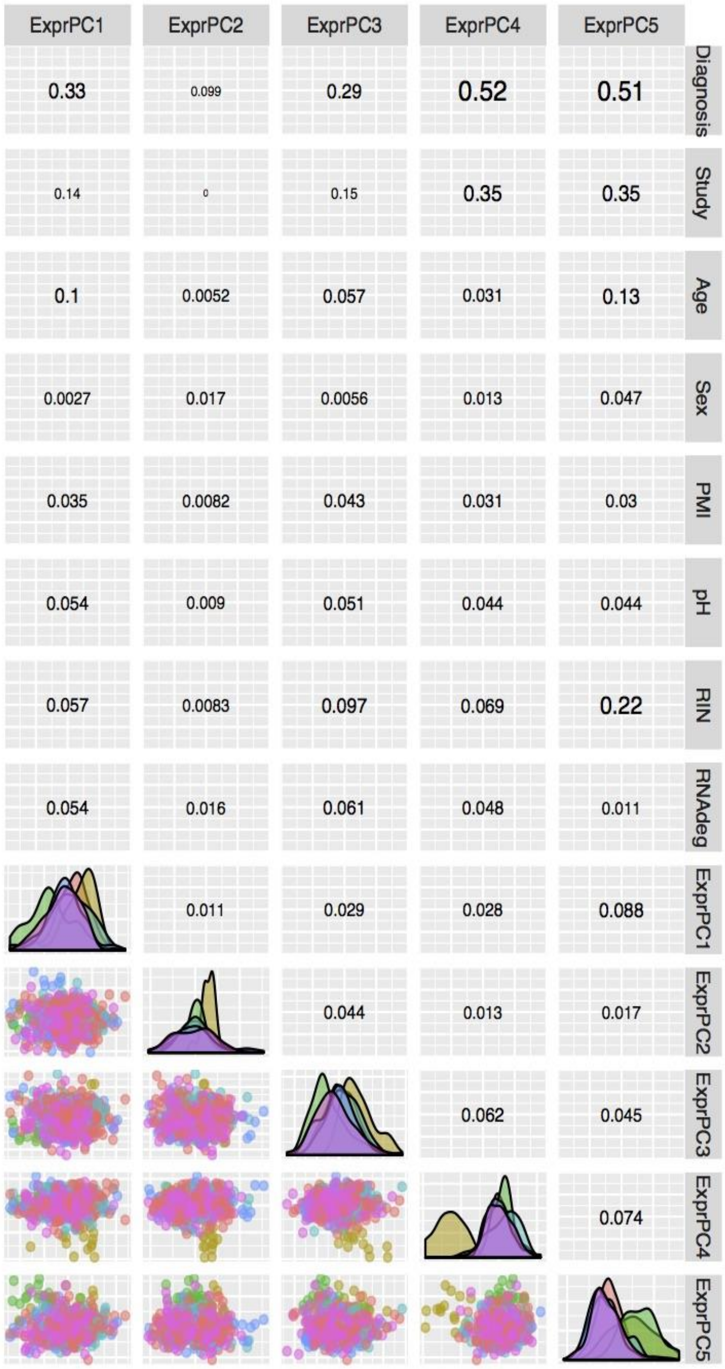
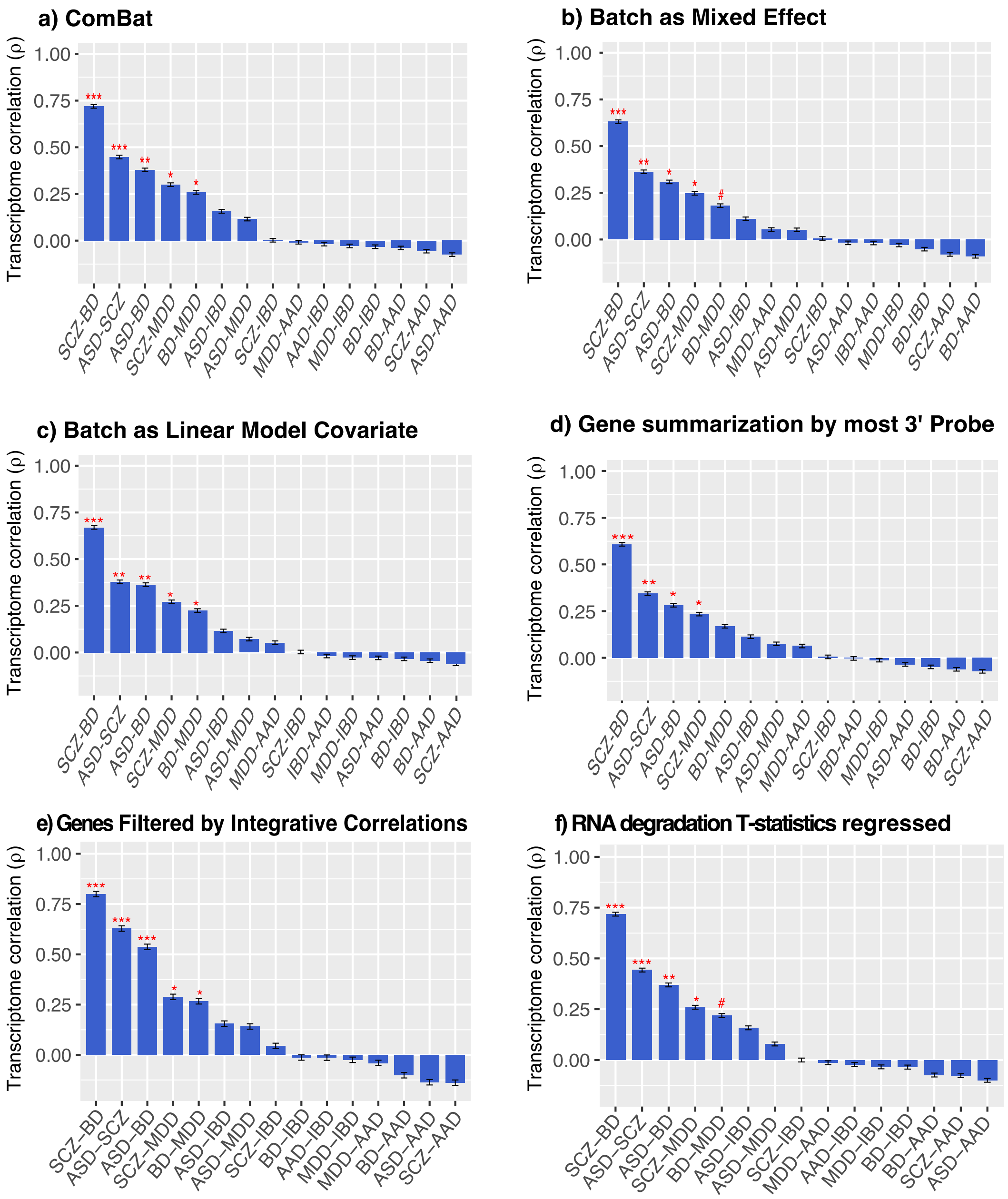


Fig. A1.2

Gene expression principal component loadings. Spearman's correlation (absolute value) is shown for the top 5 expression principal components with biological and technical covariates in the combined mega-analysis. Diagnostic group shows the highest loading on expression compared with other known covariates, including study/batch.



a) ComBat

b) Batch as Mixed Effect

c) Batch as Linear Model Covariate

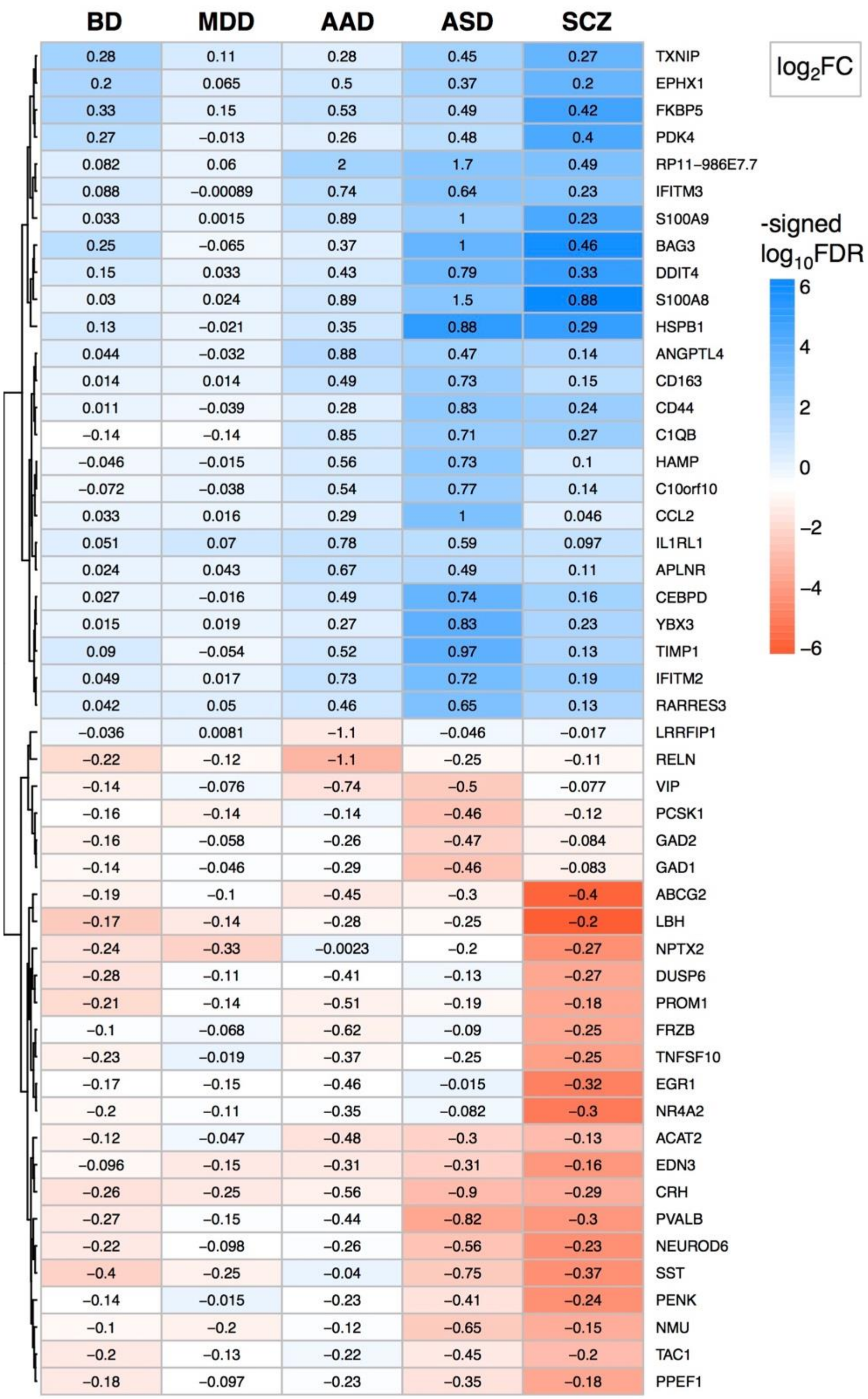
d) Gene summarization by most 3' Probe

e) Genes Filtered by Integrative Correlations

f) RNA degradation T-statistics regressed

Fig. A1.3

Transcriptome overlap across disorders is robust to methodology of batch correction, gene/probe summarization, and gene filtering. We demonstrate stability of transcriptome overlap results correcting for batch using several different methods, namely (a) ComBat, (b) including batch as a covariate in the final mixed-effect model, or (c) linear regression. In (d), we show that transcriptome overlap is largely unchanged using the 3' most probe for each gene rather than collapseRows/maxMean summarization. In (e), genes are first filtered by integrative correlations (>0.95 quantile of the null distribution), which significantly increases transcriptome overlap across disorders. f) Transcriptome overlap is shown after first regressing RNA degradation T-statistics (*74*).



BD	MDD	AAD	ASD	SCZ	
0.28	0.11	0.28	0.45	0.27	TXNIP
0.2	0.065	0.5	0.37	0.2	EPHX1
0.33	0.15	0.53	0.49	0.42	FKBP5
0.27	−0.013	0.26	0.48	0.4	PDK4
0.082	0.06	2	1.7	0.49	RP11−986E7.7
0.088	−0.00089	0.74	0.64	0.23	IFITM3
0.033	0.0015	0.89	1	0.23	S100A9
0.25	−0.065	0.37	1	0.46	BAG3
0.15	0.033	0.43	0.79	0.33	DDIT4
0.03	0.024	0.89	1.5	0.88	S100A8
0.13	−0.021	0.35	0.88	0.29	HSPB1
0.044	−0.032	0.88	0.47	0.14	ANGPTL4
0.014	0.014	0.49	0.73	0.15	CD163
0.011	−0.039	0.28	0.83	0.24	CD44
−0.14	−0.14	0.85	0.71	0.27	C1QB
−0.046	−0.015	0.56	0.73	0.1	HAMP
−0.072	−0.038	0.54	0.77	0.14	C10orf10
0.033	0.016	0.29	1	0.046	CCL2
0.051	0.07	0.78	0.59	0.097	IL1RL1
0.024	0.043	0.67	0.49	0.11	APLNR
0.027	−0.016	0.49	0.74	0.16	CEBPD
0.015	0.019	0.27	0.83	0.23	YBX3
0.09	−0.054	0.52	0.97	0.13	TIMP1
0.049	0.017	0.73	0.72	0.19	IFITM2
0.042	0.05	0.46	0.65	0.13	RARRES3
−0.036	0.0081	−1.1	−0.046	−0.017	LRRFIP1
−0.22	−0.12	−1.1	−0.25	−0.11	RELN
−0.14	−0.076	−0.74	−0.5	−0.077	VIP
−0.16	−0.14	−0.14	−0.46	−0.12	PCSK1
−0.16	−0.058	−0.26	−0.47	−0.084	GAD2
−0.14	−0.046	−0.29	−0.46	−0.083	GAD1
−0.19	−0.1	−0.45	−0.3	−0.4	ABCG2
−0.17	−0.14	−0.28	−0.25	−0.2	LBH
−0.24	−0.33	−0.0023	−0.2	−0.27	NPTX2
−0.28	−0.11	−0.41	−0.13	−0.27	DUSP6
−0.21	−0.14	−0.51	−0.19	−0.18	PROM1
−0.1	−0.068	−0.62	−0.09	−0.25	FRZB
−0.23	−0.019	−0.37	−0.25	−0.25	TNFSF10
−0.17	−0.15	−0.46	−0.015	−0.32	EGR1
−0.2	−0.11	−0.35	−0.082	−0.3	NR4A2
−0.12	−0.047	−0.48	−0.3	−0.13	ACAT2
−0.096	−0.15	−0.31	−0.31	−0.16	EDN3
−0.26	−0.25	−0.56	−0.9	−0.29	CRH
−0.27	−0.15	−0.44	−0.82	−0.3	PVALB
−0.22	−0.098	−0.26	−0.56	−0.23	NEUROD6
−0.4	−0.25	−0.04	−0.75	−0.37	SST
−0.14	−0.015	−0.23	−0.41	−0.24	PENK
−0.1	−0.2	−0.12	−0.65	−0.15	NMU
−0.2	−0.13	−0.22	−0.45	−0.2	TAC1
−0.18	−0.097	−0.23	−0.35	−0.18	PPEF1

Fig. A1.4

Top 50 most differentially expressed genes across disorders, based on $\log_2$FC. Colors indicate signed $\log_{10}$(FDR corrected p-values). Text shows effect size ($\log_2$FC).

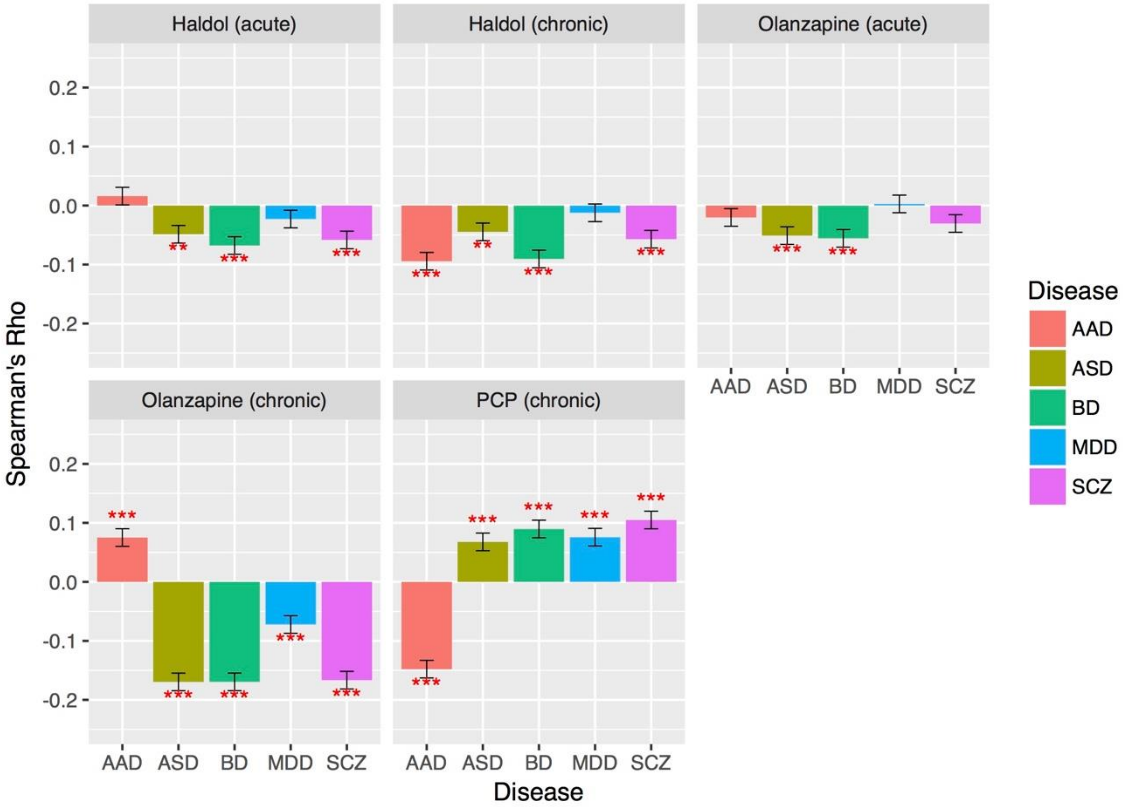
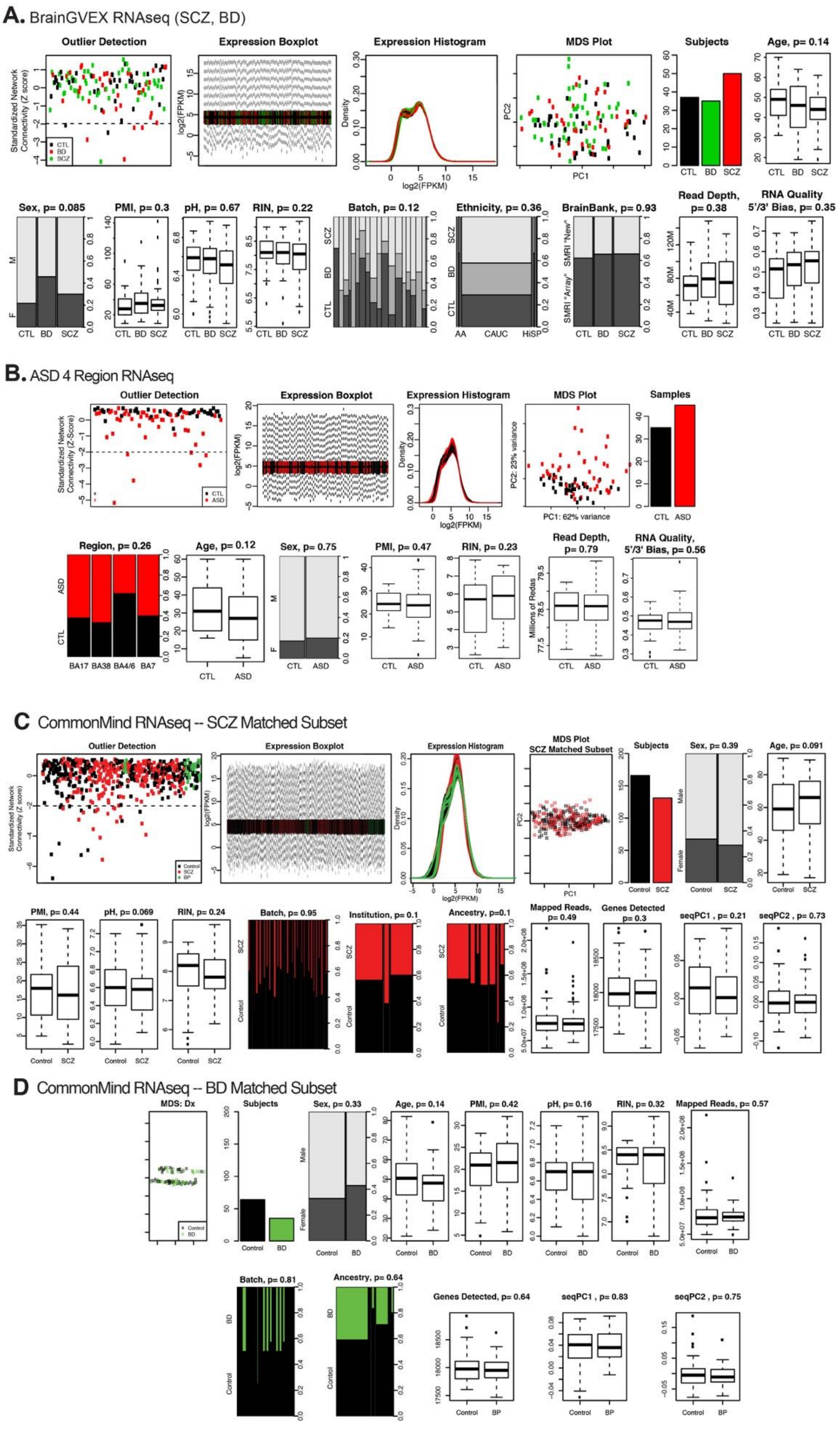


Fig. A1.5

Acute or chronic administration of antipsychotic medications to non-human primates (NHP) does not recapitulate disease specific transcriptome signatures. As described by Martin and colleagues (78), NHPs (n=7/group) were administered vehicle, acute haloperidol (8 mg/kg), acute olanzapine (3 mg/kg), chronic haloperidol (1.5 mg/kg BID x 4 weeks), chronic olanzapine (0.75 mg/kg BID x 4 weeks), or chronic PCP (1-2 mg/kg/day x 6 weeks). Gene expression results were obtained from (*78*), in which Affymetrix microarrays were used to profile anterior frontal cortex samples. Medication-induced gene expression changes ($\log_2$FC) were compared with disease-specific signatures using Spearman's correlation. Acute or chronic antipsychotic exposure shows significant negative correlation with ASD, BD, and SCZ, suggesting that disease-specific signatures are not caused by these psychiatric medications. Plot shows FDR-corrected p-values. *P < 0.05; **P < 0.01; ***P < 0.001.



A. BrainGVEX RNAseq (SCZ, BD)
Outlier Detection
Standardized Network Connectivity (Z score)
CTL
BD
SCZ
Expression Boxplot
log2(FPKM)
Expression Histogram
Density
log2(FPKM)
MDS Plot
PC1
PC2
Subjects
Age, p= 0.14
Sex, p= 0.085
PMI, p= 0.3
pH, p= 0.67
RIN, p= 0.22
Batch, p= 0.12
Ethnicity, p= 0.36
AA
CAUC
HiSP
BrainBank, p= 0.93
SMRI "Array"
SMRI "New"
Read Depth, p= 0.38
RNA Quality 5'/3' Bias, p= 0.35
B. ASD 4 Region RNAseq
Outlier Detection
Standardized Network Connectivity (Z-Score)
CTL
ASD
Expression Boxplot
Expression Histogram
MDS Plot
PC1: 62% variance
PC2: 23% variance
Samples
Region, p= 0.26
BA17
BA38
BA4/6
BA7
Age, p= 0.12
Sex, p= 0.75
PMI, p= 0.47
RIN, p= 0.23
Read Depth, p= 0.79
Millions of Redas
RNA Quality, 5'/3' Bias, p= 0.56
C CommonMind RNAseq -- SCZ Matched Subset
Outlier Detection
Control
SCZ
BP
Expression Boxplot
Expression Histogram
MDS Plot SCZ Matched Subset
Subjects
Sex, p= 0.39
Male
Female
Age, p= 0.091
PMI, p= 0.44
pH, p= 0.069
RIN, p= 0.24
Batch, p= 0.95
Institution, p= 0.1
Ancestry, p=0.1
Mapped Reads, p= 0.49
Genes Detected p= 0.3
seqPC1 , p= 0.21
seqPC2 , p= 0.73
D CommonMind RNAseq -- BD Matched Subset
MDS: Dx
Subjects
Sex, p= 0.33
Age, p= 0.14
PMI, p= 0.42
pH, p= 0.16
RIN, p= 0.32
Mapped Reads, p= 0.57
Batch, p= 0.81
Ancestry, p= 0.64
Genes Detected, p= 0.64
seqPC1 , p= 0.83
seqPC2 , p= 0.75
Control
BD
BP

Fig. A1.6

QC plots are shown for RNAseq replication datasets. (A) The BrainGVEX cohort analyzed in this study (synapse accession syn4590909) consisted of frontal cortex brain samples from subjects with SCZ (n=53), BD (n=47), and matched controls (n=53). One singular batch was removed. 7 samples were removed as outliers and 23 were removed due to covariate confounding to yield a final cohort of 122 samples. (B) The "ASD-pancortical" replication dataset (synapse accession syn4587609) consisted of brain samples across all four major cortical lobes from subjects with ASD (n=53) and matched controls (n=35 samples). Two samples were removed due to covariate confounding and 5 samples were removed as outliers, to yield a final cohort of 81. The CommonMind Consortium (synapse accession syn2759792) dataset consisted of dorsolateral prefrontal cortex brain samples from subjects with SCZ (n=262), BD (n=47), and controls (n=293). 27 samples were removed as outliers. To balance groups on demographic characteristics and other experimental covariates, we split the dataset into (C) SCZ matched (n=131 cases, n=166 controls) and (D) BD matched subsets (n=35 cases, n=65 controls).

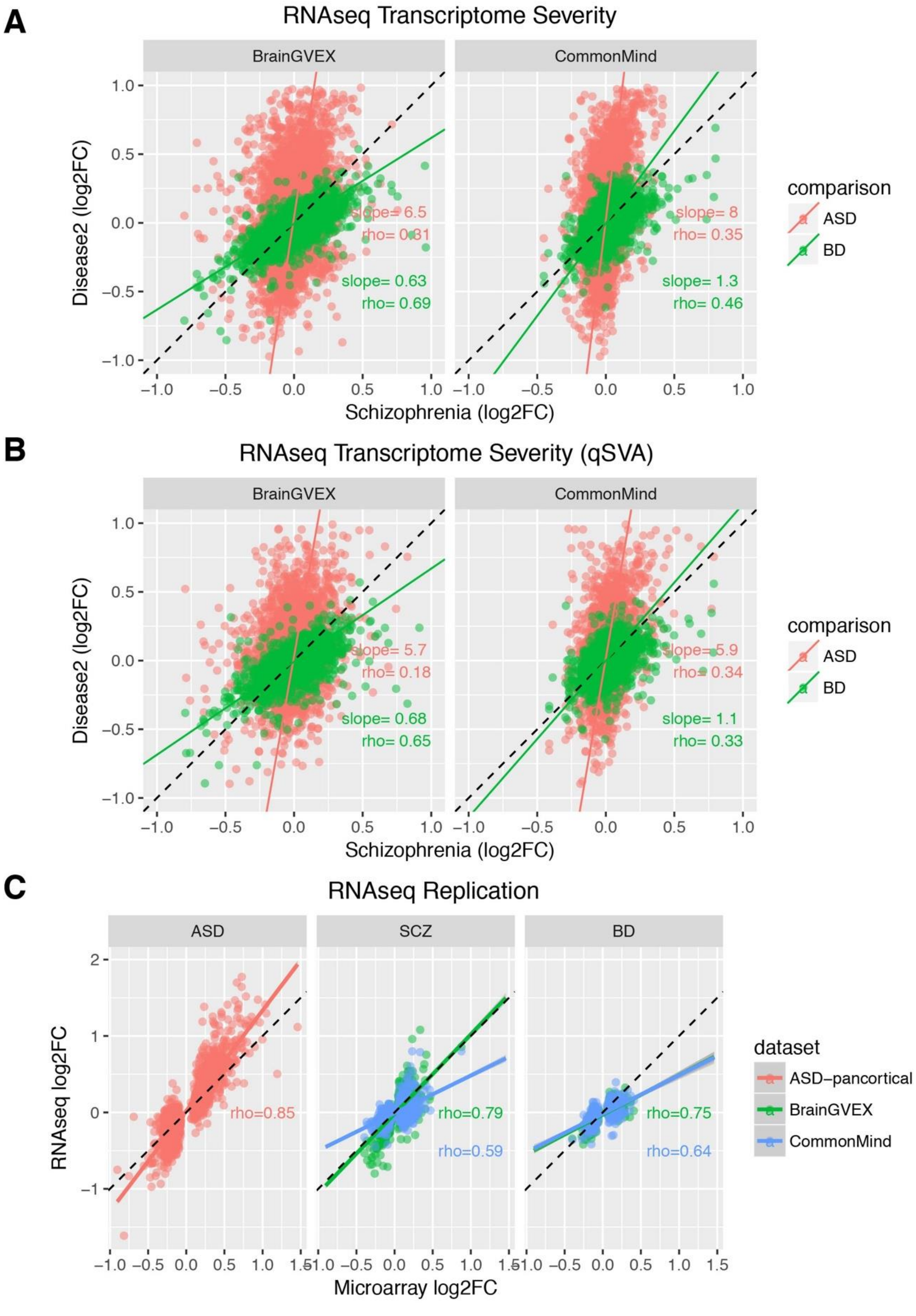
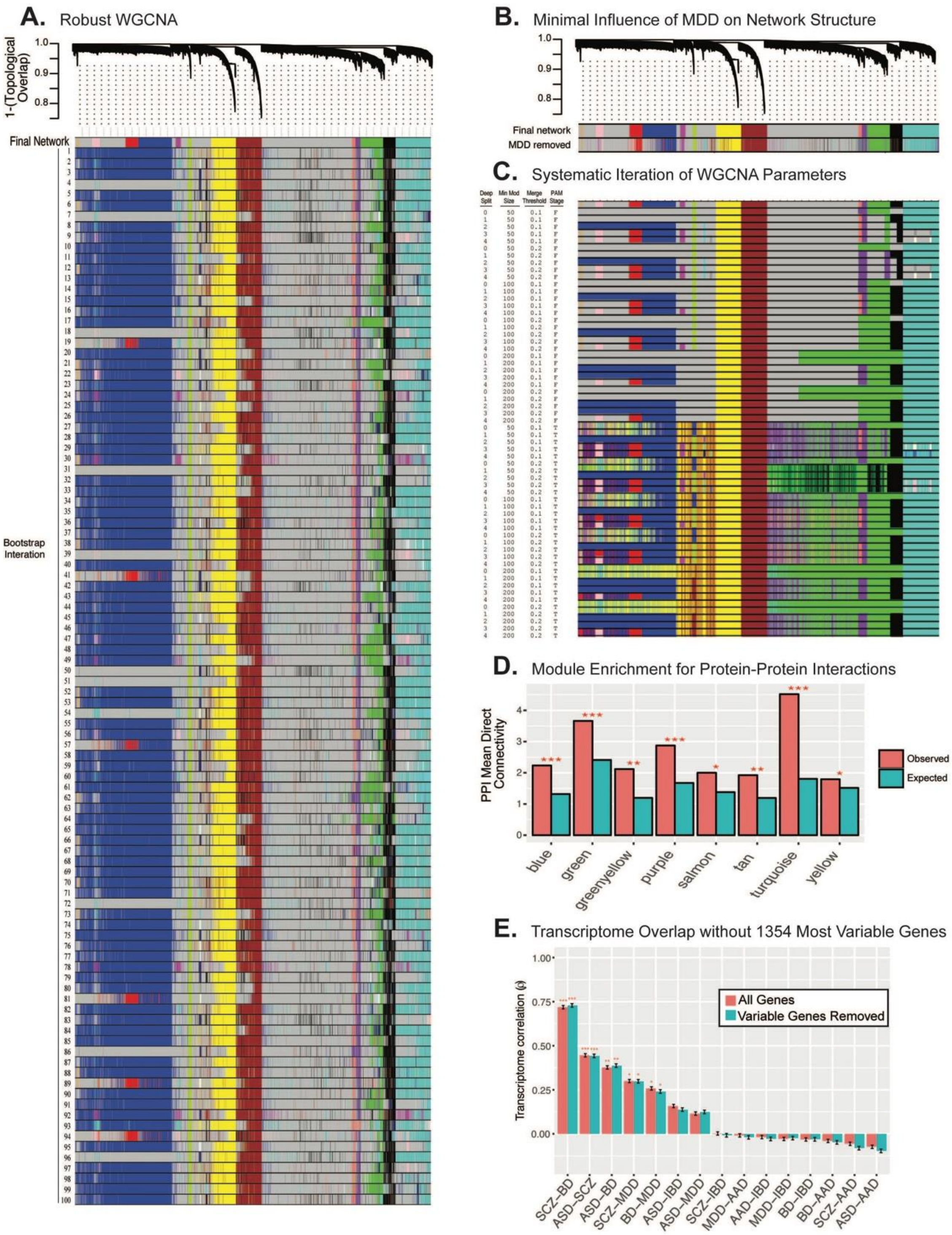

A
RNAseq Transcriptome Severity
BrainGVEX
CommonMind
Disease2 (log2FC)
Schizophrenia (log2FC)
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0
slope= 6.5
rho= 0.31
slope= 0.63
rho= 0.69
slope= 8
rho= 0.35
slope= 1.3
rho= 0.46
comparison
ASD
BD
B
RNAseq Transcriptome Severity (qSVA)
BrainGVEX
CommonMind
Disease2 (log2FC)
Schizophrenia (log2FC)
slope= 5.7
rho= 0.18
slope= 0.68
rho= 0.65
slope= 5.9
rho= 0.34
slope= 1.1
rho= 0.33
comparison
ASD
BD
C
RNAseq Replication
ASD
SCZ
BD
RNAseq log2FC
Microarray log2FC
2
1
0
−1
−1.0 −0.5 0.0 0.5 1.0 1.5
rho=0.85
rho=0.79
rho=0.59
rho=0.75
rho=0.64
dataset
ASD−pancortical
BrainGVEX
CommonMind

Fig. A1.7

Multiple RNAseq datasets were used for replication, including “BrainGVEX” (SCZ and BD) and “ASD-pancortical” (ASD) datasets through the PsychEncode Consortium. RNAseq results from subjects with SCZ, BD, and matched controls were obtained from the CommonMind Consortium. Data was processed using a standardized pipeline (see Methods) and log2FC estimates were calculated for each brain expressed gene. (A) RNAseq results replicate the gradient of transcriptomic severity observed from microarray data, as measured by the regression slope, with ASD >> SCZ ~ BD. This pattern was consistent across both BrainGVEX and CommonMind datasets. (B) qSVA-corrected RNAseq data show cross disorder transcriptomic overlap. (C) For each disorder, RNAseq results recapitulate the corresponding transcriptomic signature from microarray (FDR<0.05 genes). Spearman’s correction, all P’s<10^{-14}.



Deep Split	Min Mod Size	Merge Threshold	PAM Stage
0	50	0.1	F
1	50	0.1	F
2	50	0.1	F
3	50	0.1	F
4	50	0.1	F
0	50	0.2	F
1	50	0.2	F
2	50	0.2	F
3	50	0.2	F
4	50	0.2	F
0	100	0.1	F
1	100	0.1	F
2	100	0.1	F
3	100	0.1	F
4	100	0.1	F
0	100	0.2	F
1	100	0.2	F
2	100	0.2	F
3	100	0.2	F
4	100	0.2	F
0	200	0.1	F
1	200	0.1	F
2	200	0.1	F
3	200	0.1	F
4	200	0.1	F
0	200	0.2	F
1	200	0.2	F
2	200	0.2	F
3	200	0.2	F
4	200	0.2	F
0	50	0.1	T
1	50	0.1	T
2	50	0.1	T
3	50	0.1	T
4	50	0.1	T
0	50	0.2	T
1	50	0.2	T
2	50	0.2	T
3	50	0.2	T
4	50	0.2	T
0	100	0.1	T
1	100	0.1	T
2	100	0.1	T
3	100	0.1	T
4	100	0.1	T
0	100	0.2	T
1	100	0.2	T
2	100	0.2	T
3	100	0.2	T
4	100	0.2	T
0	200	0.1	T
1	200	0.1	T
2	200	0.1	T
3	200	0.1	T
4	200	0.1	T
0	200	0.2	T
1	200	0.2	T
2	200	0.2	T
3	200	0.2	T
4	200	0.2	T

Fig. A1.8

A number of steps were taken to ensure the robustness of the gene co-expression networks. (A) Robust, boostrapped version of WGCNA (rWGNCA) was used to reduce potential influence of outlier samples on network structure. We performed 100 iterations in which networks were created after first randomly subsetting 2/3 of the total samples. The resulting 100 networks were merged into one large, final consensus network. The individual sub-networks show highly consistent structure with each other and the final network. (B) MDD is the least heritable psychiatric disorder included in this analysis and shows significant clinical heterogeneity. To determine the specific influence of MDD samples on network architecture, we recreated co-expression networks after removing MDD samples. The resulting network structure is highly consistent, suggesting that MDD samples were not the major driver of observed results. (C) To determine the degree to which co-expression networks were contingent on the set of WGCNA parameters used, we systematically recreated networks using different parameter sets. The disease-associated modules reported in this study are preserved throughout the majority of this parameter search space. (D) To provide independent confirmation of the biological relevance of gene co-expression networks, we used DAPPLE to investigate the enrichment of protein-protein interactions (PPI) within each module. All disease-associated modules were enriched for PPI compared with a permuted null distribution, providing further evidence for a coherent biological role. (E) Finally, we assessed the influence of gene or sample-level variability on transcriptome overlap results. We used Bartlett's test to identify genes with significant within-group differences in expression variability (1354 genes at $P<0.05$, uncorrected). We then removed these 1354 genes and repeated our transcriptome overlap analyses between disease pairs. This had no appreciable effect on our results. *$P<0.05$, **$P<0.001$, ***$P<0.001$.

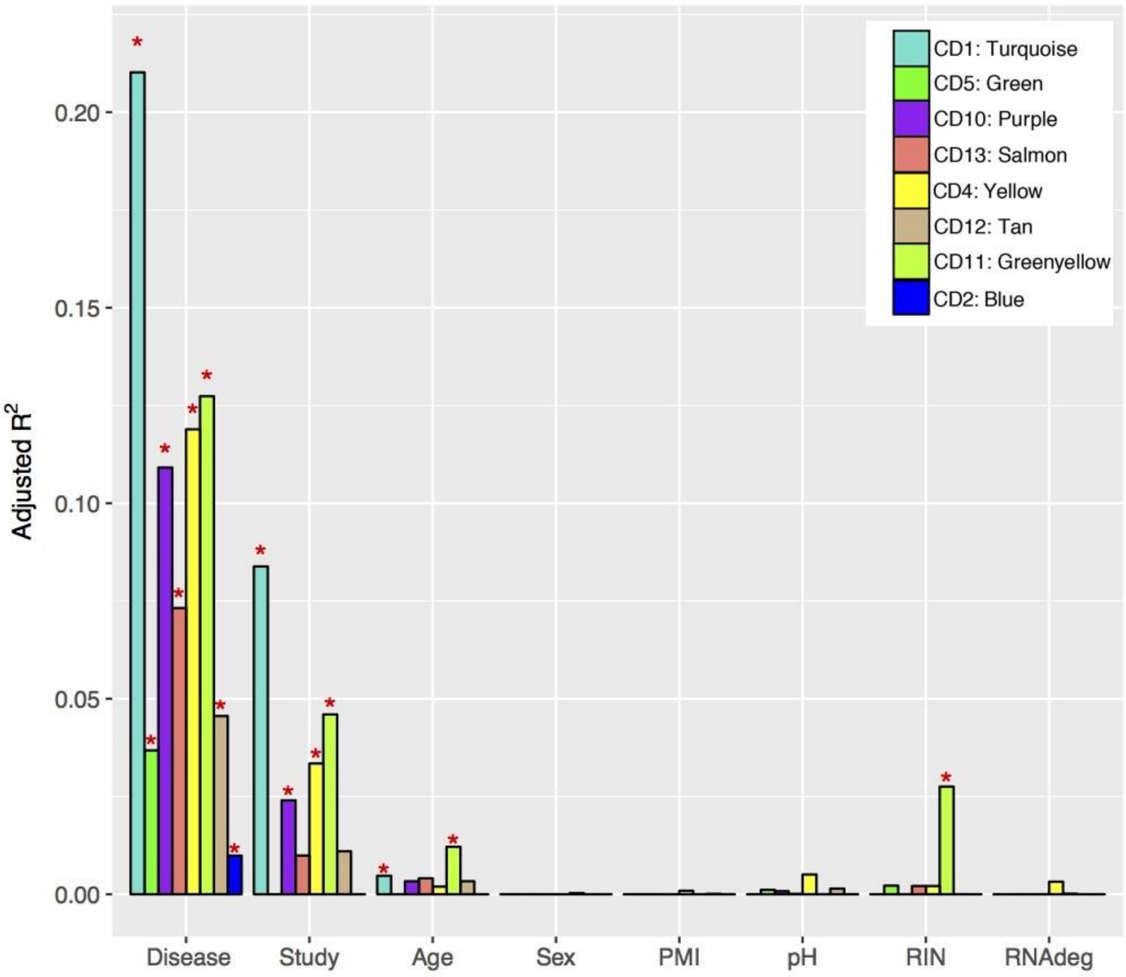


Fig. A1.9

Linear regression adjusted R^2 values are shown for each individual module eigengene-covariate association. * indicates nominal significance (ANOVA P<0.05, uncorrected for multiple comparisons). Disease status explains substantially more variance in module eigengene expression than for other biological or technical covariates. RNAdeg refers to 5'/3' bias. Full results are provided in **Data Table A1.2**.

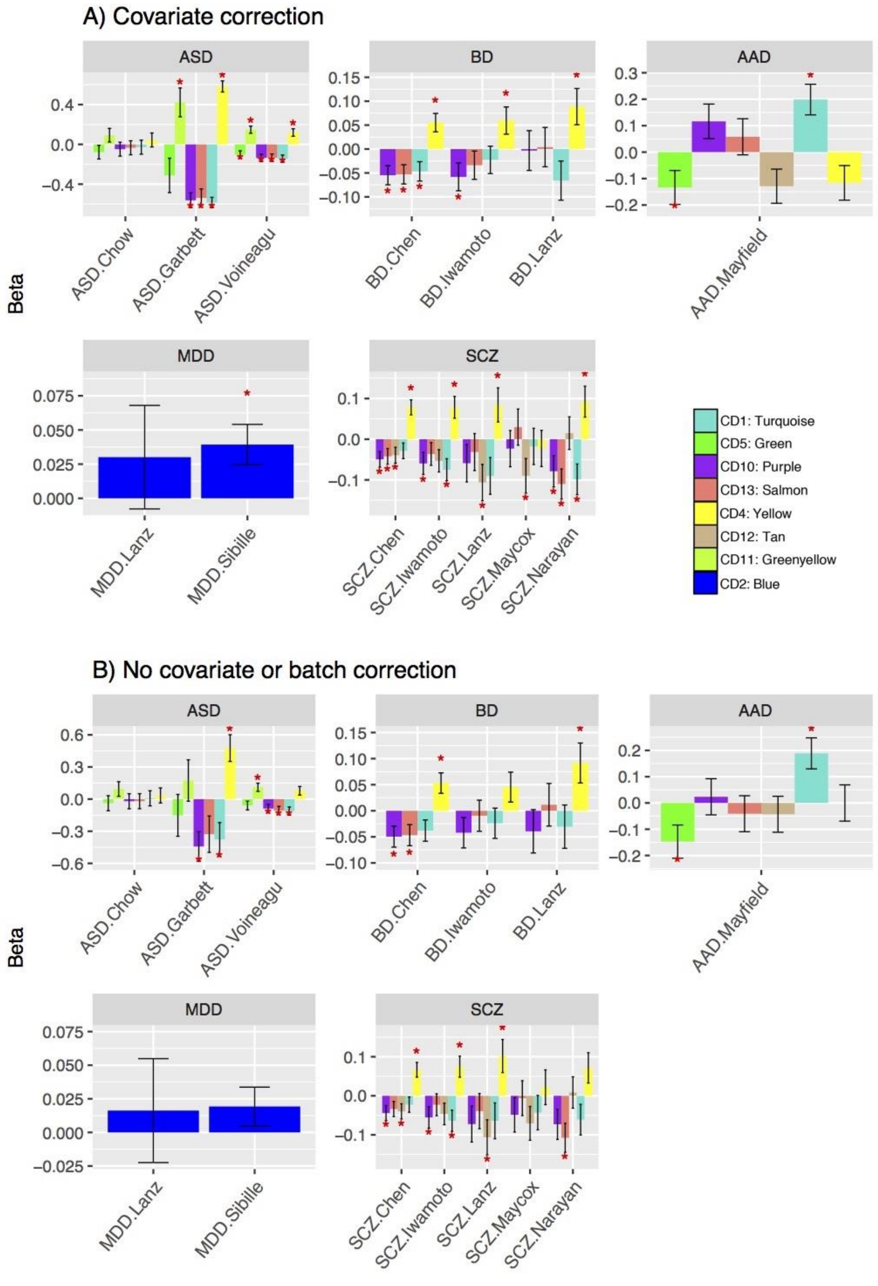
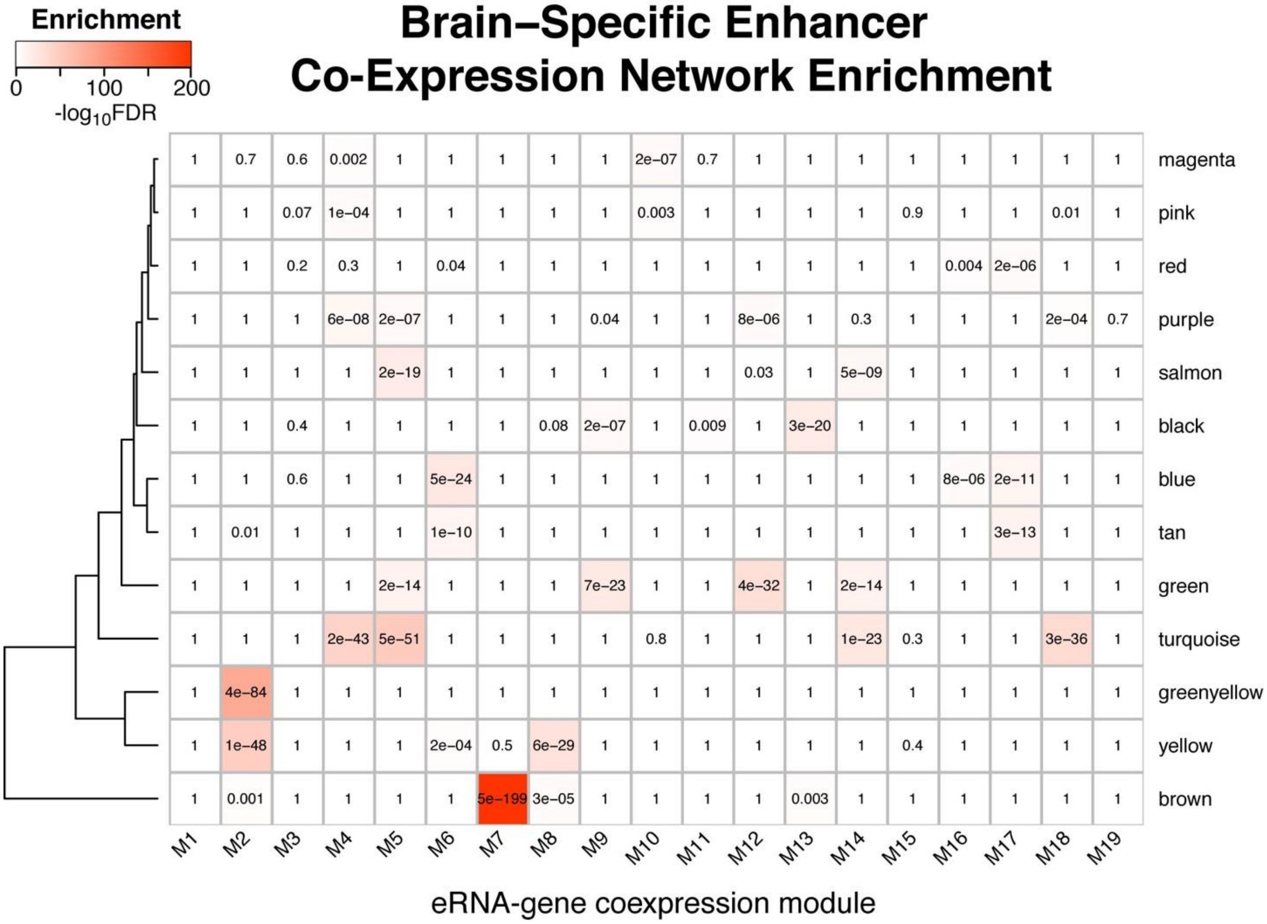


Fig. A1.10

Module stability across individual studies. A) Disease-associated modules are plotted for each individual study. Module eigengenes were calculated using covariate- and batch corrected data. (B) Disease-associated modules are shown without batch- or covariate correction.



	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16	M17	M18	M19
magenta	1	0.7	0.6	0.002	1	1	1	1	1	2e−07	0.7	1	1	1	1	1	1	1	1
pink	1	1	0.07	1e−04	1	1	1	1	1	0.003	1	1	1	1	0.9	1	1	0.01	1
red	1	1	0.2	0.3	1	0.04	1	1	1	1	1	1	1	1	1	0.004	2e−06	1	1
purple	1	1	1	6e−08	2e−07	1	1	1	0.04	1	1	8e−06	1	0.3	1	1	1	2e−04	0.7
salmon	1	1	1	1	2e−19	1	1	1	1	1	1	0.03	1	5e−09	1	1	1	1	1
black	1	1	0.4	1	1	1	1	0.08	2e−07	1	0.009	1	3e−20	1	1	1	1	1	1
blue	1	1	0.6	1	1	5e−24	1	1	1	1	1	1	1	1	1	8e−06	2e−11	1	1
tan	1	0.01	1	1	1	1e−10	1	1	1	1	1	1	1	1	1	1	3e−13	1	1
green	1	1	1	1	2e−14	1	1	1	7e−23	1	1	4e−32	1	2e−14	1	1	1	1	1
turquoise	1	1	1	2e−43	5e−51	1	1	1	1	0.8	1	1	1	1e−23	0.3	1	1	3e−36	1
greenyellow	1	4e−84	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
yellow	1	1e−48	1	1	1	2e−04	0.5	6e−29	1	1	1	1	1	1	0.4	1	1	1	1
brown	1	0.001	1	1	1	1	5e−199	3e−05	1	1	1	1	0.003	1	1	1	1	1	1

Fig. A1.11

Co-expression Enrichment for Brain Enhancer Regulation. A recent resource from Yao et al. (*79*) identified a set of robust enhancer RNAs (eRNA) expressed across human fetal and adult brain samples and used WGCNA to identify modules of eRNAs co-expressed with potential target genes. Here, we use Fisher's exact test to calculate the overlap between these eRNA-gene co-expression modules with cross disorder transcriptome modules. We find robust enrichment (FDR-corrected P<0.05) of putative eRNA regulation among all cross-disorder co-expression modules. Interestingly, no cross disorder modules showed overlap with M1 or M3, which were reported to represent fetal brain and cerebellum, respectively (*79*). Yet, all neuronal cross-disorder modules showed strong overlap with M5, which was found to be enriched for eRNA regulation in cerebral cortex.

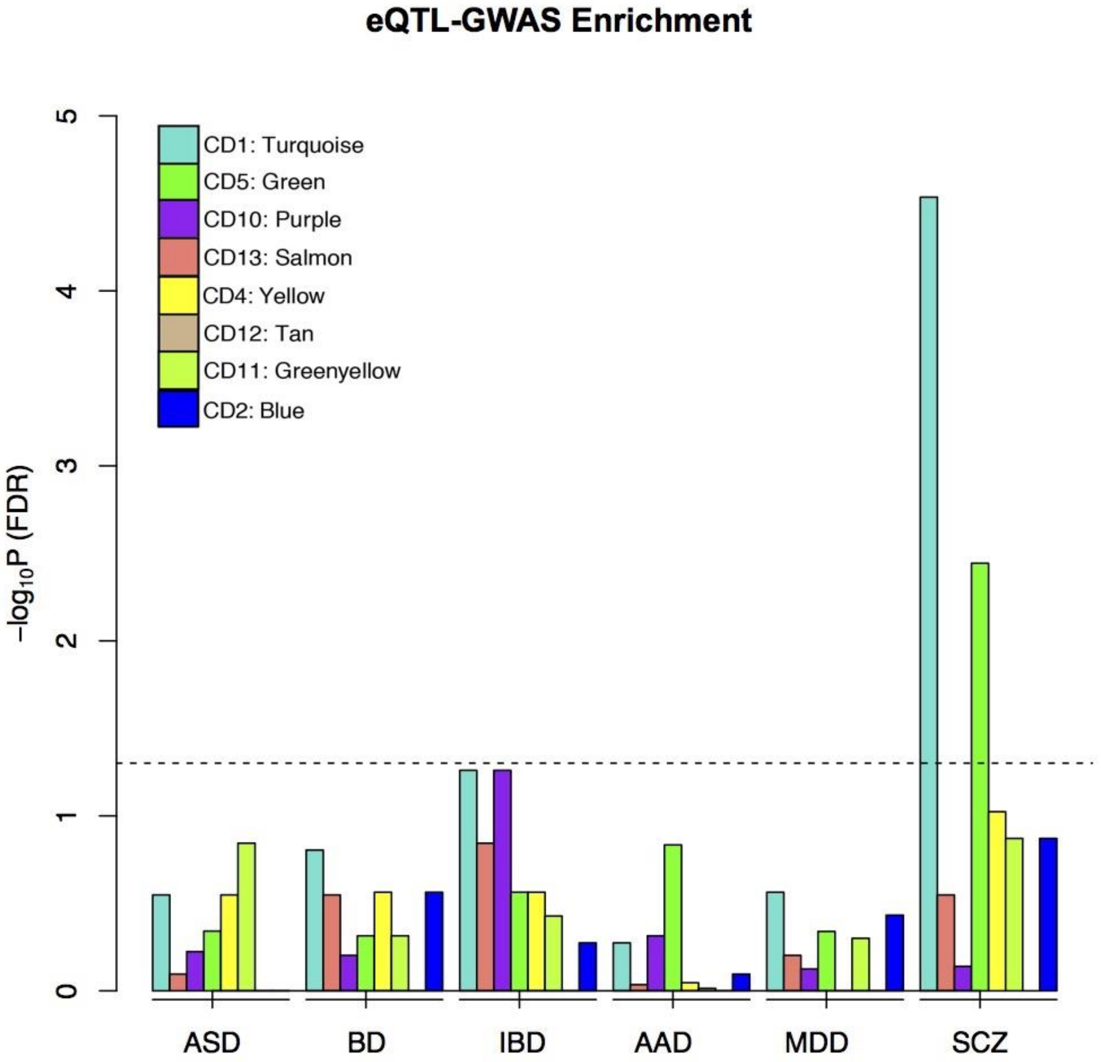


Fig. A1.12

Cerebral cortex eQTL summary statistics (all SNP-gene p-values) were downloaded from the GTEx (*80*) data browser (V6; http://www.gtexportal.org/home/)). Disease GWAS summary statistics were used as listed (**Table A1.3**). We calculated an empiric p-value for enrichment of eQTL SNPs associated with genes in a co-expression module and their significance in a GWAS dataset. For all genes in a co-expression module, we generated a list of eSNPs by choosing the most significant SNP in the eQTL summary statistics for each gene. Using this list of eSNPs, we obtained a p-value distribution in each GWAS dataset. These p-values were transformed into Z-scores using the qnorm function in R. We tested whether this Z-score distribution had a mean that was significantly lower than zero using a one sample t-test in R: t.test(Z-scores, alternative="less"). The t.test p-values for all combinations of co-expression modules and GWAS datasets were then FDR corrected. Results demonstrate that SCZ GWAS results are enriched for known brain eQTLs that regulate genes within turquoise and green (neuronal) modules.

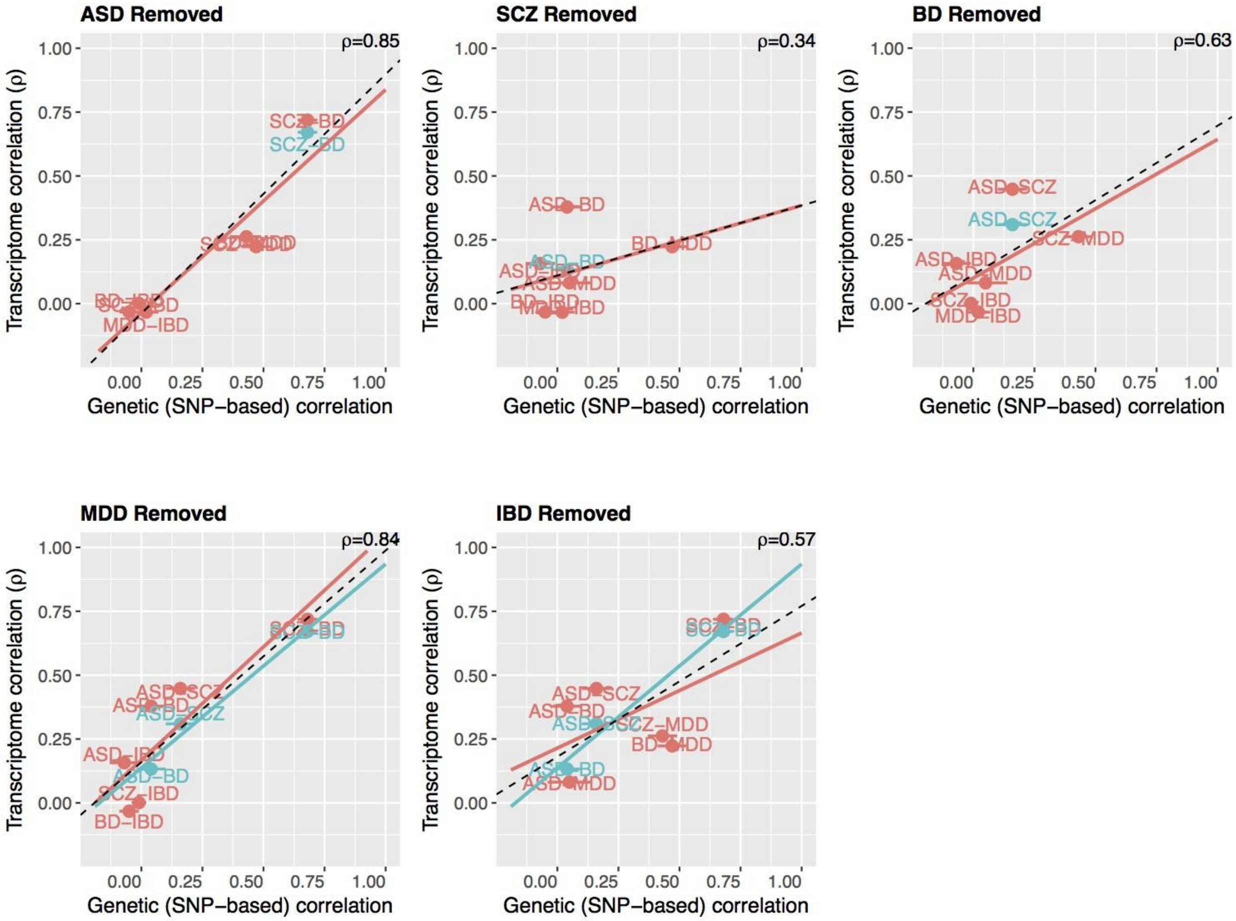


Fig. A1.13

Correlation between transcriptome similarity and genetic overlap after removing one disorder.

A1.4: Extended Tables

Table A1.1

Gene expression microarray datasets included in this study (*3-11, 75, 81-82*).

Disease	# Samples		Brain Region	Platform	Study	Data Source	Ref
	Cases	Controls					
ASD	29	29	BA9, BA41	Illumina Ref8 v3	Voineagu	GSE28521	(*3*)
	15	18	BA9/46	Illumina Ref8 v3	Chow	GSE28475	(*5*)
	6	6	BA41/42	Affy HG-U133 plus2	Garbett	mirnicslab. org	(*4*)
SCZ	51	50	Parietal cortex	Affy HuGene 1.0 ST	Chen	GSE35978	(*6*)
	15	19	BA46	Affy HG-U133 plus2	Lanz	GSE53987	(*75*)
	28	23	BA10	Affy HG-U133 plus2	Maycox	GSE17612	(*7*)
	35	34	BA46	Affy HG-U133A	Iwamoto	GSE12649	(*8*)
	30	29	BA46	Affy HG-U133 plus2	Narayan	GSE21138	(*9*)
BD	45	Included above (50)	Parietal cortex	Affy HuGene 1.0 ST	Chen	GSE35978	(*6*)
	17	Included above (19)	BA46	Affy HG-U133 plus2	Lanz	GSE53987	(*75*)
	32	Included above (34)	BA46	Affy HG-U133A	Iwamoto	GSE12649	(*8*)
MDD	17	Included above (19)	BA46	Affy HG-U133 plus2	Lanz	GSE53987	(*75*)
	70	70	BA9, BA25	Affy HG-U133 plus2	Sibille	GSE54567 GSE54568 GSE54571 GSE54572	(*10*)
AAD	17	15	Superior frontal cortex	Illumina HumanHT-12 V3	Mayfield	GSE29555	(*11*)
TOTAL	407	293					

Inflammatory Bowel Disease (Non-Brain) Comparison Datasets							
IBD	69	123	Colon punch biopsy	Illumina Human HT-12 V3	Granlund	E-MTAB-184	(*81*)
	128	73	Colon Biopsy	Agilent G4112A	Noble	GSE11223	(*82*)

Table A1.2

RNAseq replication of differential gene expression.

Disease	# DGE Genes (Microarray Discovery, FDR < 0.05)	RNAseq Replication Dataset	# DGE Genes Replicated in RNAseq (P<0.05, concordant direction)	Gene Replication Rate (%)	Overlap Odds Ratio	Overlap P
ASD	1679	ASD-Pancortical	1099	65.5%	6.4	$3.3x10^{-236}$
SCZ	1805	BrainGVEX	890	49.3%	4.5	$7.6x10^{-155}$
		CommonMind	520	28.8%	2.5	$1.8x10^{-46}$
BD	475	BrainGVEX	112	23.6%	3.9	$4.6x10^{-26}$
		CommonMind	118	24.8%	3.6	$1.5x10^{-24}$

Table A1.3

GWAS summary statistics used in this study.

Disorder / Trait	Consortium	Dataset	Date	Total Sample Size (Cases)	Ref
SCZ	PGC	SCZ2.snp.results.txt.gz	2014	82,315 (35,476)	(*40*)
BD	PGC	pgc.bip.2012-04.zip	2012	16,731 (7,481)	(*41*)
MDD	PGC	pgc.mdd.2012-04.zip	2012	18,759 (9,240)	(*42*)
ASD	iPSYCH	Data Table S5	2017	28,131 (8,605)	This study
ASD	PGC	PGC.ASD.euro.all.25Mar2015.txt.gz	2015	10,610 (5305)	(*43*)
AAD	AlcGen	Obtained directly from study authors	2011	23,347 (NA)	(*44*)
IBD	IIBDGC	EUR.IBD.gwas.assoc.txt	2015	34,652 (12,882)	(*45*)
Educational Attainment	SSGAC	EduYears_Main.txt.gz	2016	328,917	(*47*)
Subjective Well Being	SSGAC	SWB_Full.txt.gz	2016	298,420	(*46*)
Neuroticism		Neuroticism_Full.txt.gz		170,911	
Depressive Symptoms		DS_Full.txt.gz		161,460	

Please see the electronic tables associated with this document for additional data tables (Tables A1.1-5). Descriptions for these tables follow:

Additional Data Table A1.1 (separate file)

Differential gene expression summary statistics from microarray meta-analysis and RNAseq replication datasets.

Additional Data Table A1.2 (separate file)

Gene co-expression module data, including module-trait association statistics, module membership (kME) table, and enrichments for CNS cell-type markers, gene ontology pathways (gProlifeR and GO-Elite), transcription factor binding sites (TFBS), and transcription factor hub genes.

Additional Data Table A1.3 (separate file)

Compilation of genes affected by recurrent CNVs or rare, de novo variants (nonsynonymous vs silent) in ASD, SCZ, and control subjects.

Additional Data Table A1.4 (separate file)

Full results from LD score regression-based partitioned heritability analyses.

Additional Data Table A1.5 (separate file)

Summary statistics for the ASD GWAS performed on data from the iPSYCH consortium.

A1.5: Extended Bibliography

The bibliography/references for the appendix correspond with the bibliography for Chapter 2 (Section 2.6).

A2: Additional Results Accompanying Chapter 3

In this section of the appendix, I present additional results that accompany the work presented in chapter three. Everything included in this appendix section (A2) and chapter three was published in *Science*, in December 2018, with the title "Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder" (volume 362, no page numbers). Michael Gandal was the primary author for all of this work. As a co-author, in addition to my work presented in chapter three, I assisted with the interpretation and communication of results. Other co-authors included Pan Zhang, Evi Hadjimichael, Rebecca Walker, Chao Chen, Shuang Liu, Hyejung Won, Harm van Bakel, Merina Varghese, Yongjun Wang, Annie W. Shieh, Sepideh Parhami, Judson Belmont, Minsoo Kim, Patricia Moran Losada, Zenab Khan, Justyna Mleczko, Yan Xia, Rujia Dai, Daifeng Wang, Yucheng T. Yang, Min Xu, Kenneth Fish, Patrick R. Hof, Jonathan Warrell, Dominic Fitzgerald, Andrew E. Jaffe, Kevin White, Mette A. Peters, Mark Gerstein, Chunyu Liu, Lilia M. Iakoucheva, and Dalila Pinto. Daniel Geschwind was the senior author and main project director. All of these co-authors contributed to major and minor analyses for this project, and helped write, edit, and review the resulting manuscript. This work was associated with the PsychENCODE Consortium.

A2.1: Introduction

Developing more effective treatments for autism (ASD), schizophrenia (SCZ), and bipolar disorder (BD), three common psychiatric disorders that confer lifelong disability, is a major international public health priority (*3*). Studies have identified hundreds of causal genetic variants robustly associated with these disorders, and thousands more that likely contribute to their pathogenesis (*4*). However, the neurobiological mechanisms through which genetic variation imparts risk, both individually and in aggregate, are still largely unknown (*4*–*6*).

The majority of disease-associated genetic variation lies in non-coding regions (*7*) enriched for non-coding RNAs and *cis* regulatory elements that regulate gene expression and splicing of their cognate coding gene targets (*8*, *9*). Such regulatory relationships show substantial heterogeneity across human cell types, tissues, and developmental stages (*10*), and are often highly species-specific (*11*). Recognizing the importance of understanding transcriptional regulation and non-coding genome function, several consortia (*10*, *12–14*) have undertaken large-scale efforts to provide maps of the transcriptome and its genetic and epigenetic regulation across human tissues. Although some have included CNS tissues, a more comprehensive analysis focusing on the brain in both healthy and disease states is necessary to accelerate our understanding of the molecular mechanisms of these disorders (*1*, *15–17*).

We present results of the analysis of RNA-sequencing (RNA-Seq) data from the PsychENCODE Consortium (*17*), integrating genetic and genomic data from over 2000 well-curated, high-quality post-mortem brain samples from individuals with SCZ, BD, ASD, and controls (*18*). We provide a comprehensive resource of disease-relevant gene expression changes and transcriptional networks in the postnatal human brain (see Resource.PsychENCODE.org for raw data and annotations). Data was generated across eight studies (*2*, *19*, *20*), uniformly processed, and combined through a consolidated genomic data processing pipeline ((*21*); **Fig A3.1**), yielding a total of 2188 samples passing quality control (QC) for this analysis, representing frontal and temporal cerebral cortex from 1695 unique subjects across the lifespan, including 279 technical replicates (**Fig A3.2**). Extensive quality control steps were taken within and across individual studies resulting in the detection of 16,541 protein-coding and 9,233 non-coding genes with the Gencode v19 annotations ((*21*); **Fig A3.3**). There was substantial heterogeneity in RNA-Seq methodologies across cohorts, which we accounted for by including 28 surrogate variables and aggregate sequencing metrics as covariates in downstream analyses of differential expression (DE) at gene, isoform, and local splicing levels (*21*). Differential expression did not overlap with experimentally defined RNA degradation metrics in brain, indicating that results were not driven by RNA-quality confounds (**Fig A3.4**) (*22*).

To provide a comprehensive view of the transcriptomic architecture of these disorders, we characterize several levels of transcriptomic organization – gene-level, transcript isoform, local splicing, and co-expression networks – for both protein-coding and non-coding gene biotypes. We integrate results with common genetic variation and disease GWAS to identify putative regulatory targets of genetic risk variants. Although each level provides important disease-specific and shared molecular pathology that we highlight below, we find that isoform-level changes show the largest effects in disease brain, are most reflective of genetic risk, and provide the greatest disease specificity when assembled into co-expression networks.

We recognize that these analyses involve a variety of steps and data types and are necessarily multifaceted and complex. We have therefore organized the work into two major sections. The first is at the level of individual genes and gene products, starting with gene level transcriptomic analyses, isoform and splicing analyses, followed by identification of potential genetic drivers. The second section is anchored in gene network analysis, where we identify coexpression modules at both gene and isoform levels and assess their relationship to genetic risk. As these networks reveal many layers of biology, we provide an interactive web-browser to permit their in depth exploration (Resource.PsychENCODE.org).

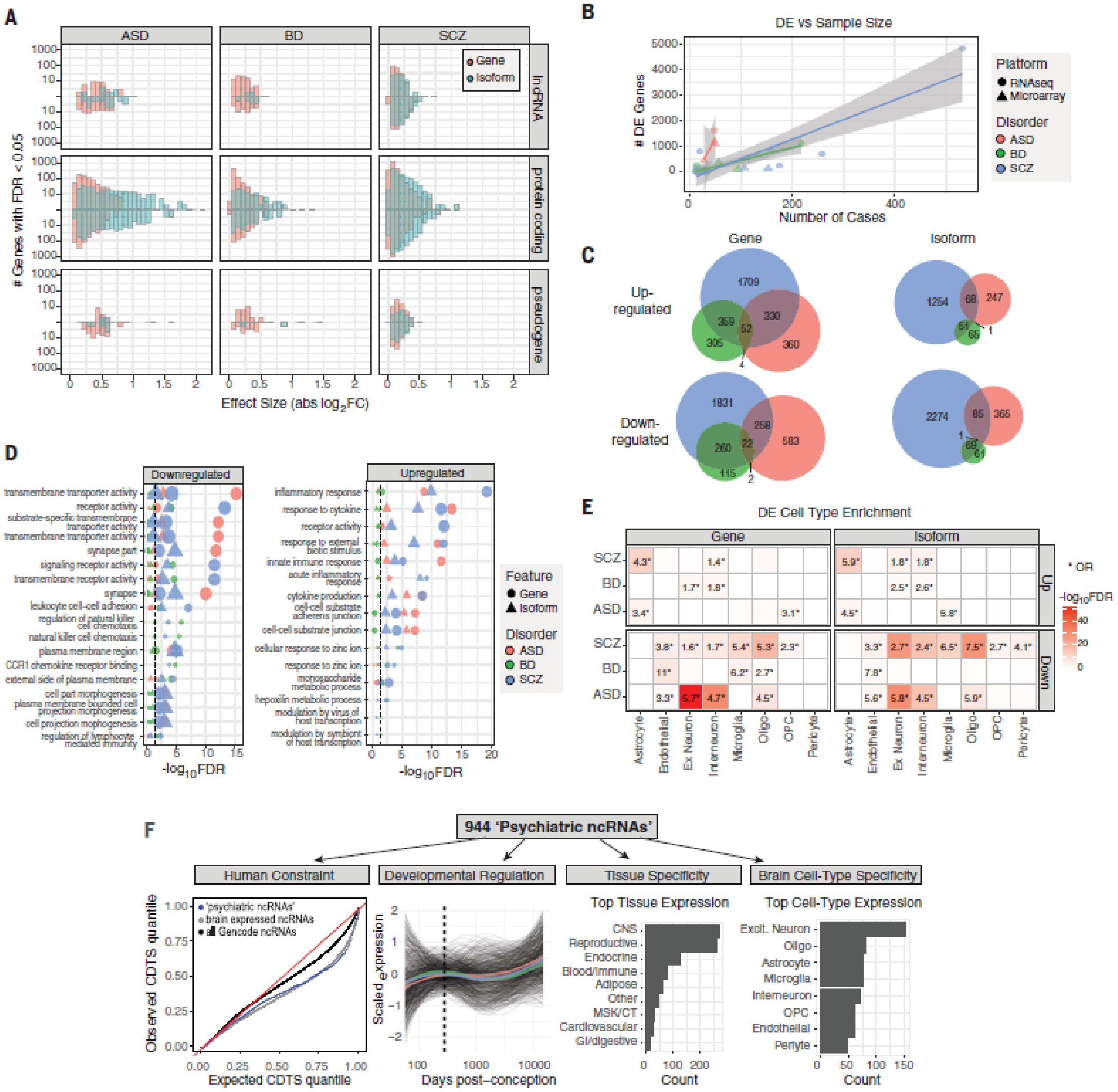


Figure A2.1. Gene and isoform expression dysregulation in psychiatric brain. A) Differential expression effect size (|log2FC|) histograms are shown for protein-coding, lncRNA, and pseudogene biotypes up or downregulated (FDR<0.05) in disease. Isoform-level changes (DTE; blue) show larger effect sizes than at gene level (DGE; red), particularly for protein-coding biotypes in ASD and SCZ. **B)** A literature-based comparison shows that the number of DE genes detected is dependent on study sample size for each disorder. **C)** Venn diagrams depict overlap among up or downregulated genes and isoforms across disorders. **D)** Gene ontology enrichments are shown for differentially expressed genes or isoforms. The top 5 pathways are shown for each disorder. **E)** Heatmap depicting cell type specificity of enrichment signals. Differentially expressed features show substantial enrichment for known CNS cell type markers, defined at the gene level from single cell RNA-Seq. **F)** Annotation of 944 unique non-coding RNAs DE in at least one disorder. From left to right: Sequence-based characterization of ncRNAs for measures of human selective constraint; brain developmental expression trajectory are similar across each disorder (colored lines represent mean trajectory across disorders); tissue, and CNS cell type expression patterns.

A2.2: Gene and isoform expression alterations in disease

RNA-Seq based quantifications enabled assessment of coding and non-coding genes and transcript isoforms, imputed using RSEM guided by Gencode v19 annotations (*21*, *23*). In accordance with previous results (*1*), we observed pervasive differential gene expression (DGE) in ASD, SCZ, and BD (n=1611, 4821, and 1119 genes at FDR<0.05, respectively; **Fig A2.1A**; **Table A3.1**). There was substantial cross-disorder sharing of this DE signal and a gradient of transcriptomic severity with the largest changes in ASD compared with SCZ or BD (ASD vs SCZ, mean $|\log_2 FC|$ 0.26 *vs* 0.10, $P<2x10^{-16}$, Kolmogorov-Smirnov (K-S) test; ASD *vs* BD, mean $|\log_2 FC|$ 0.26 *vs* 0.15, $P<2x10^{-16}$, K-S test), as observed previously (*1*). Altogether, over a quarter of the brain transcriptome was affected in at least one disorder (**Fig A2.1A-C**; complete gene list, **Table A3.1**).

DGE results were highly concordant with previously published datasets for all three disorders, although some had overlapping samples (**Fig A3.4**). We observed significant concordance of DGE effect sizes with those from a microarray meta-analysis of each disorder (ASD: ρ=0.8, SCZ: ρ=0.78, BD: ρ=0.64, Spearman ρ of $\log_2$FC, all P's<10^{-16}, (*1*)) and with previous RNA-Seq studies of individual disorders (ASD: ρ=0.96, ref (*19*); SCZ ρ=0.78, ref (*2*); SCZ ρ=0.80, ref (*24*); BD ρ=0.85, ref (*1*); Spearman ρ of $\log_2$FC, all P's<10^{-16}). These DE genes exhibited substantial enrichment for known pathways and cell type specific markers derived from single nucleus RNA-Seq in human brain (**Fig A2.1D-E**) (*21*), consistent with previously observed patterns (*1*, *19*).

Expanding these analyses to the transcript isoform-level, we observe widespread differential transcript expression (DTE) across ASD, SCZ, and BD (n=767, 3803, and 248 isoforms at FDR<0.05, respectively; **Table A3.1**). Notably, at the DTE level, the cross-disorder overlap was significantly attenuated (**Fig A2.1C**), suggesting that alternative transcript usage and/or splicing confers a substantial portion of disease specificity. In addition to greater disease specificity, isoform-level alterations in disease exhibited substantially larger effect sizes compared with gene-level changes (mean $|\log_2 FC|$ 0.25 *vs* 0.14, $P<2x10^{-16}$, K-S test), particularly for protein coding biotypes (**Fig A2.1A**), consistent with recent work demonstrating the importance of splicing dysregulation in disease pathogenesis (*25*). Furthermore, although isoform and gene-level changes were overall similar in terms of pathways and cell types affected (e.g. **Fig A2.1D**-**E**), isoform-level analysis identified DE transcripts that did not show DGE ('isoform-only DE'), including 811 in SCZ, 294 in ASD, and 60 in BD. These isoform-only DE genes were more likely to be downregulated than upregulated in disease (one sample t-test, $P<10^{-16}$), were most significantly enriched in excitatory neuron clusters (OR's > 4, Fisher's exact test, FDR's<10^{-10}), and showed significant enrichment for neuron projection development, mRNA metabolism, and synaptic pathways (FDR<$3x10^{-3}$; **Table A3.1**). To validate DTE results, we performed PCR on several selected transcripts in a subset of ASD, SCZ and control samples (*21*), and find significant concordance in fold-changes compared with those from RNA-Seq data (**Fig A3.5A-B**). Together, these results suggest that isoform-level changes are most reflective of neuronal and synaptic dysfunction characteristic of each disorder.

Though there are multiple shared pathways at the levels of DGE or DTE across disorders, there are also several distinctive features (**Fig A3.1D**-**E**). Disorder-specific pathway enrichments include decreased transmembrane transport, synapse and synaptic components, with increases in innate immune response genes in ASD; decreased chemokine signaling, regulation of lymphocyte regulated immunity and natural killer cell chemotaxis in BD; and decreased signaling receptor and transmembrane receptor activity with increases in genes involved in the inflammatory response in SCZ. With regards to cell type enrichments (**Fig A2.1E**), although there was substantial downregulation of neuronal synaptic and signaling genes, only SCZ and BD also showed increases in the expression of a distinct subset of excitatory and inhibitory neuronal genes, whereas SCZ and ASD showed upregulation of genes expressed in astrocytes. ASD was the only disorder with enrichment of microglia among upregulated features.

A2.3: Differential expression of the non-coding transcriptome

Non-coding RNAs (ncRNAs) represent the largest class of transcripts in the human genome and are associated with complex phenotypes (*26*). However, most have limited functional annotation, particularly in human brain and have not been studied in psychiatric disease. Based on Gencode annotations, we identify 944 ncRNAs exhibiting gene- or isoform-level DE in at least one disorder (herein referred to as 'neuropsychiatric (NP) ncRNAs' (*21*)), 693 of which were DE in SCZ, 178 in ASD, and 174 in BD, of which 208, 60, and 52 are annotated as intergenic long non-coding RNAs (lincRNAs), in each disorder, respectively. To place these NPncRNAs within a functional context, we examined expression patterns across human tissues, cell types, and developmental time periods, as well as sequence characteristics including evolutionary conservation, selection, and constraint. We highlight several noncoding genes exhibiting DE across multiple disorders (**Fig A3.6**) and provide comprehensive annotations for each NPncRNA (**Table A3.2**)**,** including cell type specificity, developmental trajectory, and constraint, to permit placement of these NPncRNAs within a functional context in human brain.

As a class, NPncRNAs were under greater selective constraint compared to all Gencode annotated ncRNAs (**Fig A2.1F**), consistent with the observed increased purifying selection in brain-expressed genes (*27*). We identify 74 NPncRNAs (~8%) under purifying selection in humans, with average exon-level context-dependent tolerance scores (CDTS) below the 10th percentile (*21*). Of the 944 NP ncRNAs, 212 exhibited broad and non-specific expression patterns across cell types, whereas 66 showed specific expression within a single cell type class (**Table A3.2**). These data provide a foundation for understanding cell type specific, circuit level aspects of lncRNA function in neuropsychiatric disease. Two notable examples are, *LINC00996,* which is downregulated in SCZ ($\log_2$FC -0.71, FDR<5x10^{-11}) and BD ($\log_2$FC -0.45, FDR=0.02) and restricted to microglia in brain (**Fig A3.6**), and *LINC00343,* expressed in excitatory neurons, and downregulated in BD ($\log_2$FC -0.33, FDR=0.012) with a trend in SCZ ($\log_2$FC -0.15, FDR 0.065).

A2.4: Local splicing dysregulation in disease

Isoform-level diversity is achieved by combinatorial use of alternative transcription start sites, polyadenylation, and splicing (*28*). We next used LeafCutter (*29*) to assess local differential splicing (DS) differences in ASD, SCZ and BD using *de novo* aligned RNA-seq reads, controlling for the same covariates as DGE/DTE (**Fig A3.7**). This approach complements DTE by considering aggregate changes in intron usage affecting exons that may be shared by multiple transcripts and thus, is not restricted to the specified genome annotation (*21*). Previous studies have highlighted the contribution of local DS events in ASD (*19*, *30*) and in smaller cohorts in SCZ (*2*, *24*) and BD (*31*).

We identified 515 DS intron clusters in 472 genes across all disorders (FDR<0.1), 117 of which (25%) contained one or more novel exons (**Table A3.3**; **Fig A2.2A**). Validation of DS changes for 9 genes in a subset of cases and controls (n= 5-10 in each group) by semiquantitative RT-PCR showed percent spliced-in (PSI) changes consistent with those reported by LeafCutter (**Fig A3.5C-E**). The most commonly observed local splicing change was exon skipping (41-60%), followed by alternative 5' exon inclusion (e.g. due to alternative promoter usage; 11-21%) and alternative 3' splice site usage (5-18%) (**Table A3.3**; **Fig A3.8A**). DS genes overlapped significantly with DTE results for ASD and SCZ (**Fig A3.8B**), but not BD, which likely still remains underpowered. There was significant cross-disorder correlation in PSI changes (Spearman's ρ=0.59 SCZ-BD, ρ=0.52 SCZ-ASD, all P<10^{-4}) and subsequently, overlap among DS genes (**Fig A2.2A**-**B**), although the majority of splicing changes still are disorder specific. Only two genes, *DTNA* and *AHCYL1*, were significantly DS in all three disorders (**Fig A3.9**). DS genes showed significant (FDR<0.05) enrichment for signaling, cell communication, actin cytoskeleton, synapse, and neuronal development pathways across disorders (**Figs A2.2C**, **A3.8C**), and were predominantly expressed in neuronal cell types, astrocytes (in ASD, SCZ), microglia and oligodendrocytes (in SCZ) (**Fig A2.2D**). Disorder specific pathways implicated by splicing dysfunction include plasma membrane receptor complex, endocytic vesicle, regulation of cell growth and cytoskeletal protein binding in ASD; angiotensin receptor signaling in BD; and GTPase receptor activity, neuron development and actin cytoskeleton in SCZ. We also find significant enrichment of splicing changes in targets of two RNA binding proteins that regulate synaptic transmission and whose targets are implicated in both ASD and SCZ, the neuronal splicing regulator *RBFOX1* (FDR=5.16x10^{-11}) (*32*) and the fragile X mental retardation protein (FMRP) (FDR=3.10x10^{-21}) (*33*). Notably, 48 DS genes (10%; FDR=8.8x10^{-4}) encode RNA binding proteins or splicing factors (*34*), with at least six splicing factors also showing DTE in ASD (*MATR3*), SCZ (*QKI*, *RBM3*, *SRRM2*, *U2AF1*) or both (*SRSF11*).

Many differential splicing events show predictable functional consequences on protein isoforms. Notable examples include *GRIN1* and *NRXN1*, which are known risk loci for neurodevelopmental disorders (*35*, *36*). *GRIN1* encodes the obligatory subunit of the NMDA-type glutamate ionotropic receptors, is upregulated in SCZ and BD and shows increased skipping of exon 4 in both ASD and SCZ that impacts its extracellular ligand-binding domain (**Fig A2.2E**-**G**). *NRXN1* is a heterotypic, presynaptic cell adhesion molecule that undergoes extensive alternative splicing and plays a key role in the maturation and function of synapses (*35*, *37*). We observed various DS and/or differential transcript usage (DTU) changes in *NRXN1* in ASD, SCZ and/or BD (**Fig A2.2H**-**K**). An exon skipping event in ASD disrupts a laminin domain in *NRXN1* (**Fig 3.2I-J**), while the isoform expression switch affects the expression of laminin, neurexin-like and EGF-like domains; changes which are predicted to have major effects on its function (**Fig A2.2H**). Another example is *CADPS*, which is located within an ASD GWAS risk locus and supported by Hi-C defined chromatin interactions as a putative target gene (*38*) and manifests multiple isoform and splice alterations in ASD (**Fig A3.9**; **Tables A3.1** and **A3.3**).

We found significant overlap (42%, P=3.42x10^{-27}; Fisher's exact test) of the ASD DS intron clusters and splicing changes identified in a previous study (*19*) that used a different method and only a subset of the samples in our ASD and control cohorts (**Table A3.3**). Overall, this examination of local splicing across three major neuropsychiatric disorders, coupled with the analysis of isoform-level regulation, emphasizes the need to understand the regulation and function of transcript isoforms at a cell type specific level in the human nervous system.

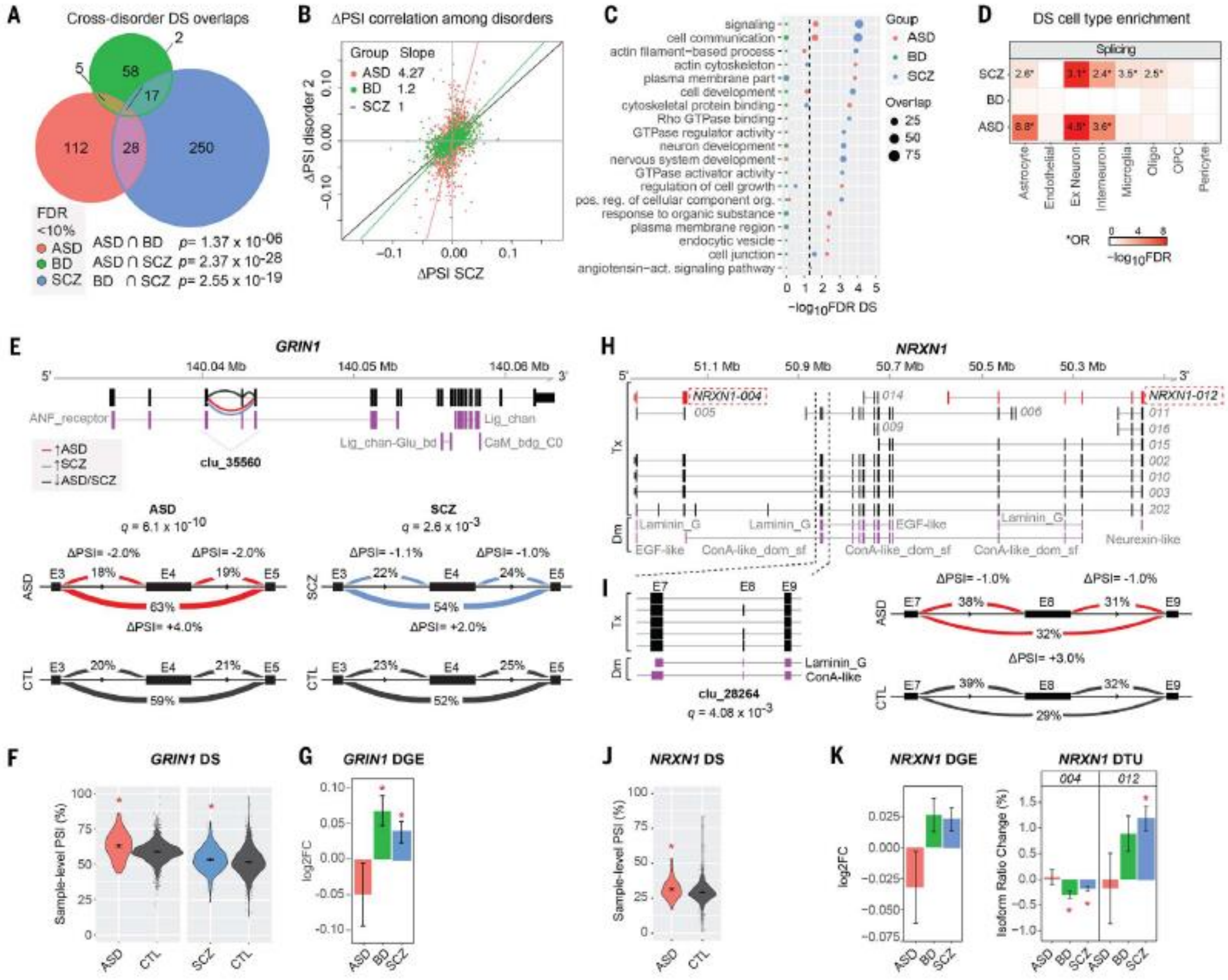


Figure A2.2. Aberrant local splicing and isoform usage in ASD, SCZ and BD. **A)** Venn diagram showing cross-disorder overlap for 472 genes with significant differentially spliced (DS) intron clusters (FDR< 10%) identified by LeafCutter. P values for hypergeometric tests of pairwise overlaps between each

disorder are shown at the bottom. **B)** Scatter plots comparing percent spliced-in (PSI) changes for all 1,287 introns in 515 significant DS clusters in at least one disorder, for significant disease pairs SCZ *vs* ASD and SCZ *vs* BD (Spearman's ρ=0.52 and ρ=0.59, respectively). Principal component regression lines are shown in red, with regressions slopes for ASD and BD delta PSI compared to SCZ in the top-left corner. **C)** Top 10 gene ontology (GO) enrichments for DS genes in each disorder (see also **Fig A3.8C**). **D)** Significant enrichment for neuronal and astrocyte markers (ASD and SCZ), as well as oligodendrocyte and microglia (SCZ) cell type markers in DS genes. *Odds Ratio (OR) is given only for FDR< 5% and OR> 1. Oligo - oligodendrocytes; OPC - oligodendrocyte progenitor cells. **E)** A significant DS intron cluster in *GRIN1* (clu_35560; chr9:140,040,354-140,043,461) showing increased exon 4 (E4) skipping in both ASD and SCZ. Increased or decreased intron usage in ASD/SCZ cases compared to controls are highlighted in red and blue, respectively. Protein domains are annotationed as ANF_receptor - Extracellular receptor family ligand binding domain; Lig_chan - Ionotropic glutamate receptor; Lig_chan-Glu_bd - Ligated ion channel L-glutamate- and glycine-binding site; CaM_bdg_C0 - Calmodulin-binding domain C0 of NMDA receptor NR1 subunit. Visualization of splicing events in cluster clu_35560 with the change in PSI (ΔPSI) for ASD (left) and SCZ (right) group comparisons. FDR-corrected p-values (q) are indicated for each comparison. Covariate-adjusted average PSI levels in ASD or SCZ (red) *vs* CTL (blue) are indicated at each intron. **F)** Violin-plots with the distribution of covariate-adjusted PSI per sample for the intron skipping E4 are shown for each disease group comparison. **G)** DGE for *GRIN1* in each disorder (*FDR< 5%). **H)** Whole-gene view of *NRXN1* highlighting (dashed lines) the intron cluster with significant DS in ASD (clu_28264; chr2:50,847,321-50,850,452), as well as transcripts NRXN1-004 and NRXN1-012 that show significant DTU in SCZ and/or BD. Protein domain mappings are shown in purple. DM - Protein domains; Tx - Transcripts. ConA-like_dom_sf - Concanavalin A-like lectin/glucanase domain. EGF-like - Epidermal growth factor-like domain; Laminin_G - Laminin G domain; Neurexin-like - Neurexin/syndecan/glycophorin C domain. **I)** Left: close-up of exons and protein domains mapped onto the DS cluster, and FDR-corrected *p*-value (*q*). Right: visualization of introns in cluster clu_28264 with their change in percent spliced in (ΔPSI). Covariate-adjusted average PSI levels in ASD (red) *vs* CTL (blue) are indicated for each intron. **J)** Violin-plots with the distribution of covariate-adjusted PSI per sample for the largest intron skipping exon 8 (E8). **K)** Bar plots for changes in gene expression and transcript usage for NRXN1-004 and NRXN1-012 (*FDR< 5%).

A2.5: Transcriptome-wide association

Seeking to integrate genetic information with gene expression data, we performed a formal transcriptome-wide association study (TWAS; (*46*)) to directly identify those genes whose *cis*-regulated expression is associated with disease (*21*). Compared with the PRS-based approach, TWAS is restricted to *cis* effects on expression and genes with evidence of heritable expression patterns in our dataset. TWAS and related methods have the advantage of collapsing signals onto specific genes, reducing multiple comparisons and increasing power for association testing (*46*, *47*). Further, by imputing the *cis*-regulated heritable component of brain gene expression into the association cohort, TWAS enables direct prediction of the transcriptomic effects of disease-associated genetic variation, identifying potential mechanisms through which variants may impart risk. However, the limited size of brain eQTL datasets to date has necessitated use of non-CNS tissues to define TWAS weights (*46*), limiting identification of brain relevant genetic regulation. Given the substantial enrichment of psychiatric GWAS signal within CNS expressed regulatory elements (*39*), we reasoned that our dataset would provide substantial increased power and specificity.

We identify 14,750 genes with heritable *cis*-regulated expression in brain in the PsychENCODE cohort, enabling increased transcriptomic coverage for detection of association signal (**Fig A2.3**). In ASD, TWAS prioritizes 12 genes across 3 genomic loci (Bonferroni-corrected *P*<0.05; **Fig A2.3**). This includes the 17q21.31 region, which showed multiple PRS associations as described above, but did not reach genome-wide significance in the largest GWAS to date (*38*), highlighting the complementarity of the TWAS approach. Of the seven TWAS-significant genes at 17q21, conditional analysis prioritizes one – *LRRC37A,* which is further supported by a Hi-C interaction in fetal brain (*38*). *LRRC37A* is intriguing due to its primate-specific evolutionary expansion, loss-of-function intolerance, and expression patterns in brain and testis (*45*). However, it is also possible that common variants in GWAS are indirectly tagging the known common inversions or other recurrent structural variants previously identified at this locus (*45*). TWAS additionally prioritizes *XKR6* and *PINX1* as well as *PLK1S1* and *NKX2-2* at ASD loci on chromosomes 8 and 20, respectively (**Fig A2.3**; **Table A3.4**; (*21*)).

In BD, TWAS prioritizes 17 genes across 14 distinct loci (Bonferroni-corrected *P*<0.05; **Fig A2.3**; **Table A3.4**), none of which were DE. These included *VPS45*, *TMEM258*, *CKMT1A*, *HLF*, *ILF3*, *LMAN2L*, *DCLK3, BMPR1B*, *SNAP91*, and *SYS1-DBNDD2*. At loci with multiple hits, we applied conditional and colocalization analyses (*21*) to further finemap these regions, permitting prioritization of single top candidate genes – *CDKN2C*, *UBE2Q2L*, and *HAPLN4*. The two isoforms showing PRS associations in BD (*NCALD*, *SNF8*) were not significant in TWAS, likely due to lack of a nearby genome-wide significant locus, or due to isoform-specific regulation, suggesting those expression changes may be driven by *trans*-acting factors.

Finally, TWAS identifies 193 genes in SCZ, of which 107 remain significant following conditional analysis at each gene within multi-hit loci. Excluding the MHC region, there remained 164 significant genes representing 78 genome-wide significant GWAS loci (**Fig A2.3**; **Table A3.4**). A previous TWAS study in SCZ primarily based on non-neural tissue prioritized 157 genes, of which 37 coincide with the current results, a highly significant overlap (OR 60.7, $p<10^{-42}$, Fisher's exact test). Moreover, 60 TWAS prioritized genes overlapped with the list of 321 'high confidence' SCZ risk genes identified in the companion manuscript (*18*), identified using gene regulatory networks and a deep learning approach (OR 34.7, $p<10^{-60}$, Fisher's exact test). Twenty one genes prioritized by TWAS were also concordantly DE in SCZ brain in the same direction as predicted by TWAS (**Table A3.4**).

Overall, these analyses prioritize 125 candidate risk genes whose *cis* expression regulation is associated with disease. Most genes show disease-specific effects, as only three genes showed overlap between SCZ and BD TWAS, including *VPS45*, *SNAP91*, *DCLK3*, while none overlapped with ASD. Considering genes independently identified by each method separately at genome-wide significance, only 2 genes are identified by both PRS association and TWAS, 1 in ASD (*LRRC37A*) and 1 in SCZ (*LRRC37A2*), which may represent structural variation at this locus (*45*). When we restrict our analysis to those genes that are captured by TWAS and replicated by association with the PRS (Likelihood Ratio Test; p <0.01), 11 genes are identified in SCZ (*JKAMP*, *SETD6*, *TMEM214*, *FTSJ2*, *BTN2A2*, *CLEC18B*, *CACNA1D*, *HIST1H4L*, *HLA-DOA*, *TRIM27*, *LRRC37A2*, *RP11-350N15.5*), and one, the *LRRC37A* locus, for ASD.

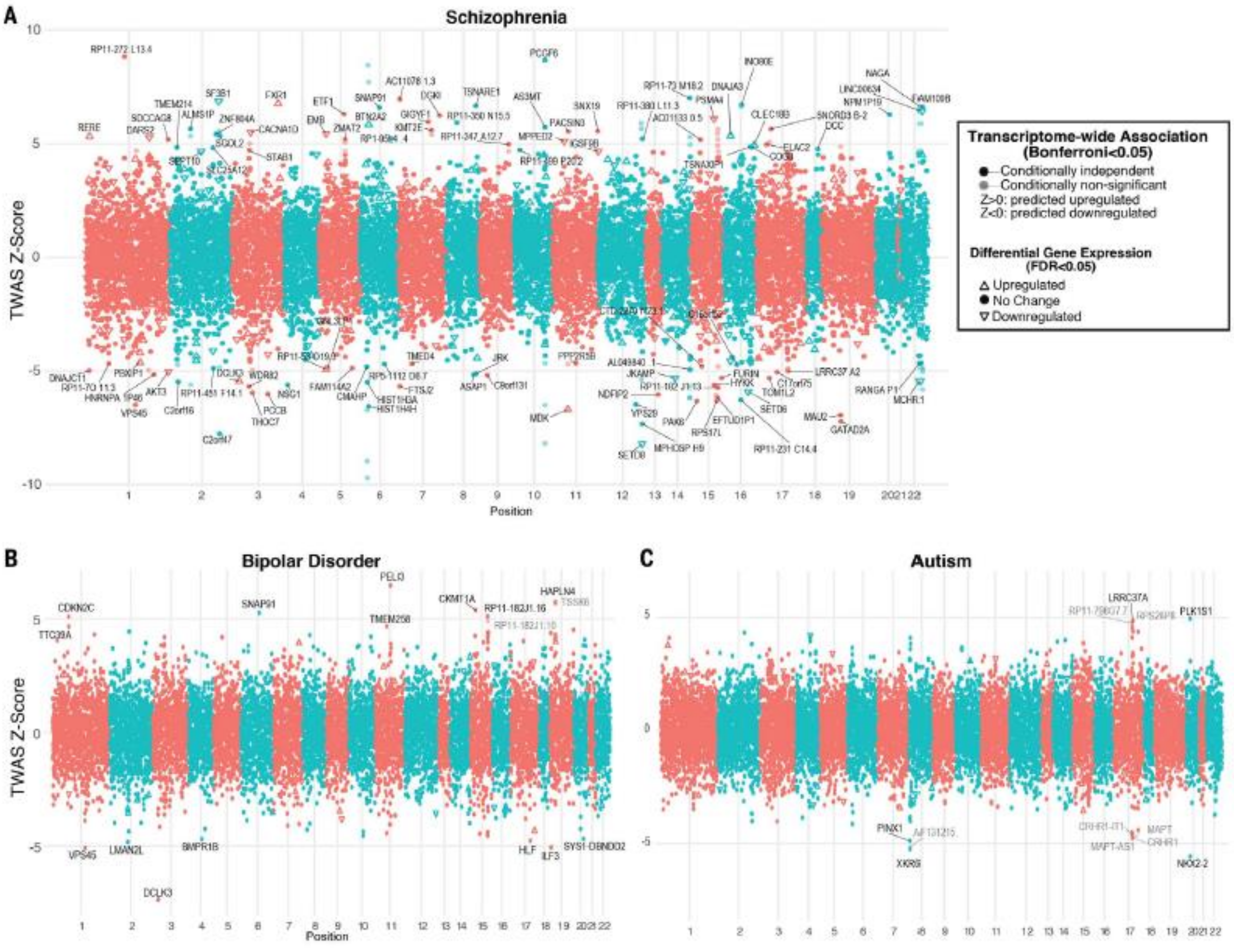


Figure A2.3. Transcriptome-wide association. Results from TWAS prioritize genes whose *cis*-regulated expression in brain is associated with disease. Plots show conditionally-independent TWAS prioritized genes, with lighter shade depicting marginal associations. The sign of TWAS Z-scores indicates predicted direction of effect. Genes significantly up or downregulated in disease brain are shown with arrows, indicating directionality. **A)** In SCZ, 193 genes (164 outside of MHC) are prioritized by TWAS at Bonferroni-corrected P<0.05, including 107 genes with conditionally independent signals. Of these, 23 are also differentially expressed in SCZ brain with 11 in the same direction as predicted. **B)** Seventeen genes are prioritized in BD, of which 15 are conditionally independent. Three TWAS associations overlap between SCZ and BD: *SNAP91*, *DCLK3*, *VPS45*. C) In ASD, TWAS prioritizes 12 genes, of which 5 are conditionally independent.

A2.6: Co-expression networks refine shared cross-disorder signals

To place transcriptomic changes within a systems-level context and more fully interrogate the specific molecular neuropathology of these disorders, we performed weighted gene correlation network analysis (WGCNA) to create independent gene and isoform-level networks

(*15*, *48*, *49*), which we then assessed for disease association and GWAS enrichment using stratified LD score regression ((*21*); see Resource.PsychENCODE.org for interactive visualization). Although calculated separately, gene and isoform-level networks generally reflect equivalent biological processes, as demonstrated by hierarchical clustering (**Fig A2.4A**). However, the isoform-level networks captured greater detail and a larger proportion were associated with disease GWAS than gene-level networks (61% *vs* 41% with nominal GWAS enrichment, P=0.07, χ^2; **Fig A2.4A**). Consistent with expectations, modules showed enrichment for gene ontology pathways and we identified modules strongly and selectively enriched for markers of all major CNS cell types (**Fig A2.4A**-**B**; **Fig A3.12**), facilitating computational deconvolution of cell type specific signatures (*15*, *48*, *50*). For ease of subsequent presentation, we group gene-isoform module pairs that co-cluster, have overlapping parent genes, and represent the same biological processes.

The large sample sizes, coupled with the specificity of isoform-level quantifications, enabled refinement of previously identified gene networks related to ASD, BD and SCZ (*1*, *2*, *15*, *16*, *19*, *51*). Of a combined 90 modules, including 34 gene- (geneM) and 56 isoform-level (isoM) modules, 61 (68%) showed significant association with at least one disorder, demonstrating the pervasive nature of transcriptome dysregulation in psychiatric disease. Five modules are shared across all three disorders, 3 up and two downregulated; 22 modules are shared by 2 of the 3 disorders, and 36 demonstrate more specific patterns of dysregulation in either ASD, SCZ or BD (**Fig A2.4**; **Table A3.5**). It is notable that of these 61 co-expression modules with a disease-association, 41 demonstrate cell type enrichments, consistent with the strong cell type disease-related signal observed using both supervised and unsupervised methods in our companion paper (*18*). This demonstrates the importance of cell type specific changes in the molecular pathology of these major psychiatric disorders; the cell type relationships defined by the disease modules substantially enhance our knowledge of these processes, as we outline below.

The five modules shared between ASD, BD and SCZ can be summarized to represent 3 distinct biological processes. Two of these processes are upregulated, including an inflammatory NFkB signaling module pair (geneM5/isoM5; further discussed in neural-immune section below), and a module (geneM31) enriched primarily for genes with roles in the postsynaptic density, dendritic compartments, and receptor mediated presynaptic signaling that are expressed in excitatory neurons, and to a lesser extent, inhibitory neurons (**Fig A2.4C**). Remarkably, *DCLK3*, one of the hubs of geneM31, is a genome-wide significant TWAS hit in both SCZ and BD. The third biological process, geneM26/isoM22 (**Fig A2.4C**), is downregulated, and enriched for endothelial and pericyte genes, with hubs that represent markers of the blood-brain barrier, including *ITIH5*, *SLC38A5*, *ABCB1,* and *GPR124*, a critical regulator of brain-specific angiogenesis *(52, 53)*. This highlights specific, shared alterations in neuronal-glial-endothelial interactions across these neuropsychiatric disorders.

In contrast to individual genes or isoforms, no modules were significantly associated with PRS scores after multiple-testing correction. However, 19 modules were significantly (FDR < 0.05) enriched for SNP-heritability based on published GWAS ((*21*)**; Fig A2.4A**; **Fig A3.13**). A notable example is geneM2/isoM13, which is enriched for oligodendrocyte markers and neuron projection developmental pathways and is downregulated in ASD and SCZ, with a trend in BD (**Fig A2.4C**). This module pair showed the greatest overall significance of enrichment for SCZ and educational attainment GWAS, and was also enriched in BD GWAS to a lesser degree, suggesting that the processes represented by geneM2/isoM13 genes play a causal role in SCZ and BD disease risk. As additional causal evidence, this module is enriched for genes harboring ultra-rare variants identified in SCZ (*54*) (**Fig A3.13**). Finally, we also observe pervasive and distinct enrichments for syndromic genes and rare variants identified through whole exome sequencing in individuals with neurodevelopmental disorders (**Table A3.5**; **Fig A3.13**).

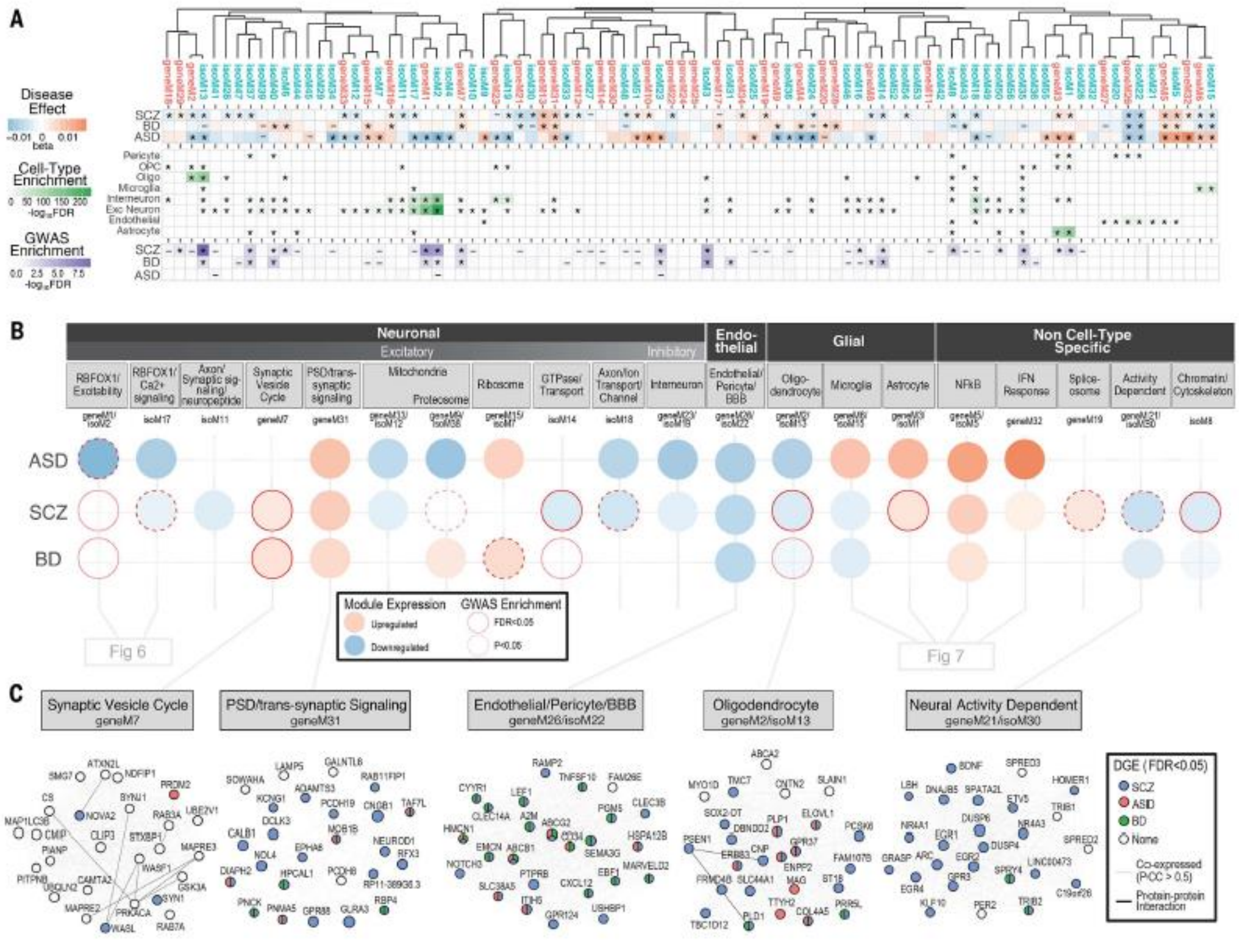


Figure A2.4. Gene and isoform co-expression networks capture shared and disease-specific cellular processes and interactions. A) Gene and isoform co-expression networks demonstrate pervasive dysregulation across psychiatric disorders. Hierarchical clustering shows that separate gene- and isoform-based networks are highly overlapping, with greater specificity conferred at the isoform level. Disease associations are shown for each module (linear regression β value, * FDR<0.05, – P<0.05). Module cell type enrichments (*FDR < 0.05) are shown for major CNS cell types defined from PsychENCODE UMI single cell clusters. Enrichments are shown for GWAS results from SCZ (*58*), BD (*90*), and ASD (*38*), using stratified LD score regression (* FDR<0.05, – P<0.05). **B)** Co-expression modules capture specific cellular identities and biological pathways. Colored circles represent module differential expression effect size in disease, with red outline representing GWAS enrichment in that disorder. Modules are organized and labeled based on CNS cell type and top-gene ontology enrichments. However, we recognize that these annotations are imprecise with respect to complex neurobiological processes, such as those dysregulated in disease. **C)** Examples of specific modules dysregulated across disorders, with top 25 hub genes shown. Edges represent co-expression (Pearson correlation > 0.5) and known protein-protein interactions. Nodes are colored to represent disorders in which that gene is differentially expressed (*FDR<0.05).

A2.7: Neuronal isoform networks capture disease specificity

Multiple neuronal and synaptic signaling pathways have been previously demonstrated to be downregulated in a diminishing gradient across ASD, SCZ, and BD brains without identification of clear disease-specific signals for these neuronal-synaptic gene sets (*1*, *2*, *16*, *19*, *55*, *56*). We do observe neuronal modules broadly dysregulated across multiple disorders, including a neuronal/synaptic module (isoM18) with multiple isoforms of the known ASD risk gene, *ANK2,* as hubs. However, the large sample size, coupled with the specificity of isoform-level qualifications, enabled us to identify synaptic modules containing unique isoforms with distinct disease associations and to separate distinct signals from excitatory and inhibitory neurons (**Fig A2.4B**).

A particularly salient example of differential module membership and disease association of transcript isoforms is *RBFOX1,* a major neuronal splicing regulator implicated across multiple neurodevelopmental and psychiatric disorders (*16*, *32*, *57*, *58*). Previous work has identified downregulated neuronal modules in ASD and SCZ containing *RBFOX1* as a hub (*1*, *16*). Here, we identify two neuronal modules with distinct *RBFOX1* isoforms as hub genes (**Fig A2.5A**). The module pair geneM1/isoM2, downregulated only in ASD (**Fig A2.5B**), contains the predominant brain-expressed *RBFOX1* isoform (*44*) and includes several cation channels (e.g., *HCN1*, *SCN8A*). The second most abundant *RBFOX1* isoform is in another module, isoM17, which is downregulated in both ASD and SCZ (**Fig A2.5B**). Experiments in mouse indicate that *RBFOX1* has distinct nuclear and cytoplasmic isoforms with differing functions, the nuclear isoform primarily regulating pre-mRNA alternative splicing, and the cytoplasmic isoform binding to the 3’ UTR to stabilize target transcripts involved in regulation of neuronal excitability (*28*, *32*, *57*, *59*). Here, we find that isoM17 shows greater enrichment for nuclear *RBFOX1* targets (**Fig A2.5C**), whereas isoM2 shows stronger overlap with cytoplasmic targets (*32*). Consistent with a predicted splicing-regulatory effect, isoM17 shows greater enrichment for genes exhibiting DS in ASD and SCZ (**Fig A2.5D**). In accordance with a predicted role in regulating excitability, isoM2 shows strong and selective enrichment for epilepsy risk genes (**Fig A2.5E**). Moreover, the two modules show differential association with common genetic risk (**Fig A2.5E**), with isoM2 exhibiting GWA signal enrichment across SCZ, BD, and MDD. This widespread enrichment of neurodevelopmental and psychiatric disease risk factors -- from rare variants in epilepsy to common variants in BD, SCZ, and MDD – is consistent with a model where broad neuropsychiatric liability emanates from myriad forms of dysregulation in neuronal excitability, all linked via *RBFOX1*. These results highlight the importance of further studies focused on understanding the relationship between human *RBFOX1* transcript diversity and functional divergence, as most of what is known is based on mouse, and the human shows far greater transcript diversity (*32*, *57*, *60*).

Previous transcriptional networks related to ASD, BD and SCZ did not separate inhibitory and excitatory neuron signals (*1*). The increased resolution here allowed us to identify several modules enriched in inhibitory interneuron markers (**Fig A2.4B**), including geneM23/isoM19, which is downregulated in ASD and SCZ, with a trend toward downregulation observed in BD; downsampling in the SCZ dataset suggests that the lack of significance in BD may be due to smaller sample size (**Fig A3.14**). This module pair contained as hubs the two major GABA synthesizing enzymes (*GAD1*, *GAD2)*, multiple GABA transporters (*SLC6A1*, *SLC24A3*), many other known interneuron markers (*RELN*, *VIP*), as well as *DLX1* and the lncRNA *DLX6-AS1*, both critical known regulators of inhibitory neuron development (*61*). This inhibitory neuron-related module is not enriched for common or rare genetic disease-associated variation, although other studies have found enrichment for SCZ GWAS signal among interneuron markers defined in other ways (*62*).

Several neuronal modules that distinguish between the disorders differentiate BD and SCZ from ASD, including the module pair geneM21/isoM30 (**Fig A2.4C**), which captures known elements of activity-dependent neuronal gene regulation, whose hubs include classic early-response (*ARC*, *EGR1*, *NPAS4*, *NR4A1)* and late-response genes (*BDNF, HOMER1*) (*40*).

Although these modules were not significantly downregulated in ASD, subsampling indicates that the differences between disorders could be driven by sample size (**Fig A3.14**). These genes play critical roles in regulating synaptic plasticity and the balance of excitatory and inhibitory synapses (*40*). Remarkably, a nearly identical module was recently identified as a sex-specific transcriptional signature of major depression and stress susceptibility (*63*). Since psychiatric drug use is more prevalent in SCZ and BD than ASD, and the geneM21/isoM30 module pair are altered more substantially in these disorders, we explored whether these modules may be affected by medication exposure. Indeed, geneM21/isoM30 was associated with genes downregulated by chronic high-doses (but not low-doses) of haloperidol, as well as genes upregulated by the antidepressant fluoxetine (**Fig A3.11A**). Furthermore, geneM21/isoM30 expression was negatively correlated with the degree of lifetime antipsychotic exposure in the subset of patients for whom these data were available (P=0.001, Pearson; **Fig A3.11B**). As such, it will be worthwhile to determine whether this module is a core driver of the therapeutic response, as has been suggested (*64*). Other neuronal modules distinguished SCZ and BD from ASD (**Fig A2.4B**), including geneM7, enriched for synaptic and metabolic processes with the splicing regulator *NOVA2* (**Fig A2.4C**). This neuronal module was significantly enriched for both BD and SCZ GWAS signals, supporting a causal role for this module.

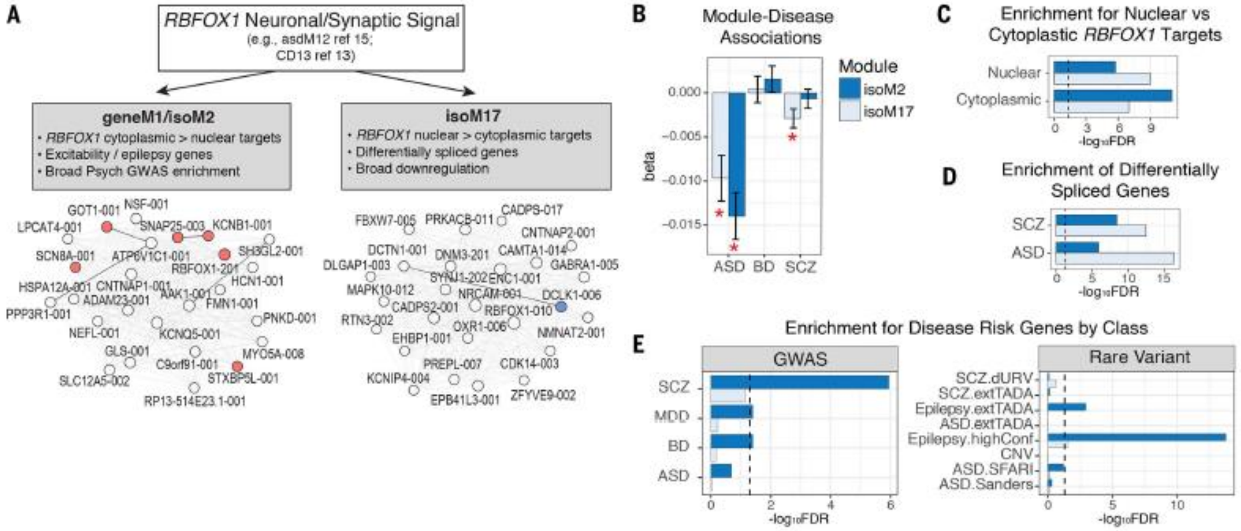


Figure A2.5. Two *RBFOX1* isoform modules capture distinct biological and disease associations. A) Previous studies have identified *RBFOX1* as a critical hub of neuronal and synaptic modules downregulated across multiple psychiatric disorders (*1*, *16*, *19*, *32*). Here, we identify two pairs of modules with distinct *RBFOX1* isoforms as hub genes. Plots show the top 25 hub genes of modules isoM2 and isoM17, following the same coloring scheme as **Fig A2.4C**. **B)** Distinct module-eigengene trait associations are observed for isoM2 (downregulated in ASD only) compared with isoM17, which is downregulated in ASD and SCZ. **C)** Modules show distinct enrichments for nuclear and cytoplasmic *RBFOX1* targets, defined experimentally in mouse (*32*)**. D)** Genes harboring differential splicing events observed in ASD and SCZ show greater overlap with isoM17, consistent with its association with nuclear *RBFOX1* targets. **E)** Modules show distinct patterns of genetic association. isoM2 exhibits broad enrichment for GWAS signal in SCZ, BD, and MDD, as well as for epilepsy risk genes, whereas isoM17 shows no apparent genetic enrichment. GWAS enrichments show FDR-corrected P-values calculated using stratified-LDSC, and rare-variant associations were calculated using logistic regression, controlling for gene length and GC content (*21*).

A2.8: Distinct trajectories of neural-immune dysregulation

Previous work has identified differential activation of glial and neural-immune processes in brain from patients with psychiatric disorders (*16*, *51*, *56*, *65*–*68*), including upregulation of astrocytes in SCZ and BD (*1*, *56*) and both microglia and astrocytes in ASD (*19*, *69*). Evidence supports hyperactive complement-mediated synaptic pruning in SCZ pathophysiology, presumably through microglia (*5*), although post-mortem microglial upregulation was observed only in ASD (*19*, *69*). We examined whether our large cohort including ~1000 control brains, capturing an age range from birth to 90 years, would enable refinement of the nature and timing of this neuroinflammatory signal and potential relationship to disease pathogenesis (**Fig A2.6A**).

Four modules were directly related to neural-immune processes (**Fig A2.6A**-**C**), two of which are gene/isoform module pairs that correspond clearly to cell type specific gene expression; one representing microglia (geneM6/isoM15) and the other astrocytes (geneM3/isoM1), as they are strongly and selectively enriched for canonical cell type specific marker genes (**Fig A2.6C-E**). Two additional immune-related modules appear to represent more broadly expressed signaling pathways: interferon response (geneM32) and NFkB (geneM5/isoM5). The interferon response module (geneM32) contains critical components of the IFN-stimulated gene factor 3 (ISGF3) complex that activates the transcription of downstream interferon-stimulated genes (ISGs), which comprise a striking 59 of the 61 genes in this module (*70*). The NF*k*B module pair (geneM5/isoM5) includes four out of five of the NFkB family members (*NFkB1*, *NFkB2*, *REL*, *RELA*), as well as many downstream transcription factor targets and upstream activators of this pathway.

The dynamic trajectories of these processes in cases with respect to controls reveal distinct patterns across disorders that coincide with disease course (**Fig A2.6F**). The IFN-response and microglial modules are most strongly upregulated in ASD, peaking during early development, coincident with clinical onset. In contrast, in SCZ and BD, the microglial module is actually downregulated and driven by a later dynamic decrease, dropping below controls after age 30. The NF*k*B module, which is upregulated across all three disorders, maximally diverges from controls during early adulthood, coincident with typical disease onset in SCZ and BD (~25). Accordingly, this NF*k*B module contained *C4A* – the top GWAS-supported, and strongly upregulated, risk gene for SCZ (*5*). This pattern is clearly distinct from ASD, which shows a dynamic trajectory, but remains upregulated throughout (**Fig A2.6F**).

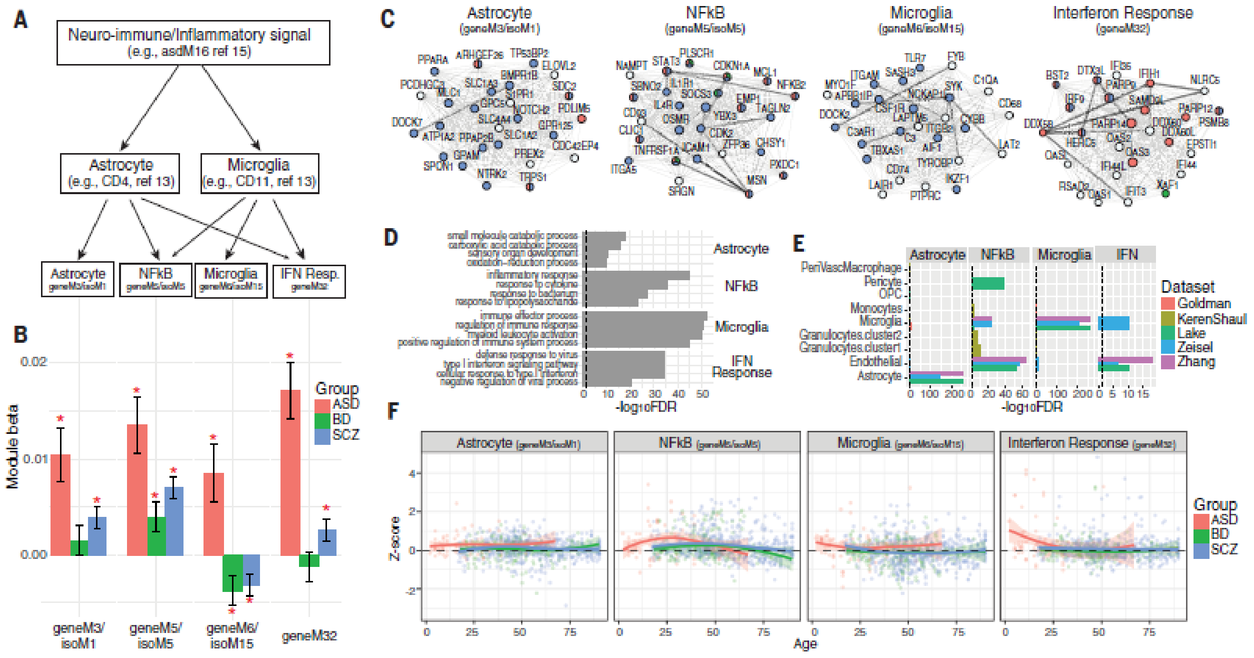


Figure A2.6. Distinct neural-immune trajectories in disease. A) Coexpression networks provide substantial refinement of the neuro-immune/inflammatory processes upregulated in ASD, SCZ, and BD. Previous work has identified specific contributions to this signal from astrocyte and microglial populations (*1*, *19*). Here, we further identify additional, distinct interferon (IFN)-response and NFkB signaling modules. **B)** Eigengene-disease associations are shown for each of 4 identified neural-immune module pairs. The astrocyte and IFN-response modules are upregulated in ASD and SCZ. NFkB signaling is elevated across all three disorders. The microglial module is upregulated in ASD and downregulated in SCZ and BD. **C)** Top hub genes for each module are shown, along with edges supported by co-expression (light grey; Pearson correlation > 0.5) and known protein-protein interactions (dark lines). Nodes follow same coloring scheme as in Fig 5C. Hubs in the astrocyte module (geneM3/isoM1) include several canonical, specific astrocyte markers, including *SOX9*, *GJA1*, *SPON1*, and *NOTCH2*. Microglial module hub genes include canonical, specific microglial markers, including *AIF1*, *CSF1R*, *TYROBP*, *TMEM119*. The NFkB module includes many known downstream transcription factor targets (*JAK3*, *STAT3*, *JUNB*, *FOS*) and upstream activators (*IL1R1*, 9 TNF receptor superfamily members) of this pathway. **D)** The top 4 GO enrichments are shown for each module. E) Module enrichment for known cell type-specific marker genes, collated from sequencing studies of neural-immune cell types (*91–95*). F) Module eigengene expression across age demonstrates distinct and dynamic neural-immune trajectories for each disorder.

A2.9: Non-coding modules and lncRNA regulatory relationships

Given that many lncRNAs are predicted to have transcriptional regulatory roles, we next assessed whether mRNA-based co-expression networks could provide additional functional annotation for ncRNAs. As a subset of lncRNAs are thought to function by repressing mRNA targets (*71*), we applied csuWGCNA (*72*) to identify potential regulatory relationships (*21*). We identified 39 modules (csuM) using csuWGCNA, all preserved in the signed networks with strong cell type and GWAS enrichments, which captured 7186 negatively correlated lncRNA-mRNA pairs within the same module (**Fig A3.15**). We provide a table of putative mRNA targets for thesebrain expressed lncRNAs, including 209 exhibiting DE in ASD, 122 in BD and 241 in SCZ (**Table A3.6**).

A salient example of the power of this approach for functional annotation is *LINC00473,* a hub of the neuronal activity dependent gene regulation module (geneM21/isoM30; **Fig A2.4C**). Expressed in excitatory neurons and downregulated in SCZ ($\log_2$FC -0.16, FDR<0.002), *LINC00473* is regulated by synaptic activity and downregulates immediate early gene expression (*73*), consistent with its hub status in this module. Similarly, we identify the lncRNA *DLX6-AS1,* a known development regulator of interneuron specification (*61*), as the most central hub gene in the interneuron module (geneM23/isoM19), which is downregulated in ASD and SCZ. This interneuron module also contains *LINC00643* and *LINC01166*, two poorly annotated, brain enriched lncRNAs. *LINC00643* is downregulated in SCZ ($\log_2$FC -0.06, FDR=0.04) whereas *LINC01166* is significantly downregulated in BD ($\log_2$FC -0.17, FDR<0.05) with trends in ASD and SCZ (FDR's < 0.1). Our data suggest a role for these lncRNAs in interneuron development, making them intriguing candidates for follow-up studies. Using fluorescence *in situ* hybridization (FISH), we confirmed that both *LINC00643* and *LINC1166* are expressed in GAD1+ GABAergic neurons in area 9 of adult brain, present both in the cell nucleus and cytoplasm (**Fig A2.7A**; **Fig A3.16**), although expression was also detected in other non GAD1+ neurons as well.

Multiple ncRNAs including *SOX2-OT, MIAT*, and *MEG3* are enriched in oligodendrocyte modules (geneM2/isoM13/csuM1; **Fig A2.4C**) that are downregulated in both SCZ and ASD. *SOX2-OT* is a heavily spliced, evolutionarily-conserved lncRNA exhibiting predominant brain expression and a hub of these oligodendrocyte modules, without previous mechanistic links to myelination (*74*, *75*). The lncRNAs *MIAT* and *MEG3* are negatively correlated with most of the hubs in this module, including *SOX2-OT* (**Fig A3.15**). *MIAT* is also known to interact with *QKI,* an established regulator of oligodendrocyte-gene splicing also located in this module (*76*, *77*). These analyses predict critical roles for these or these often overlooked non-coding genes in oligodendrocyte function (*76*, *77*) and potentially in psychiatric conditions.

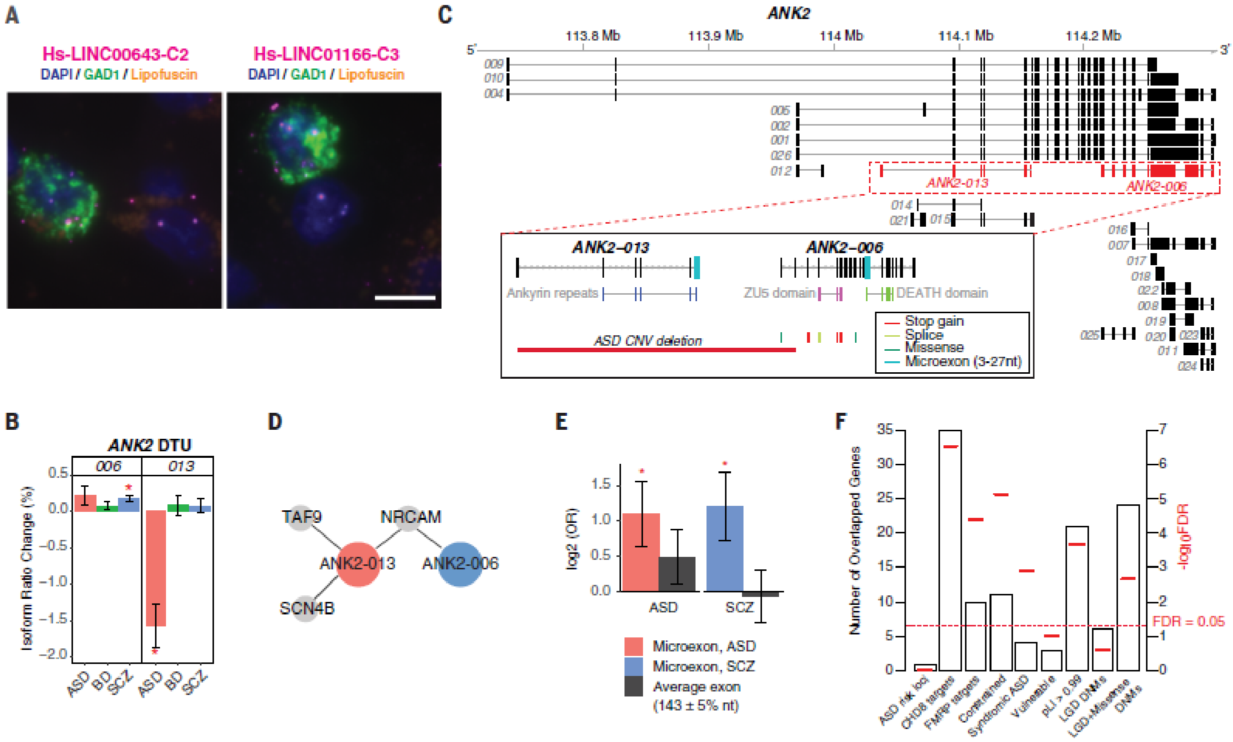


Figure 3.7. LncRNA annotation, *ANK2* isoform switching & micro-exon enrichment. A) FISH images demonstrate interneuron expression for two poorly annotated lincRNAs – *LINC00643* and *LINC01166* – in area 9 of adult human prefrontal cortex. Sections were labeled with *GAD1* probe (green) to indicate GABAergic neurons and lncRNA (magenta) probes for *LINC00643* (left) or for *LINC01166* (right). All sections were counterstained with DAPI (blue) to reveal cell nuclei. Lipofuscin autofluorescence is visible in both the green and red channels and appears yellow/orange. Scale bar, 10 µm. FISH was repeated at least twice on independent samples (**Table A3.9** (*21*)) with similar results (see also **Fig A3.16**). **B**) *ANK2* isoforms *ANK2-006* and *ANK2-013* show significant DTU in SCZ and ASD, respectively (*FDR<0.05). **C**) Exon structure of *ANK2* highlighting (dashed lines) the *ANK2-006* and *ANK2-013* isoforms. Inset, these isoforms have different protein domains and carry different microexons. *ANK2-006* is hit by multiple ASD DNMs while *ANK2-013* could be entirely eliminated by a *de novo* CNV deletion in ASD. **D**) Disease-specific co-expressed PPI network. Both ANK2-006 and ANK2-013 interact with NRCAM. The ASD-associated isoform ANK2-013 has two additional interacting partners, SCN4B and TAF9. **E**) As a class, switch isoforms are significantly enriched in microexon(s). In contrast, exons of average length are not enriched among switch isoforms. Y-axis displays odds ratio on $\log_2$ scale. P-values are calculated using logistic regression and corrected for multiple comparisons. **F**) Enrichment of 64 genes with switch isoforms in: ASD risk loci (*80*); CHD8 targets (*96*); FMRP targets (*33*); Mutationally constraint genes (*97*); Syndromic and highly ranked (1 and 2) genes from SFARI Gene database; Vulnerable ASD genes (*98*); Genes with probability of loss-of-function intolerance (pLI) > 0.99 as reported by the Exome Aggregation Consortium

(*99*); Genes with likely-gene-disruption (LGD) or LGD plus missense *de novo* mutations (DNMs) found in patients with neurodevelopmental disorders (*21*).

A2.10: Isoform network specificity and switching

To more comprehensively assess whether aspects of disease specificity are conferred by alternative transcript usage or splicing, versus DE, we surveyed genes exhibiting DTU across disorders (*21*). We identified 134 such ‘switch isoforms’, corresponding to 64 genes displaying different DTU between ASD and SCZ (**Table A3.7**). As an example, isoforms of *SMARCA2, a* member of the BAF-complex strongly implicated in several neurodevelopmental disorders including ASD (*78*), are up and downregulated in ASD and SCZ, respectively (**Fig A3.17**). Conversely, the isoforms of *NIPBL,* a gene associated with Cornelia de Lange Syndrome (*79*) are down and upregulated in ASD and SCZ, respectively (**Fig A3.17**). Such opposing changes in isoform expression of various genes may represent differences in disease progression or symptom manifestation in diseases as ASD and SCZ, mediated by genetic risk variants that create subtle differences in isoforms within the same gene that exhibit distinct biological effects in each disorder. A remarkable example is the ASD risk gene *ANK2* (*80*), whose two alternatively spliced isoforms, *ANK2-006* and *ANK2-013*, are differentially regulated in SCZ and ASD (**Fig A2.7B**). These switch isoforms show markedly different expression patterns, belonging to different co-expression modules, geneM3/isoM1 (**Fig A2.6C**) and isoM18, which are enriched in astrocyte and neuronal cell types, respectively (**Fig A2.4A**; **Fig A3.12**). The protein domain structure of these transcripts is also non-overlapping, with *ANK2-006* carrying exclusively ZU5 and DEATH domains, and *ANK2-013* carrying exclusively ankyrin repeat domains (**Fig A2.7C**). Both isoforms are impacted by a *de novo* ASD CNV, and *ANK-006* also carries *de novo* mutations from neurodevelopmental disorders. Both isoforms bind to the neuronal cell adhesion molecular *NRCAM*, but *ANK2-013* has two additional, unique partners – *TAF9* and *SCN4B* (**Fig A2.7D**), likely cell type specific interactions that suggest distinct functions of the isoforms of this genes in different neural cell types and diseases.

Several studies have demonstrated that genes carrying microexons are preferentially expressed in brain and their splicing is dysregulated in ASD (*30*, *81*, *82*). This PsychENCODE sample provided the opportunity to assess the role of microexons in a far larger cohort and across several disorders. Indeed, we find that switch isoforms with microexons (3-27 bp) are significantly enriched in both ASD (FDR=0.03) and SCZ (FDR=0.03, logistic regression) (**Fig A2.7E**; (*21*)). Genes with switch isoforms are also enriched for the regulatory targets of two ASD risk genes, *CHD8* and *FMRP*, as well as highly mutationally constrained genes (pLI>0.99), syndromic ASD genes, and in genes with *de novo* exonic mutations in ASD, SCZ and BD (**Fig A2.7F**; **Table A3.7**; (*21*)). These data confirm the importance of microexon regulation in neuropsychiatric disorders beyond ASD, and their potential role in distinguishing among biological pathways differentially affected across conditions. This role for microexons further highlights local splicing regulation as a potential mechanism conferring key aspects of disease specificity, extending the larger disease signal observed at the isoform-level in co-expression and differential expression analyses.

A2.11: Discussion

We present a large-scale RNA-Seq analysis of the cerebral cortex across three major psychiatric disorders, including extensive analyses of the non-coding and alternatively spliced transcriptome, as well as gene- and isoform-level co-expression networks. The scope and complexity of these data do not immediately lend themselves to simple mechanistic reduction. Nevertheless, at each level of analysis, we present concrete examples that provide proofs-of-principle and starting points for investigations targeting shared and distinct disease mechanisms to connect causal disease drivers with brain-level perturbations.

Broadly, we find that isoform-level changes exhibit the largest effect sizes in disease brain, are most enriched for genetic risk, and provide the greatest disease specificity when assembled into co-expression networks. Remarkably, disturbances in the expression of distinct isoforms of more than 50 genes are differentially observed in SCZ and ASD, which in the case of the ASD risk gene *ANK*2, is predicted to affect different cell types in each disorder. Moreover, we observe disease-associated changes in the splicing of dozens of RNA-binding proteins and splicing factors, most of whose targets and functions are unknown. Similarly, nearly 1000 ncRNAs are dysregulated in at least one disorder and most of these ncRNAs show significant CNS enrichment, but until now, have limited functional annotation.

As with any case/control association study, multiple potential factors contribute to gene expression changes in post-mortem human brain, many of which may represent reactive processes. At each step of analysis, we have attempted to mitigate the contribution of these factors through known and hidden covariate correction, assessment of age trajectories, and via enrichment for causal genetic variation. Supporting the generalizability of our findings, we find highly significant correlations of the $\log_2$FC between randomly split halves of the data (**Fig A3.3**). This likely varies by transcript class, and some of the modest correlations are likely due to low abundance genes, such as ncRNAs, which we prefer to include, while recognizing the inherent tension between expression level and measurement accuracy. We provide access to this extensive resource, both in terms of raw and processed data and as browsable network modules (Resource.PsychENCODE.org).

Several broad shared patterns of gene expression dysregulation have been observed in post mortem brain in previous studies, most prominently, a gradient of downregulation of neuronal and synaptic signaling genes, and upregulation of glial-immune or neuroinflammatory signals. Here, we are able to substantially refine these signals, by distinguishing both up and downregulated neuron-related processes that are differentially altered across these three disorders. Furthermore, we extend previous work that identified broad neuroinflammatory dysregulation in SCZ, ASD, and BD, by identifying specific pathways involving IFN-response, NFkB, astrocytes and microglia that manifest distinct temporal patterns across conditions. A module enriched for microglial-associated genes, for example, shows a clear distinction between disorders, with strong upregulation observed on ASD and significant downregulation in SCZ and BD. Overall, these results provide substantially increased specificity to the observations that ASD, BD, and SCZ are associated with elevated neuroinflammatory processes (*68*, *83*–*85*).

This work highlights isoform-level dysregulation as a critical, and relatively underexplored, proximal mechanism linking genetic risk factors with psychiatric disease pathophysiology. In contrast to local splicing changes, isoform-level quantifications require imputation from short-read RNA-Seq data guided by existing genomic annotations. Consequently, the accuracy of these estimates is hindered by incomplete annotations, as well as by limitations of short-read sequencing, coverage, and genomic biases like GC content (*86*, *87*). This may be particularly problematic in brain where alternative splicing patterns are more distinct than in other organ systems (*81*). We present experimental validations for several specific isoforms, but try to focus on the class of dysregulated isoforms, and the modules and biologically processes they represent, rather than individual cases which may be more susceptible to bias. Longer read sequencing, which provides a more precise means for isoform quantification, will be of great utility as it becomes more feasible at scale.

By integrating transcriptomic data with genetic variation, we identify multiple disease-associated co-expression modules enriched for causal variation, as well as new mechanisms potentially underlying specific disease loci in each of the diseases. In parallel, by performing a well-powered brain-relevant TWAS in SCZ, and to a lesser extent in BD and ASD, we are further able to elucidate candidate molecular mechanisms through which disease-associated variants may act. TWAS prioritizes dozens of new candidate disease genes, including many dysregulated in disease brain. Similar to the eQTLs identified in the companion paper (*18*), the majority of these new loci do not overlap with disease GWAS association signals. Rather, most are outside of the LD block and quite distal to the original association signal, highlighting the importance of orthogonal functional data types, such as transcriptome or epigenetic data (*17*, *47*, *81*, *88*, *89*), in deciphering the underlying mechanisms of disease-associated genetic effects.

A large proportion of disease-associated co-expression modules are enriched for cell type specific markers, as is overall disease DE signal, indicating that transcriptomic alterations in disease are likely driven substantially by (even subtle) shifts in cell type proportions, or cell type specific pathways, consistent with our previous observations (*1*) and those in the companion PsychENCODE manuscript (*18*). Functional genomic studies often remove such cell type-specific signals, through use of large numbers of expression-derived principle components or surrogate variables as covariates, to remove unwanted sources of variation and maximize detection of *cis* eQTLs (*44*). We retain the cell type-specific signals as much as possible, reasoning that cell type-related alterations may directly inform the molecular pathology of disease in psychiatric disorders, in which there is no known microscopic or macroscopic pathology. This rationale is supported by the consistent observation of the dynamic and disease-specific microglial upregulation observed in ASD, and the shared astrocyte upregulation in SCZ and ASD. This approach, however, reduces the ability to detect genetic enrichment from GWAS, as current methods predominately capture *cis*-acting regulatory effects. The modesty of genetic enrichments among disease-associated transcriptomic alterations may also indicate that gene expression changes reflect an indirect cascade of molecular events triggered by environmental as well as genetic factors, or that genetic factors may act earlier such as during development.

Finally, these data, while providing a unique, large-scale resource for the field, also suggest that profiling additional brains, especially from other implicated brain regions from patients will continue to be informative. Similarly, these data suggest that isoform level analyses including the identification of isoform-specific PPI and cell type specificity, while posing major challenges for high-throughput studies, are likely to add substantial value to understanding brain function and neuropsychiatric disorders. Finally, as GWAS studies in ASD and BD increase in size and subsequently in power, their continued integration with these transcriptome data will likely prove critical in identifying the functional impact of disease-associated genetic variation.

A2.12: Materials and Methods

Please see the Appendix (section A3.1) for all materials and methods.

A2.14: Bibliography

The bibliography/references for the appendix sections A2 and A3 correspond with the bibliography for Chapter 3 (Section 3.4).

A3: Supplementary Materials for Chapter 3 and Appendix Section A2

A3.1: Extended Materials and Methods

Data Generation

The data generated for this manuscript represent Freeze 1 and 2 of the PsychENCODE consortium dataset. Post-mortem human brain samples were collected as part of eight studies, detailed below and in **Fig A3.1**. RNA-Seq and genotype array data was generated by each site and then processed together through a unified pipeline by a central data analysis core. For this capstone analysis, we restricted analysis to frontal and temporal cortex brain samples from postnatal timepoints. We provide a description of each individual study below, derived from the PsychENCODE website. All data are available at doi.org/10.7303/syn12080241.

Study 1 - BrainGVEX

For the BrainGVEX study, RNA-Seq data was generated from 427 post-mortem prefrontal cortex samples from subjects with schizophrenia (n=95), bipolar disorder (n=73), and non-psychiatric controls (n=259). RNA samples were collected from the Stanley Medical Research Institute (SMRI) as part of the “Array Collection”, “Consortium Collection”, “New Collection” and ”Extra Collection”. Array collection and consortium collection samples were from the superior frontal gyrus (Brodmann’s area (BA) 9) whereas those from extra and new collections were from the mid frontal gyrus (BA46). Another 184 controls were obtained as fresh-frozen brain tissue from the Banner Sun Health Research Institute (BSHRI). All BSHRI samples were from the frontal cortex. RNA was extracted from BSHRI samples by first homogenizing 20-50 mg of tissue in QIAzol (Qiagen) using the Lysin Matrix D and FastPrep-24 system (MPBiomedicals). Total RNA were then isolated using the miRNeasy Kit (Qiagen) according to manufacturer’s instructions. RNA integrity was assessed with Agilent Technologies RNA 600 nano kit. Samples with RNA

Integrity Number (RIN) lower than 5.5 were excluded from the study. RNA sequencing libraries were prepared using TruSeq Stranded Total RNA sample prep kit with RiboZero Gold HMR (Illumina). Libraries were multiplexed (3 per lane) for paired-end 100 bp sequencing on Illumina HiSeq2000 with read depth >70 million reads on average. Genotyping was performed using two different platforms. 144 samples (SMRI Consortium and Array Collections) were genotyped using the Affymetrix GeneChip Mapping 5.0K Array. Genotypes were called with the BRLMM-p algorithm (Affymetrix) on all arrays simultaneously (*100*). The remaining samples (SMRI New and Extra Collection, and BSHRI samples) were genotyped on the Human PsychChip platform, which is a custom version of the Illumina Infinium CoreExome-24 v1.1 BeadChip (#WG-331-1111). However, PsychChip data were not yet available for this study.

Study 2 - BrainSpan

For the BrainSpan study, RNA-Seq data was generated from 606 brain samples from 41 unique individuals. RNA was extracted using RNeasy Plus Mini Kit (Qiagen) for mRNA. Either approximately 30 mg of pulverized tissue (12 PCW – 40 Y specimens) or entire amount of dissected brain piece (8 – 9 PCW, smaller than 30 mg) was processed. Tissue was pulverized with liquid nitrogen in a chilled mortar and pestle and transferred to a chilled safe-lock microcentrifuge tube (Eppendorf). Per tissue mass, equal mass of chilled stainless steel beads (Next Advance, cat# SSB14B) along with two volumes of lysis buffer were added. Tissue was homogenized for 1 min in Bullet Blender (Next Advance # SSB14B) at speed 6 and incubated at 37°C for 5 min. Lysis buffer up to 0.6 ml was again added, tissue homogenized for 1 min and incubated at 37°C for 1 min. Extraction was further carried out according to manufacturer's protocol. Genomic DNA was removed by a proprietary column provided in RNeasy Plus Mini Kit (Qiagen) or by DNase treatment using TURBO DNA-free Kit (Ambion/ Life technologies). 260:A280 ratio and RNA Integrity Number (RIN) were determined for each sample with NanoDrop

(Thermo Fisher Scientific) and Agilent 2100 Bioanalyzer system, respectively. The mRNA-sequencing (mRNA-Seq) sample preparation Kit (Illumina) was used to prepare cDNA libraries per manufacturer instructions with some modifications. Briefly, polyA RNA was purified from 1 to 5 µg of total RNA using Oligo (dT) beads. Quaint-IT RiboGreen RNA Assay Kit (Invitrogen) was used to quantitate purified mRNA with the NanoDrop 3300. Following mRNA quantitation, 2.5 µl spike-in master mixes, containing five different types of RNA molecules at varying amounts (2.5 × 10-7 to 2.5 × 10-14 mol), were added per 100 ng of mRNA. Spike-in RNAs were synthesized by the External RNA Control Consortium (ERCC) by *in vitro* transcription of *de novo* DNA sequences or DNA derived from B. subtilis or the deep-sea vent microbe M. jannaschii and were a generous gift of Dr. Mark Salit at The National Institute of Standards and Technology (NIST). Each sample was tagged by adding two spike-in RNAs unique to the region from which the sample was taken. Further, three common spike-in RNAs with gradient concentrations were added to each sample, to enable the assessment of sequencing quality. Spike-in sequences are available at http://archive.gersteinlab.org/proj/brainseq/spike_in/spike_in.fa. The mixture of mRNA and spike-in RNAs was subjected to fragmentation, reverse transcription, end repair, 3' end adenylation, and adapter ligation to generate libraries of short cDNA molecules, followed by PCR amplification. The PCR enriched product was assessed for its size distribution and concentration using Bioanalyzer DNA 1000 Kit. Genotype data were not used in this study.

Study 3 - CommonMind

Full details of the CommonMind study have been published (*2*), although the data here were processed separately according to the uniform RNA-Seq pipeline described below. Samples were acquired through brain banks at three institutions:The Mount Sinai NIH Brain Bank and Tissue Repository, University of Pennsylvania Brain Bank of Psychiatric illnesses and Alzheimer's Disease Core Center, and the University of Pittsburgh NIH NeuroBioBank Brain and Tissue

Repository. Details about brain banks, inclusion/exclusion criteria, and sample collection and processing have been previously described (https://www.synapse.org/#!Synapse:syn2759792/wiki/71104). RNA-Seq data from 613 total human post-mortem dorsolateral prefrontal cortex (DLPFC) brain samples were obtained from 603 subjects with schizophrenia (n=263), bipolar disorder (n=47), affective disorder (8), and neurotypical controls (n=285), where 10 neurotypical controls were sequenced as technical replicates. Subjects with affective disorder were not used in this study. Total RNA was extracted from 50 mg of homogenized dorsolateral prefrontal cortex tissue using RNeasy kit. Samples with RIN < 5.5 (n=51) were excluded. The remaining samples had a mean RIN of 7.7. RNA-Seq library preparation was performed using ribosomal RNA depletion, with the RiboZero Magnetic Gold Kit. Samples were barcoded, multiplexed (n=10/lane), and sequenced across two lanes as 100 bp paired end sequencing on the Illumina HiSeq 2500 with an average of 85 million reads. Data are provided for those samples that passed all of the following QC filters: samples were required to have had a minimum of 50 million total reads and less than 5% rRNA alignment. For genotyping, DNA was isolated from approximately 10 mg dry homogenized tissue coming from the same dissected samples as the RNA isolation using the Qiagen DNeasy Blood and Tissue Kit according to manufacturer’s protocol. Genotyping was performed using the Illumina Infinium HumanOmniExpressExome platform (Catalog #: WG-351-2301). All data were checked for discordance between nominal and genetically-inferred sex using Plink software to calculate the mean homozygosity rate across X-chromosome markers and to evaluate the presence or absence of Y-chromosome markers. In addition, pairwise comparison of samples across all genotypes was done to identify potentially duplicate samples (genotypes > 99% concordant) or related individuals using Plink.

Study 4 - Yale-ASD

For the Yale-ASD study, RNA-Seq data was generated from 45 brain samples from 37 unique individuals, including 9 with ASD and 28 controls. Total RNA was extracted using mirVana kit (Ambion) with some modifications to the manufacturer's protocol. Approximately 60 mg of tissue was pulverized with liquid nitrogen in a pre-chilled mortar and pestle and transferred to a chilled safe-lock microcentrifuge tube (Eppendorf). Per tissue mass, equal mass of chilled stainless steel beads (Next Advance, catalog # SSB14B) along with one volume of lysis/binding buffer were added. Tissue was homogenized for 1 min in Bullet Blender (Next Advance) and incubated at 37°C for 1 min. Another nine volumes of the lysis/binding buffer were added, homogenized for 1 min, and incubated at 37°C for 2 min. One-tenth volume of miRNA Homogenate Additive was added and extraction was carried out according to the manufacturer's protocol. RNA was treated with DNase using TURBO DNA-free Kit (Ambion/ Life Technologies) and RNA integrity was measured using Agilent 2200 TapeStation System. Barcoded libraries for RNA-Seq were prepared with 5 ng of RNA using TruSeq Stranded Total RNA with Ribo-Zero Gold kit (Illumina) per manufacturer's protocol. Paired-end sequencing (100bp x 2) was performed on HiSeq 2000 sequencers (Illumina) at Yale Center for Genome Analysis. Genotype data was not yet available as part of Freeze 1 or 2 of the PsychENCODE dataset.

Study 5 - UCLA-ASD

For the UCLA-ASD study, RNA-Seq data was generated from 253 brain samples from 97 unique individuals, across prefrontal cortex (BA9/46), temporal cortex (BA41/42/22), and cerebellum. Full details of the UCLA-ASD study have been published (*19*). Brain samples were obtained from the Harvard Brain Bank as part of the Autism Tissue Project (ATP). Frozen brain regions were dissected on dry ice in a dehydrated dissection chamber to reduce degradation effects from sample thawing or humidity. Approximately 50-100 mg of tissue across the cortical region of interest was isolated from each sample using the miRNeasy kit with no modifications (Qiagen). For each RNA sample, RNA quality was quantified using the RNA Integrity Number (RIN) on an Agilent Bioanalyzer. Strand-specific, rRNA-depleted RNA-Seq libraries were prepared using TruSeq Stranded Total RNA sample prep kit with RiboZero Gold (Illumina) kits. Libraries were randomly pooled to multiplex 24 samples per lane using Illumina TruSeq barcodes. Each lane was sequenced five times on an Illumina HiSeq 2500 instrument using high output mode with standard chemistry and protocols for 50 bp paired-end reads to achieve a target depth of 70 million reads. Genotyping data was generated at the UCLA Neurogenomics Core (UNGC) on the Illumina Omni 2.5 8v1 platform (Human Exome). Illumina Genome Studio files were clustered using Illumina's standard HapMap cluster file. SNP genotypes were exported from the Illumina GenomeStudio Software as forward strand in PLINK format. SNP marker names were updated with a conversion file from Illumina which converts local marker name to rsID (plink --update-map --update-name). All quality filtering was performed using PLINK v1.07. SNPs missing more than 99.99% data were excluded (--geno 0.9999). Individuals missing > 5% data, SNPs missing > 5% data, and SNPs with HW p<0.0000001 were also excluded. The order of filtering was performed according to PLINK default procedures (plink --mind 0.05 --geno 0.05 --hwe 0.0000001).

Study 6 - CMC_HBCC

Brain specimens for the CMC_HBCC study were obtained from the the NIMH Human Brain Collection Core (HBCC; https://www.nimh.nih.gov/labs-at-nimh/research-areas/research-support-services/hbcc/human-brain-collection-core-hbcc.shtml) under protocols approved by the CNS IRB (NCT00001260), with the permission of the next-of-kin through the Offices of the Chief Medical Examiners in the District of Columbia, Northern Virginia, and Central Virginia. All specimens were characterized neuropathologically, clinically and toxicologically. A clinical diagnosis was obtained through family interviews and review of medical records by two psychiatrists based on DSM-IV criteria. Non-psychiatric controls were defined as having no history of a psychiatric condition or substance use disorder. Brain samples were dissected at the NIMH Human Brain Collection Core and shipped to Icahn School of Medicine at Mount Sinai (ISMMS) for sample preparation and RNA-sequencing. Samples for the study were dissected from either the left or right hemisphere of fresh frozen coronal slabs cut at autopsy from the dorsolateral prefrontal cortex. Total RNA from 468 HBCC samples was isolated from approximately 100 mg homogenized tissue from each sample by TRIzol/chloroform extraction and purification with the Qiagen RNeasy kit (Cat#74106) according to manufacturer's protocol. Samples were processed in randomized batches of 12. The order of extraction was assigned randomly with respect to diagnosis and all other sample characteristics. The mean total RNA yield was 24.2 ug. The RNA Integrity Number (RIN) was determined by fractionating RNA samples on the 4200 Agilent TapeStation System. Sixty nine samples with RIN <5.5 were excluded from the study. An additional 12 samples were removed post sequencing due to evidence of sample swap or contamination, resulting in a final dataset of 387 samples (70 BD, 97 SCZ, 220 neurotypical controls) with a mean RIN of 7.5 and a mean ratio of 260/280 of 2.0. RNA sequencing raw and quantified expression data is provided for these 387 samples from 387 unique individuals. Data was generated, QCed, processed and quantified as follows: All samples submitted to the New York Genome Center for RNA-Seq were prepared for sequencing in randomized batches of 94. The sequencing libraries were prepared using the KAPA Stranded RNA-Seq Kit with RiboErase (KAPA Biosystems). rRNA was depleted from 1ug of RNA using the KAPA RiboErase protocol that is integrated into the KAPA Stranded RNA-Seq Kit. The insert size and DNA concentration of the sequencing library was determined on Fragment Analyzer Automated CE System (Advanced Analytical) and Quant-iT PicoGreen (Thermo Fisher Scientific) respectively. A pool of 10 barcoded libraries were layered on a random selection of two of the eight lanes of the Illumina flow cell at appropriate concentration and bridge amplified to ~ 250 million raw clusters. One-hundred base pair paired end reads were obtained on a HiSeq 2500. Genotyping was performed using Illumina_1M, Illumina_h650, and Illumina_Omni5 platforms.

Studies 7+8 - BipSeq & LIBD_szControl

Post-mortem tissue homogenates of dorsolateral prefrontal cortex (DLPFC) approximating BA46/9 in postnatal samples and the corresponding region of PFC in fetal samples were obtained from all subjects. Total RNA was extracted from ~100 mg of tissue using the RNeasy kit (Qiagen) according to the manufacturer’s protocol. The poly-A containing RNA molecules were purified from 1 µg DNAse treated total RNA and sequencing libraries were constructed using the Illumina TruSeq© RNA Sample Preparation v2 kit. Sequencing indices/barcodes were inserted into Illumina adapters allowing samples to be multiplexed across lanes in each flow cell. These products were then purified and enriched with PCR to create the final cDNA library for high throughput sequencing using an Illumina HiSeq 2000 with paired end 2x100bp reads. Further details are available in (*101*). SNP genotyping with HumanHap650Y_V3, Human 1M-Duo_V3, and Omni5 BeadChips (Illumina, San Diego, CA) was carried out according to the manufacturer’s instructions with DNA extracted from cerebellar tissue. Genotype data were processed and normalized with the crlmm R/Bioconductor package separately by platform.

RNA-sequencing Data Processing Pipeline

All sample FASTQ files were run through a unified RNA-Seq processing pipeline (**Fig A2.1**) run at the University of Chicago on an OpenStack cloud system and modeled after the long-rna-seq-pipeline used by the ENCODE Consortium. Fastqs were trimmed for adapter sequence and low base call quality (Phred score < 30 at ends) using cutadapt (v1.12). Trimmed reads were then aligned to the GRCH37.p13 (hg19) reference genome via STAR (2.4.2a) using comprehensive gene annotations from Gencode (v19). BAM files were produced in both genomic and transcriptome coordinates and sorted using samtools (v1.3). Gene and isoform-level quantifications were calculated using RSEM (v1.2.29). Quality control metrics were calculated using RNA-SeQC (v1.1.8), featureCounts (v1.5.1), PicardTools (v1.128), and Samtools (v1.3.1). Pipeline source code can be found on Synapse at doi:10.7303/syn12026837. RNA-Seq data was processed in two batches: Freeze 1 consisted of re-processed RNA-Seq data from the following studies: BrainGVEX, BrainSpan, CMC, UCLA-ASD, Yale-ASD, iPSC; Freeze 2 consisted of BipSeq, LIBD_szControl, CMC_HBCC, EpiGABA. Data from EpiGABA and iPSC studies was not used in this Capstone project.

Genotyping and QTL Pipeline

Genotype calls were generated at each data production site separately, as described above, and centralized for imputation. Genotype imputation and QTL analyses were performed as described in our companion manuscript (*18*) and on the PsychENCODE website using a uniform genotype QC and imputation pipeline for all studies. To generate high-quality observed genotypes (removing low quality and rare variants), initial QC was performed using Plink to remove SNPs with zero alternate alleles, MAF <1%, genotyping call rate < 0.95, Hardy-Weinberg p-value < $1x10^{-6}$, individuals with genotyping call rate < 0.95, and to correct strand flips. Parallel haplotype pre-phasing and imputation were done using Beagle2, Minimac3 with the HRC reference panel for imputation. Calculation of gene-level expression QTLs (eQTL) and isoform-level expression QTLs (isoQTL) was done using QTLtools, as described in our companion manuscript (*18*). Imputation of *C4A* structural variation for each genotyped sample of European ancestry was performed using Beagle5 with a custom HapMap3 CEU reference panel as described (*5*). Inferred copy number of C4 structural elements (C4A, C4B, C4L, and C4S) based on the imputed C4 alleles was then associated with normalized C4A expression using a linear model.

RNA-Seq Quality Control and Normalization

Expected counts were compiled from gene and isoform-level RSEM quantifications and imported into R for downstream analyses. Genes were filtered to include those with TPM > 0.1 in at least 25% of samples. We removed all transcripts derived from mitochondrial DNA and Y-chromosome pseudoautosomal regions (“ENSR”) as well as transcripts with immunoglobulin (IG or TR) biotypes or those shorter than 250 bp. Downstream analyses were performed on the resulting 25,774 transcribed genes based on GENCODE V19 annotations. We restricted our analysis to frontal and temporal cortex brain samples obtained from subjects at postnatal time points (**Fig A3.2**). We removed samples with an ambiguous diagnosis or a diagnostic label other than ASD, SCZ, BD, or CTL (n=11). We removed samples with unspecified or ambiguous age (n=2), or sex (n=2) as well as samples with less than 10 million total reads. Each individual study was then assessed for outlier samples (**Fig A3.2C**), defined as those with standardized sample network connectivity Z scores < -2, as published, which were removed (*102*). We further removed 8 samples whose documented sex was discordant from that predicted by gene expression, based on hierarchical clustering of samples using expression of *XIST* and the first principal component of genes on the Y chromosome.

Covariate Selection

We compiled a set of 187 RNA-Seq quality control metrics as the aggregate sample-level outputs from RNA-SeQC, cutadapt, featureCounts, PicardTools (CollectAlignmentSummaryMetrics, CollectInsertSizeMetrics, CollectRnaSeqMetrics, MarkDuplicates), and STAR (**Fig A3.2**). As many of these metrics were highly overlapping, we summarized these measures by the top 29 principal components which collectively explained 99% of the total variance. To determine which covariates to include in the final differential expression model, we performed multivariate adaptive regression as implemented in the *earth* package in R. This builds a model in two phases using a forward pass to capture maximal amount of variance explained by an underlying set of covariates, followed by a backward (pruning) pass to remove potential redundant terms. The superset of potential covariates available for all samples included: diagnosis, age, study/batch, sex, PMI, RIN, libraryPrep, sequencing platform, strand specificity, brain bank, brain region, ethnicity, along with all 29 seqPCs. For continuous variables, we also included squared terms. These covariates with input into the *earth* model along with gene expression data (limma voom normalized, centered, and scaled). The model was run using linear predictors and otherwise default parameters. As the model fits a maximum of 1000 features (genes) simultaneously, we performed 1000 permutations randomly subsetting 1000 genes at a time. From this, we chose as a set of known covariates those present in at least half of the resulting pruned models, which consisted of: diagnosis, age, age^2, study/batch, sex, PMI, RIN, RIN^2, brain bank, brain region, seqPCs (1-3, 5-8, 10-14, 16, 18-25, 27-29) and seqPC3^2.

The above set contained known covariates (or those derived from known sequencing quality metrics) that contributed uniquely to variance in gene expression. However, these do not capture potential underlying hidden factors or confounders that may also influence gene expression. To ensure that DGE signal is not being driven by such hidden confounding factors, we performed surrogate variable analysis (SVA) on gene expression measurements (*43*). To determine the optimal number of SVs to include in our final model, we randomly split our dataset into equal halves and calculated differential expression for each gene and disorder using a fixed number of SVs (**Fig A3.3A**). We then compared the replicability of differential expression ($\log_2$FC) effect sizes between the two split halves of the dataset, quantified using spearman's correlation. This analysis was repeated 1000 times each for a fixed number of SVs increasing from 0 to 25.

We found that including 4 SVs in addition to the final set of known covariates above maximized this split-dataset replicability (**Fig A3.3**). As such, our final model used for all differential gene expression, isoform expression, and splicing analyses consisted of: diagnosis, age, age^2, study/batch, sex, PMI, RIN, RIN^2, brain bank, brain region, seqPCs (1-3, 5-8, 10-14, 16, 18-25, 27-29), seqPC3^2, and SVs (1-4).

Differential Gene and Transcript Expression/Usage

Count level quantifications were corrected for library size using TMM normalization in *edgeR* and were transformed as $\log_2$(CPM+0.5). DGE was then calculated using a linear mixed-effects model using the *nlme* package in R. The covariates specified in the previous section were included as fixed effects in the model. In addition, we included a random effect term for each unique subject to account for subject overlap across sequencing studies. Resulting P-values were FDR-corrected using the Benjamini-Hochberg method, to control for multiple comparisons. Differential transcript expression (DTE) was calculated similarly as for DGE except that the transcript-level quantifications from RSEM were used as inputs for the linear mixed-effects model. Finally, differential transcript usage (DTU) was calculated similarly as for DGE except that isoform percentage data reported by RSEM was used as inputs for the linear mixed-effects model.

To ensure the robustness of DGE results, we compared $\log_2$FC effect size measurements for genes identified as significantly differentially expressed in several previous studies profiling gene expression using cortical brain samples from ASD, SCZ, and BD (**Fig A3.4A**-**D**). Finally, to ensure that differential gene expression in disease was not being driven by subtle differences in RNA quality or degradation, we compared differential expression T-statistics with those experimentally derived from brain tissue samples allowed to degrade for fixed intervals of time (**Fig A3.4E**) (*22*). We did not observe substantial concordance between these RNA degradation metrics and psychiatric disease DGE summary statistics.

Enrichment Analysis of Gene Sets

Enrichment for Gene Ontology (GO; biological process, molecular function and cellular component) and KEGG pathways was performed using the *gProfileR* R v0.6.4 package (*103*). Only pathways containing less than 1000 genes were assessed. Background was restricted to brain expressed genes. An ordered query was used, ranking genes by $\log_2$FC for DE analyses or by kME for coexpression module enrichment analyses. P-values were FDR corrected to account for multiple comparisons.

Enrichment analyses were also performed using several established, hypothesis-driven gene sets including: high confidence ASD risk loci (*80*); CHD8 targets (*96*); FMRP targets (*33*); mutationally constrained genes (*97*); syndromic and highly ranked (1 and 2) genes from the SFARI GENE database; 'vulnerable' ASD genes (*98*); genes with probability of loss-of-function intolerance (pLI) > 0.99 as reported by the Exome Aggregation Consortium (*104*). Statistical enrichment analyses were performed using logistic regression, correcting for both gene length and GC content. All results were FDR-corrected for multiple comparisons.

Cell Type Enrichment Analyses

Cell type enrichment analyses were performed using uniformly processed human brain single-cell RNA-Seq datasets, compiled by the companion manuscript (*18*) which combined multiple published datasets (*91*, *105*, *106*) with newly generated data from PsychENCODE. Clustering was performed separately for single-cell datasets using TPM and UMI quantifications. See (*18*) for further details. Enrichment was performed for cell type specific marker genes using Fisher's exact test, followed by FDR-correction for multiple testing.

For neural-immune modules (**Fig A2.6)**, we additionally assessed several mouse experimentally derived cell type specific expression datasets. These included: a translating ribosome affinity purification (TRAP) dataset profiling 24 genetically identified populations of CNS cell types in mouse using microarray (*107*); a large-scale single-cell RNA-Seq dataset of mouse somatosensory cortex and hippocampus (*94*); a MARS-seq dataset of FACS-sorted CD45+ cells from mouse brain tissue, representing the major CNS immune cell populations (*93*); and a single cell RNA-Seq dataset of cells derived from meninges and choroid plexus in mouse (*92*).

Differential Local Splicing (DS) analysis

Local splicing analysis used LeafCutter (*29*), which detects splicing variation using the sequencing reads that span an intron (or spliced reads) to quantify intron usage across samples, without relying on existing reference annotations and without estimation of isoform abundance or exon inclusion levels. The same BAM alignment files to the hg19 genome assembly produced by STAR (version 2.4.2a) (*108*) for the DGE/DTE analyses were used as input for leafcutter intron clustering. The BAM files included the XS strand tags to all canonically spliced alignments based on their intron motifs (parameters: alignSJoverhangMin =8, outSAMstrandField =intronMotif). We used LeafCutter to first call clusters of variable spliced introns across all our samples and then to identify differential splicing between each disorder (ASD, SCZ, and BD) and the control (CTL) group by jointly modeling intron clusters using the Dirichlet-Multinomial generalized linear model (GLM) (*29*). We controlled for the same technical, biological covariates and hidden confounds as described above in the DGE/ DTE analyses, except that we did not incorporate a random term for individuals (random effects are not supported by the Dirichlet-Multinomial GLM). Accordingly, we also removed tissue sample replicates that were sequenced in more than one study, randomly retaining only one sample in our analysis. The dataset for LeafCutter analysis numbered 944 controls, 79 ASD, 531 SCZ and 217 BD samples (1,771 total).

We used LeafCutter to call intron clusters as follows: overlapping introns, defined as spliced reads, were clustered and filtered to keep intron clusters supported by at least 50 split reads across all 1,771 samples, retaining introns of up to 100 kb and accounting for at least 1% of the total number of reads in the entire cluster. This yielded 37,215 clusters encompassing 120,921 introns in 17,342 genes that were used for further analysis. This intron count file was then used in the differential splicing (DS) analysis.

DS intron clusters were identified in pairwise analyses comparing each psychiatric disorder (ASD, BD, SCZ) to the common set of 944 control samples. After discarding introns that were not supported by at least one read in 5 or more samples, clusters were analyzed for DS if at least 3 samples in each comparison group (*i.e*. cases or controls) had an overall coverage of 20 or more reads. *P*-values were corrected for multiple testing using the Benjamini-Hochberg (BH) method and used to select clusters with significant splicing differences (FDR *q*< 0.1).

Percent-spliced-in (PSI) values were corrected for covariates using the *quantify_PSI* function provided in the LeafCutter *psi* branch (https://github.com/davidaknowles/leafcutter/tree/psi). Violin plots of intron PSI values were prepared using ggplot2. Principal component analysis (PCA) plots were evaluated before and after covariate-correction (**Fig A3.6**). Schematic visualization of significant intron clusters was done using the *leafviz* R shiny package [https://davidaknowles.github.io/leafcutter/articles/Visualization.html]. All DS events were further annotated using *leafviz* and custom R code, manually inspected, and classified into single- or multi-exon skipping events (changes in cassette splicing), alternative 5' and 3' exon usage, and alternative 5' (donor) or 3' (acceptor) splice site usage. Intron clusters that did not match any of these categories were classified as complex events involving multiple changes. DS intron clusters were mapped onto transcripts using *gViz* (v3.7) (*109*) and *ensembldb* (v3.7) (https://github.com/jotsetung/ensembldb) bioconductor R packages. SMART (*110*) and PFAM (*111*) protein domains were mapped onto transcript structures using the *proteinToGenome* function of *ensembldb*.

DeltaPSI (∆PSI) Correlation Across Disorders

To determine significance for the correlation of ∆PSI across disorders for significant intron clusters identified by LeafCutter, we permuted the case/control status within each disorder 3,000 times and repeated the LeafCutter analysis with the same GLM described above. In each permutation we assessed the ∆PSI correlation between disorders using Spearman's correlation (ρ) to yield a null distribution of *ρ* values that was used to assess the significance of the observed correlations.

Cross-disorder DS Overlaps

For cross-disorder DS overlaps we selected all genes associated with significant intron clusters identified by LeafCutter at FDR <10%. Venn diagrams area-proportional to the number of genes with significant DS clusters in each disorder were then created using the *eulerr* R package. Hypergeometric p-values for pairwise overlaps between disorders were calculated using the *phyper* function in R and setting the size of the 'universe' to all genes with intron clusters meeting the LeafCutter clustering criteria.

Functional Enrichment of DS Genes

Gene set enrichment for Gene Ontology (GO) biological process, molecular function and cellular component aspects was performed using the *gProfileR* v0.6.4 package in R (*103*) with moderate hierarchical filtering and using an ordered query, after ranking genes in increasing order of the LeafCutter p-value (i.e. most significant at the top). In case a gene had multiple significant intron clusters, the most significant cluster was used for the ranking. The custom background set for each disorder consisted of all 10,677 genes with intron clusters that were evaluated in pairwise LeafCutter analyses between each disorder and control groups, as described above. Visualization of enriched GO terms used custom ggplot2 functions.

Gene-set enrichment for RBFOX1 targets (*32*), FMRP targets (*33*), and a curated list of genes coding for RNA binding proteins (RBPs) (*34*), used the same custom background of genes with LeafCutter intron clusters as in the GO enrichment analysis, and was assessed using Fisher's exact test and correcting for multiple testing by FDR. The curated list of RBP genes included those with high confidence for RNA binding (*112–114*) and those annotated as RNA-binding in Ensembl including known and potentially auxiliary splice factors (*34*).

The comparison of LeafCutter DS events with Parishak *et al*. 2016 (*19*) MATS events was based on genomic coordinates overlap using BEDtools, irrespective of the event type assigned by each algorithm.

Microexon Enrichment

Transcripts that carry at least one exon of 3-27 nucleotides in length (*i.e.* microexon) (*30*) were extracted from the Gencode V19 database. Statistical enrichment analyses were performed using logistic regression, correcting for both gene and transcript length on linear and log10 scales. As an additional control, transcripts that carry exon(s) of average length (143±5% nucleotides) were also extracted and their overlap with switch transcripts was also tested with logistic regression. All results were FDR-corrected for multiple comparisons by the Benjamini-Hochberg method.

Construction of Disease-specific Isoform-level Co-expressed PPI Networks

Pairwise spearman correlation coefficients (SCC) between transcript of interest and all other transcripts were calculated using either ASD samples or SCZ samples. To obtain age-balanced datasets, only samples from donors of age 17-67 years old were used. A cutoff of SCC > 0.5 was used to filter for the co-expressed partners of the transcript of interest in either ASD samples or SCZ samples. PPI data was compiled from well-characterized PPI databases, including Bioplex (*115*), HPRD (*116*), Inweb (*117*), HINT (*118*), Biogrid (*119*), GeneMANIA (*120*), STRING (*121*) and CORUM (*122*). Only physical interactions and co-complex associations were kept. Co-expressed partners which are also supported by PPI were used to construct the co-expressed PPI network.

ncRNA Annotation

To identify ncRNAs that may be relevant to neuropsychiatric disorders, we compiled a list of non-protein-coding genes exhibiting differential gene (DGE) or transcript expression (DTE) at FDR < 0.05 in at least one disorder (**Table A3.2**). As ncRNA designation can change based on genomic annotation, we filtered out genes that were designated as protein coding in the most recent version of Gencode v27, yielding a total of 944 unique ncRNAs, many of which were differentially expressed across more than one disorder or at both gene and transcript-level features. Differentially expressed ncRNAs were annotated according to sequence and expression characteristics. Human tissue-specific expression was assessed using data from GTEX v6. Median RPKM values per tissue were obtained and averaged into broad categories (**Fig A2.1F**). To identify ncRNAs broadly expressed across human tissues, we ran an ANOVA on log2(RPKM +1) values across tissues, and selected those with uncorrected P > 0.05. Brain-specific expression was defined as $RPKM_{brain}$ / sum($RPKM_{all\ tisues}$) > 0.8. CNS cell type specificity was assessed in a similar fashion using single-nucleus RNA-Seq from the Lake dataset (*91*). Expression counts were CPM normalized and then averaged together across defined cell clusters. We ran an ANOVA on log2(CPM + 1) values across cell clusters, and report those ncRNAs with P>0.05 as “broadly expressed” with regard to cell type. Cell type specificity was quantified by $CPM_{max\ cluster}$ / sum($CPM_{all\ cell\ clusters}$) > 0.8.

Evolutionary conservation was assessed using phastCons and phyloP scores (*123*, *124*). Both methods assign a score to each base in the human genome, quantifying its degree of conservation across selected species. Whereas phastCons base scores are smoothed according to scores of neighboring bases, phyloP evaluates each base independently. We downloaded phastCons and phyloP per-base scores for hg19 from UCSC, computed from 17-way (primate), 30-way (mammalian), and 100-way (vertebrate) Multiz alignments, to calculate a mean base score for each ncRNA across 1) the gene, and 2) its exonic regions only. The per-exon scores were averaged over all exons belonging to a gene to produce a more robust metric for gene conservation.

Context-dependent tolerance (CDTS) scores were used to quantify patterns of human selective constraint (*27*). CDTS scores are computed for each 10bp window in high-confidence regions of the genome, which we intersected with exonic coordinates for ncRNAs using Bedtools (*125*). To produce a per-gene score, we first computed the mean across all 10bp windows intersecting a single exon, then averaged the mean exon scores across all exons for a gene.

Signed Gene and Isoform Co-Expression Network Analysis

To place results from individual genes within their systems-level network architecture, we performed Weighted Gene Co-Expression Network Analysis (WGCNA) separately for gene- and isoform-level quantifications (*49*). All covariates except for diagnostic group were first regressed from our expression dataset. Network analysis was performed with the WGCNA package using signed networks. A soft-threshold power of 7 was used for all studies to achieve approximate scale-free topology (R^2>0.8). Networks were constructed using the blockwiseModules function. The network dendrogram was created using average linkage hierarchical clustering of the topological overlap dissimilarity matrix (1-TOM). Modules were defined as branches of the dendrogram using the hybrid dynamic tree-cutting method. Modules were summarized by their first principal component (ME, module eigengene) and modules with eigengene correlations of >0.9 were merged together. A robust version of WGCNA (rWGCNA) was run to reduce the influence of potential outlier samples on network architecture (*126*). Module robustness was ensured by randomly resampling (2/3 of the total) from the initial set of samples 100 times followed by consensus network analysis, a meta-analytic approach, to define modules using a consensus quantile threshold of 0.2. Modules were defined using biweight midcorrelation (bicor), with a minimum module size of 50, deepsplit of 4, merge threshold of 0.1, and negative pamStage. Module (eigengene)-disease associations were evaluated using a linear mixed-effects model, using a random effect of subject, to account for subject overlap across datasets. Significance values were FDR-corrected to account for multiple comparisons. Results from module-eigengene association tests are shown in **Fig A2.4**. Genes within each module were prioritized based on their module membership (kME), defined as correlation to the module eigengene. The top 'hub' genes for several of the modules are shown in **Figs A2.4**-**A2.6** and through an interactive portal on our companion website (Resource.PsychENCODE.org).

The robustness of all network modules were tested as described previously (*48*). In brief, each module's density (defined as the average intramodular topological overlap) was compared to the density of modules of equivalent size selected randomly from the same network (n = 5,000 permutations). Density p-values were determined for each initial module by calculating the percentage of trials in which the density of the "random" modules exceeded the density of the initial module. All modules have density p-values less than 0.05.

csuWGCNA

We also used a modified version of WGCNA named *Combination of Signed and Unsigned WGCNA* (csuWGCNA), which captures strong and moderate negative correlations in the coexpression network (*72*). Current versions of WGCNA allow unsigned, signed and signed hybrid options for network types, but have disadvantages when trying to capture moderate negatively correlated features such as lncRNA-mRNA regulatory relationships. Signed and signed hybrid networks down-weight negatively correlated pairs in network. Unsigned networks highlight strong positive and negative correlations, but has worse performance on identifying functionally-related gene pathways than its signed counterparts (*127*). To address these limitations, we modified two functions for picking soft thresholding power and calculating the network adjacency. The core modification of csuWGCNA is in its definition of adjacency, $a_{ij} = ((1 + |cor(x_i , x_j)|)/2)^{\beta}$, which integrates the advantages of signed networks ($a_{ij} = |(1 + cor(x_i , x_j))/2|^{\beta}$) and unsigned networks ($a_{ij} = |cor(x_i , x_j)|^{\beta}$). Using this adjacency function, csuWGCNA then constructs a topological overlap matrix (TOM) and follows the procedure described above for clustering, tree cutting, and network module detection. Using this method, csuWGCNA can detect modules containing genes with negative correlations, which may be more useful when lncRNAs and miRNAs are included in the network (**Fig A3.14**).

Assessment of Psychiatric Medications

To assess the potential impact of medications on differential expression and co-expression results, we analyzed several published datasets of animal models exposed to multiple classes of psychiatric medications. These included: 1) a published RNA-Seq dataset of the DLPFC from non-human primates exposed for six months to haloperidol, clozapine, or placebo (*2*); 2) a published microarray dataset (GSE66276) of cortex from mice exposed to the SSRI fluoxetine for 21 days (*128*); 3) a microarray dataset (GSE66276) of rats exposed lithium, lamotrigine or placebo for 21 days. All datasets were reprocessed and analyzed as described below.

The antipsychotic dataset consisted of ~6 year old rhesus macaques treated with medications or placebo orally for six months, including high doses of haloperidol (4 mg/kg/d; n=7), low doses of haloperidol (0.14 mg/kg/d; n=10), clozapine (5.2 mg/kg/d; n=9), or vehicle (n=8). DLPFC tissue was extracted and RNA-Seq was run using rRNA-depleted libraries. Genes were kept that had expression greater than 0.1 cpm (counts per million) in at least half of samples. Limma voom normalization using TMM normalization factors was used for subsequent differential gene expression analysis, including the covariates: age, sex, sequencing batch factors, RNA quality statistics (RIN and RNA concentration) and sequencing statistics. In accordance with results from (*2*), none of the groups (clozapine, haloperidol_low_dose, and haloperidol_high_dose) had any genes significantly differentially expressed from placebo after FDR-correcting for multiple comparisons. We therefore used an unadjusted p-value threshold of 0.01 for downstream analyses, resulting in 133, 120, and 188 genes for clozapine, haloperidol_low_dose, and haloperidol_high_dose, respectively. Genes were grouped based on direction of effect (up or downregulated) and mapped to human orthologues using Ensembl. Overlap with PsychENCODE disease gene sets (DE and DS genes, gene and isoform-level coexpression modules) was assessed using Fisher's exact test followed by FDR correction. Although some gene sets showed nominal overlap with antipsychotic genes, no enrichments were significant after correction for multiple comparisons (**Fig A3.11**).

In the SSRI dataset, thirty mouse strains were treated for 21 days with fluoxetine (18 mg/kg per day) or vehicle, cortical RNA was extracted and profiled with an Affymetrix expression microarray (GeneChip Mouse Genome 430 2.0 array). Raw microarray data was normalized using the RMA function from the 'affy' package in R. Batch correction was performed using ComBat, and differential expression was detected using the lmFit and eBayes functions from the 'limma' R package (covariates for the linear model: fluoxetine treatment, strain, and RNA degradation score). In our analysis, only two genes were found to be significantly differentially expressed (downregulated) in the fluoxetine group after correction for multiple comparisons (FDR p-value < 0.05): SST, a hormone regulating factor, and FDFT1, an enzyme involved in cholesterol biosynthesis. For downstream enrichment analyses, we used the relaxed threshold of p < 0.01 (uncorrected), corresponding to 558 genes.

In the third dataset, rats (n=5 per group) were administered lithium in chow (0.2%) or lamotrigine via subcutaneous injection (30 mg/kg) and compared to a vehicle chow group or vehicle injection group. All regimens were administered once daily for 21 days and tissues were collected from frontal cortex, striatum, and hippocampus for analysis via an Affymetrix expression microarray (Affymetrix Rat Genome 230 2.0 Array). Differential expression analysis was performed as described above. For lamotrigine, no genes were differentially expressed following FDR-correction, so for downstream enrichment analyses we used the relaxed threshold of p < 0.01 corresponding to 121 genes. For lithium, 2338 genes at FDR-corrected p < 0.05 were used for downstream enrichment analyses.

Assessment of Non-linear Age Effects

To assess the influence of age on the magnitude of differential expression, and to account for potentially non-linear effects of age, we performed a local regression analysis using the locfit package in R. For each gene expression measure, a local regression function was fit to model the effect of age on expression in control samples, as follows: fit = locfit(Expr ~ Age, data=df[df$Group=="CTL",])

For each non-control sample, expression was then converted to a z-score using the interpolated mean expression in controls at the same age. We then assessed the correlation between z-transformed expression and age within each disease group (ASD, SCZ, BD) separately, to identify those genes whose magnitude of differential expression was associated with age. Several examples are shown in **Fig A3.10**.

GWAS Datasets

We performed a number of GWAS enrichment analyses as described in the following sections. In each analysis, we used summary statistics from the largest publicly available GWAS in SCZ (*58*), ASD (*38*), and BD (*90*). Additional secondary analyses were performed using a variety of relevant traits, including major depressive disorder (MDD; ref (*129*)), neuroticism (*130*), educational attainment (*131*), diabetes (*132*), as well as previous GWA studies of SCZ (*133*), ASD (*134*), and BD (*135*).

GWAS Enrichment in DE Genes and Modules

We used stratified LD score regression (s-LDSR) (*39*) to investigate whether differentially expressed or spliced genes, and/or co-expression modules, are enriched for disease-associated genetic variation using the summary statistics described above. SNPs were assigned to these custom gene categories if they fell within ±10 kb of a gene in the set. These categories were added to a 'full baseline model' that includes 53 functional categories capturing a broad set of genomic annotations, as published (*39*). Enrichment was calculated as the proportion of SNP heritability accounted for by each module divided by the proportion of total SNPs within the module. Significance was assessed using a block jackknife procedure, followed by FDR correction of P values.

Polygenic Risk Score Calculation

Polygenic risk scores (PRS) were calculated using the same GWAS summary statistics as above, for SCZ (*58*), BD (*90*) and ASD datasets (*38*). Samples were restricted to those of European ancestry based on clustering with samples from HapMap3 (*136*). Genotypes were additionally filtered as follows, using plink: plink --bfile PECDC_EUR --geno 0 --maf 0.05 --hardy --hwe 1e-40 --make-bed –out PECDC_EUR_PRSfilter. To calculate PRS, we used LDpred (*137*) with the 1000 Genomes phase 3 European subset as a reference panel. The first five genotype principal components (gPC1-5, as calculated with plink) were included in the PRS calculation, to account for ancestry and technical effects. We then compared PRS for each diagnostic group with the strict set of non-psychiatric controls, contrasting baseline and full models. PRS significance was measured with a likelihood ratio test and Nagelkerke's pseudo-R^2.

```
mod.baseline=glm(dx~study+sex+age+gPC1+gPC2+gPC3+gPC4+gPC5,family=binomial)
mod.full=glm(dx~PRS+study+sex+age+gPC1+gPC2+gPC3+gPC4+gPC5,family=binomial)
adjustedR2=NagelkerkeR2(mod.full)$R2-NagelkerkeR2(mod.baseline)$R2
prs.significance=lrtest(mod.baseline, mod.full)
```

The default LDpred GWAS p-value thresholds were used (.001, .003, .01, .03, .1, .3, 1, and Inf). Maximal Nagelkerke pseudo-R^2 values were achieved for prediction of psychiatric diagnosis using thresholds of 0.001 for ASD, 0.01 for BD, and 1 for SCZ.

Association between PRS and measures of gene, isoform, or module (eigengene) expression was performed as described above, except using linear regression analogs. Associations were repeated for each p-value threshold in the 3 GWAS studies and the resulting association p-values were then FDR-corrected for multiple testing. Full results are compiled in **Table A3.4**.

Transcriptome-wide Association Study (TWAS)

To identify genes whose *cis*-regulated expression is associated with disease, we performed a transcriptome wide association study (TWAS) to identify putative molecular (e.g., *cis*-eQTL) phenotypes in brain underlying disease GWAS associations (*46*). TWAS was implemented using the fusion package (https://github.com/gusevlab/fusion_twas; (*46*)) with custom SNP-expression weights generated from brain using our dataset of 1321 unique individuals of European ancestry with imputed genotypes. Using the AI-REML algorithm implemented in GCTA (*138*) by the fusion package, we first identified the subset (n=14,750) of total expressed genes found to have significant *cis* SNP-heritability in our dataset (*cis*- h^2_g P<0.05 within 1 Mb window around the gene body). SNP-expression weights were calculated in a 1Mb region around all heritable genes using expression measurements adjusted for diagnosis, study, age, age^2, RIN, RIN^2, sex, tissue, PMI, 20 ancestry PCs, and 100 hidden covariates (*139*). Accuracy of five expression prediction models were tested (best *cis*-eQTL, best linear unbiased predictor, Bayesian linear mixed model, Elastic-net regression, LASSO regression) using the most accurate model for final weight calculations as implemented in fusion. TWAS disease-association statistics were computed using these custom weights, LD structure calculated from our PsychENCODE samples' genotypes, and disease GWAS summary statistics described above. For each disease, TWAS association statistics were Bonferroni-corrected for multiple comparisons. Full results are compiled in **Table A3.4**.

Rare Variant Enrichment Analyses

Gene and isoform co-expression modules were also assessed for enrichment of rare variants identified in disease, compiled from several datasets. These included: 71 risk loci harboring rare *de novo* variants associated with ASD through the transmission and *de novo* association test (TADA) (*80*); Syndromic and highly ranked (1 and 2) genes from SFARI Gene database; genes harboring recurrent *de novo* copy-number variants associated with ASD or SCZ, as defined in (*1*); genes harboring an excess of rare exonic variants in ASD, SCZ, intellectual disability (ID), developmental delay (DD), and epilepsy as assessment through an extended version of TADA (extTADA) (*140*); genes harboring disruptive and damaging ultra-rare variants (dURVs) in SCZ (*54*); a list of high confidence epilepsy risk genes, compiled in (*141*). For binary gene sets, enrichment among gene and isoform modules was calculated using logistic regression, correcting for linear- and log-transformed gene and transcript lengths as well as GC content. For dURVs, a two step procedure was used, first creating a logistic regression model for dURV genes identified in controls and a second model for those affected in cases and controls. A likelihood ratio test was used to calculate significance. Finally, for the extTADA datasets, the posterior-probability (PP) was used in the logistic regression model in place of a binary annotation. P-values were FDR-corrected for multiple comparisons. Results are shown in **Fig A3.13** and compiled in **Table A3.5**.

Experimental Validation

Initial optimization of the PCR conditions for all splicing and isoform primers used cDNA samples derived from total brain or cortex RNA (Clontech), and were performed on a Mastercycler Nexus Gradient Thermal Cycler (Eppendorf) and amplicons resolved on precast 96-well 2% agarose E-Gels (Invitrogen) stained with SYBR safe.

Splicing validation

For differential splicing (DS) analysis, selected exon-skipping events were validated by semiquantitative RT–PCR in ASD, BD, SCZ, and control samples. Total RNA (1-2 µg) was treated with 1 unit of Baseline-ZERO DNase (Lucigen), cleaned up with 1.8x AMPure XP (Beckman Coulter), and reverse-transcribed using SuperScript III reverse transcriptase and random hexamer primers (Invitrogen). After clean-up with 1.8x AMPure XP, DS events were PCR amplified from 20 ng of cDNA for 30 cycles in 25 µl volume containing exon-specific primers at a concentration 0.5 µM each, and ChoiceTaq Blue MasterMix (Denville) according to manufacturer instructions. Exon-specific PCR primers (**Table A3.8**) were designed in the flanking exons of each skipping event using Primer3 (*142*) and BLAST (*143*). PCR products were cleaned up with 1.8x AMPure XP (Beckman Coulter) and analyzed on DNA 1000 chips on an Agilent 2100 Bioanalyzer system. Peaks corresponding to the amplicon including or excluding the skipped exon were quantified using the Bioanalyzer Expert software, and percent spliced in (PSI) ratios were calculated by dividing the molarity of the lower band (exon skipped) by the sum of the molarity of the lower and upper band (exon included). The ΔPSI between cases and control for each event was calculated as the difference between the average PSI in cases and average PSI in controls. Sample details and primers are reported in **Table A3.8**.

Isoform Validation

For DTE analysis, selected isoforms were validated by semiquantitative RT–PCR using a similar approach as for DS. Each isoform was PCR amplified from 20 or 40 ng of cDNA for 30 or 35 cycles in a 25 µL volume containing isoform-specific primers at a concentration 0.5 µM each and ChoiceTaq Blue MasterMix with DNA polymerase (Denville), or 0.4 µM each and LongAmp Hot Start Taq DNA polymerase (New England Biolabs) (**Table A3.8**), according to manufacturer instructions. Isoform-specific PCR primers (**Table A3.8**) were designed using Primer3 (*142*) and BLAST (*143*), and based on GENCODE v19 annotations. PCR products were resolved on 1.5 or 2% agarose gels, counterstained with GelStar Nucleic Acid Gel Stain (Lonza) for visualization, and *GAPDH* and *ACTB* were used as loading controls. Gels were quantified using ImageLab (BioRad). The intensity of each isoform was first normalized to the average expression levels of *GAPDH* and *ACTB* in each sample. The intensity ratio between cases and controls for each isoform was then calculated by dividing the average intensity of cases by the average intensity of controls. The log2 intensity ratios were then compared to the log2 ratio differences from the DTE analysis. Sample details and primers are reported in **Table A3.8**.

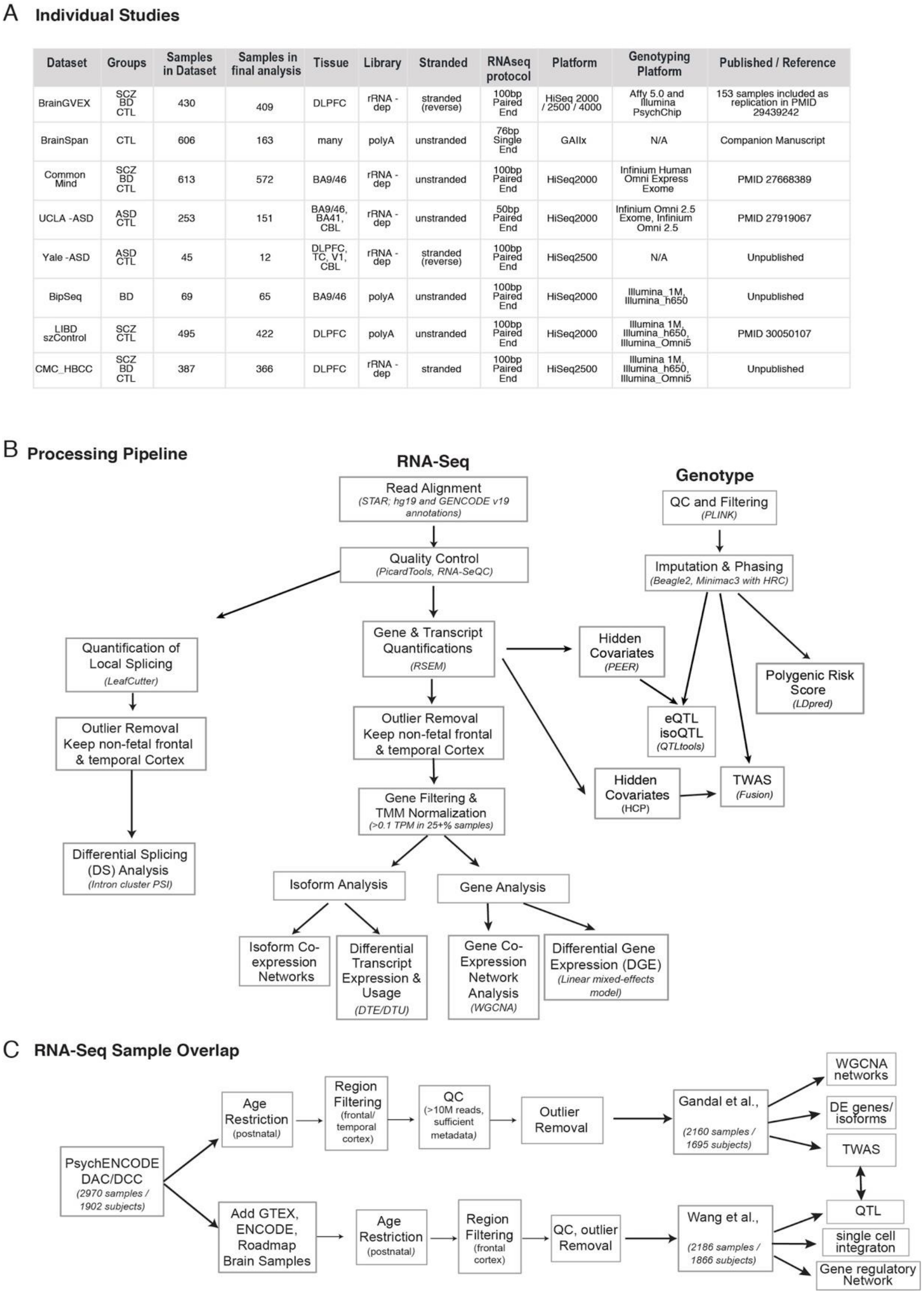
Fluorescent *in situ* hybridization (FISH)

Fresh-frozen tissue blocks from the Brodmann's area 9 of the prefrontal cortex of five neurologically normal control donors were obtained from the Mount Sinai Neuropathology Research Core and Brain Bank and stored at -80ºC. Clinical information on the subjects is summarized in **Table A3.9A**. The blocks were embedded in O.C.T. compound, frozen at -20ºC, 10 µm-thick sections were cut using a cryostat (Leica), and the sections were collected onto Superfrost Plus slides. The slides were stored in an airtight box at -80ºC until FISH was conducted.

The *in situ* hybridization probes for detecting human *GAD1*, *LINC00643*, and *LINC01166* as well as the positive and negative control probes were designed by Advanced Cell Diagnostics (ACD; see **Table A3.9B** for RNAscope probe information). For the assay, we used RNAscope Multiplex Fluorescent Reagent Kit v2 (ACD), that provided the hydrogen peroxide, protease IV, amplification reagents (Amp1-3), HRP reagents, and wash buffer for probe hybridization. DAPI, TSA buffer (ACD) and TSA Plus fluorophores (PerkinElmer) were used for detection of the signal. We used a modified version of the manufacturer's protocol for sample preparation, probe hybridization, and signal detection. Briefly, the fresh frozen sections on slides were retrieved from -80ºC and immediately fixed by immersion in freshly prepared cold 4% paraformaldehyde for 2 h. After fixation, the sections were rinsed briefly with phosphate buffered saline (PBS) and then dehydrated in an ethanol series (5 min each in 50%, 70%, and two changes of 100% ethanol) at room temperature (RT). The sections were air-dried for 5 min and a hydrophobic barrier was created around the section using an Immedge pen (Vector Laboratories). After the barrier had completely dried, the sections were treated with hydrogen peroxide for 10 min at RT, washed twice with PBS, treated with protease IV for 15 min at RT, and washed twice with PBS. The LINC-C2 or LINC-C3 probes for detecting lncRNAs were diluted at 1:50 in the GAD1-C1 probe. The sections were then hybridized with the probes at 40ºC for 2 h in the HybEZ Hybridization System (ACD), washed twice with wash buffer, and stored overnight at RT in 5x SSC buffer. The next day, the slides were rinsed twice with wash buffer, followed by the three amplification steps (AMP 1, AMP 2, and AMP 3 at 40ºC for 30, 30, and 15 min respectively, with two washes of 2 min each with wash buffer after each amplification step). The signal was developed by treating the sections in sequence with the HRP reagent corresponding to each channel (e.g. HRP-C1) at 40ºC for 15 min, followed by the TSA Plus fluorophore assigned to the probe channel (fluorescein for GAD1-C1 probe and cyanine 5 or Cy5 for LINC-C3 probes, prepared at a dilution of 1:750) at 40ºC for 30 min, and HRP blocker at 40ºC for 15 min, again with two wash steps after each of the incubation steps. As autofluorescence due to lipofuscin was detected in both the green and the red channels whereas the far red channel was relatively free of background, the highly expressed GAD1-C1 probe was assigned to the green fluorescein channel, the red cyanine 3 channel was left empty and the lncRNAs were probed on separate sections in the far red Cy5 channel. The sections were treated with TrueBlack Lipofuscin Autofluorescence Quencher (Biotium) for 30 s, rinsed twice with PBS, counterstained with DAPI for 30 s, mounted using ProLong Gold mounting medium (Thermo Fisher Scientific) and slides were stored at 4 ºC until ready for imaging. Two experiments were performed with two to three biological replicates each, and using positive and negative control probes to test for RNA quality and background signal respectively.

Layer III of area 9 was identified using a 5x/0.16 N.A. objective and the sections were imaged using a 63x/1.4 N.A. or 100x/1.4 N.A. oil DIC Plan Apochromat objectives on an AxioImager.M2 microscope (Carl Zeiss), equipped with a motorized stage (MBF Biosciences) and an Orca-R^2 digital camera (Hamamatsu), and operated using Neurolucida software (version 11.11.3 64-bit, MBF Biosciences). Camera exposure times were set for each of the four channels (red for lipofuscin, blue for DAPI, green for fluorescein, and magenta for Cy5) and were kept similar among the cases imaged in each experiment in order to enable comparison. The images at 100x magnification were presented as maximum intensity projections of Z-stacks imaged at 0.5 µm intervals. Adobe Photoshop was used for adjusting brightness/contrast and sharpness of the images.

A3.2: Extended Figures



A Individual Studies

Dataset	Groups	Samples in Dataset	Samples in final analysis	Tissue	Library	Stranded	RNAseq protocol	Platform	Genotyping Platform	Published / Reference
BrainGVEX	SCZ BD CTL	430	409	DLPFC	rRNA - dep	stranded (reverse)	100bp Paired End	HiSeq 2000 / 2500 / 4000	Affy 5.0 and Illumina PsychChip	153 samples included as replication in PMID 29439242
BrainSpan	CTL	606	163	many	polyA	unstranded	76bp Single End	GAIIx	N/A	Companion Manuscript
Common Mind	SCZ BD CTL	613	572	BA9/46	rRNA - dep	unstranded	100bp Paired End	HiSeq2000	Infinium Human Omni Express Exome	PMID 27668389
UCLA -ASD	ASD CTL	253	151	BA9/46, BA41, CBL	rRNA - dep	unstranded	50bp Paired End	HiSeq2000	Infinium Omni 2.5 Exome, Infinium Omni 2.5	PMID 27919067
Yale -ASD	ASD CTL	45	12	DLPFC, TC, V1, CBL	rRNA - dep	stranded (reverse)	100bp Paired End	HiSeq2500	N/A	Unpublished
BipSeq	BD	69	65	BA9/46	polyA	unstranded	100bp Paired End	HiSeq2000	Illumina_1M, Illumina_h650	Unpublished
LIBD szControl	SCZ CTL	495	422	DLPFC	polyA	unstranded	100bp Paired End	HiSeq2000	Illumina 1M, Illumina_h650, Illumina_Omni5	PMID 30050107
CMC_HBCC	SCZ BD CTL	387	366	DLPFC	rRNA - dep	stranded	100bp Paired End	HiSeq2500	Illumina 1M, Illumina_h650, Illumina_Omni5	Unpublished

B Processing Pipeline

RNA-Seq
Read Alignment (STAR; hg19 and GENCODE v19 annotations)
Quality Control (PicardTools, RNA-SeQC)
Quantification of Local Splicing (LeafCutter)
Outlier Removal Keep non-fetal frontal & temporal Cortex
Differential Splicing (DS) Analysis (Intron cluster PSI)
Gene & Transcript Quantifications (RSEM)
Outlier Removal Keep non-fetal frontal & temporal Cortex
Gene Filtering & TMM Normalization (>0.1 TPM in 25+% samples)
Isoform Analysis
Gene Analysis
Isoform Co-expression Networks
Differential Transcript Expression & Usage (DTE/DTU)
Gene Co-Expression Network Analysis (WGCNA)
Differential Gene Expression (DGE) (Linear mixed-effects model)
Genotype
QC and Filtering (PLINK)
Imputation & Phasing (Beagle2, Minimac3 with HRC)
Hidden Covariates (PEER)
eQTL isoQTL (QTLtools)
Polygenic Risk Score (LDpred)
Hidden Covariates (HCP)
TWAS (Fusion)

C RNA-Seq Sample Overlap

PsychENCODE DAC/DCC (2970 samples / 1902 subjects)
Age Restriction (postnatal)
Region Filtering (frontal/ temporal cortex)
QC (>10M reads, sufficient metadata)
Outlier Removal
Gandal et al., (2160 samples / 1695 subjects)
Add GTEX, ENCODE, Roadmap Brain Samples
Age Restriction (postnatal)
Region Filtering (frontal cortex)
QC, outlier Removal
Wang et al., (2186 samples / 1866 subjects)
WGCNA networks
DE genes/ isoforms
TWAS
QTL
single cell integraton
Gene regulatory Network

Fig. A3.1. Dataset composition, analysis and integration pipeline

A) Description of individual studies contributing to this PsychENCODE analysis. B) Analysis pipeline through which all samples were uniformly processed. C) Comparison of samples overlapping between this manuscript and our companion paper (*18*).

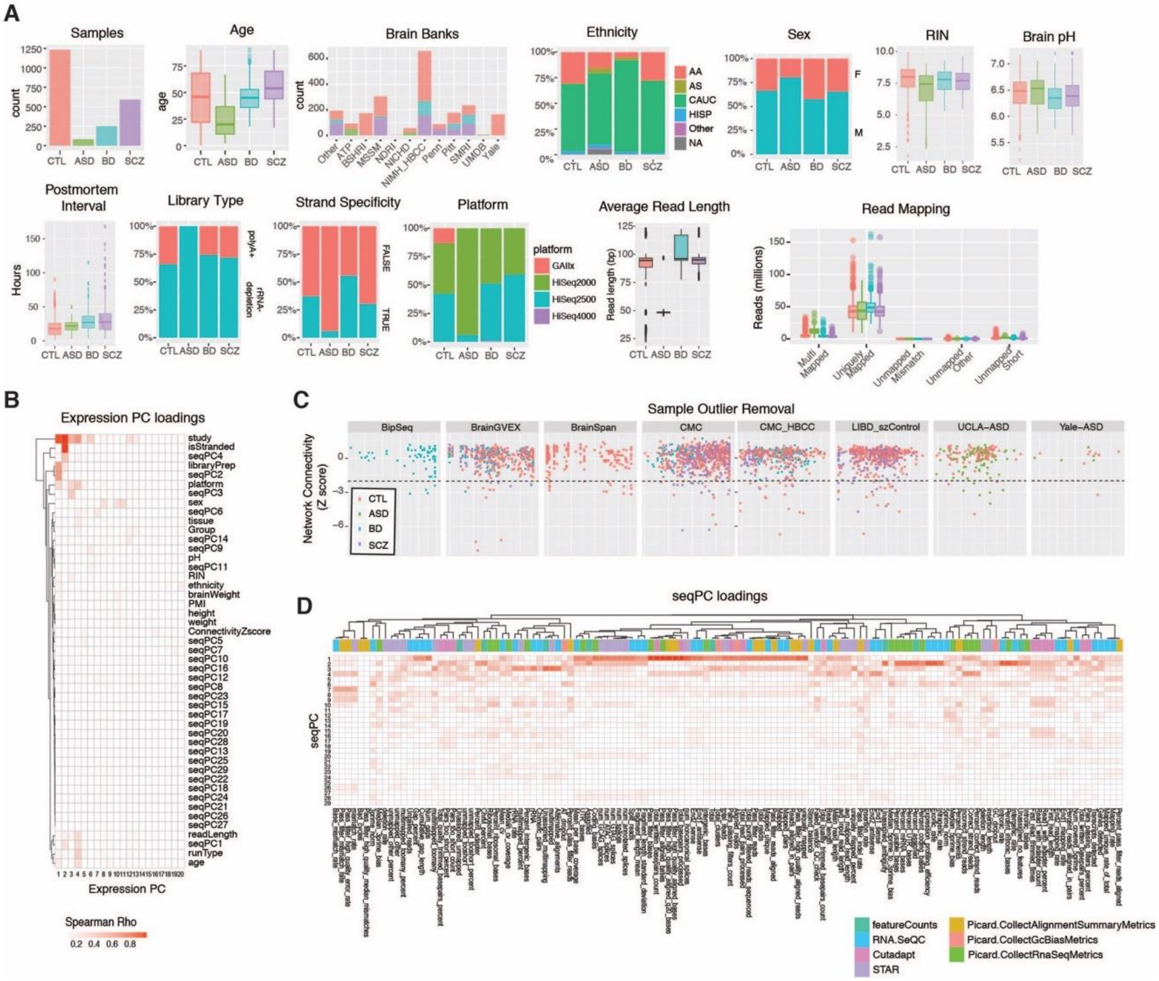


Fig. A3.2. Dataset demographics and quality control

A) Sample information, subject demographics, and sequencing characteristics are shown for each group. This study only used frontal and temporal cortex samples from subjects at postnatal time points. B) Spearman's ρ values are shown for correlations between dataset covariates with the top 20 expression PCs. C) Sample outlier removal was performed individually for each study before combining data, based on Z-scores of standardized network connectivity (**Methods**). D) Sequencing surrogate variables ('seqPCs') were calculated as the top 29 principal components of the matrix of sequencing QC metrics. Loadings are shown between seqPCs and individual metrics, colored by the source of the QC metrics.

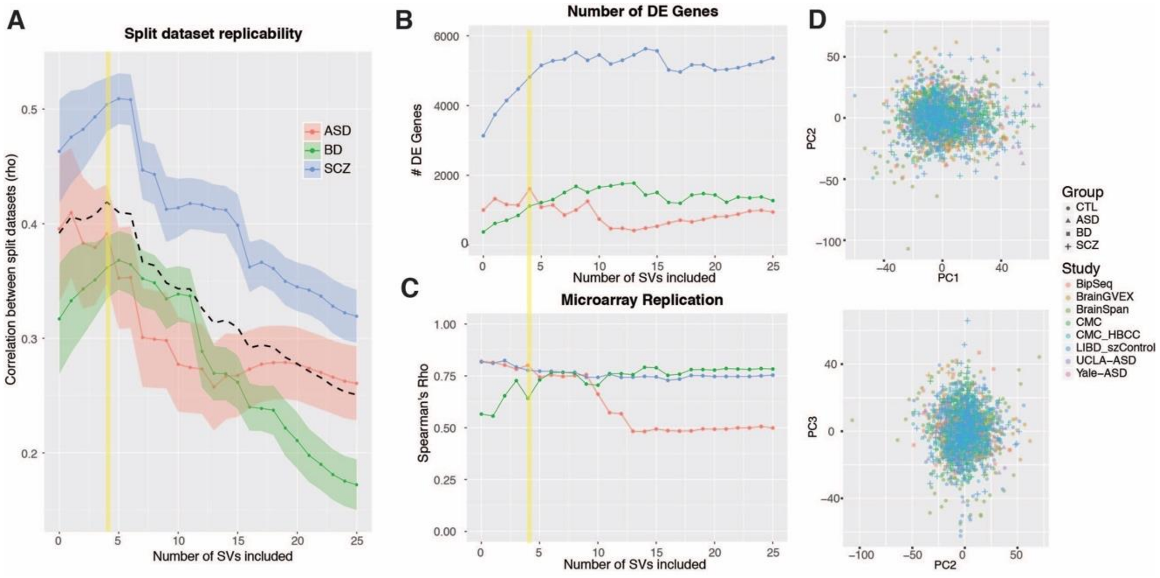


Fig. A3.3. Selection of Covariates

To capture the full range of factors influencing gene expression in our dataset, the final differential expression model included known covariates, aggregate sequencing metrics (seqPCs), and surrogate variables (SVs) calculated using SVA to correct for unmeasured sources of variation. A) To determine the appropriate number of SVs to use, we randomly split our dataset into two halves and calculated differential gene expression for each disorder using a fixed number of SVs ranging from 0 to 25. We compared DGE for each disorder between the split datasets using spearman's correlation of $\log_2$FC effect sizes for all brain-expressed genes (N=25,774 genes). We then repeated this analysis 1000 times and compared results across the range of SVs included. Addition of 4 SVs yielded the greatest cross-dataset replicability. B) Here, we plot the number of genes considered differentially expressed as a function of the number of SVs included in the differential expression model. C) DE results from this study are compared with published microarray datasets for each disorder (*1*) as a function of the number of SVs included. Spearman's correlation is shown for DGE $\log_2$FC effect sizes for genes previously identified as DE (FDR<0.05) in the microarray dataset, as described in **Fig A3.4**. D) Multidimensional scaling plots are shown for the top 3 PC's of the covariate-corrected dataset, colored by study/batch and diagnosis.

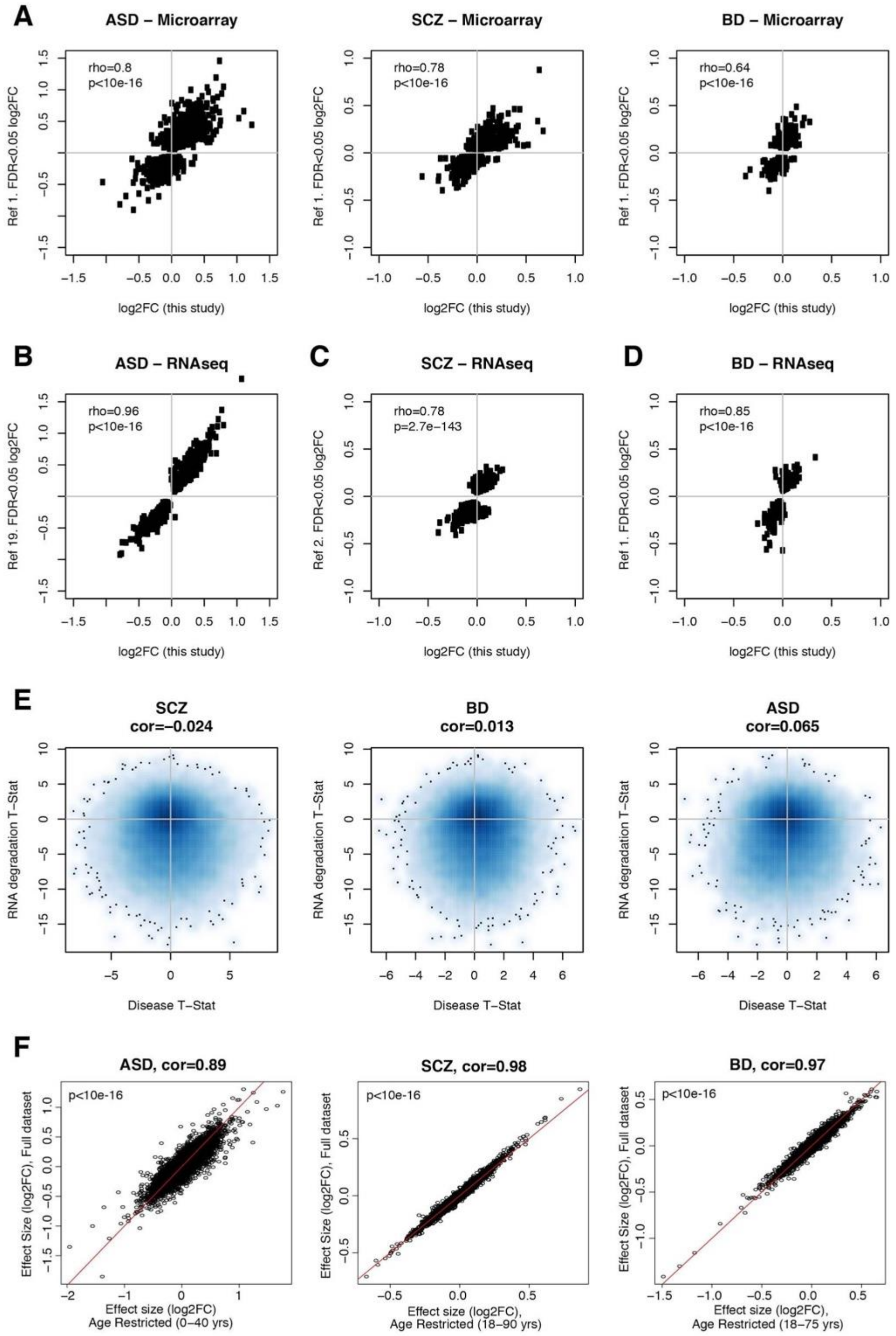
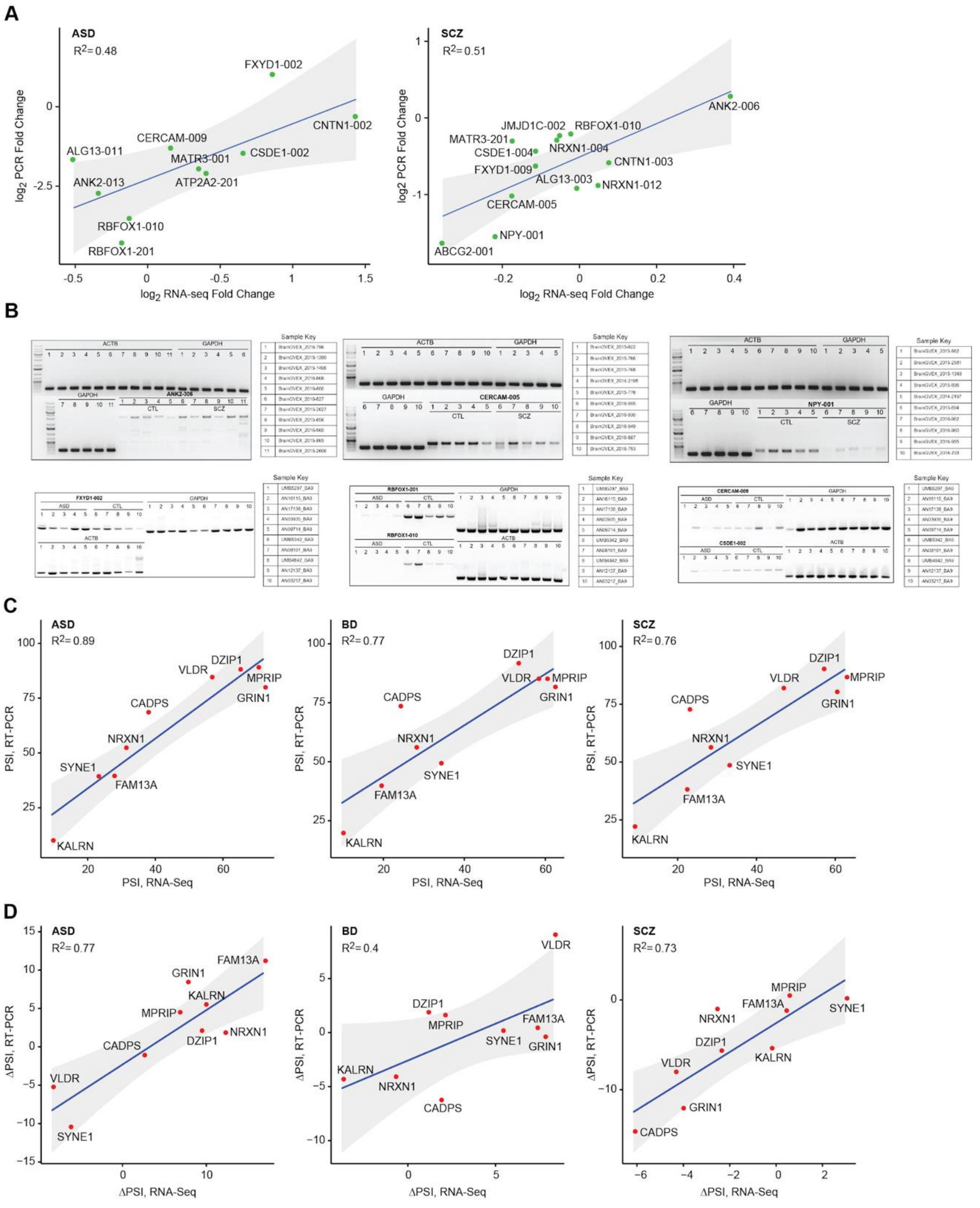


Fig. A3.4. Validation of DGE Results

Differential gene expression results from this study were compared with several published microarray and RNA-Seq datasets. A) Log_2FC effect sizes are plotted in comparison to a microarray meta-analysis of ASD, SCZ, and BD for genes identified as DGE (FDR<0.05) (*1*). We see substantial concordance of gene-level effect sizes across studies and platforms. Similar concordance is observed in comparison to results from RNA-Seq studies in B) ASD (*19*), C) SCZ (*2*), and D) BD (*1*). There is some overlap in samples across studies, due to the limited availability of post-mortem brain tissue from subjects with psychiatric disease. E) To ensure that differential gene expression in disease was not being driven by differences in RNA quality or degradation, we compared differential expression T-statistics with those experimentally derived from brain tissue samples allowed to degrade for fixed intervals of time (*22*). We did not observe substantial concordance between these RNA degradation metrics and psychiatric disease DGE summary statistics. F) Age balancing of case-control comparisons (0-40 years for ASD/CTL; 18-90 years for SCZ/CTL; 18-75 years for BD/CTL) does not substantially alter disease DGE signal.


A
ASD
R^2= 0.48
FXYD1-002
CNTN1-002
CERCAM-009
ALG13-011
MATR3-001
CSDE1-002
ANK2-013
ATP2A2-201
RBFOX1-010
RBFOX1-201
0
-2.5
-0.5
0
-0.5
1
1.5
$\log_2$ PCR Fold Change
$\log_2$ RNA-seq Fold Change
SCZ
R^2= 0.51
1
0
-1
ANK2-006
JMJD1C-002
RBFOX1-010
MATR3-201
CSDE1-004
NRXN1-004
CNTN1-003
FXYD1-009
ALG13-003
NRXN1-012
CERCAM-005
NPY-001
ABCG2-001
-0.2
0
0.2
0.4
$\log_2$ PCR Fold Change
$\log_2$ RNA-seq Fold Change
B
ACTB
GAPDH
ANK2-006
CTL
SCZ
1 2 3 4 5 6 7 8 9 10 11
Sample Key
1 BrainGVEX_2016-796
2 BrainGVEX_2015-1399
3 BrainGVEX_2015-1406
4 BrainGVEX_2016-868
5 BrainGVEX_2016-800
6 BrainGVEX_2016-827
7 BrainGVEX_2015-2627
8 BrainGVEX_2015-856
9 BrainGVEX_2016-960
10 BrainGVEX_2015-865
11 BrainGVEX_2015-2606
ACTB
GAPDH
CERCAM-005
CTL
SCZ
Sample Key
1 BrainGVEX_2015-922
2 BrainGVEX_2015-766
3 BrainGVEX_2015-768
4 BrainGVEX_2014-2195
5 BrainGVEX_2015-776
6 BrainGVEX_2016-955
7 BrainGVEX_2016-930
8 BrainGVEX_2016-949
9 BrainGVEX_2016-887
10 BrainGVEX_2016-793
ACTB
GAPDH
NPY-001
CTL
SCZ
Sample Key
1 BrainGVEX_2015-862
2 BrainGVEX_2015-2581
3 BrainGVEX_2015-1343
4 BrainGVEX_2015-896
5 BrainGVEX_2014-2197
6 BrainGVEX_2015-894
7 BrainGVEX_2016-962
8 BrainGVEX_2016-960
9 BrainGVEX_2016-955
10 BrainGVEX_2016-793
FXYD1-002
ASD
CTL
GAPDH
ACTB
Sample Key
1 UMB5297_BA9
2 AN16115_BA9
3 AN17138_BA9
4 AN03935_BA9
5 AN09714_BA9
6 UMB5342_BA9
7 AN08161_BA9
8 UMB4842_BA9
9 AN12137_BA9
10 AN03217_BA9
RBFOX1-201
RBFOX1-010
ASD
CTL
GAPDH
ACTB
Sample Key
1 UMB5297_BA9
2 AN16115_BA9
3 AN17138_BA9
4 AN03935_BA9
5 AN09714_BA9
6 UMB5342_BA9
7 AN08161_BA9
8 UMB4842_BA9
9 AN12137_BA9
10 AN03217_BA9
CERCAM-009
CSDE1-002
ASD
CTL
GAPDH
ACTB
Sample Key
1 UMB5297_BA9
2 AN16115_BA9
3 AN17138_BA9
4 AN03935_BA9
5 AN09714_BA9
6 UMB5342_BA9
7 AN08161_BA9
8 UMB4842_BA9
9 AN12137_BA9
10 AN03217_BA9
C
ASD
R^2= 0.89
DZIP1
VLDR
MPRIP
GRIN1
CADPS
NRXN1
SYNE1
FAM13A
KALRN
100
75
50
25
20
40
60
PSI, RT-PCR
PSI, RNA-Seq
BD
R^2= 0.77
DZIP1
VLDR
MPRIP
GRIN1
CADPS
NRXN1
SYNE1
FAM13A
KALRN
PSI, RT-PCR
PSI, RNA-Seq
SCZ
R^2= 0.76
DZIP1
VLDR
MPRIP
GRIN1
CADPS
NRXN1
SYNE1
FAM13A
KALRN
PSI, RT-PCR
PSI, RNA-Seq
D
ASD
R^2= 0.77
FAM13A
GRIN1
KALRN
MPRIP
CADPS
DZIP1
NRXN1
VLDR
SYNE1
15
10
5
0
−5
−10
−15
0
10
ΔPSI, RT-PCR
ΔPSI, RNA-Seq
BD
R^2= 0.4
VLDR
DZIP1
MPRIP
FAM13A
SYNE1
GRIN1
KALRN
NRXN1
CADPS
5
0
−5
−10
0
5
ΔPSI, RT-PCR
ΔPSI, RNA-Seq
SCZ
R^2= 0.73
MPRIP
FAM13A
SYNE1
NRXN1
DZIP1
KALRN
VLDR
GRIN1
CADPS
0
−10
−6
−4
−2
0
2
ΔPSI, RT-PCR
ΔPSI, RNA-Seq

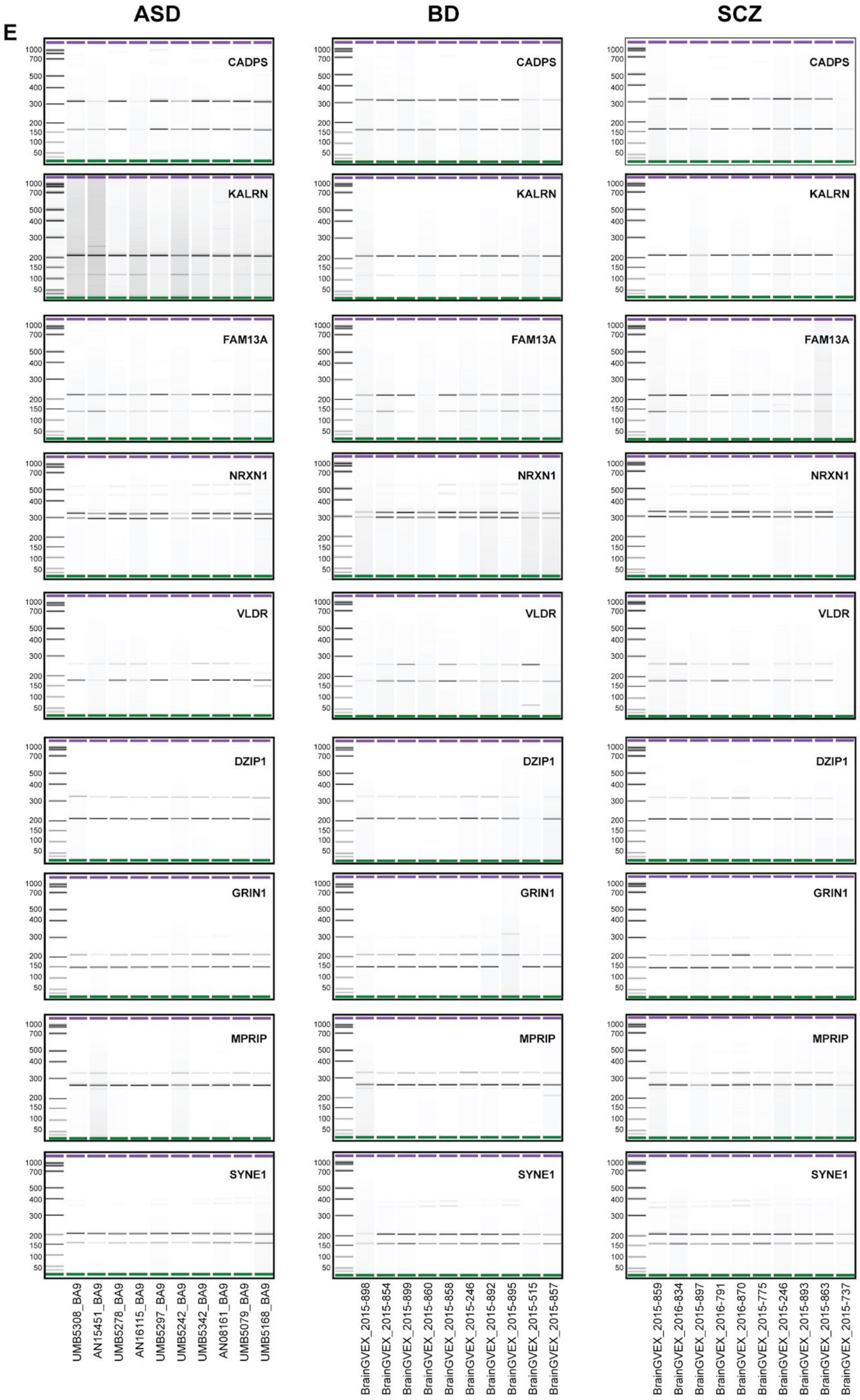

E
ASD
BD
SCZ
CADPS
KALRN
FAM13A
NRXN1
VLDR
DZIP1
GRIN1
MPRIP
SYNE1
1000
700
500
400
300
200
150
100
50
UMB5308_BA9
AN15451_BA9
UMB5278_BA9
AN16115_BA9
UMB5297_BA9
UMB5242_BA9
UMB5342_BA9
AN08161_BA9
UMB5079_BA9
UMB5168_BA9
BrainGVEX_2015-898
BrainGVEX_2015-854
BrainGVEX_2015-899
BrainGVEX_2015-860
BrainGVEX_2015-858
BrainGVEX_2015-246
BrainGVEX_2015-892
BrainGVEX_2015-895
BrainGVEX_2015-515
BrainGVEX_2015-857
BrainGVEX_2015-859
BrainGVEX_2016-834
BrainGVEX_2015-897
BrainGVEX_2016-791
BrainGVEX_2016-870
BrainGVEX_2015-775
BrainGVEX_2015-246
BrainGVEX_2015-893
BrainGVEX_2015-863
BrainGVEX_2015-737

Fig. A3.5. Validation of differential transcript expression and differential splicing

A) Comparison of fold changes obtained from RSEM-based isoform quantification of RNA-Seq data (*21*) to semiquantitative PCR results for 10 isoforms tested in ASD and control samples (left) and 13 isoforms tested in SCZ and control samples (right). Fold changes were calculated between cases and control samples. B) Representative 1.5 to 2% agarose gel images obtained for isoform validation. C) Scatter plots comparing the average percent spliced-in (PSI) of exon-skipping events called by LeafCutter from RNA-Seq data (*21*) to semi-quantitative PCR. A total of 9 genes were tested in 5 cases and 5 controls in ASD, BD and SCZ. An additional 5 cases and 5 controls were tested for *FAM13A* and *SYNE1* in BD and SCZ to resolve outliers. D) Same as C, but now comparing the change in average PSI (ΔPSI) between cases and controls in each disorder. E) Representative Agilent 2100 Bioanalyzer gel images (DNA 1000 chips) obtained for splicing validation. A-E) Gene or isoform names are indicated at each point. Regression lines with 95% confidence intervals are shown in blue and grey, respectively and the corresponding R^2 values are shown at the top-left in each plot.

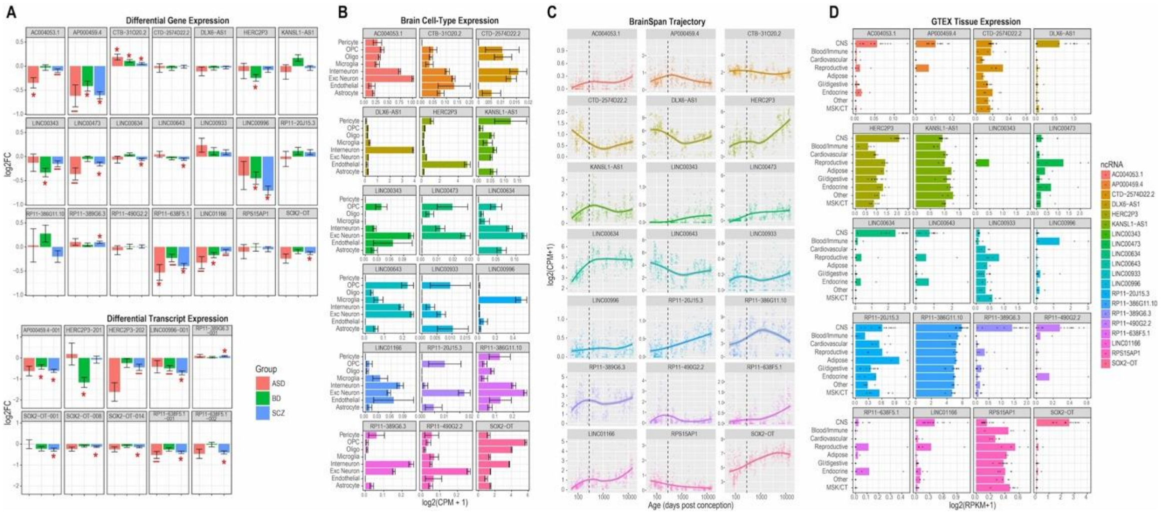


Fig. A3.6. Annotation of individual ncRNAs

We highlight several individual ncRNAs differentially expressed in psychiatric disease or identified as hubs of relevant co-expression modules. A) We show differential gene expression (DGE; top) and differential transcript expression (DTE; bottom) in SCZ, BD, and ASD. *FDR<0.05, -- FDR<0.1. B) Human brain cell type expression patterns are shown for each ncRNA using data from Ref (*91*). Plots show mean expression for cells identified in specific clusters. C) Developmental expression trajectory is shown for each ncRNA using data from BrainSpan (*144*). Plots show expression as a function of age (days post-conception) on a $\log_{10}$ scale, with the dotted line denoting birth. D) Human tissue-specific expression levels are shown for each ncRNA using data from GTEX (*81*).

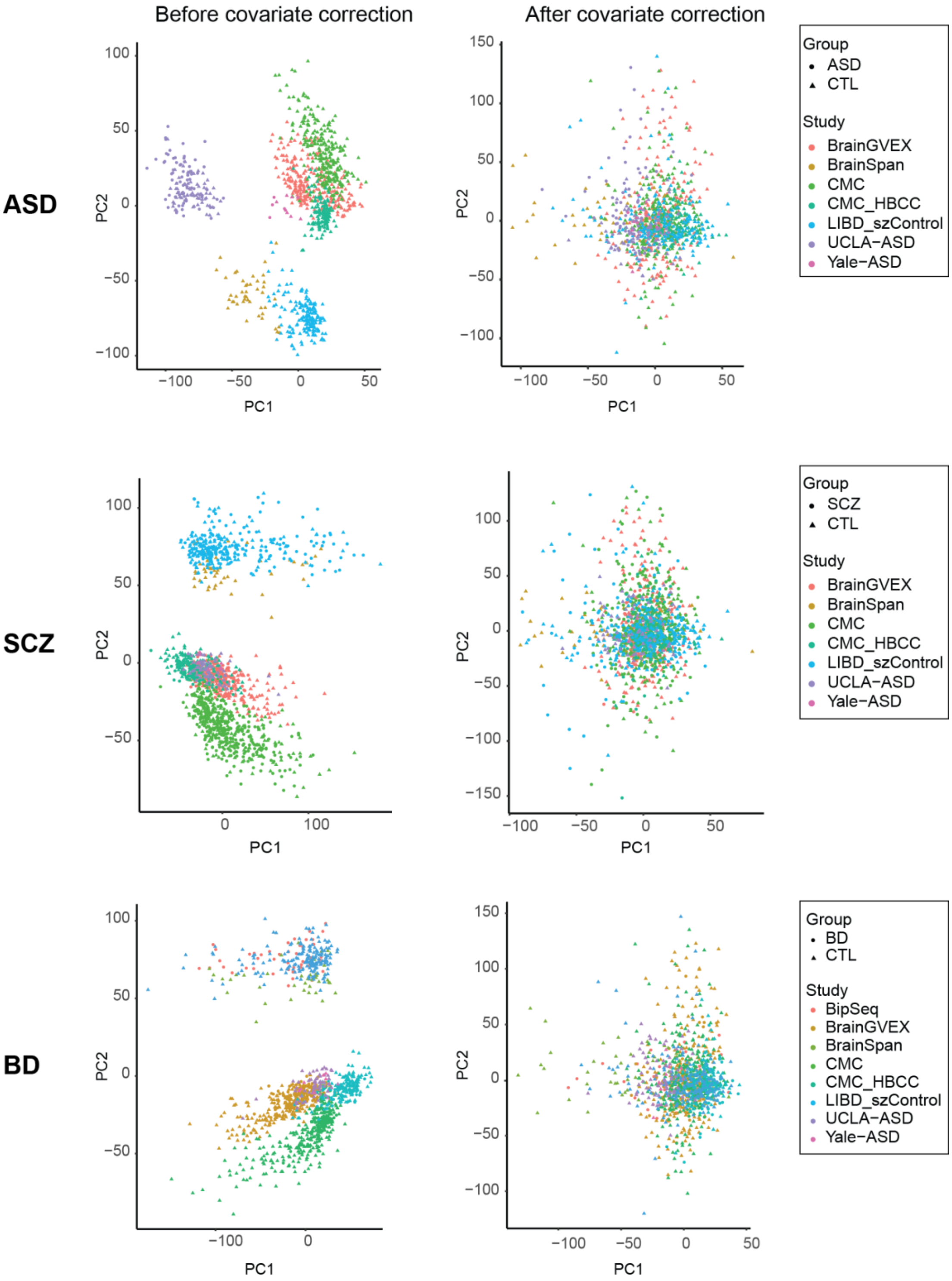


Fig. A3.7. Covariate correction of DS

Top two principal components (PCs) of percent spliced-in (PSI) values are shown before (left) and after (right) covariate correction for the ASD, BD and SCZ datasets. Points are colored according to study origin and shape denotes disorder (circle) or control (CTL) status (triangle). See inset legends for further details.

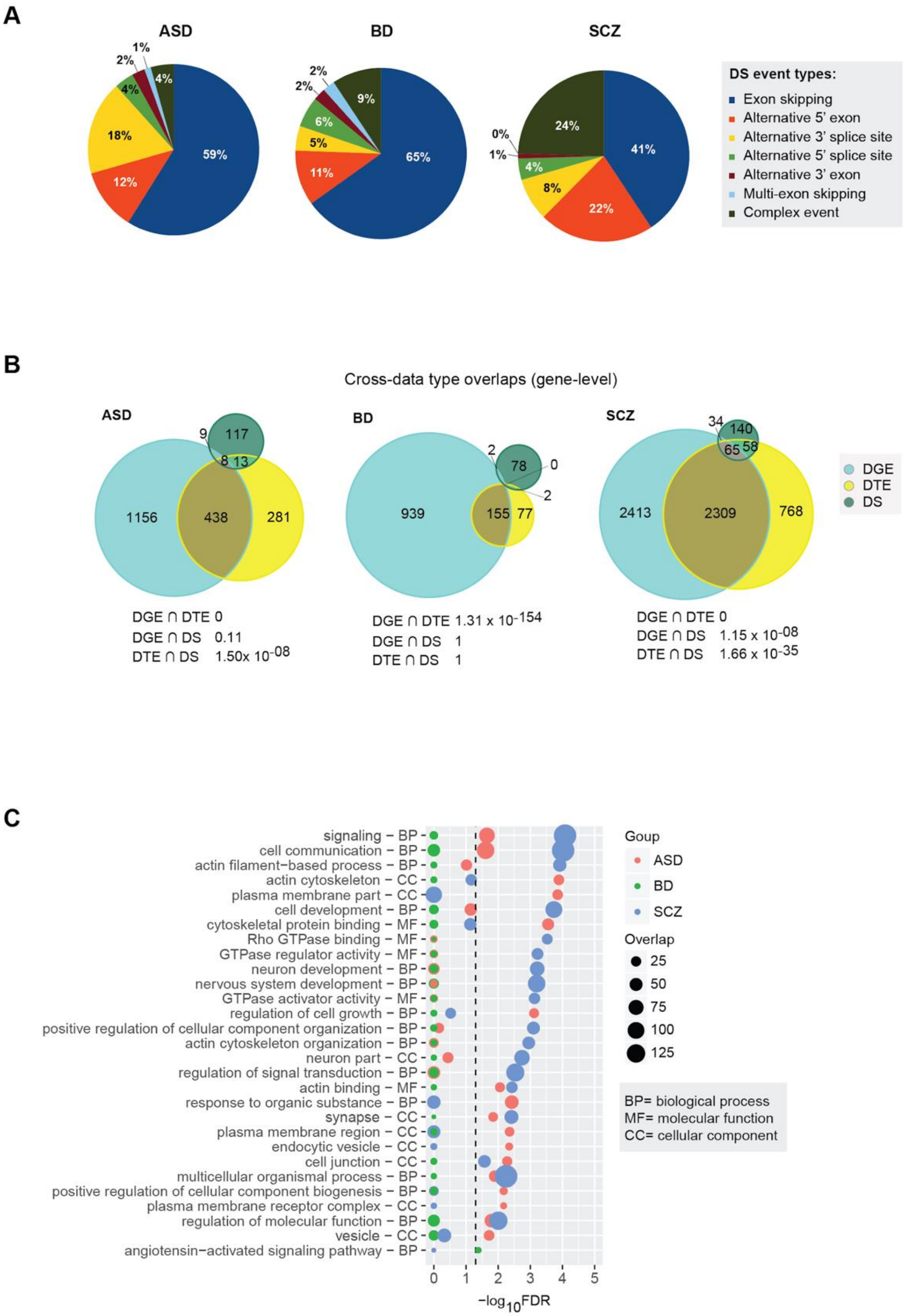


Fig. A3.8. Annotation of DS events, Cross data DGE-DTE-DS overlaps

A) Pie charts with breakdown of DS event types identified in each disorder. B) Venn diagrams showing overlaps between genes with significant DGE, DTE or DS changes for each disorder. P values for hypergeometric tests of pairwise overlaps between data types are shown at the bottom of the venn diagrams for each disorder. C) Top 20 gene ontology (GO) enrichments for DS genes in each disorder.

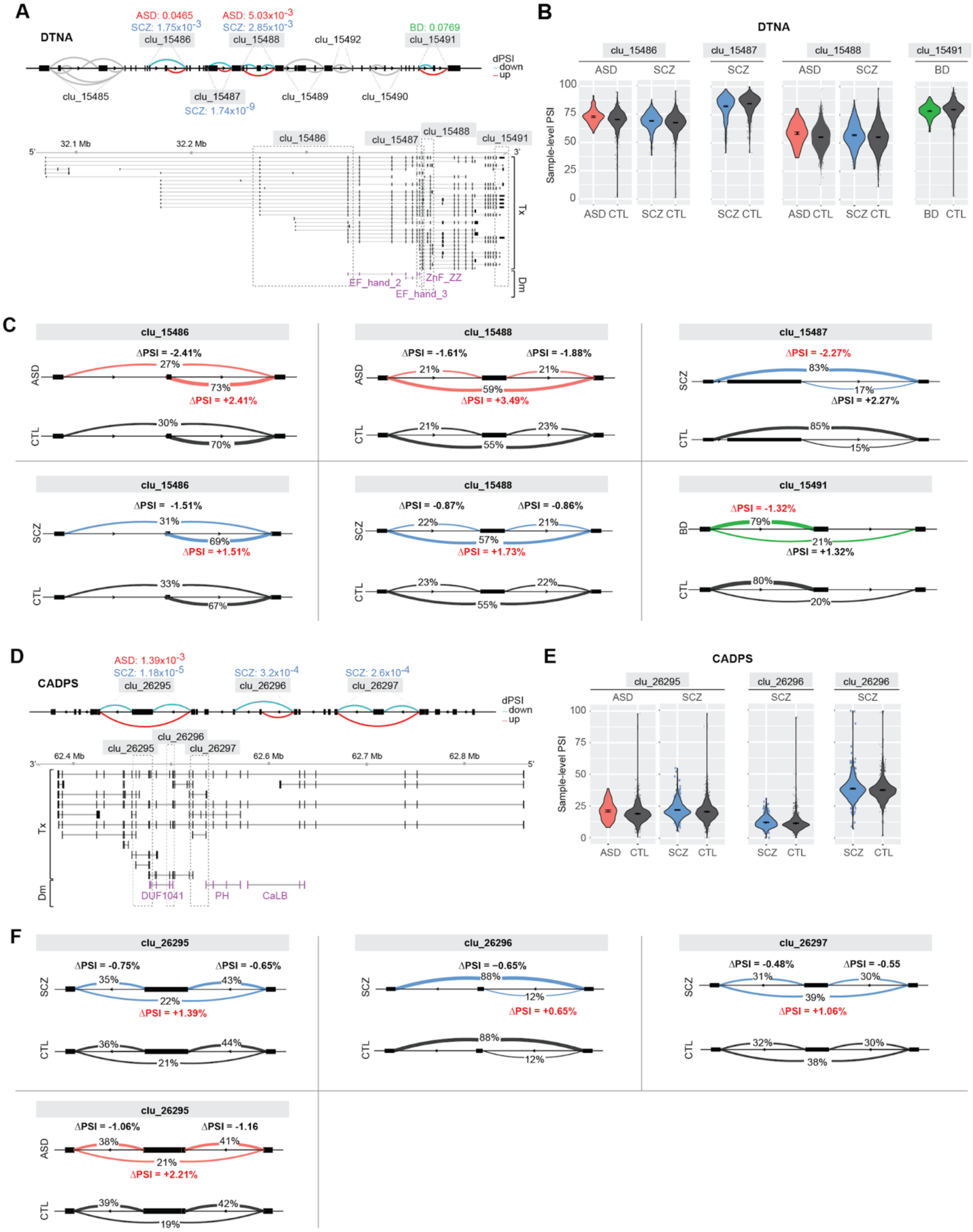


Fig. A3.9. Additional differential local splicing examples

A) Top: Significant differentially spliced (DS) intron clusters in *DTNA* for ASD, SCZ and BD. Increased or decreased intron usage in cases compared to controls (CTL) are shown in red and blue, respectively. Bottom: Overview of known isoforms (GENCODE v19) and protein domains for *DTNA*. Locations of significant DS clusters are indicated by dotted lines. Protein domains (purple) are annotationed as

EF_hand_2 - EF hand domain 2; EF_hand_3 - EF hand domain 3; ZnF_ZZ - Zinc-binding domain, present in Dystrophin, CREB-binding protein. B) Violin-plots with the distribution of covariate-adjusted percent spliced in (PSI) *per* sample for the intron with the maximum change in PSI for each cluster and disorder. C) Visualization of introns in each significant cluster for each disorder, with their change in PSI (PSI). Covariate-adjusted average PSI levels in disorder *vs* CTL are indicated for each intron. D) Same as A), but for *CADPS*. Protein domains (purple) are annotated as PH - Pleckstrin homology domain; CaLB - C2 domain (Calcium/lipid-binding domain, CaLB) superfamily; DUF1041 - Domain of unknown function. E) Same as B), but for *CADPS*. F) Same as C, but for *CADPS*.

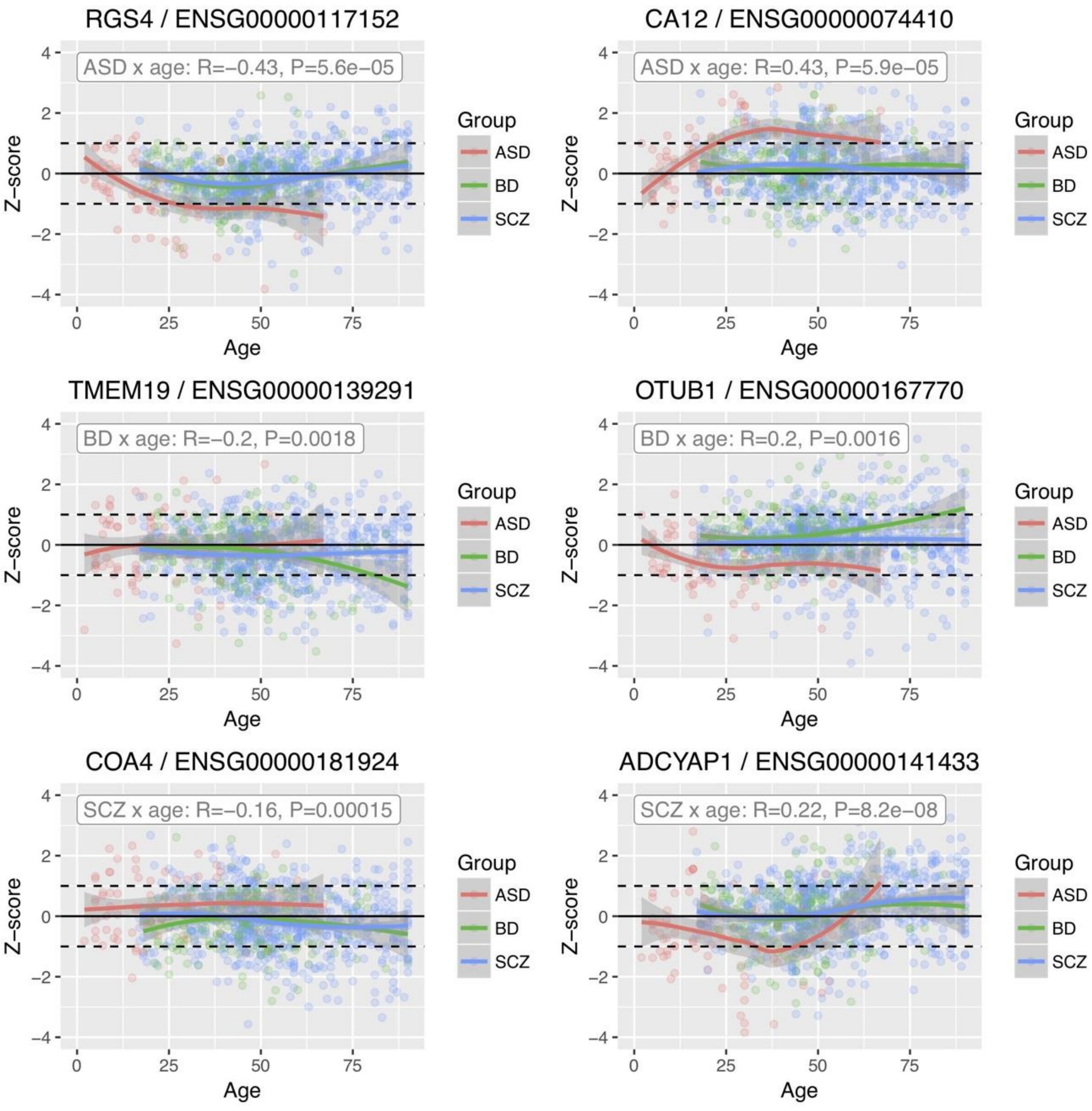


Fig. A3.10. Age effects on differential gene expression

Examples are shown for genes whose magnitude of differential expression in a given disorder is significantly associated with age. For each gene, a local regression function was fit to model the effect of age on expression in control samples, and expression in cases was then converted to a z-score relative to the local mean in controls. We then assessed the correlation between z-transformed expression and age within each disease group (ASD, SCZ, BD) separately, to identify those genes whose magnitude of differential expression was associated with age. We find that 143 of the 4821 DGE genes in SCZ show a nominal increase in effect size magnitude as a function of age, consistent with a reactive interpretation. In ASD, 85 of 1611 DE genes showed this same pattern and in BD there were 29 of the 1119 DE genes.

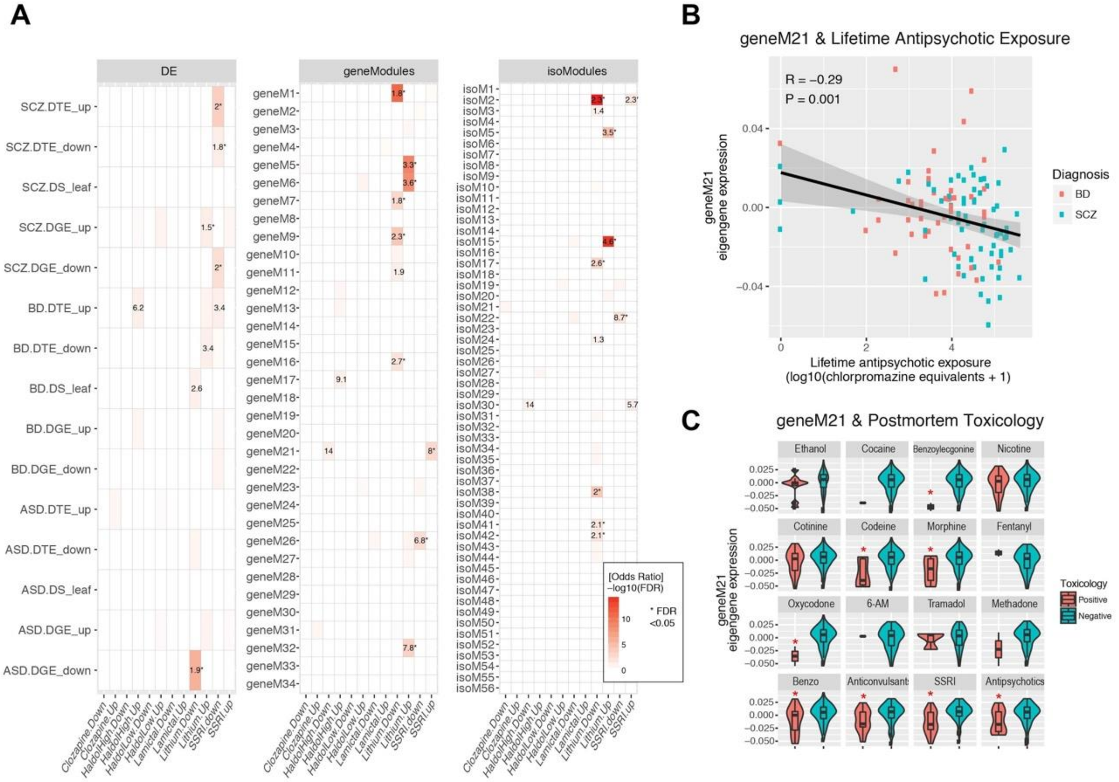


Fig. A3.11. Assessment of psychiatric medication effects

A) We investigated whether antipsychotic medications could explain differential gene expression and module associations identified in SCZ, BD, and ASD. We used three experimental datasets: (1) an RNA-Seq dataset from DLPFC of nonhuman primates exposed for 6 months to clozapine, haloperidol (low dose), or haloperidol (high dose) compared to placebo; (2) a microarray dataset from mouse brain following chronic exposure to the SSRI fluoxetine; (3) a microarray dataset from rat cortex following chronic exposure to the mood stabilizers lithium or lamotrigine compared with vehicle (*21*). Overlap of DE genes and modules with genes up or downregulated by medications (at nominal significance thresholds, except for lithium) was assessed by Fisher's exact test. Plot shows odds-ratios of enrichment for P<0.05 significant associations, with * denoting FDR<0.05 associations. With the exception of lithium, medications show minimal overlap with disease-associated transcriptomic changes. The one exception was for the activity dependent module pair, geneM21/isoM30, which did seem to be associated with SSRIs and high dose haloperidol. B) To investigate this relationship further, we compared geneM21 eigengene expression with medication history in those subjects where this information was available. There was a significant negative correlation between geneM21 expression and lifetime antipsychotic exposure (chlorpromazine equivalents, log scale). C) A subset of samples also had results from post-mortem toxicology testing. We found broadly decreased levels of geneM21 eigengene expression in those subjects who tested positive for a host of psychiatric medications, including antipsychotics (*FDR<0.05).

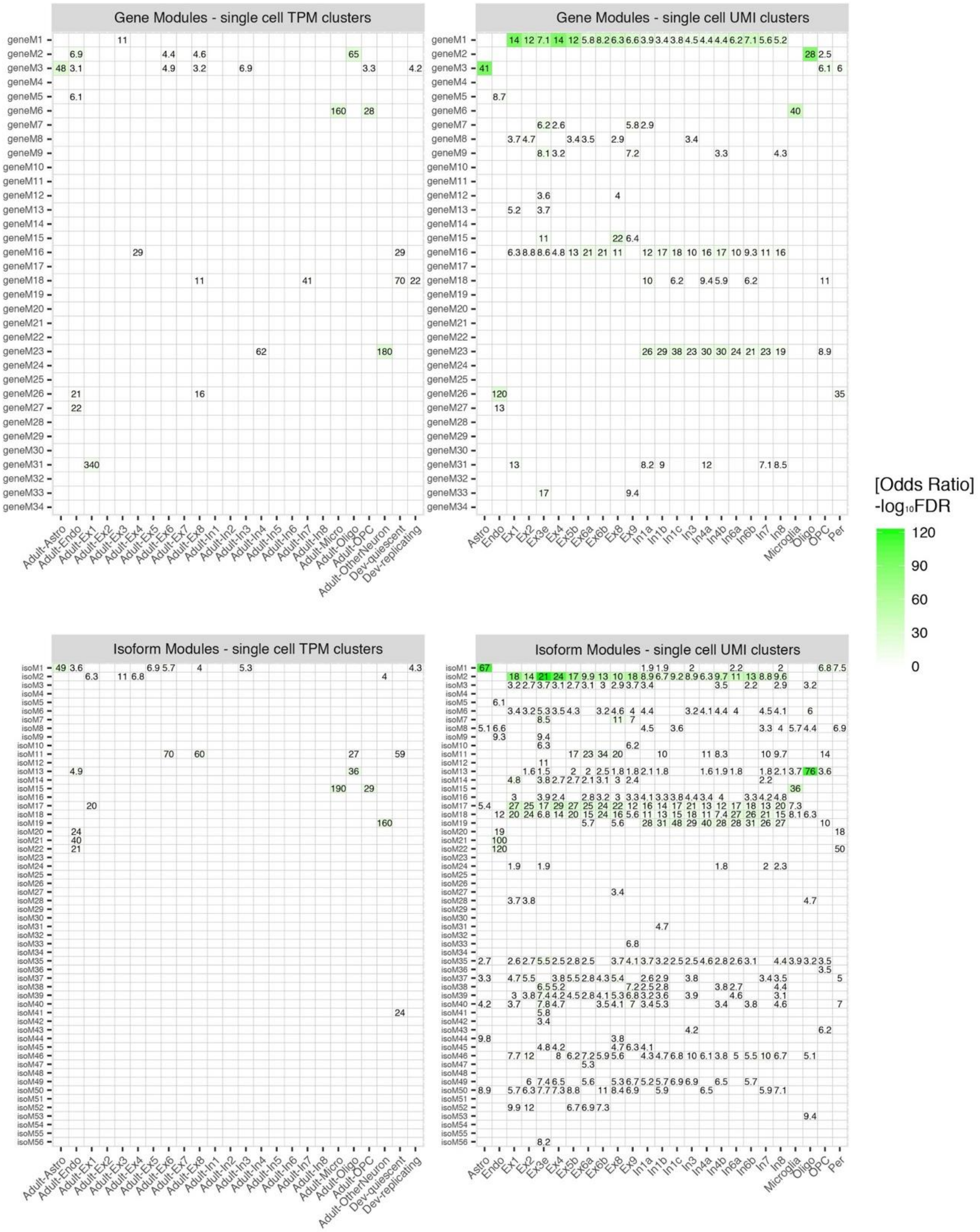


Fig. A3.12. Co-expression network cell type enrichments

Plots show enrichment of gene and isoform-level co-expression modules for established markers of CNS cell types from human brain single-cell RNA-Seq clusters, as compiled in the companion manuscript (*18*). Clusters were defined separately for TPM- and UMI- based scRNA-Seq quantifications. Text denotes odds ratios of enrichment for significant associations (FDR<0.05). The UMI dataset is from adult human brain, whereas the TPM dataset includes two fetal cell types. (Ex# - excitatory neuron cluster; In# - interneuron cluster; Per- pericyte; OPC-oligodendrocyte progenitor cell; Oligo - oligodendrocyte; Micro - microglia; End - endothelial; Ast - astrocyte).

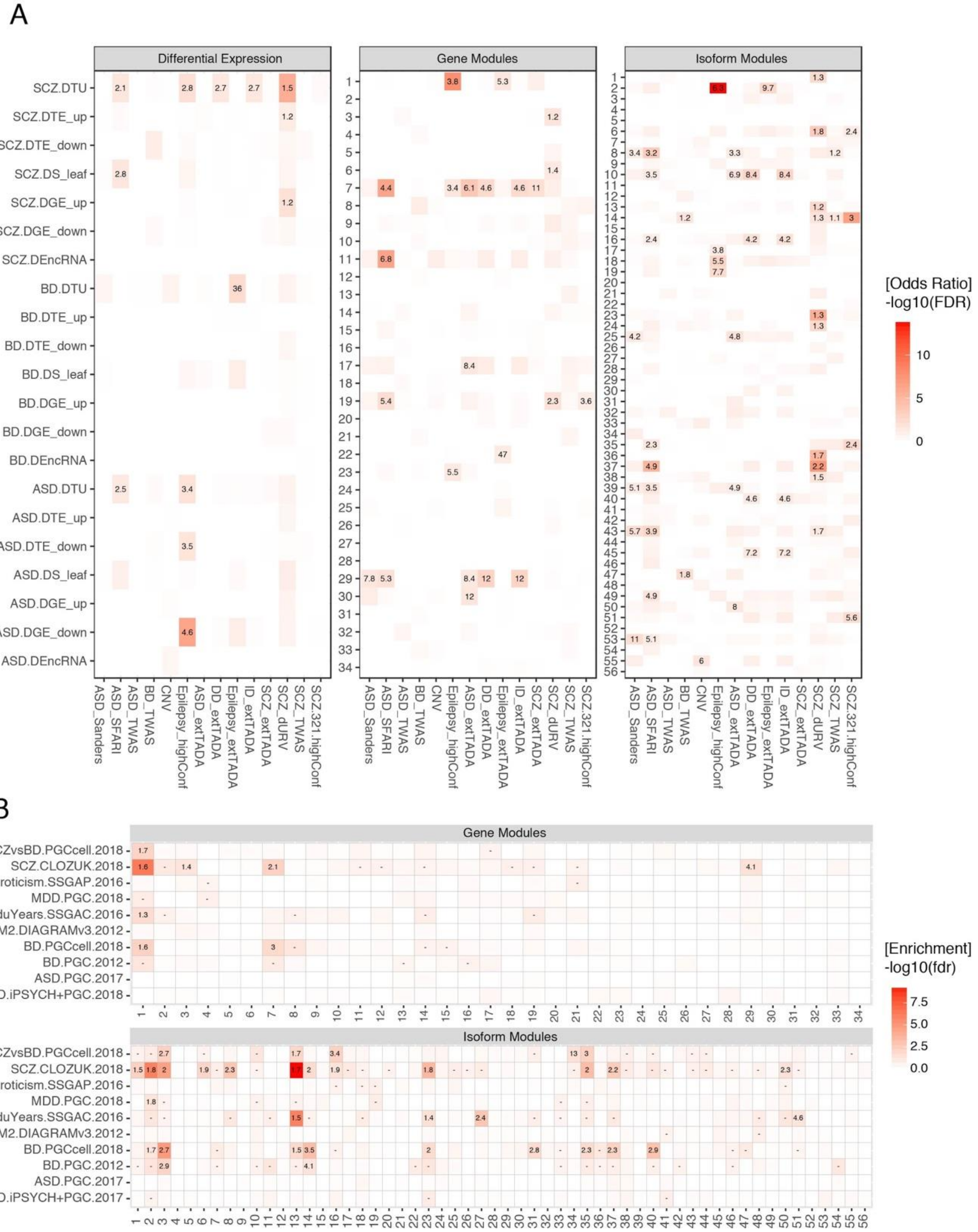


Fig. A3.13. Genetic enrichment analyses

A) Enrichment for several sets of disease risk genes was assessed among DE features, gene, and isoform-coexpression modules, including those harboring rare *de novo* variants identified in each disorder, as well as in related neurodevelopmental and psychiatric traits. TWAS signal for each disorder was also included as was the list of 321 "high confidence" SCZ risk genes identified in the companion manuscript (*18*). Enrichment was calculated using logistic regression, controlling for gene and transcript length as well as GC content (*21*). Risk gene sets include: 71 risk loci harboring rare *de novo* variants associated with ASD through the transmission and *de novo* association test (TADA; "ASD_Sanders") (*80*); Syndromic and highly ranked (1 and 2) genes from SFARI Gene database ("ASD_SFARI"); genes harboring recurrent *de novo* copy-number variants associated with ASD or SCZ, as defined in (*1*) ("CNV"); genes harboring an excess of rare exonic variants in ASD, SCZ, intellectual disability (ID), developmental delay (DD), and epilepsy as assessment through an extended version of TADA ("extTADA") (*140*); genes harboring disruptive and damaging ultra-rare variants in SCZ (*54*) ("SCZ_dURVs"); a list of high confidence epilepsy risk genes, compiled from (*141*). B) Enrichment of GWAS signal among gene and isoform co-expression modules, using stratified LD score regression (s-LDSR) with summary statistics from several psychiatric, cognitive, and behavioral traits (*21*). Cells are labeled with GWAS enrichment, for those with FDR < 0.05. Cells labeled with "-" are nominally (P<0.05) significant but do not pass FDR-correction.

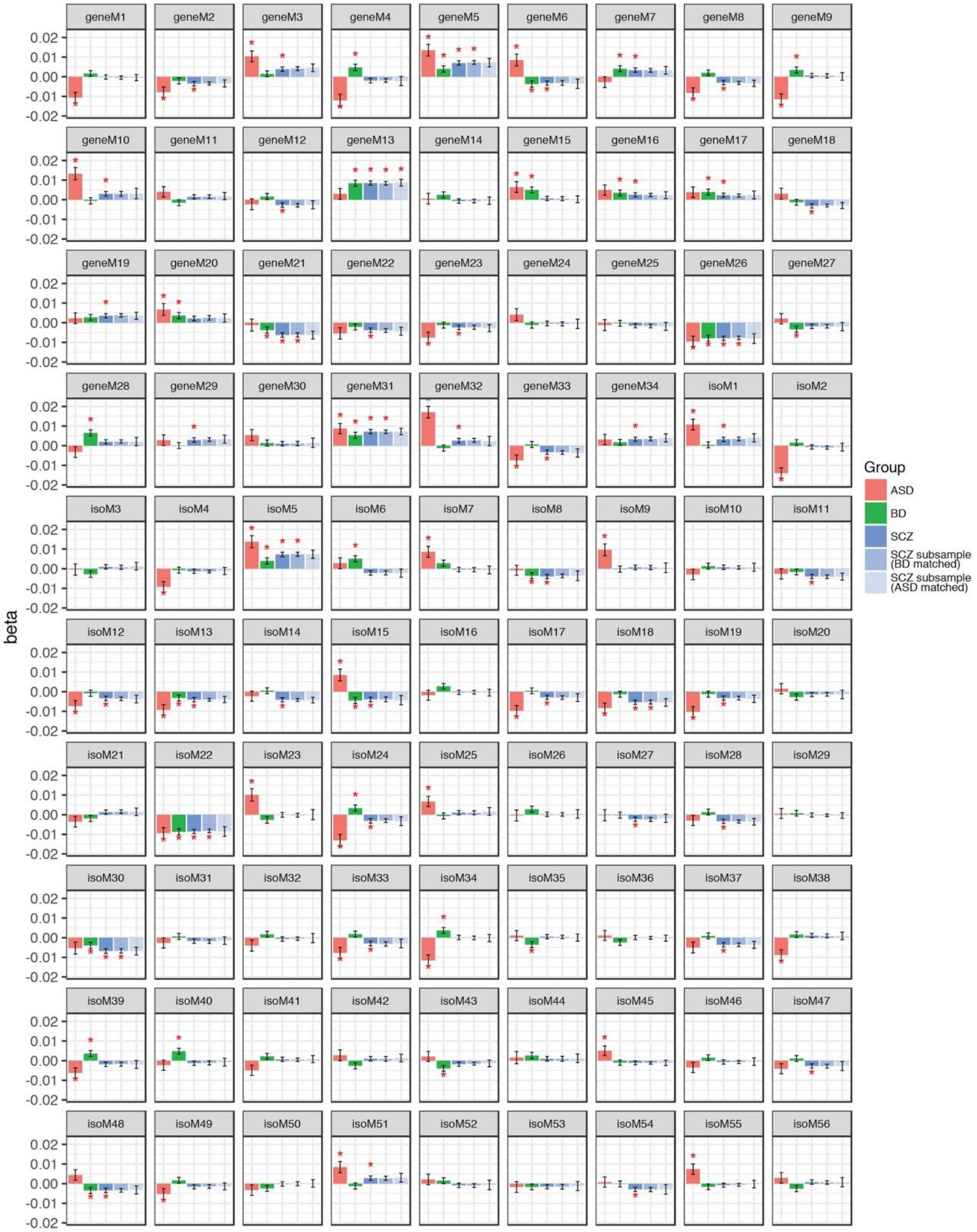


Fig. A3.14. Module-trait associations after SCZ downsampling

To determine whether differences in module associations observed across disorders was due to the larger sample size of the SCZ dataset, we repeated our module-trait association analyses using a randomly subsampled SCZ dataset to match the sample size of ASD and BD datasets. We repeated this 100 times and reran our module-level associations using these matched sample sizes. Plots show module-trait association β values with standard errors. *P<0.05.

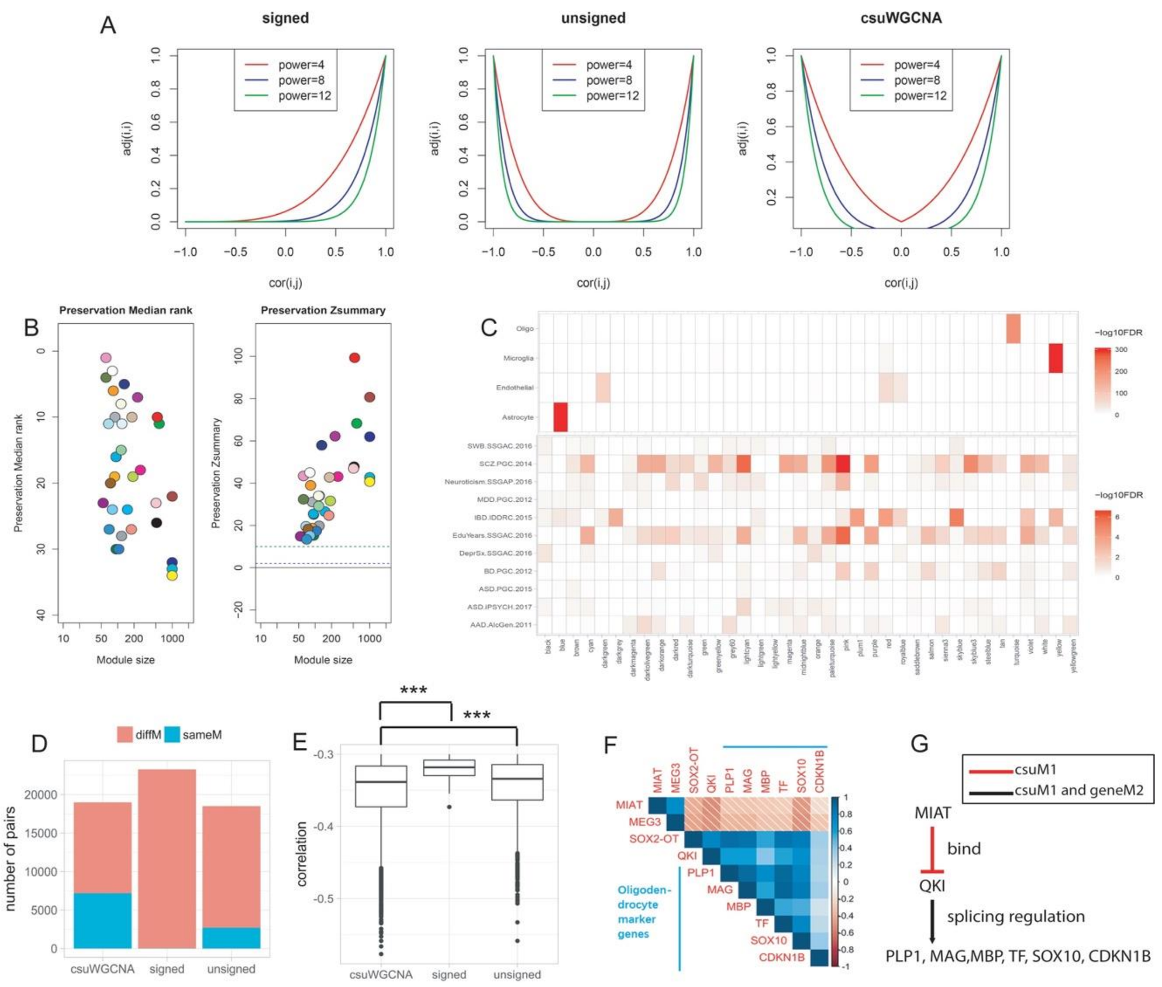


Fig. A3.15. csuWGCNA identifies putative lncRNA negative regulatory relationships

A) Network adjacency (y-axis) versus correlation (x-axis) in the signed network, unsigned network, and csuWGCNA network. The color of the line denotes the soft threshold power setting. Note that correlation=-1 leads to adjacency = 0 in the signed network and adjacency =1 in the unsigned and csuWGCNA network. B) All modules detected by csuWGCNA were well preserved in the signed networks (Zsummary > 10 indicates high preservation). C) The enrichment of cell type and GWAS signal in csuWGCNA modules. D) csuWGCNA captures more negative lncRNA-gene pairs (cor<-0.3) in the same module than signed and unsigned WGCNA (csuWGCNA=7186, signed=20, unsigned=2701). E) csuWGCNA captures stronger negative relationships than signed and unsigned network types (Welch two sample t-test, $p<10^{-6}$ and $p<10^{-11}$, respectively). F) The lncRNAs *MIAT* and *MEG3* are negatively correlated with most of the hubs in oligodendrocyte modules, including *SOX2-OT* and oligodendrocyte marker genes (*PLP1*, *MAG*, *MBP*, *TF*, *SOX10*, and *CDKN1B*). The blue color indicates negative correlations and the red indicates positive correlations. G) Putative target relationships for the lncRNA *MIAT*. The red line indicates a negative relationship only detected in csuM1, and the black line indicates positive relationships detected in both csuM1 and geneM2.

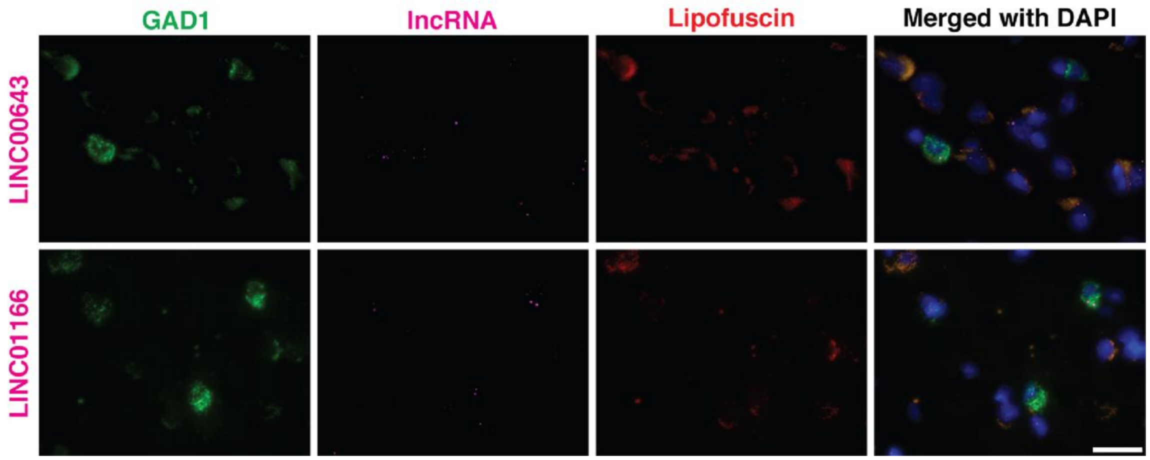


Fig. A3.16. *LINC00643* and *LINC01166* expression in human prefrontal cortex

Sections from human prefrontal cortex (area 9) were labeled with *GAD1* probe (green) and lncRNA (magenta) probes for *LINC00643* (upper panel) or for *LINC01166* (lower panel). All sections were counterstained with DAPI (blue) to reveal cell nuclei. Lipofuscin autofluorescence is visible in both the green and red channels and appears yellow/orange in the merged image. The lncRNAs are present both in GABAergic interneurons and cells without *GAD1* signal. Scale bar, 25 μm.

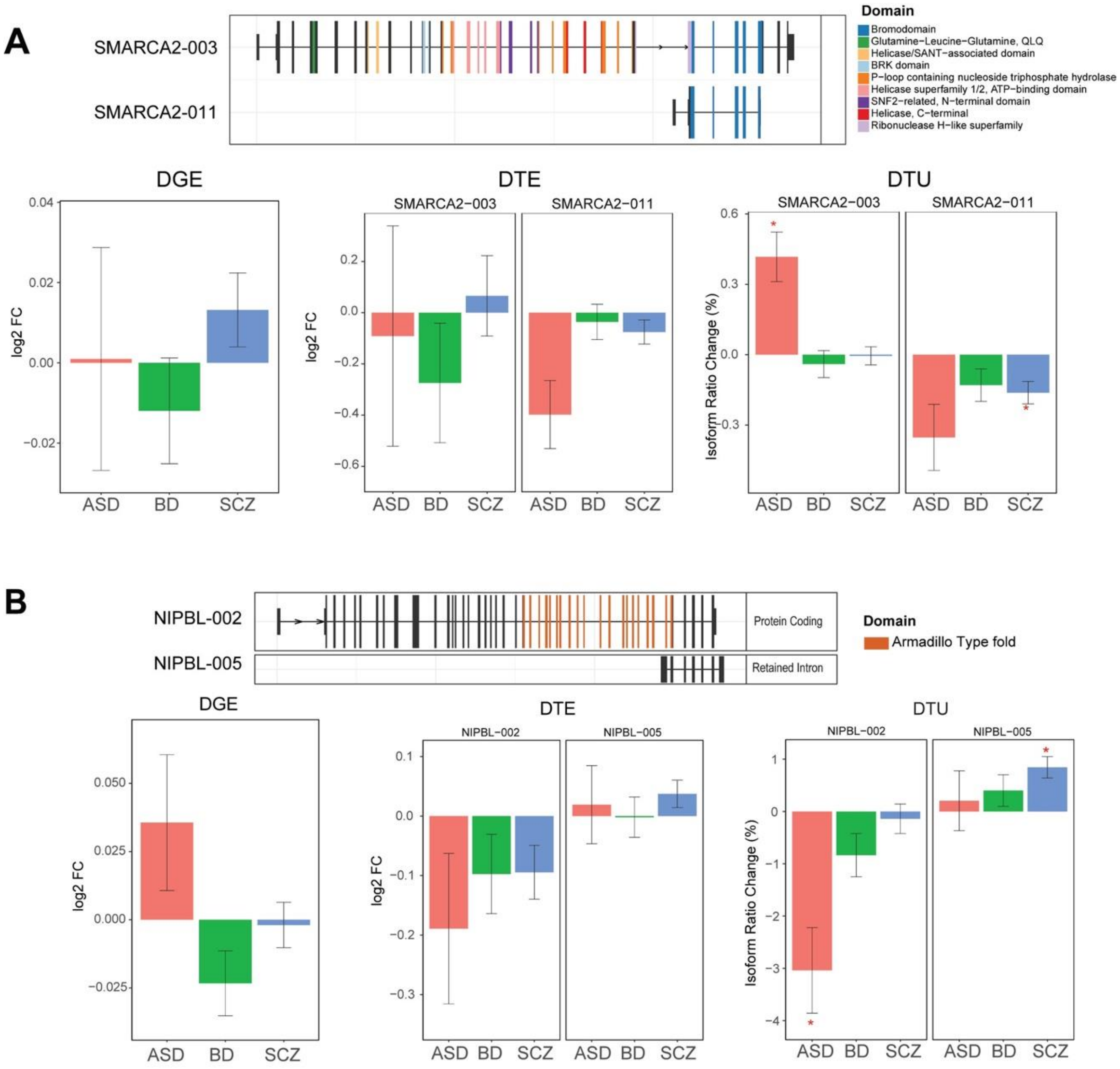


Fig. A3.17. Additional switch isoforms

A) The isoform ratio of two *SMARCA2* isoforms, *SMARCA2-003* and *SMARCA2-011*, are up and downregulated in ASD and SCZ, respectively. B) The isoform ratio of two *NIPBL* isoforms, *NIPBL-002* and *NIPBL-005*, are down and upregulated in ASD and SCZ, respectively. *FDR < 0.05

A3.3: Extended Tables

Please see the electronic tables associated with this document for additional tables (Tables A2.1-9). Descriptions for these tables follow:

Table A3.1 (separate file)

Differential gene and isoform expression summary statistics and DE enrichment analyses

Table A3.2 (separate file)

Annotation of psychiatric ncRNAs

Table A3.3 (separate file)

Differential splicing summary statistics, annotation and disease overlaps

Table A3.4 (separate file)

TWAS summary statistics and PRS associations with gene and isoform expression

Table A3.5 (separate file)

Gene and isoform co-expression module annotation

Table A3.6 (separate file)

csuWGCNA network annotation and putative lncRNA-mRNA targets

Table A3.7 (separate file)

Switch isoform and microexon characterization

Table A3.8 (separate file)

Splicing and isoform validation primers and samples

Table A3.9 (separate file)

RNAscope - Tissue samples and RNA FISH probes

A3.4: Extended Bibliography

The bibliography/references for the appendix sections A2 and A3 correspond with the bibliography for Chapter 3 (Section 3.4).

A4: Supplementary Materials for Chapter 4

A4.1: Extended Materials and Methods

Linear model design

To select the biological and technical covariates to use in downstream linear mixed-effects models, the EARTH[39] package in R was used. This package applies the Multivariate Adaptive Regression Splines (MARS) technique to build regression models. The covariates assessed were subject, region, brain bank, diagnosis, sex, age, PMI, sequencing batch, ancestry genotype, and RIN, as well as STAR and Picard Tools RNA-seq quality measures (all listed in Table A4.1). For 4 subjects with no recorded PMI, the average of the rest of the subjects' PMI was used. Before input into the EARTH algorithm, STAR and Picard Tools quality measures were filtered such that collinearity with any other biological or technical covariate was eliminated (only one covariate was kept for every identified collinear pair, with collinearity defined as an adjusted $R^2 > 0.95$ between the two covariates). All continuous covariates were centered and scaled for input into the EARTH algorithm and for remaining analyses. A cross-validated approach was used to run EARTH: it was run 10 times with 90% of samples, and then the resulting linear model was tested with the remaining 10% of samples. The median R^2 across all genes/isoforms was used to assess the performance of each cross-validated EARTH model. Using this metric, the following covariates from the highest performing EARTH model were selected for the gene and isoform linear mixed models used in subsequent transcriptomic analyses:

Gene Model: subject, diagnosis, region, sequencing batch, sex, ancestry, age, age^2, PMI, RIN, picard_gcbias.AT_DROPOUT, star.deletion_length, picard_rnaseq.PCT_INTERGENIC_BASES, picard_insert.MEDIAN_INSERT_SIZE, picard_alignment.PCT_CHIMERAS, picard_alignment.PCT_PF_READS_ALIGNED, star.multimapped_percent, picard_rnaseq.MEDIAN_5PRIME_BIAS, star.unmapped_other_percent, picard_rnaseq.PCT_USABLE_BASES, picard_alignment.PCT_CHIMERAS^2, star.uniquely_mapped_percent^2.

Isoform Model: subject, diagnosis, region, sequencing batch, sex, ancestry, age, age^2, PMI, RIN, picard_rnaseq.PCT_MRNA_BASES, picard_gcbias.AT_DROPOUT, picard_rnaseq.PCT_UTR_BASES, star.multimapped_toomany_percent, picard_rnaseq.MEDIAN_CV_COVERAGE, picard_insert.MEDIAN_INSERT_SIZE, picard_rnaseq.PCT_INTERGENIC_BASES, picard_rnaseq.PF_BASES.

For both models, ‘subject’ was input as a random effects term (specifically, a random intercept), and diagnosis and region were combined to create one ‘diagnosis x region’ term (eg. ASD_BA17, ASD_BA9, Control_BA17, Control_BA9, etc.). This was done to facilitate region-specific contrasts in downstream analyses. The rest of the covariates were input as fixed effects into the linear mixed models. The ‘variancePartition’[40] R library was used to visualize the percent of variance explained by each model covariate across all genes/isoforms.

Comparing region-specific ASD effects to whole cortex ASD effects

To test if region-specific ASD dysregulation was significantly greater in magnitude than the whole cortex dysregulation in the Parikshak et al.[5] modules, a permutation approach was utilized. The region-specific ASD signed -$\log_{10}$(p-value) of each module was tested against a permuted distribution (10,000 permutations) of this statistic generated from randomly assigning cortical regions to samples. Regions were randomly assigned within subjects so that regional sample size was consistent for every permutation and subject variability was controlled. A region was considered significantly more dysregulated than the whole-cortex if the one-tailed p-value derived from comparing the true region-specific ASD signed -$\log_{10}$(p-value) to the permuted distribution was less than 0.05. The same approach was implemented to test if the number of region-specific DE ASD genes was significantly greater than the number of whole cortex DE ASD genes, with the number of region-specific DE ASD genes replacing the region-specific ASD signed -$\log_{10}$(p-value) as the statistic of interest.

To compare region-specific ASD gene dysregulation effect sizes to the whole cortex ASD effect, we calculated the principal components regression slope comparing the whole cortex ASD $\log_2$ Fold Change (FC/effect) to the region-specific ASD $\log_2$ FC for the 4,223 genes identified as DE in ASD across the whole cortex. We then generated a bootstrapped distribution (1,000 bootstraps) for each of the 11 region-specific slopes (sampling with replacement from the region of interest for each 'diagnosis x region' group) to calculate a 95% confidence interval for these slopes. Sample size was kept consistent for each bootstrap with the number of samples from each 'diagnosis x region' group.

ARI gene group formation and functional annotation

To evaluate the ARI genes across the whole-cortex, instead of only in the regional pairs in which they were identified, the ARI genes from regional pairs containing either BA17 or BA39-40 were assembled into two groups: the union (without duplicates) of ARI genes with higher Control expression in BA39-40 and BA17 relative to other regions (posteriorly ASD-downregulated ARI genes), or the union (without duplicates) of ARI genes with higher Control expression in the remaining cortical regions relative to BA39-40 and BA17 (posteriorly ASD-upregulated ARI genes). Genes which were sorted into both groups (eg. highest expression in BA39-40 v. BA44-45 in one regional comparison, and highest expression in BA7 v. BA17 in another) were removed. Additionally, for each remaining ARI gene, the median Control gene expression in BA17 and BA39-40 (from the regressed gene expression dataset used for the permutation analysis, using all Control samples) was compared to the median across all remaining regions. Only ARI genes with higher median expression in their respective group (eg.

higher median expression in BA17 and BA39-40 in the posteriorly ASD-downregulated ARI gene group) were retained. For each gene in each of the two groups, the linear contrast comparing BA17 and BA39-40 gene expression to all other cortical regions was assessed in Controls with the same linear model workflow and normalized, outlier-removed gene expression dataset used to identify DE genes and isoforms described before. The beta values and p-values from this analysis are shared in Table A4.4 and for top attenuated transcription factors (TFs) in Figure 4.2c-d.

To functionally characterize the ARI gene groups, we performed cell-type and gene ontology enrichment, identified transcription factors present, and calculated transcription factor binding site enrichment. Cell-type enrichment was conducted with EWCE,[41] with broad (Level 1) neural cell-type gene markers acquired from Lake et al. Nat Biotechnol 2018[42] (frontal and visual cortex samples combined). To obtain cell-type specificity scores, first genes were filtered such that the gene needed to have a mean UMI of 0.005 across all cells. Then, gene UMI averages were taken across all Level 2 cell-types, and these averages were used to generate the cell-type specificity scores utilized by EWCE to calculate cell-type enrichment in the ARI gene groups. This approach was taken to reduce bias introduced by differing numbers of cells across Level 2 cell-types when calculating Level 1 cell-type specificity scores. 100,000 bootstraps were generated to determine cell-type enrichment with EWCE. gProfileR[43] was used for gene ontology enrichment, with FDR-adjustment for p-values, strong hierarchical filtering, and a required overlap size of 10 genes. For the ARI downregulated gene group, a max set size of 2500 was enforced, whereas no max set size was enforced for the ARI upregulated gene group. Only 'BP' (biological process) terms were included in Figure 4.2 and Table A4.4. Transcription factor binding site enrichment was also conducted with gProfileR,[43] with a Bonferroni-adjustment for p-values and strong hierarchical filtering. To identify transcription factors within the ARI gene groups, AmiGo 2[44] was used to acquire all genes in GO:0003700

(DNA-binding transcription factor activity) in the Homo sapiens organism (Gene Ontology Consortium,[45,46] accessed May 7, 2020).

WGCNA network formation and module identification

Weighted Gene Co-Expression Network Analysis (WGCNA)[10] was conducted to sort observed gene and isoform expression dysregulation into empirically-informed networks which could provide precise functional insight into affected neural cell-types and biological processes. Regressed gene and isoform expression datasets containing only the random effect of subject, the fixed biological effects (diagnosis, region, age, age^2, sex, and ancestry), and the model residual were used for WGCNA signed network generation. Regression was performed as described previously for the previously identified co-expression modules. A soft-threshold power of 6 was chosen for gene network generation, whereas a power of 10 was selected for isoform network generation. These values were selected to optimize induced scale-free topology in the gene and isoform networks ($R^2 > 0.8$). For the gene-level WGCNA, a robust version of WGCNA (rWGCNA)[21] was implemented to mitigate the influence of potential sample outliers in network formation. Subjects within each diagnosis group were randomly selected (with replacement) for inclusion in the adjacency matrix (formulated using the bi-midweight correlation of genes) and subsequent TOM matrix generation, 100 times. These TOMs were merged into one consensus TOM through first using a quantile scale of 0.95 to calibrate each TOM, and then taking the median across all TOMs to create the consensus TOM. To identify modules from the consensus TOM, the 'cutTreeHybrid' function was used with average linkage hierarchical clustering of the consensus TOM, a deep split of 4, cut height of 0.9999, a negative PAMstage, and minimum module size of 50. Modules within a cut height of 0.1 were merged.

Since rWGCNA could not be implemented for the isoform expression data due to memory allocation limitations, the ‘blockwiseModules’ function was used with 4 blocks (26,000 or less isoforms per block) to generate the isoform network and identify modules. The same module identification parameters (except for the soft power threshold) used for the gene network were also used for the isoform network. To test the robustness of the isoform network, a permutation approach was utilized.[22,47] For each module, this method tests if the mean connectivity within the module (also defined as the module’s density, or the average intramodular topological overlap) is significantly different from that of modules of equivalent size randomly selected from the same network (n=5,000 permutations). One-tailed p-values were calculated through comparing the permuted distribution to the true mean connectivity for each module, and only modules with p-values < 0.05 were retained. When merging modules from all blocks for the isoform network, a merge cut height of 0.2 was used.

Module eigengenes (MEs) were calculated for all modules using the regressed gene and isoform expression dataset used to generate the networks. We only retained isoform modules which were non-redundant with gene modules forward for further analysis. To achieve this, isoform and gene MEs were clustered using the ‘cutTreeHybrid’ WGCNA[10] function using average linkage hierarchical clustering of the bi-midweight correlation of the MEs, a deep split of 4, a negative PAMstage, a minimum module size of 1, and a cut height of 0.9999. Any isoform modules which clustered with gene modules were labeled as overlapping with the gene modules, with the exception of Isoform_M26_skyblue3, which upon visual inspection was suitably distant from the other gene modules within its cluster to be considered distinct. To determine if any of these other overlapping isoform modules were distinct enough from the gene modules to be retained for further analysis, for each of the conserved isoform modules an over-representation analysis (ORA) was conducted with each of the gene modules in its identified cluster. Any isoform modules which had no significant overlap (p > 0.01) were retained for further analysis, of which only two were identified - Isoform_M55_blue2 and Isoform_M61_navajowhite1. In total, 39 distinct isoform modules were carried forward for further analysis out of the original 61 identified isoform modules.

Module functional characterization

Gene and the distinct isoform MEs were assessed with the gene and isoform linear mixed models containing all of the biological covariates from the full models described previously (the technical covariates were not included, since these covariates were previously removed from the regressed expression data used to generate the MEs). The same limma[35] workflow was implemented as described before for calculating DE genes and isoforms. Whole cortex and region-specific ASD and dup15q effects were also ascertained as described previously for the DE gene and isoform analysis. A covariate required an FDR-adjusted p-value < 0.05 to be considered associated with any ME. To determine if any region-specific ASD effects in the gene MEs were significantly greater than the whole cortex ASD effect, a permutation approach was used which was synonymous to the previously described method used with the Parikshak et al. MEs. Regionally-variable modules are those with any region having a region-specific ASD effect significantly greater than the whole cortex ASD effect ($p < 0.05$).

To further functionally characterize modules, we calculated enrichments for neural cell-types, neuronal subtypes, gene ontology terms, protein-protein interactions, the ARI gene groups, gene biotypes, relevant GWAS, ASD and epilepsy associated rare variants, and gene modules previously associated with ASD published in Parikshak et al. Nature 2016[5] and Gandal et al. Science 2018b.[1] While all of these enrichment analyses were performed for the gene modules, only a subset were performed for the isoform modules (neural cell-types, gene ontology terms, gene biotypes, psychiatric GWAS, and ASD and epilepsy associated rare variants).

Neural cell-type enrichment was performed with EWCE[41] as previously described for the ARI gene groups. For neuronal subtype enrichment, medial temporal gyrus single neuron RNA-seq from the Allen Brain Map[13,41] was used to define neuronal subtype specific markers for enrichment analysis with EWCE.[41] EWCE was implemented as previously described for the ARI gene groups, with the Allen Brain Map neuronal cells being grouped into cortical layer groups (eg. Exc L2, Inh L2-3), for cell-type enrichment. For gene ontology terms, the Metascape[48] web portal was used with default functions ('Express Analysis'). Only 'GO Biological Process' terms with an FDR-adjusted p-value < 0.05 were examined for each module. PPI annotations and enrichments were calculated with STRING,[49] run with default settings in June 2019. A direct connection FDR-corrected p-value < 0.05 was needed for a module to be considered significantly enriched with PPIs. ARI gene group enrichment was calculated with ORA, with an FDR-corrected p-value < 0.05 and OR > 1 being required for a significant enrichment.

Gene biotype enrichment was determined with a permutation approach. The number of each unique gene biotype was first acquired for each module. Then, for each permutation (10,000 in total) gene biotypes were samples across all genes without replacement and randomly assigned. The number of each unique gene biotype in each module was collected for each permutation. A distribution could then be created for each unique gene biotype in each module across the 10,000 permutations. Both over- and under-enrichment of each unique gene biotype in each module was determined directly with this distribution (one-tailed p-value). An FDR-corrected p-value < 0.05 was required for a significant enrichment.

For the psychiatric GWAS enrichments, partitioned heritability was calculated with stratified LD Score regression[50] (run with recommended settings) using 10 kb windows around genes (matched genes were used for isoform modules). An FDR-corrected p-value <

0.1 was required for a significant GWAS enrichment (the threshold for significance was relaxed since many of the best available GWAS datasets utilized are underpowered, particularly the ASD GWAS). We selected the most recent and best powered GWAS which were relevant and interesting for comparison with these gene and isoform modules, including GWAS conducted for ASD,[11] ADHD,[51] BD,[52] MDD,[53] SCZ,[54] Educational Attainment,[55] Intelligence,[56] and IBD.[57] Logistic regression was used for rare variant enrichment, controlling for both gene length and GC content, with an FDR-corrected p-value < 0.05 being required for a significant enrichment. Syndromic and highly ranked (1 and 2) ASD SFARI[12] gene and high-confidence Epilepsy (compiled by D. Polioudakis et al. Neuron 2019)[58] gene associations were examined. Finally, ORA was used to assess previous module enrichment, with an FDR-corrected p-value < 0.05 and OR >1 indicating a significant positive overlap.

Neuronal density and cortical layer 4 association with ASD dysregulation

A linear model was used to compare region-specific macaque NeuN density[15] to region-specific ASD effects (model beta) in the regionally-variable gene MEs. Macaque brain areas were matched to Brodmann areas (shared in Table A4.7), with six regions matching between this dataset and the macaque dataset. FDR-corrected p-values < 0.1 were considered significant neuronal density associations (the FDR threshold was relaxed, since only 6 regions/points were available for every comparison). A leave-one-out cross-validation was performed to assess individual regional contributions to neuronal density associations, in which a single region was withheld and linear model statistics were re-calculated. In addition to neuronal density, we also examined the association between cortical layer 4 thickness[18] (von Economo and BigBrain estimates, as shared in the publication) and region-specific ASD effects in the regionally-variable gene MEs. All 11 regions were matched to layer 4 thickness measures (this key is shared in Table A4.7). This comparison was also performed with a linear model, with FDR-corrected p-values < 0.05 considered significant layer 4 thickness associations.

snRNA-seq

Matched control and ASD samples for co-variates (i.e age, sex, manner of death) were processed in the same nuclear isolation batch to minimize potential batch effects. 50 mg of sectioned brain tissue was homogenized in 2.5 mL of RNAase-free homogenization buffer (250mM sucrose, 5mM MgCl2, 25mM KCL, 10mM Tris pH8, 1 uM DTT, 0.2U RNaseIN, 1% BSA, 0.01% Triton X-100, 0.001% Digitonin in RNAse-free water) using glass dounce homogenizer on ice. The homogenate was filtered and subjected to a two layer micro-iodixanol nuclei centrifugal gradient (50%/30%) for 13500g for 20 minutes at 4°C. Supernatant was carefully removed and the nuclei containing pellet were resuspended in RNase-free PBS pH7.4, 5mM MgCl2, 1% BSA, 0.2U RNaseIN. The nuclear suspension was filtered twice through a 30 um cell strainer. Nuclei were counted using a hemocytometer and diluted to 1,000 nuclei/uL before performing single-nucleus isolation on the 10X Genomics controller.; The 10X capture and library preparation protocol was used without modification. Single-nucleus libraries from individual samples were pooled and sequenced on the NovaSeq 6000 machine (average depth 60,000 reads/nucleus).

Raw snRNA-seq data processing was performed with 10X Genomics CellRanger software, Seurat,[59] and Liger.[60] CellRanger was used with default parameters, except we utilized the human pre-mRNA reference file (ENSEMBL GRCh38)[27] to insure capturing intronic reads originating from pre-mRNA transcripts abundant in the nuclear fraction. Individual libraries were analyzed in Seurat for quality control metrics and filtering. Individual libraries were filtered to retain nuclei with at least 500 genes expressed and less than 5% of total UMIs originating from mitochondrial RNAs. Individual matrices were combined, UMIs were normalized to the total

UMIs per nucleus and log transformed. Nuclei for all ASD and control subjects from both the PFC and OCC were used for clustering with integrative non-negative matrix factorization with K=40 and lamba= 5.0 followed by quantile normalization and louvain clustering in Liger. We then visualize integrated cells in two-dimensional space with Uniform Manifold Approximation and Projection (UMAP).

Cell-type deconvolution

Selected datasets

Bulk RNA-seq: For the 808 samples from 11 regions, we used the residuals of the regression of the bulk data against technical covariates. While we ran the deconvolutions for all samples, we explicitly excluded the Dup15q samples from the downstream comparisons between the ASD and CTL cohorts.

Single-nuclei RNA-seq: The single-nuclei dataset used in this analysis was obtained from the frontal cortex (FC) and primary visual cortex (V1C) of 4 individuals (2 ASD and 2 CTL) overlapping with the bulk tissue cohort, comprising four FC libraries and four V1C libraries (described in the preceding section). Cell type assignments for the bulk-tissue-overlapping snRNA-seq libraries were obtained by looking at the expression of canonical markers from years of culminated mouse and human studies and recent single-cell atlases. Specifically, we utilized the Hodge 2019 (Allen Institute)[13] and Lake 2018[18] papers to establish frontal cortex and V1 specific signatures found previously. Based on both the expression level of the gene, but also the percentage of cells within a cluster that expressed said gene, we identified 35 cell types/states. BA17 specific neurons are superficial neurons in V1 that express SYT2 and RORB higher than frontal cortex, and Layer 4 V1 Ext neurons express PHACTR2 and EYA4 over other areas.

Here are examples of core genes used to establish excitatory neuronal types: SLC17A7, RBFOX3. For Excit L2/3: “LAMP5",“CUX2”,“GLRA3", “CUX1”, “LHX2",“CBLN2”, “RASGRF2", “COL5A2”, “LMO3", “SATB2”. For Excit 3/4: “RORB”, “PCP4”, “LMO3", “CUX2”, “SATB2", “NEFM”, “PHACTR2",“EYA4” . For Excit L5/5B: “BCL11A”,“CRYM”,“FOXP2",“BCL11B”,“FEZF2", “RORB”, “DKK3", “TLE4”, “SEMA3E”, “LMO4", “CCK”,“ETV1", “NEFH”, “CNTN6", “FOXO1”, “OPN3", “LIX1”, “SYT9", “S100A10”, “LDB2", “CRIM1”, “PCP4", “SATB2”, “CRYM”. For Excit L6: “GLRA3", “LMO3”, “BHLHE22", “RORB”, “NNAT”, “FOXP2”, “ETV1", “FEZF2”, “TLE4", “GRIK4”, “NTNG2", “OPRK1", “NR4A2”, “BCL11B”, “THY1”.

Data processing for deconvolution analyses

We applied the following processing steps to the bulk tissue RNA-seq and snRNA-seq data prior to running CIBERSORTx.

1. **Bulk RNA-seq:** For the post-regression residual matrices, gene names were converted from ENSEMBL IDs to HGNC symbols. Any genes that were not mapped (by *biomaRt* in R) were removed. Gene expression values were transformed from their log-base-2 values to non-log expression values, as required by CIBERSORTx.
2. **snRNA-seq:**
    a. Expression analysis: The input data format consisted of standard *CellRanger* output matrices. The matrices were read into *Seurat* (https://cran.r-project.org/web/packages/Seurat) objects in R. Lenient QC cutoffs of “Number of RNA features > 500” and “Percent Mitochondrial genes < 10” were chosen to filter out cells. The *Liger* cell-type-cluster labels were associated with cells, and any cells without identified cell types were removed from the analysis. We again used *scran* but modified the previous pipeline slightly: (I) removed cell types that have < 10 cells within the sample; (II) found the size factors using pool sizes of 20, 30,

40, 50, and 60; and (III) further removed any cells that produced negative size factors. Finally, the counts for each cell were converted to a weighted counts-per-million scale as $Weighted\_CPM(cell\ i, gene\ g) = 10^6 . Size\_factor(cell\ i) . \frac{Counts(cell\ i, gene\ g)}{\sum_{g=1}^{N_{genes}} Counts(cell\ i, gene\ g)}$, and N_{genes} = Total number of genes.

b. Pooling of cells: We combined cells from the FC and V1C regions into a single reference matrix for the deconvolution. Each of the 4 FC and 4 V1C libraries was run through the preprocessing steps separately and combined subsequently. The merging of matrices was carried out by concatenating *pandas* data frames in Python, using a join on overlapping gene names. It is worth noting that this merging process results in only the intersecting genes across all datasets being included, and due to the removal of rows corresponding to non-overlapping genes, the resulting cell expression vectors may not be normalized to 10^6. The numbers of cells included in the final deconvolution analysis are 145,373 cells.

CIBERSORTx parameters

We downloaded the *fractions* module of CIBERSORTx (doi.org/10.1038/s41587-019-0114-2) from the website, in the form of a Singularity image of the Docker file. We ran the program in the mode that accepts a full single cell count matrix as input, thereby implicitly generating a reference matrix (the parameter *single_cell* is set to *TRUE*), and set the batch-correction mode to the ‘S’ mode (the parameter *rmbatchSmode* is set to *TRUE*).

Testing for differences in cell fractions between ASD and CTL groups

To evaluate whether differences, between the ASD and CTL groups, in the cell fractions of particular cell types in each region were statistically significant we calculated p-values from the Wilcoxon Rank-Sum Test and two-sided Kolmgorov-Smirnov (KS) Test on the distributions of the two groups. These were calculated using the *scipy.stats.ranksums* and *scipy.stats.ks_2samp*, respectively in python's *scipy* library. For multiple-hypothesis-testing correction, we performed Bonferroni correction for each region: that is, we divided the p-values for each cell type and each region by the number of cell types considered (and not by the product of the number of cell types and the number of regions). Wilcoxon rank-sum test p-values that were lower than the Bonferroni corrected significance threshold (0.05/35 = 0.0014) were considered to represent significant differences in cell-type proportion.

Testing for differences in cell fractions across regions

ANOVA was used to test for differences in cell fractions across all eleven regions separately in ASD and CTL groups. Significant differences were those lower than the Bonferroni significance threshold, corrected across all cell-types within diagnosis groups (0.05/35 = 0.0014). ASD differences were considered attenuated if the ASD p-value was greater than that of the CTL group.

A4.2: Extended Figures

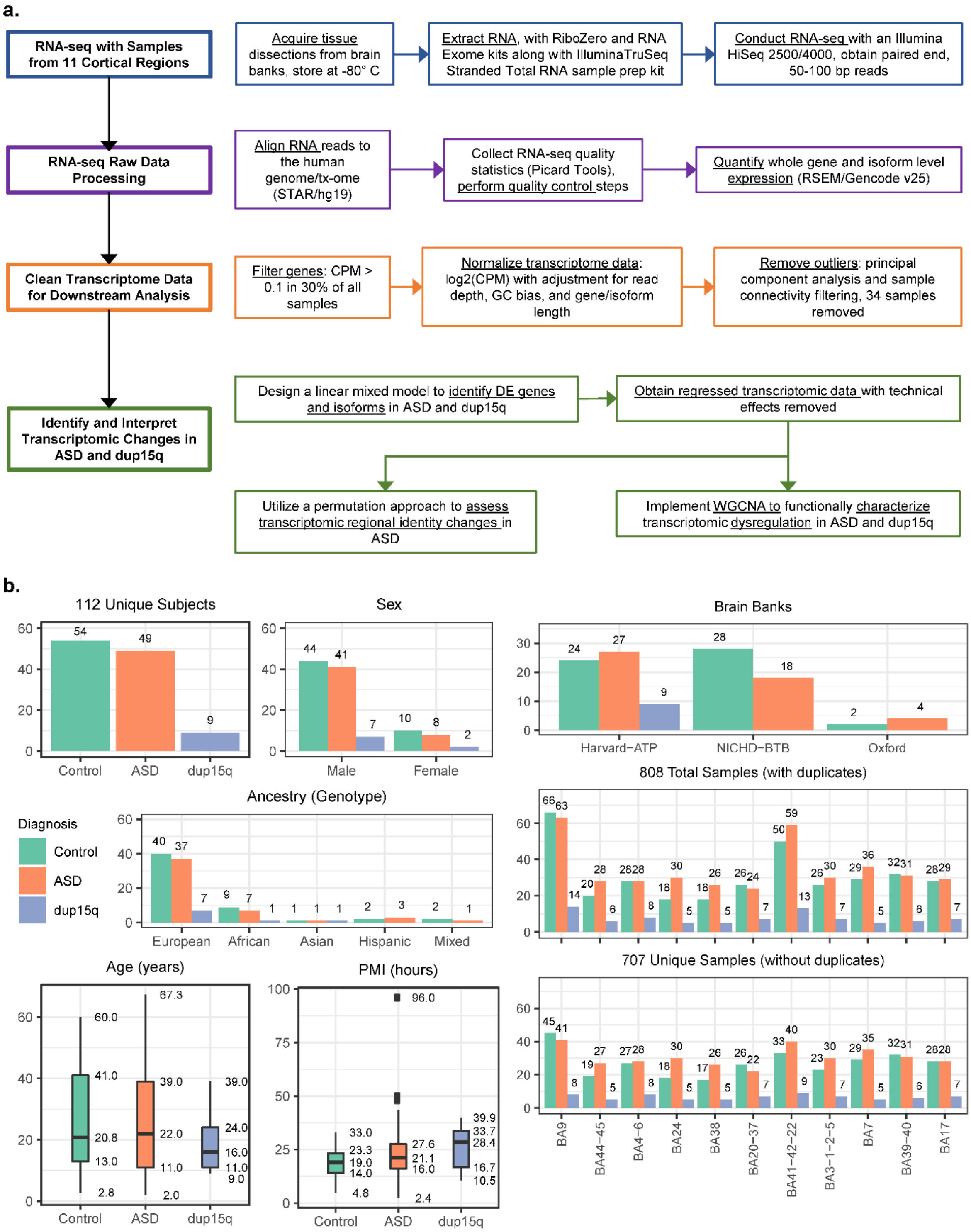
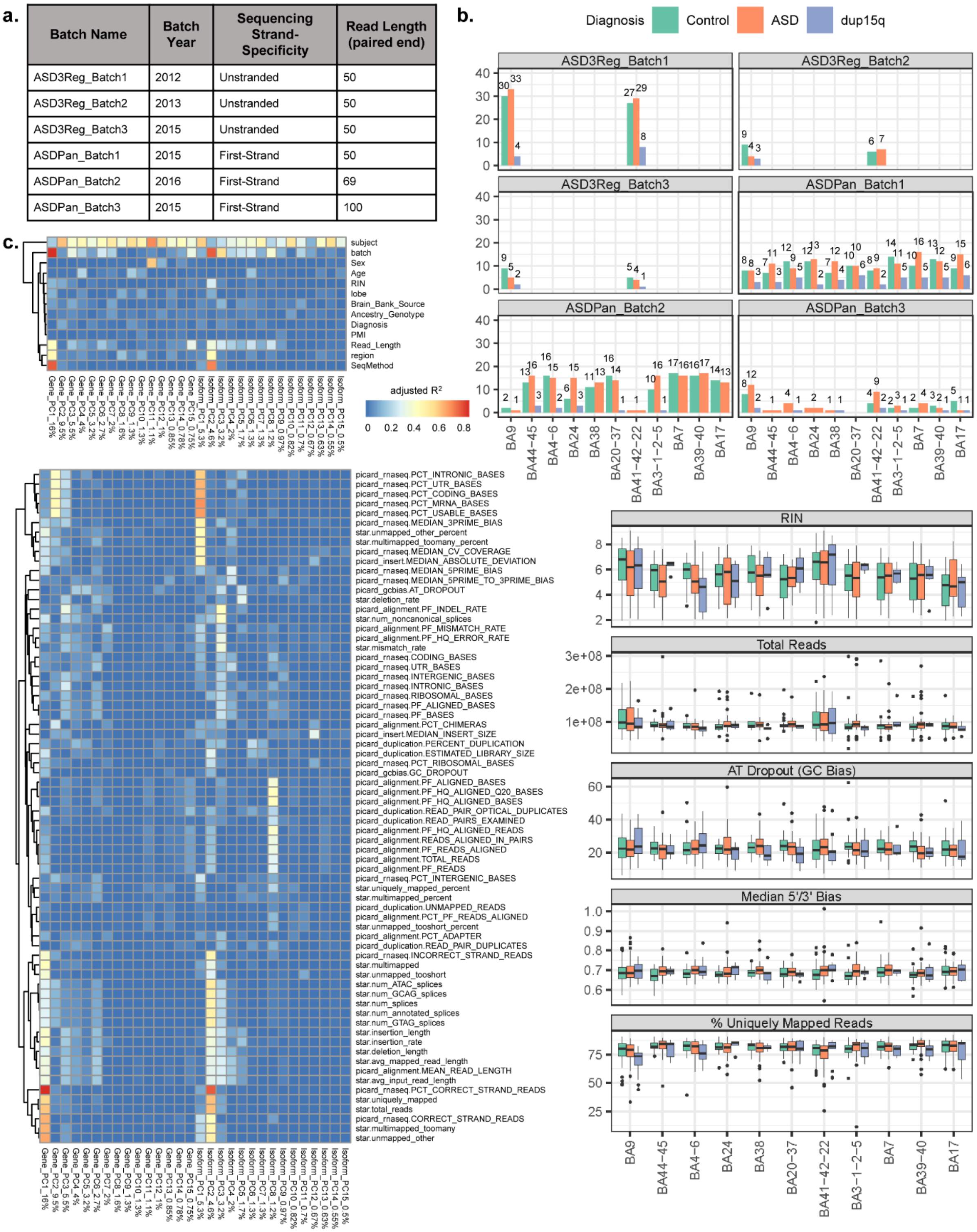


Figure A4.1 | Experiment workflow and sample overview. a. Overview of experiment workflow. b. Summary of sample composition (biological data, brain bank source, and PMI).



Batch Name	Batch Year	Sequencing Strand-Specificity	Read Length (paired end)
ASD3Reg_Batch1	2012	Unstranded	50
ASD3Reg_Batch2	2013	Unstranded	50
ASD3Reg_Batch3	2015	Unstranded	50
ASDPan_Batch1	2015	First-Strand	50
ASDPan_Batch2	2016	First-Strand	69
ASDPan_Batch3	2015	First-Strand	100

Figure A4.2 | Quality control measures. a. Sequencing batch parameters. b. Sequencing batches (top) and sequencing statistics (bottom) by region and diagnosis. c. Top 15 expression PCs (gene and isoform, with % of variance explained denoted) association with meta data (top) and sequencing statistics (bottom).

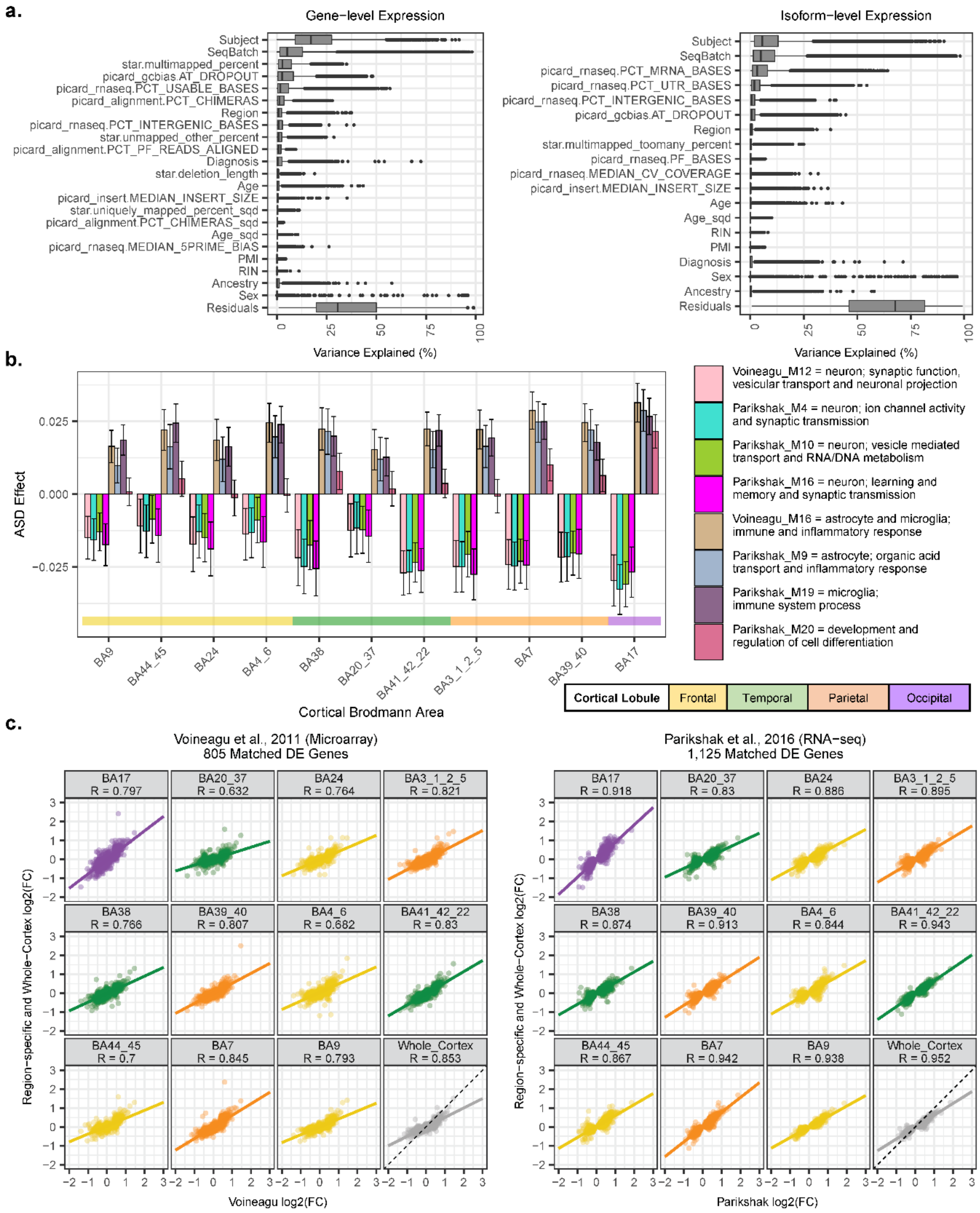


Figure A4.3 | Model covariates and previous studies across 11 cortical regions. a. For the covariates selected for the gene (left) and isoform (right) linear mixed models, % of expression variance explained across all genes/isoforms. b-c. For the Voineagu et al. and Parikshak et al. studies, b. ASD associated gene module ASD effect (standard error bars and cortical lobes indicated) and c. ASD log2 FC of DE genes identified in these studies, compared to this dataset (Spearman's correlation rho, R, is plotted along with the linear least squares regression best fit line).

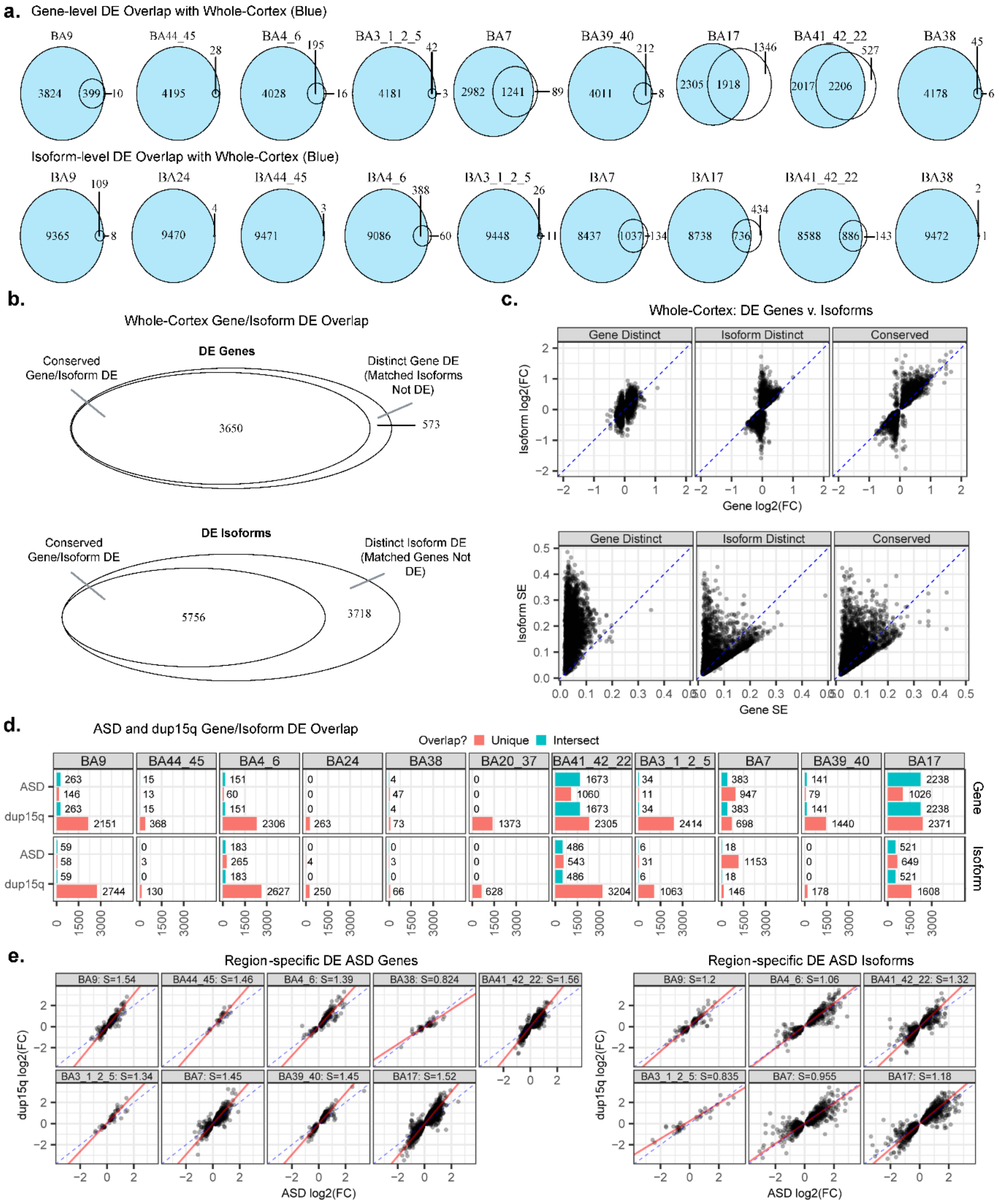


Figure A4.4 | Transcriptomic changes across 11 cortical regions. a. Overlap of Whole-Cortex DE ASD genes and isoforms (blue) with other cortical region DE genes (no color). Regions with no third numeric label on the right completely overlap with the Whole-Cortex DE genes. b. For the Whole-Cortex DE, overlap of genes and isoforms. Regions not shown have no unique DE. c. log2(FC) (top) and standard error (SE, bottom) of the Whole-Cortex ASD DE overlapping and distinct genes and isoforms. d. Overlap in DE ASD and dup15q genes and isoforms. e. For regions with DE ASD genes (left) and isoforms (right), ASD log2(FC) v. dup15q log2(FC) for specific regions (with principal components regression slope, S).

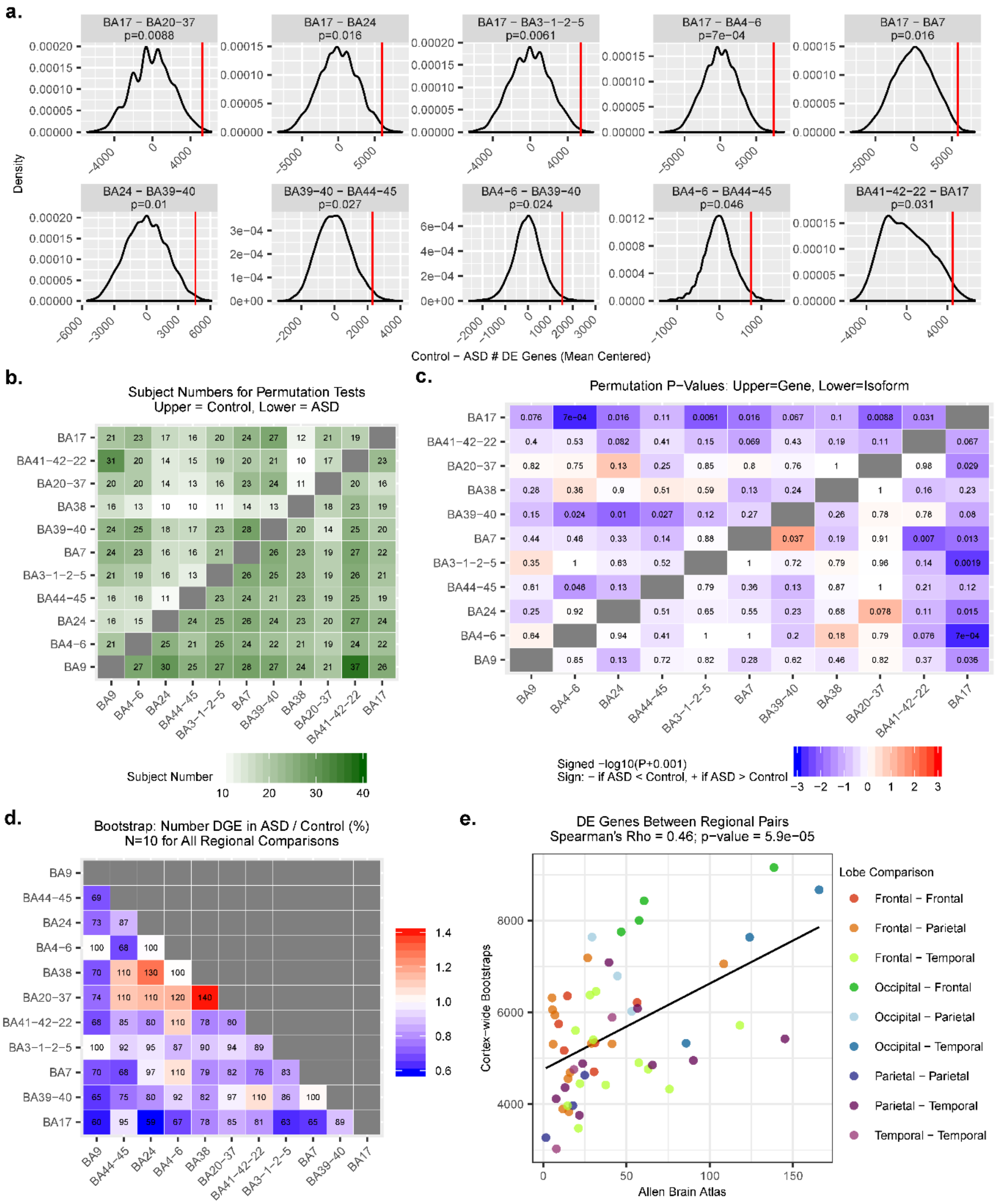


Figure A4.5 | Transcriptomic regional identity attenuation in ASD. a. Mean-centered distribution of 10,000 permutations for the significantly attenuated regional comparisons in ASD. Red bar = true difference in the number of DE genes between controls and ASD. b. Sample size for all regional comparisons. c. Permutation p-values for all regional comparisons. d. For 10,000 regional comparison bootstraps, ratio of DE genes in ASD compared to controls. e. Number of DE genes between pairs of regions in this study (mean across bootstraps in controls, y-axis) compared to the Allen Brain Atlas (ref. 10, mean across matched regions, x-axis; see Methods for matched regions). This Allen Brain Atlas dataset, with only 2 unique brains, is the best publicly available dataset for comparison (linear least squares regression best fit line plotted).

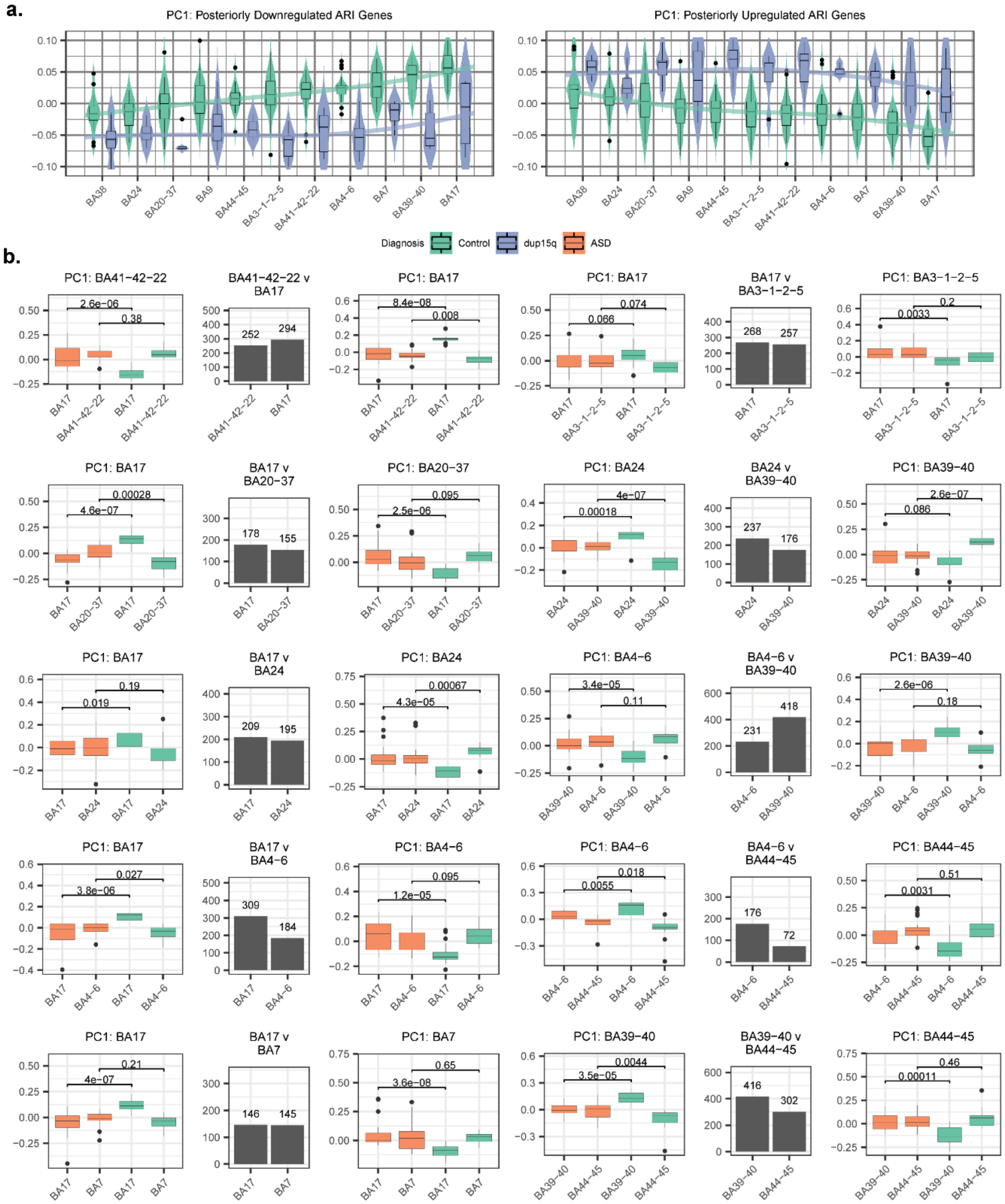


Figure A4.6 | Additional ARI gene dysregulation. a. First principal component (PC1) of posteriorly downregulated (1,881, left) and upregulated (1,695, right) ARI genes identified in ASD, plotted in Controls and dup15q (loess regression line plotted). b. For each significantly attenuated regional comparison, the identified attenuated regional identity (ARI) genes. At center, number of ARI genes with greater neurotypical expression in each pair of regions. On either side of the barplot, the PC1 of the genes with greater neurotypical anterior (left) or posterior (right) expression is plotted across the pair of regions in Controls and ASD. The Wilcoxon signed-rank test (unpaired) p-value is shown.

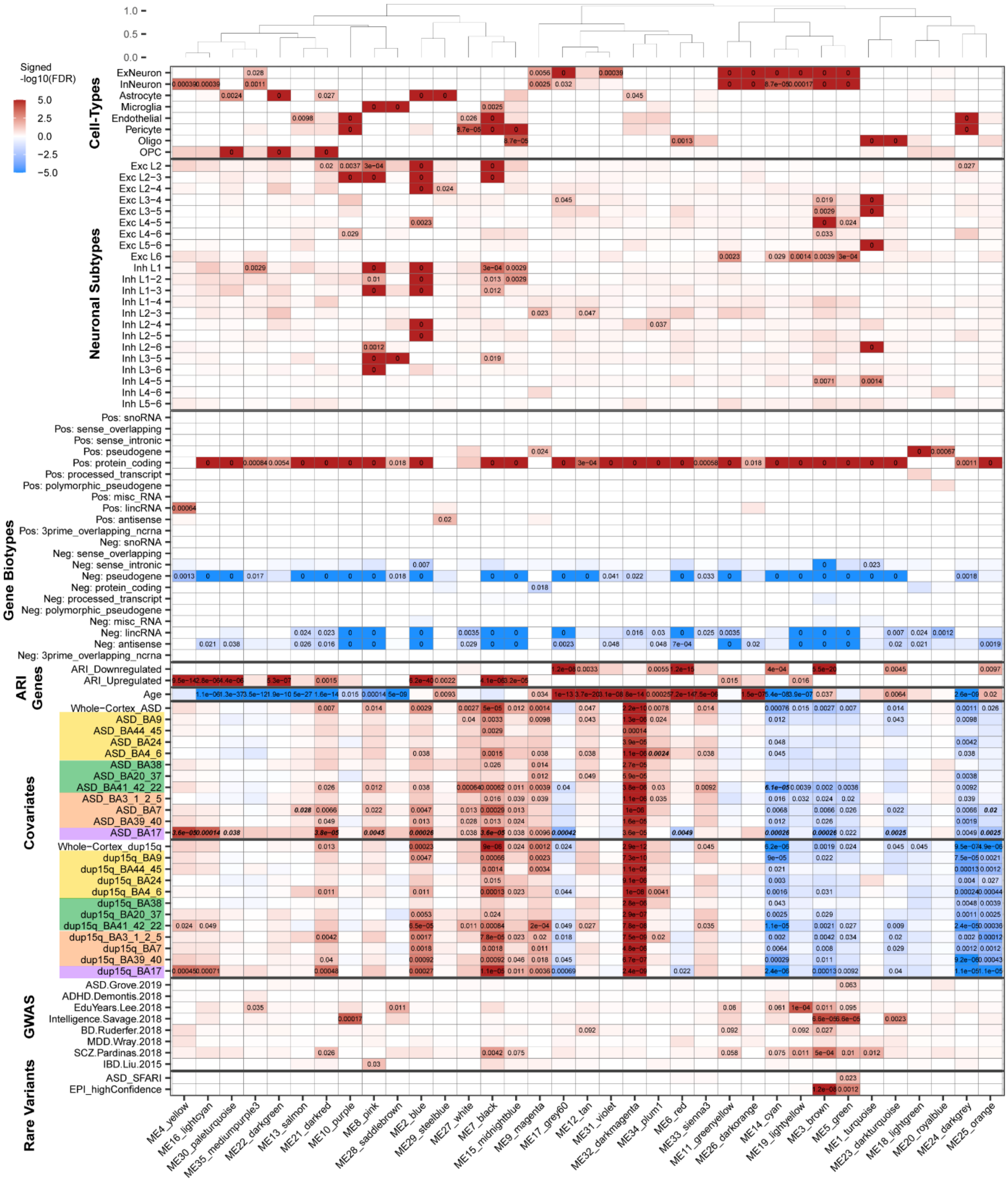


Figure A4.7 | Gene-level co-expression network analysis module associations. Top: average-linkage hierarchical clustering of module eigengene biweight midcorrelations. Significant FDR corrected p-values are indicated (FDR < 0.05; for GWAS, FDR < 0.1). Any signed –log10(p) colors greater or less than 5/-5 are set at a max/min of 5/-5 . For ASD, dup15q, and Age covariates, FDR p-value from the linear mixed model testing the association of these covariates with module eigengenes is depicted. For the ASD and dup15q region-specific comparisons, cortical lobule colors are indicated (Fig. 1a), and bold-italic FDR p-values indicate that these regions are effected with significantly greater magnitude than the ASD whole-cortex (Methods). For gene biotypes, both positive and negative enrichment is shown (Methods). Positive enrichment is shown for cell-type, neuronal subtype (ref: Hodge et al, Nature 2019), ARI gene, GWAS, and rare variant enrichment (Methods).

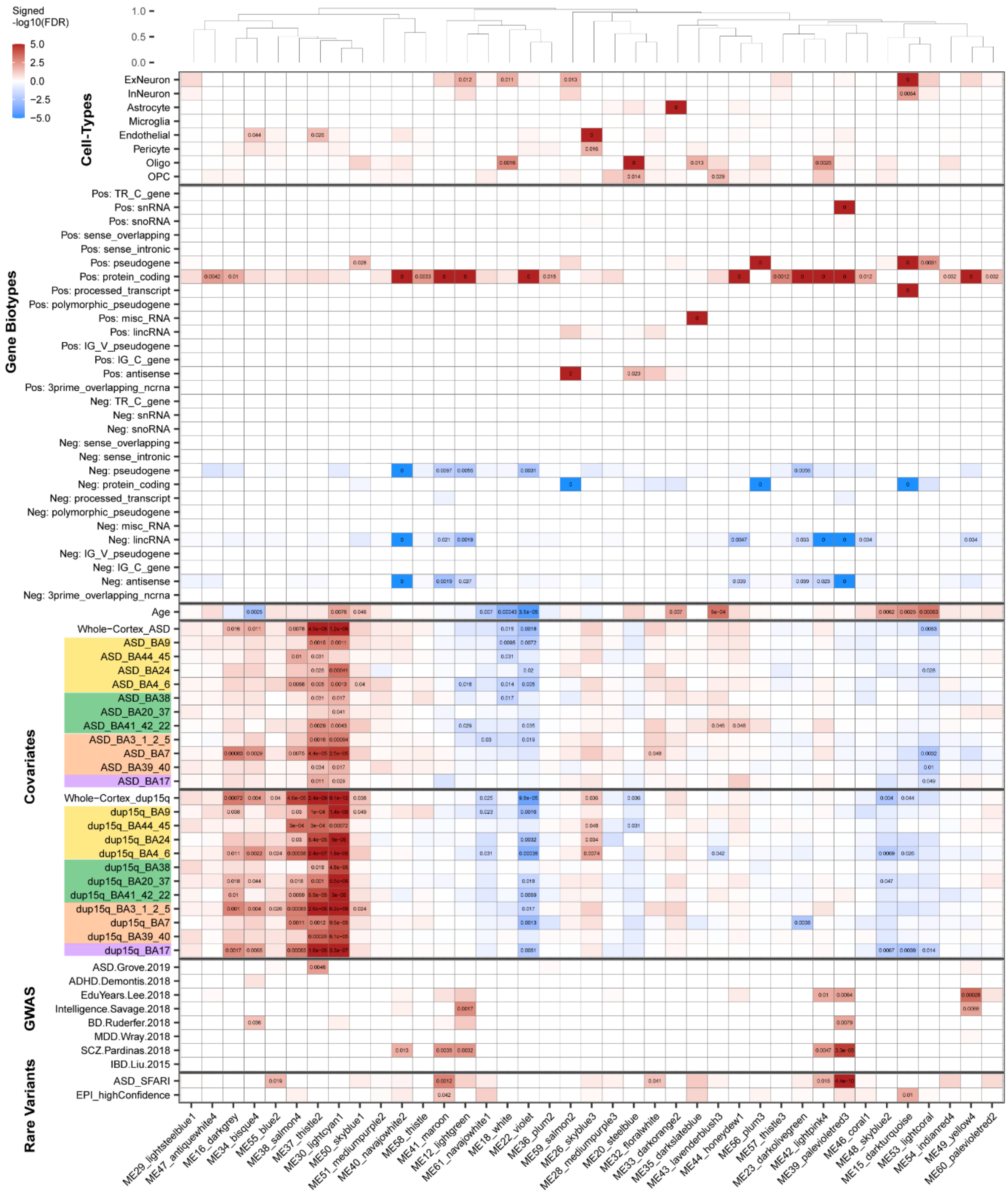


Figure A4.8 | Isoform-level co-expression network analysis module associations. Top: average-linkage hierarchical clustering of module eigengene biweight midcorrelations. Significant FDR corrected p-values are indicated (FDR < 0.05; for GWAS, FDR < 0.1). Any signed –log10(p) colors greater or less than 5/-5 are set at a max/min of 5/-5 . For ASD, dup15q, and Age covariates, FDR p-value from the linear mixed model testing the association of these covariates with module eigengenes is depicted. For the ASD and dup15q region-specific comparisons, cortical lobule colors are indicated (Fig. 1a). For gene biotypes, both positive and negative enrichment is shown (Methods). Positive enrichment is shown for cell-type, GWAS, and rare variant enrichments (Methods).

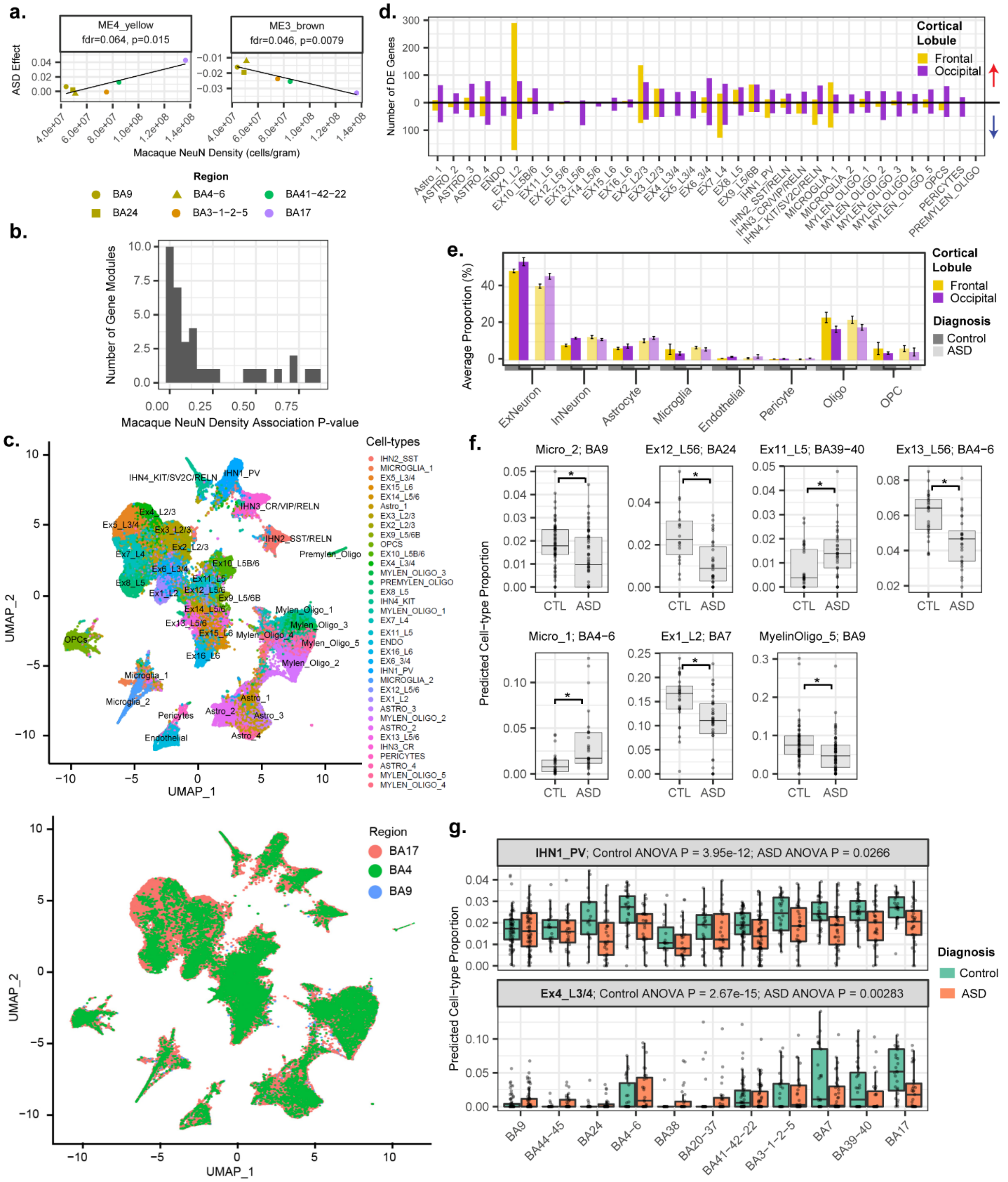
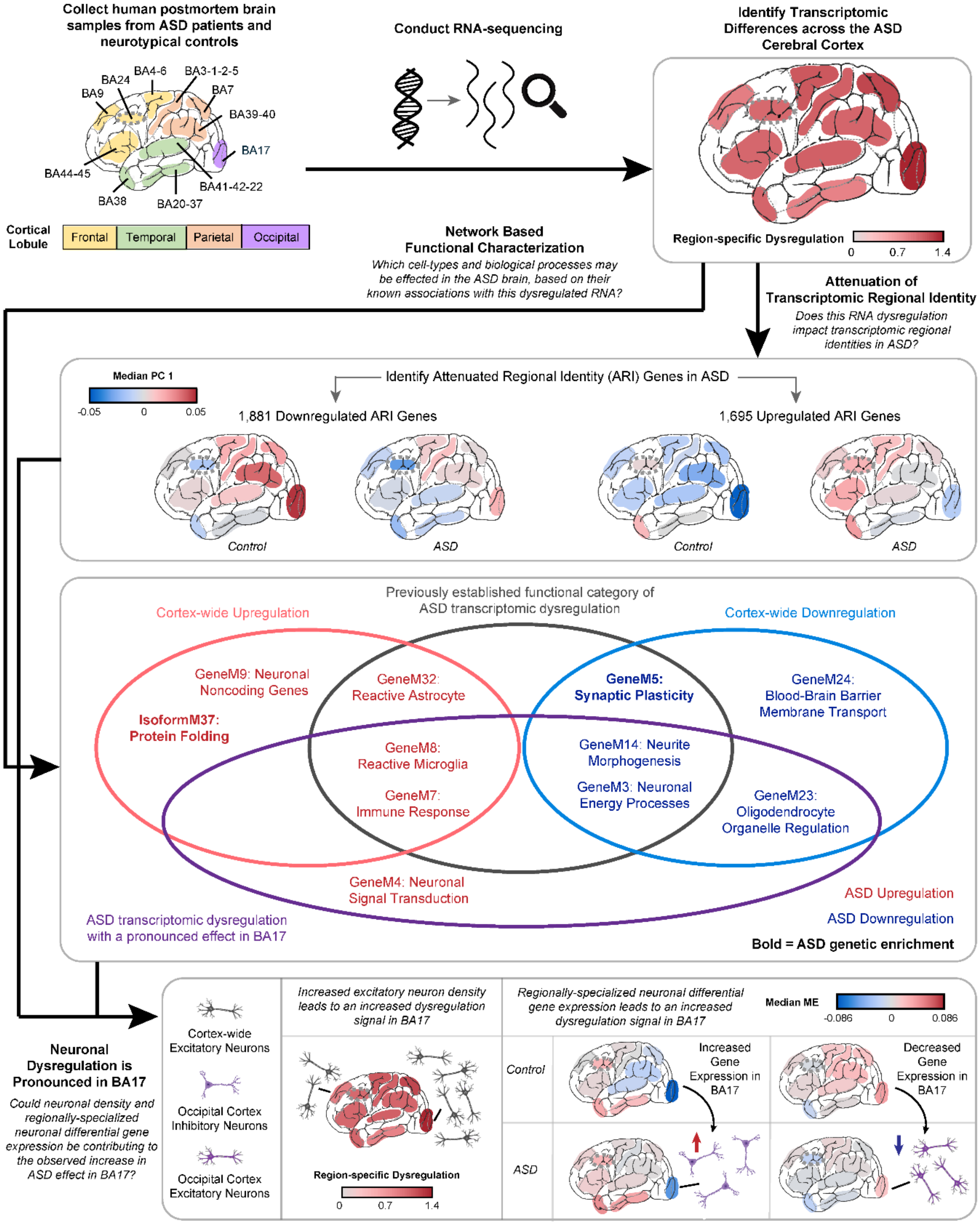


Figure A4.9 | Neuronal Density Associations, snRNA-seq, and Cell-type Deconvolution. a. Macaque neuronal density v. module eigengene ASD effect for modules featured in Fig. 4c-d (linear least squares regression). Both p-value and FDR corrected p-value are plotted. b. P-value histogram of all gene modules' linear least squares regression with macaque region-specific neuronal density. c. UMAP plots of snRNA-seq with cell sub-types (top) and brain regions (bottom) depicted. d. Number of genes differentially expressed in ASD in each cell subtype. Upregulated genes are above 0 (red arrow) and downregulated genes are below 0 (blue arrow). e. Average proportion of each broad cell-type in each diagnosis x cortical lobule group, derived directly from the snRNA-seq data. f. Additional significant (Bonferroni corrected p-value < 0.05) cell-type proportion differences in ASD from cell-type deconvolution. Region and cell-type are indicated in the title of each plot. g. For two example cell-types, cell-type proportion attenuation in ASD across regions. ANOVA p-values stratified by diagnosis are shown.


Collect human postmortem brain samples from ASD patients and neurotypical controls
BA24
BA4-6
BA3-1-2-5
BA9
BA7
BA39-40
BA17
BA44-45
BA41-42-22
BA38
BA20-37
Cortical Lobule
Frontal
Temporal
Parietal
Occipital
Conduct RNA-sequencing
Identify Transcriptomic Differences across the ASD Cerebral Cortex
Region-specific Dysregulation
0
0.7
1.4
Network Based Functional Characterization
Attenuation of Transcriptomic Regional Identity
Median PC 1
-0.05
0
0.05
Identify Attenuated Regional Identity (ARI) Genes in ASD
1,881 Downregulated ARI Genes
1,695 Upregulated ARI Genes
Control
ASD
Control
ASD
Previously established functional category of ASD transcriptomic dysregulation
Cortex-wide Upregulation
Cortex-wide Downregulation
GeneM9: Neuronal Noncoding Genes
IsoformM37: Protein Folding
GeneM32: Reactive Astrocyte
GeneM5: Synaptic Plasticity
GeneM24: Blood-Brain Barrier Membrane Transport
GeneM8: Reactive Microglia
GeneM7: Immune Response
GeneM14: Neurite Morphogenesis
GeneM3: Neuronal Energy Processes
GeneM23: Oligodendrocyte Organelle Regulation
GeneM4: Neuronal Signal Transduction
ASD transcriptomic dysregulation with a pronounced effect in BA17
ASD Upregulation
ASD Downregulation
Bold = ASD genetic enrichment
Neuronal Dysregulation is Pronounced in BA17
Cortex-wide Excitatory Neurons
Occipital Cortex Inhibitory Neurons
Occipital Cortex Excitatory Neurons
Region-specific Dysregulation
0
0.7
1.4
Median ME
-0.086
0
0.086
Control
ASD
Increased Gene Expression in BA17
Decreased Gene Expression in BA17

Which cell-types and biological processes may be effected in the ASD brain, based on their known associations with this dysregulated RNA?

Does this RNA dysregulation impact transcriptomic regional identities in ASD?

Could neuronal density and regionally-specialized neuronal differential gene expression be contributing to the observed increase in ASD effect in BA17?

Increased excitatory neuron density leads to an increased dysregulation signal in BA17

Regionally-specialized neuronal differential gene expression leads to an increased dysregulation signal in BA17

Figure A4.10 | Results Summary. Overview of RNA-sequencing experiment and results. Region-specific dysregulation scale in the top right corner and the leftmost portion of the bottom panel depict the region-specific slopes compared to the whole cortex effect from Fig 1d. Median PC 1 of the ARI dysregulated genes is plotted in the middle panel. In the right portion of the bottom panel, the median ME of GeneM4 (left) and GeneM3 (right) is depicted.

A4.3: Extended Tables

Table A4.1: Metadata and sequencing quality metrics

Metadata and sequencing quality metrics for all samples, along with top gene and isoform expression principal component associations with metadata and sequencing quality metrics.

Table A4.2: DE gene overlap analyses

Differential gene expression overlap data, both with previous publications (Voineagu et al. 2011 and Parikshak et al. 2016) and within this dataset (across regions, diagnoses, and whole gene v. isoform datasets). Regional ASD DE gene permutation data is also included.

Table A4.3: DE gene results

Linear mixed model statistics for biological covariates (diagnosis, region, age, sex), for all genes and isoforms assessed.

Table A4.4: Transcriptomic regional identity analysis results

Statistics from transcriptomic regional identity analysis, including permutation data (for both gene and isoform level expression), bootstrap data, and attenuated regional identity (ARI) gene data. A region matching key for comparing Allen Grain Atlas regions to Brodmann areas is also included.

Table A4.5: Genes matched to WGCNA modules

For all genes and isoforms, WGCNA module assignment, kME values, and gene/isoform annotation.

Table A4.6: Functional characterization of WGCNA modules

Functional characterization data for all gene and isoform modules, along with linear mixed model effects for biological covariates (diagnosis, region, age, and sex) in all modules.

Table A4.7: Supporting data for analyses of regionally-variable ASD transcriptomic dysregulation

Supporting data for the analysis of regionally-variable ASD transcriptomic dysregulation, including: neuronal density and cortical L4 thickness associations with modules, brain area matching keys for these associations, module overlap with low integrity RNA genes, snRNA-seq cell-type proportions across regions (including bootstrapped proportion statistics), snRNA-seq DE gene data, and cell-type deconvolution results and statistics.

A4.4: Extended Bibliography

Bibliography extends from chapter four (section 4.6).

39. Milborrow. Derived from mda:mars by T. Hastie and R. Tibshirani., S. *earth: Multivariate Adaptive Regression Splines*. http://CRAN.R-project.org/package=earth (2011).

40. Hoffman, G. E. & Schadt, E. E. variancePartition: interpreting drivers of variation in complex gene expression studies. *BMC Bioinformatics* **17**, 483 (2016).

41. Skene, N. G. & Grant, S. G. N. Identification of Vulnerable Cell Types in Major Brain Disorders Using Single Cell Transcriptomes and Expression Weighted Cell Type Enrichment. *Front. Neurosci.* **10**, 16 (2016).

42. Lake, B. B. *et al.* Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat. Biotechnol.* **36**, 70–80 (2018).

43. Raudvere, U. *et al.* g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research* vol. 47 W191–W198 (2019).

44. Carbon, S. *et al.* AmiGO: online access to ontology and annotation data. *Bioinformatics* vol. 25 288–289 (2009).

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

46. The Gene Ontology Consortium & The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research* vol. 47 D330–D338 (2019).

47. Oldham, M. C. *et al.* Functional organization of the transcriptome in human brain. *Nature Neuroscience* vol. 11 1271–1282 (2008).

48. Zhou, Y. *et al.* Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat. Commun.* **10**, 1523 (2019).

49. Szklarczyk, D. *et al.* STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **43**, D447–52 (2015).

50. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics* vol. 47 1228–1235 (2015).

51. Demontis, D. 73 NEW ADHD RISK LOCI IDENTIFIED IN A GWAS META-ANALYSIS OF 126,000 ADHD CASES AND 900,000 CONTROLS. *European Neuropsychopharmacology* vol. 29 S100–S101 (2019).

52. Bipolar Disorder and Schizophrenia Working Group of the Psychiatric Genomics Consortium. Electronic address: douglas.ruderfer@vanderbilt.edu & Bipolar Disorder and Schizophrenia Working Group of the Psychiatric Genomics Consortium. Genomic Dissection of Bipolar Disorder and Schizophrenia, Including 28 Subphenotypes. *Cell* **173**, 1705–1715.e16 (2018).

53. Wray, N. R. *et al.* Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat. Genet.* **50**, 668–681 (2018).

54. Pardiñas, A. F. *et al.* Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nat. Genet.* **50**, 381–389 (2018).

55. Lee, J. J. *et al.* Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat. Genet.* **50**, 1112–1121 (2018).

56. Savage, J. E. *et al.* Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nat. Genet.* **50**, 912–919 (2018).

57. Liu, J. Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47**, 979–986 (2015).

58. Polioudakis, D. *et al.* A Single-Cell Transcriptomic Atlas of Human Neocortical Development during Mid-gestation. *Neuron* vol. 103 785–801.e8 (2019).

59. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).

60. Welch, J. D. *et al.* Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell* **177**, 1873–1887.e17 (2019).

61. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).

62. Feng, H., Zhang, X. & Zhang, C. mRIN for direct assessment of genome-wide and gene-specific mRNA integrity from large-scale RNA-sequencing data. *Nat. Commun.* **6**, 7816 (2015).