UCLA
UCLA Electronic Theses and Dissertations

Title
Transcriptional Regulators in Normal Human Brain Development and Autism

Permalink
https://escholarship.org/uc/item/6rb4f60t

Author
Parikshak, Neelroop

Publication Date
2015

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA

Los Angeles

Transcriptional Regulators in Normal Human Brain Development and Autism

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

by

Neelroop Narendra Parikshak

2015

© Copyright by

Neelroop Narendra Parikshak

2015

ABSTRACT OF THE DISSERTATION

Transcriptional Regulators in Normal Human Brain Development and Autism

by

Neelroop Narendra Parikshak
Doctor of Philosophy in Neuroscience
University of California, Los Angeles, 2015
Professor Daniel H. Geschwind, Chair

Autism spectrum disorder (ASD) is a group of etiologically and phenotypically heterogeneous neurodevelopmental disorders defined by deficits in social communications and mental flexibility. It is established that variation in hundreds of genetic loci contributes substantially to ASD risk. This has raised the question of whether these mutations, which are found across disparate genes, affect similar biological pathways, perturb brain development at a particular time point, or disrupt a specific brain system. Addressing this is critical to develop a molecular and neurobiological understanding of ASD. However, searching for such convergence is made challenging by the genetic complexity of ASD and the molecular, cellular, and circuit-level complexity of the brain.

Here, I apply comprehensive profiling of RNA levels (the transcriptome) by RNA sequencing (RNA-seq) to characterize the role of genes in normal human brain development and ASD. My overarching hypothesis is that there exist molecular regulators of transcription which are particularly susceptible to mutations in ASD, and that their regulatory targets are affected in ASD brain. My general approach is to organize the hundreds to thousands of disparate changes in transcript levels into more tractable and biologically meaningful gene sets or modules that are differentially expressed, co-expressed, or co-regulated. I then integrate these gene sets or modules with molecular and phenotypic information from whole genome studies and targeted experimental studies in order to systematically reveal new insights about ASD neurobiology and highlight specific genes and pathways worth investigating further. I first assess the role of ASD risk genes in normal brain development and then apply RNA-seq to measure transcriptomic changes in postmortem ASD brain to evaluate whether convergent neurobiological pathways are affected.

I find robust evidence that developmentally co-expressed, co-regulated, and physically interacting genes are affected in ASD during normal brain development. Rare, highly deleterious variants predominantly exert their effect by disrupting major transcriptional and chromatin regulators in early fetal development, while less deleterious inherited variants affect late prenatal and early postnatal cellular and circuit maturation through alterations in synaptic function.

Additionally, across most individuals with ASD, I find strongly shared changes in synaptic and neuronal genes at both a gene expression and transcript splicing level in cortex. Moreover, shared cortical changes are also seen in a genetically defined subtype of autism, duplication 15q syndrome (dup15q). In contrast, ASD-associated changes in cerebellum are weaker. Co-expression network analysis identifies specific cell types and circuits that are affected, and highlights specific transcriptional regulators likely to play a role in ASD pathology. Taken together, these results identify roles for transcriptional regulators in ASD and define the potential consequences of their dysregulation.

The dissertation of Neelroop Narendra Parikshak is approved.

Eleazar Eskin

Steve Horvath

Nelson B. Freimer

X. William Yang

Daniel H. Geschwind, Committee Chair

University of California, Los Angeles

2015

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

There are no figures or tables in this section.

GLOSSARY

Adjacency matrix – a pairwise matrix of node-node connectivity that quantifies all possible edge strengths in a network. Adjacency matrices can be combined across multiple types of data and may be mathematically transformed to improve network clustering and predictive power

Autism Spectrum Disorder – abbreviated ASD, a collection of heterogeneous neurodevelopmental disorders that share a deficit in social communication and mental flexibility relative to the general population. ASD is frequently (>30%) comorbid with epilepsy and intellectual disability (low IQ). Additional comorbidities include gastrointestinal distress, sensory hypersensitivity, attention deficit hyperactivity disorder, and other neuropsychiatric conditions.

Binary network – a network where edges are 1s and 0s, either by the nature of the edge measurement (e.g. physically interacting or not) or by thresholding an otherwise continuous measurement to assign 1 only if the values pass the cut-off

Causal anchor – a network node that is not affected by variation in other nodes, and can therefore be used to orient edges and transform an undirected correlational network to a directed causal network. In gene networks, genotypes can be used as causal anchors to understand the direction of causation between other variables, such as gene expression or methylation levels.

ChIP-seq – chromatin immunoprecipitation followed by high-throughput sequencing; allows elucidation of binding sites of a protein on DNA in a genome-wide manner

CLIP-seq – cross-linking immunoprecipitation followed by high-throughput sequencing; allows elucidation of binding sites of a protein on RNA in a genome-wide manner

Differential gene expression analysis – abbreviated DGE, an approach commonly used in transcriptomic studies that serially compares thousands of genes between groups (e.g. disease and controls) to evaluate the mean difference and its significance for each gene independently.

Differential splicing analysis – abbreviated DS, an approach where, instead of evaluating gene expression level differences, transcript splicing events are quantified and compared between conditions

Edges – the relationships between nodes in a network delineating some measure of shared function (e.g. co-expression, computationally predicted binding sites, physical interaction)

Eigengene – a module-level summary of expression, calculated by taking the first principal component of the expression levels of genes in the module. In co-expression networks, genes in a module are highly correlated by definition, leading this one vector to explain a high proportion of gene expression variation in the module

Expression quantitative trait locus analysis – abbreviated eQTL, an association analysis of genome-wide SNPs on genome-wide expression levels in a population that identifies the causal effect of changes in genetic loci on gene expression

Gene network - a graph consisting of genes as nodes connected by edges that reflect a measure of shared function between the connected genes

Gene set enrichment – an analysis approach that assesses the statistical significance of the overlap between two gene sets, usually one set is an annotated reference and the other is a set of interest that is unannotated

Genetic architecture – the relationship between the allele frequency and effect size across all genetic loci contributing to a given trait

High-throughput sequencing – also called “next generation” sequencing or “massively parallel high-throughput sequencing of short reads” this approach refers to fragmenting long sequences of nucleotides (DNA or RNA), clonally amplifying fragments in a manner that is tractable by imaging technology, and reading out the bases comprising each fragment in parallel. This yields millions to billions of short read sequences, which are either assembled into genomes or transcriptomes, or aligned to existing reference genomes or transcriptomes to understand molecular changes in a genome-wide manner.

Hub –genes within a module that have high intramodular connectivity relative to other genes

Module – a highly inter-connected subset of genes in a gene network, for example, genes in a transcriptomic network sharing highly similar patterns of gene expression. Modules are also known as clusters, cliques, or communities.

Molecular systems or omic approach - systems biology methods that include high-throughput quantification, analysis, and interpretation of the genome, transcriptome, epigenome, proteome, and other ‘omens’ as well as the relationships between omic levels

Mutual information – a nonlinear measure of dependence between two variables that may capture patterns linear measures, such as Pearson correlation, cannot accurately detect

Negative symptoms – a mental state defined by a loss of normal emotional responses including a lack of motivation, an inability to experience pleasure, and reduced expression through speech

Nodes – the molecular entities that comprise a network, e.g. genes in a gene network

Psychosis – a mental state defined by a loss of contact with reality characterized by exaggerations or distortions of normal perception

RNA-seq – extraction of RNA followed by construction of cDNA libraries that undergo high-throughput sequencing; allows elucidation of transcript levels in a genome-wide manner

Seeded (prior-based) network – network analysis approach where edges are "grown" around selected genes of interest or "seed" genes. Network structure and modules are dependent on these initial genes of interest.

Selective vulnerability – the relative susceptibility of brain regions, cell populations, or time points to genetic or environmental insults that can be leveraged to identify vulnerable and protective molecular pathways

Signed network – a network where the sign is taken into consideration in addition to the magnitude of the correlation, e.g. high positive correlations are assigned high edge values, but high negative correlations are assigned low edge values

Small N, large p – small sample sizes (N) but many features or parameters (p). This is the case in statistical and big data analysis when the number of features or predictors (p) is equal to or much larger than the number of samples (N), and requires special considerations to prevent overfitting statistical models.

Systems neuroscience – area of neuroscience that focuses on short- and long- range circuits. This area also applies many systems biology methods, but at a higher level of organization in the nervous system than the molecular systems approach

Topological overlap – a transformation of edge relationships in a network that makes network edges reflect shared neighbourhoods between nodes instead of direct pairwise relationships

Unseeded (genome-wide) network – network analysis approach where unbiased genome-wide data are clustered into modules and genes of interest are studied for their position in these modules. Network structure and modules are independent of genes of interest.

Unsigned network – a network where any high magnitude association is assigned a high edge value (e.g. the absolute value of the correlation)

Unsupervised analysis – a prior-free analysis approach that uses the intrinsic variation in data to define shared patterns (e.g. hierarchical clustering). This can identify novel clusters or groupings of data points.

Weighted network – a network where the edges retain continuous values, with higher values reflecting an increased strength or probability of connectivity

ACKNOWLEDGEMENTS

I would like to thank the key individuals who provided me with the intellectual, financial, and emotional resources to pursue this work. I am most grateful for the guidance provided to me by Dan Geschwind, my primary mentor (and an interviewer for my admission to the MD/PhD program at UCLA), who gave me the flexibility to pursue interesting work but provided the guidance necessary to complete my graduate work in a timely manner. I am particularly thankful for the invaluable opportunities that he has provided. These include the writing of several manuscripts and reviews, the chance to participate or speak at multiple international conferences, and the privilege to work with the best minds in our field.

I would further like to express my gratitude to other individuals who supported my career to this point, including Kelsey Martin who served as my PhD mentor in the MD/PhD program, Steve Horvath and Peter Langfelder who fielded occasional nonsensical questions I might have had about statistics and network analysis, and my committee members, Eleazar Eskin, Nelson Freimer, and William Yang, who provided valuable guidance within each of their areas of expertise. I am very fortunate to be at UCLA, where this diverse group of mentors is willing to take their time and train future researchers. Finally, Lauren Kawaguchi, Jenifer Sakai, and Margaret Chu ensured the timely execution of all things during my graduate work infinitely easier.

I am also extremely grateful to my peers in the lab, particularly all of the co-authors mentioned in the Chapter attributions below, and throughout this thesis. In particular, though, I would like to thank Jeremy Miller for having guided me in my initial endeavors with network analysis, T. Grant Belgard for helping me refine my bioinformatics skills, and Jason L. Stein, Vivek Swarup, and Michael J. Gandal for helping me tackle challenging problems as they arose.

I also thank Jennifer K. Lowe, Luis de la Torre Ubieta, Hyejung Won, and Donna M. Werling for invaluable discussions regarding genetics and neurobiology.

I would like to acknowledge that Chapter 1 contains significant content from, "Systems biology and gene networks in neurodevelopmental and neurodegenerative disorders," a manuscript currently under review at *Nature Reviews Genetics*, written in collaboration with Daniel H. Geschwind. The results in Chapter 2 primarily come from published work from "Integrative Functional Genomic Analyses Implicate Specific Molecular Pathways and Circuits in Autism" published on November 12, 2013 in *Cell*, Volume 155, Issue 5, p 1008-1021. doi: 10.1016/j.cell.2013.10.031. Cell Press©. This work was co-authored by Rui Luo, Alice Zhang, Hyejung Won, Jennifer K. Lowe, Vijayendran Chandran, Steve Horvath, and Daniel H. Geschwind, all of who have consented to the inclusion of this work in this thesis. In this work, I performed network analysis, curated and analyzed gene sets, interpreted data, and wrote the manuscript with Daniel H. Geschwind, my thesis advisor. Results in Figures 4 comes partially from transcription factor binding analyses performed by Alice Zhang and Vijayendran Chandran, and results in Figures 5 and 6 were computed by Rui Luo. The work in Chapter 3 is an extension of the work in Chapter 2, and was spurred by discussions with Daniel H. Geschwind about evaluating the reproducibility of our findings.

The work in Chapter 4 is from a manuscript in preparation, with the working title "Dysregulation of the transcriptome in autism spectrum disorder". This work was done in collaboration with Vivek Swarup, T. Grant Belgard, Jennifer K. Lowe, Virpi Leppä, Steve Horvath, and Daniel H. Geschwind. All of who have consented to its inclusion in this thesis. I compiled all results and wrote the manuscript with Daniel H. Geschwind. T. Grant Belgard and I performed the dissections; T. Grant Belgard, Vivek Swarup, and I collaborated on the preprocessing of RNA-seq data; Vivek Swarup performed initial splicing analysis; and T. Grant Belgard and Virpi Leppa performed initial analysis of the dup15q samples. I performed all final statistical analyses, assembled figures, and wrote the initial draft of the manuscript, which is what is included in this thesis, with guidance from Daniel H. Geschwind. Importantly, this is an incomplete work and requires further and more detailed analysis and experiments in several areas discussed at the end of Chapter 4.

I gratefully acknowledge sources of financial support for this work. The first year (2010-2011) was primarily supported by the Medical Scientist Program at UCLA, the second year (2011-2012) primarily by a Neurobehavioral Genetics training grant (T32MH073526), and the last two-and-a-half (2012-2015) by a Ruth L. Kirchstein National Research Service Award from the National Institute of Mental Health (F30MH099886). The projects in this work were funded partly by grants to the BrainSpan consortium (NIMH grant 5RC2 MH089921), an R01 to Daniel H. Geschwind (5R01MH060233), and an Autism Center for Excellence network grant (5R01MH100027). I would also like to acknowledge the resources that made this work possible: The Simons Simplex Consortium (www.sfari.org), the Autism Genetics Resource and Exchange (www.agre.org), the University of Maryland Brain and Tissue Bank (www.medschool.umaryland.edu/btbank), and the Autism Tissue Program at the Harvard Brain Bank (www.autismtissueprogram.org).

Finally, I would like to thank my family, who supported me in both my professional and non-professional endeavors throughout this time. Without the influence and encouragement of my parents, Niranjana and Narendra, and my brother, Harshraj, I would not be interested in neuroscience. Moreover, without their ongoing support I would be much less likely to succeed in my efforts. Despite the individual hard work and effort I have put into my graduate work, I believe my efforts build on those who came before me and are supported by those around me, and therefore apologize to whomever I may have failed to acknowledge in their contribution to this work.

VITA

EDUCATION

08/08 to present	David Geffen School of Medicine at UCLA, Los Angeles, California Medical Scientist Training Program Medical Student at David Geffen School of Medicine Doctoral Training Program via the Neuroscience Interdepartmental Program
08/03 to 05/07	Rice University BA in Mathematics BA in Biochemistry and Cell Biology with Honors

ACADEMIC AND PROFESSIONAL HONORS

Pre-doctoral Fellowship - National Institutes of Mental Health F30 NRSA (2012-2015)
Pre-doctoral Training Award - Neurobehavioral Genetics T32 Training Grant (2011-2012)
Undergraduate Research Award - Recipient, Summer Undergraduate Research Program in Neuroengineering, UCLA (2005)
Pre-college National Awards: National AP Scholar (2003), National Merit Finalist (2003)

RESEARCH ARTICLES

1. Suthana, NA, **Parikshak, NN,** Ekstrom AD, Ison, M, Knowlton, B, Bookheimer SY, and Fried, I. Pattern separation in human hippocampal neurons is associated with better memory. (*under revision*)
2. O'Dushlaine C, Rossin E, Lee P, Duncan L, Parikshak NN, Newhouse S, et al. Psychiatric genome wide association study analyses implicate neuronal, immune and histone pathways. *Nature Neuroscience*, in press for 2/2015.
3. Irimia M, Weatheritt RJ, Ellis J, **Parikshak NN**, Gonatopoulos-Pournatzis T, et al. (2014). A Highly Conserved Program of Neuronal Microexons Is Misregulated in Autistic Brains. *Cell*, in press. http://dx.doi.org/10.1016/j.cell.2014.11.035
4. Stein JL, de la Torre-Ubieta L, Tian Y, **Parikshak NN**, Hernández IA, Marchetto MC, Baker DK, Lu D, Hinman CR, Lowe JK, et al. (2014). A Quantitative Framework to Evaluate Modeling of Cortical Development by Neural Stem Cells. *Neuron 83*, 69–86.
5. **Parikshak NN,** Luo R, Zhang A, Won H, Lowe JK, Chandran V, Horvath S, Geschwind DH. (2013). Integrative Functional Genomic Analyses Implicate Specific Molecular Pathways and Circuits in Autism. *Cell*, 155(5), 1008–1021. doi:10.1016/j.cell.2013.10.031 PMCID: PMC3107252
6. Fogel, BL, Wexler EM, Wahnich, A, Friedrich, T, Vijayendran, C, Gao, F, **Parikshak, N**, Konopka, G, Geschwind DH (2012). RBFOX1 regulates both splicing and transcriptional networks in human neuronal development. *Human Molecular Genetics*.
7. Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ, Ercan-Sencicek AG, Dilullo NM, **Parikshak NN**, Stein JL, Walker MF, Ober GT, Teran NA, Song Y, El-Fishawy P, Murtha RC, Choi M, Overton JD, Bjornson RD, Carriero NJ, Meyer KA, Bilguvar K, Mane SM, Šestan N, Lifton RP, Günel M, Roeder K, Geschwind DH, Devlin B, and State MW (2012). De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature*. DOI: 10.1038/nature10945.
8. Lu PH, Thompson PM, Leow A, Lee GJ, Lee A, Yanovsky I, **Parikshak N**, Khoo T, Wu S, Geschwind DH, Bartzokis G. Apolipoprotein E genotype is associated with temporal and hippocampal atrophy rates in healthy elderly adults: a tensor-based morphometry study. *Journal of Alzheimer's Disease: JAD*, *23*(3), 433–442. PMID: 21098974; PMCID: PMC3107252.
9. Raji CA, Ho AJ, **Parikshak N**, Becker JT, Lopez OL, Kuller LH, Hua X, Leow AD, Toga AW, Thompson PM. Brain Structure and Obesity. *Hum Brain Mapp.* 2009 Aug 6. PMID: 19662657; PMCID: PMC2826530.

10. Leow AD, Yanovsky I, **Parikshak N**, et al. Alzheimer's disease neuroimaging initiative: a one-year follow up study using tensor-based morphometry correlating degenerative rates, biomarkers and cognition. *Neuroimage*. 2009;45(3):645-655. PMID: 19280686; PMCID: PMC2696624.
11. Morra JH, Tu Z, Apostolova LG, Green AE, Avedissian C, Madsen SK, **Parikshak N**, et al. Automated 3D mapping of hippocampal atrophy and its clinical correlates in 400 subjects with Alzheimer's disease, mild cognitive impairment, and elderly controls. *Hum Brain Mapp*. 2009. PMID: 19172649; PMCID: PMC2733926.
12. Morra JH, Tu Z, Apostolova LG, Green AE, Avedissian C, Madsen SK, **Parikshak N**, et al. Automated mapping of hippocampal atrophy in 1-year repeat MRI data from 490 subjects with Alzheimer's disease, mild cognitive impairment, and elderly controls. *Neuroimage*. 2009;45(1 Suppl):S3-15. PMID: 19041724; PMCID: PMC2733354.
13. Chou Y, Leporé N, Avedissian C, Madsen SK, **Parikshak N**, et al. Mapping correlations between ventricular expansion and CSF amyloid and tau biomarkers in 240 subjects with Alzheimer's disease, mild cognitive impairment and elderly controls. *Neuroimage*. 2009;46(2):394-410. PMID: 19236926; PMCID: PMC2696357.
14. Aganj I, Sapiro G, **Parikshak N**, Madsen SK, Thompson PM. Measurement of cortical thickness from MRI by minimum line integrals on soft-classified tissue. *Hum Brain Mapp*. 2009. PMID: 19219850; PMID: 19219850.
15. Hua X, Leow AD, **Parikshak N**, et al. Tensor-based morphometry as a neuroimaging biomarker for Alzheimer's disease: an MRI study of 676 AD, MCI, and normal subjects. *Neuroimage*. 2008;43(3):458-469. PMID: 18691658; PMCID: PMC3197851.
16. Morra JH, Tu Z, Apostolova LG, Green AE, Avedissian C, Madsen SK, **Parikshak N**, et al. Validation of a fully automated 3D hippocampal segmentation method using subjects with Alzheimer's disease mild cognitive impairment, and elderly controls. *Neuroimage*. 2008;43(1):59-68. PMID: 18675918; PMCID: PMC2624575.

BOOK CHAPTERS AND REVIEW ARTICLES

1. **Parikshak, NN**, Geschwind, DH. Systems biology in neurobiological disorders: from omics to networks. *Nature Reviews Genetics. Invited review, under revision.*
2. Stein, JL, **Parikshak, NN**, Geschwind, DH. Rare inherited variation in autism: beginning to see the forest and a few trees. *Neuron*. 2013;77(2):209-11. PMCID: PMC3691080
3. **Parikshak NN**, Geschwind DH. Overview of Neuroscience and Genomics (Chapter 95). In *Genomic and Personalized Medicine, Second Edition: V1-2* 2nd ed, Academic Press. Willard, H.P.D. & Ginsburg, G.S. eds., 2012.

PRESENTATIONS AT NATIONAL/INTERNATIONAL MEETINGS

1. Poster: "Integrative functional genomic analyses implicate specific molecular pathways in autism." 2014 Gordon Research Conference on Fragile X and Autism Related Disorders. (June 2014). Contributors: R. Luo, S.H. Horvath, and D.H. Geschwind.
2. *Invited talk:* "Gene expression from neurotypical and autistic cortex – insights into the molecular underpinnings of autism" 2014 Gordon Research Conference on Fragile X and Autism Related Disorders. (June 2014). Contributor: D.H. Geschwind.
3. Poster: "Integrative functional genomic analyses implicate specific regulators in molecular pathways in autism." Cell Symposium on The Networked Brain. (November 2013). Contributors: S. Horvath and D.H. Geschwind.
4. *Invited talk:* "Dissecting Complexity in Neuropsychiatric Genetics with Network Inference." Presentation as part of session: "Are we at a turning point in psychiatric genetics?" 2012 Annual Meeting, American College of Neuropsychopharmacology 2012 Annual Meeting. (December 2012). Contributor: D.H. Geschwind.

5. Poster: “Weighted correlation network analysis of miRNA and mRNA in human brain.” Meeting on Regulatory and Noncoding RNA, Cold Spring Harbor Laboratory. (August 2012). Contributors: S. Horvath and D.H. Geschwind.

CHAPTER 1:
Molecular systems biology, transcriptomics, and neuroscience

"The main thesis is that developmental reactions, *as they occur in organisms submitted to natural selection*, are in general canalized. That is to say, they are adjusted so as to bring about one definite end-result regardless of minor variations in conditions during the course of the reaction."

― Conrad H. Waddington (emphasis in italics is his), Canalization of Development and the Inheritance of Acquired Characters, *Nature*, 1942.

1.1: Introduction

The human brain is a remarkably complex organ; it comprises on the order of 100 billion neurons and an equivalent number of glia that communicate by approximately 1 trillion synapses (Azevedo et al., 2009). Moreover, these neurons and glia are not homogeneous, there are hundreds of cell types that vary in cellular morphology, synaptic connectivity, electrophysiological properties, and molecular make-up (Masland, 2004; Nelson et al., 2006; Stevens, 1998). The intricate process by which these cells organize into circuits is exemplified by the development of the neocortex, during which molecular programs encoded by the genome regulate the biological processes of cellular proliferation, differentiation, and migration to define cell fate and produce this computational organ (Angevine et al., 1970; Bystron et al., 2008; Greig et al., 2013).

In humans, despite individual variability at genetic loci and in specific cell-cell connections, microcircuits, and even some long-range circuits (Meredith et al., 2011; Saenz et al., 2008), the brain is a largely similar structure with a highly consistent functional organization across most individuals. Additionally, despite the considerable genetic and phenotypic variation across humans, global behavioral and cognitive functioning throughout life follows a stereotypical pattern that can be called typical development and aging (Figure 1.1).

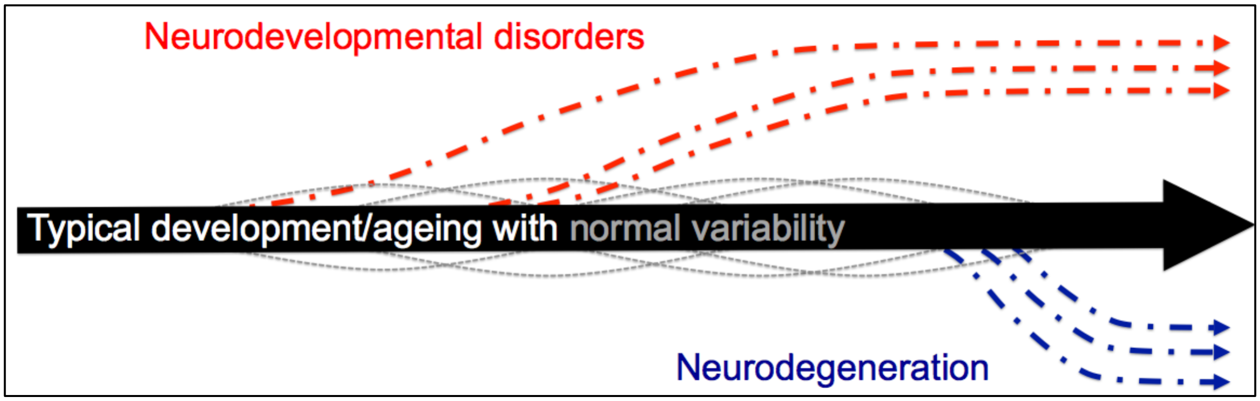

Neurodevelopmental disorders
Typical development/ageing with normal variability
Neurodegeneration

Figure 1.1 Diagram of normal and abnormal development and aging of the brain. Both brain development and aging exhibit a typical trajectory (black arrow) with normal variation (grey dotted lines) in phenotypes. Variability across individuals may manifest at a molecular, cellular, circuit-level, cognitive, or behavioural level. Genetic and environmental factors can alter this trajectory substantially, causing disorders and diseases that manifest as abnormal phenotypic trajectories (red for neurodevelopmental, blue for neurodegeneration).

This minor variation but overall consistency is reminiscent of canalization, a theory of development in naturally selected organisms proposed by Conrad H. Waddington (Waddington, 1942). Applied to the brain, the theory of canalization predicts robustness to minor genetic, environmental, or stochastic perturbations during normal development, allowing for a certain degree of variability but leading to the same broad phenotypic outcomes: typical human behavior and cognition. When a severe insult perturbs brain development, aberrant phenotypes that are seen as disorders or diseases manifest. Therefore, in order to understand abnormal human brain development from a molecular perspective, it is critical to understand the robust spatial and temporal patterns that are characteristic of typical development, and then understand how they might be affected by genetic or environmental insults during development. In this chapter, I review the utility of transcriptomics, co-expression networks, and other areas of high-throughput biology for understanding normal brain function and its perturbations. These approaches lay the foundation for the remainder of my work in later chapters.

1.2: Background

The genome of an individual organism is largely similar across all of its cells (with the exception of somatic mosaicism (De, 2011; Poduri et al., 2013) though the following points apply just the same). Cellular diversity – the cell's structure, function, and responses to stimuli – is manifested through diverse transcriptional and epigenetic programs that are initiated and maintained by transcription factors (TFs), chromatin regulators, and other regulatory molecules that alter methylation, chromatin marks, and chromatin folding over the lifespan of the cell. How a given gene is expressed at the RNA level is an important intermediate step in this genome to function relationship, and RNA transcript levels can serve as a valuable marker for understanding how genes and environment affect the molecular composition of a cell, tissue, and organism to result in phenotypes (Figure 1.2).

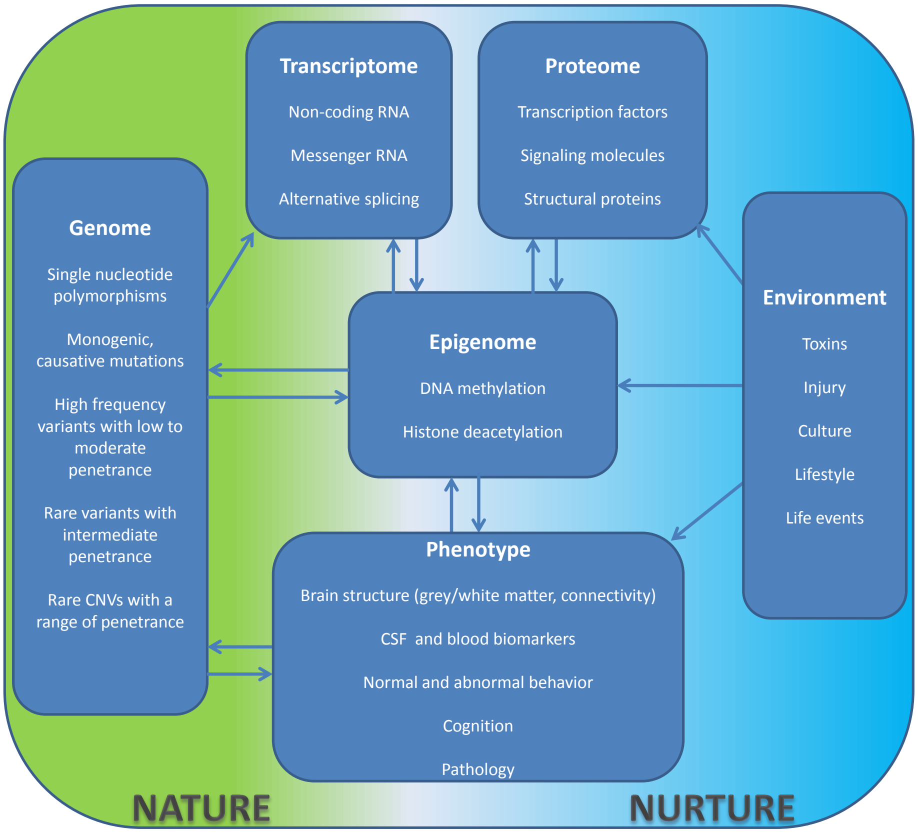


Figure 1.2 Diagram of measurable molecular levels and their distribution across the nature versus nurture spectrum. Each box delineates a category of what might be measured, with specific examples. Genetic factors are at the most extreme nature end while environmental factors at the most extreme nurture end. Arrows delineate the possible flow of information between these multiple levels of measurement. The transcriptome is an intermediate in this hierarchy, and can capture changes in the genome, epigenome, proteome, environment, and many phenotypes.

The measurement of gene expression levels has been an extremely valuable approach in multiple areas of biology and encompasses detection of individual transcripts using the Northern blot, resolution of the spatial distribution of transcripts with *in situ* hybridization (ISH), and more precise quantification of transcript levels with quantitative real-time PCR (qRT-PCR). Recently, the genomic era has produced unbiased, quantitative, and high-throughput approaches to simultaneously quantify hundreds to thousands of transcripts, together called the transcriptome, with microarray technology (Geschwind and Gregg, 2002) and massively parallel high-throughput sequencing technologies (Wang et al., 2009b).

Multiple studies have demonstrated that protein-coding gene expression, transcript splicing, and non-coding RNA levels provide a global molecular phenotype that reflects the state of the cellular or tissue system being analysed (Carter et al., 2004; Guttman et al., 2009; Khalil et al., 2009; Wang et al., 2008). In human biology, investigations in cancer biology have been particularly successful in utilizing transcriptomics to understand differential gene expression between conditions, stratify heterogeneity in disease, and identify the consequence of genetic variation on gene expression (Li et al., 2013; Rhodes and Chinnaiyan, 2005; Vaske et al., 2010). This, along with several exciting studies in human brain over the past decade (Oldham et al., 2008; Ramasamy et al., 2014; Rhinn et al., 2013; Rosen et al., 2011; Torkamani et al., 2010; Voineagu et al., 2011; Zhang et al., 2013), has suggested that similar methods, and high-throughput molecular biology in general, are poised to revolutionize our understanding of the human brain (Geschwind and Konopka, 2009; Grant, 2003).

However, application of transcriptomics in disease-relevant neuroscience research continues to lag behind other areas such such as immune and cancer biology. This relative immaturity can be attributed to several major challenges encountered in studying the brain: 1) complexity of molecular phenotypes due to cell-type and regional heterogeneity; 2) extensive temporal changes occurring throughout nervous system development and maturation; 3) a dearth of human tissue and model systems that have definitive human relevance (the “translational” and “evolutionary” problems) and 4) poor prior knowledge of appropriate phenotypes for disorders and diseases.

Though these hurdles are not unique to the brain, neuroscience has historically struggled with each of these points due to the extent they affect the ability to ask questions about brain function. Foundational aspects of each point are not agreed upon: the definition of a cell type in brain remains controversial, the relationships of human disease phenotypes to developmental trajectories are relatively unknown, the model systems in many neurobiological studies are often chosen as a matter of convenience and history, and most phenotypes are based on clinical and behavioral symptomatology rather than biological mechanism or etiology (Casey et al., 2013; Geschwind, 2008; Insel et al., 2010). However, these challenges have been steadily addressed and even leveraged to understand neurobiology by applying the appropriate study designs and analytical methods.

1.2.a: Spatial and temporal heterogeneity in the brain

In order to understand why spatial heterogeneity is a challenge and how this challenge can be addressed, it is important to recognize the immense cellular diversity of the human brain and its consequences. The staggering diversity of the cell type morphology in the nervous system was first appreciated over a century ago by Ramon y Cajal. Decades later, studies in the retina and brain connected morphologically distinct mammalian CNS cell types to molecular markers (Barnstable and Dräger, 1984; Hockfield and McKay, 1985). This regional and cellular heterogeneity poses distinct obstacles for transcriptomic studies in the CNS (Coppola and Geschwind, 2006; Mirnics and Pevsner, 2004), as isolation of individual circuits of cell types in brain requires knowledge of the poorly defined molecular identity of these components. The diversity of cell-types also complicates comparisons across macroscopic brain structures, which can have vastly different molecular architecture. However, such differences are necessary to make in order to understand disorders of neural circuitry that lack a focal neuropathological locus in the brain, such as ASD.

For neurodevelopmental disorders, the specific brain regions, cell types, or time points that are most affected are particularly poorly defined or unknown. In this context, whole tissue profiling still has a major role: even if changes in cell populations at a gross anatomical level drive molecular changes and obscure underlying mechanistic changes, these alterations can point to selective vulnerability and unidentified circuits. In contrast, most neurodegenerative diseases have known neuropathological and brain-imaging changes accompanied by the well-defined death of selective cellular populations and infiltration of inflammatory cells. Therefore, transcriptional changes in later neurodegeneration often reflect changes in cell type composition, strongly obscuring the key disease-initiating molecular pathway changes within cells(Mirnics and Pevsner, 2004).

Understanding transcriptomic changes in the context of this spatial complexity in human CNS disorders, and the regional vulnerability that is a fundamental characteristic of neurodegenerative disorders necessitates a thorough knowledge of the molecules that distinguish different brain regions, circuits, and cells in normal brain development and aging. Extensive work by the Allen Brain Institute and others has mapped out the mouse, non-human primate, and primate spatial heterogeneity using genetically targeted cell sorting or microdissection coupled with transcriptomics (Srinivasan et al., 2012; Zhang et al., 2014), or large-scale *in situ* hybridization (ISH) (Lein et al., 2006; Thompson et al., 2014; Zeng et al., 2012), and other complementary methods (Sunkin et al., 2013). The utility of cell type specific gene expression for understanding neurodevelopmental and neurodegenerative disease is highlighted by several recent publications (Dalal et al., 2013; Dougherty et al., 2013; Heiman et al., 2014; Xu et al., 2014).

These studies have provided a critical first pass at the major cell-type and neuroanatomical markers in brain, but the vast majority of cell-types and their markers have yet to be elucidated. For example, there is immense morphological and functional diversity in cortical GABAergic interneurons (DeFelipe et al., 2013) and more than 70 cell types in the mammalian retina, many of which are morphologically indistinguishable (Siegert et al., 2009). Even the current framework for understanding well-defined excitatory neuronal populations is coarse. In fact, each cortical layer is anticipated to have ~100 distinct cell types (Masland, 2004) with up to 1000 total cell types estimated in the cerebral cortex (Stevens, 1998). If the relationship between functional cell type and molecular identity from the retina generalize (Siegert et al., 2009), there are likely hundreds of cell-type specific molecular barcodes left to be defined for the human cortex.

1.2.b: Challenges due to the unique cytoarchitecture and development of the brain

The molecular architecture of the brain also changes throughout development and aging, delineating periods of cellular growth, migration, differentiation, and maintenance that are vulnerable to genetic and environmental insults (Andersen, 2003). Cell markers at one stage of development may not apply to another stage (Bystron et al., 2008; Molyneaux et al., 2007), and some major cell-type features, such as neurotransmitter phenotype, may require maintained expression of specific transcription factors (TFs) (Deneris and Hobert, 2014). Furthermore, comparison of the transcriptome across neurodevelopmental stages reveals striking changes in gene expression and alternative splicing of most genes in the human genome (Colantuoni et al., 2011; Johnson et al., 2009; Kang et al., 2011), with a key inflection point at birth that is marked by large-scale changes in gene expression and epigenetic programs (Colantuoni et al., 2011; Kang et al., 2011; Lister et al., 2013).

Recent work suggests substantial transcriptomic differences during the development of three major populations of pyramidal neurons in mouse cortex, the callosal projection neurons, corticothalamic projection neurons, and subcerebral projection neurons (Molyneaux et al., 2015), highlighting the importance of resolving both spatial and temporal specificity. Importantly, this dynamic regulation is not restricted to prenatal and childhood periods. Relative to rodent or primate brains, the human brain exhibits prolonged development, with biological processes such as synaptic pruning and stabilization extending into the third decade of life (Changeux and Danchin, 1976; Geschwind and Rakic, 2013).

Given the massive changes in gene expression that occur over development, network approaches (Hawrylycz et al., 2012; Kang et al., 2011; Miller et al., 2014; Oldham et al., 2008) play an increasingly important role in organizing transcriptomic data and relating genes and pathways to neuroanatomical regions and critical time points that are of particular importance to neurodevelopmental disorders.

Additionally, the shift from microarrays to high-throughput sequencing has revealed the neurodevelopmental importance of noncoding and regulatory regions that express lncRNAs (Fogel et al., 2012; Konopka et al., 2009) or are modified at the chromatin level (Lessard et al., 2007; Tuoc et al., 2013; Yamada et al., 2014). Additionally, until recent genome-wide experimental analysis of mammalian telencephalic enhancers (Attanasio et al., 2013), there was scant knowledge of the scope and architecture of the developmental regulatory networks underlying forebrain development (Pattabiraman et al., 2014; Visel et al., 2013). Many of these noncoding regulatory regions exhibit signatures of accelerated evolution (Pollard et al., 2006;

Prabhakar et al., 2006), providing a better framework is critical for understanding human or primate specific cortical specializations (Capra et al., 2013; Geschwind and Rakic, 2013) which are essential to understanding reveal the mechanistic drivers of human brain evolution.

Neurons also undergo rapid state dependent changes at a faster time scale related to activity-dependent transcription and translation (Ebert and Greenberg, 2013) and neuronal activity can induce widespread expression of noncoding enhancer regions of the genome (Kim et al., 2010) as well as chromatin-level changes that are associated with the recruitment of activity-dependent transcription factors (Malik et al., 2014). Further complexity is highlighted by the role of locally regulated translation of sub-cellular transcriptomes (Crino and Eberwine, 1996), which has been demonstrated to play a critical role in synaptic function and plasticity (Wang et al., 2010). Deeper characterization of these events at a high spatiotemporal resolution will be necessary, and integrating them with transcriptional data, experimental perturbations, and genetic variation across individuals will help establish a mechanistic foundation for understanding their dysregulation in disease**.**

1.2.c: An overview of transcriptional networks

The architecture of gene expression networks was initially investigated in yeast(Langfelder and Horvath, 2008) and across species in an evolutionary context (Stuart, 2003). Multiple metabolic and protein interaction network studies demonstrated that biological network architecture can be modeled by an approximate scale free topology, which was applied to characterize gene co-expression networks (Barabasi, 2009; Barabási and Oltvai, 2004; Zhang and Horvath, 2005).

From a practical perspective, network based methods reduce the dimensionality of genome-wide RNA or protein expression patterns (Carter et al., 2013; Shirasaki et al., 2012), using correlation, mutual information, or other metrics to organize thousands of genes corresponding to millions of relationships into a relatively small collection of modules (Allen et al., 2012; Margolin et al., 2006; Zhang and Horvath, 2005). Modules, also known as cliques or communities, correspond to groups of genes sharing expression patterns across the experimental observations. Modular organization reflects the higher-order structure of biological relationships, while local modular organization can identify network hubs within modules (Liu et al., 2011). These network hubs may be key drivers or representatives of the biological processes represented by individual modules (Horvath et al., 2006; Oldham et al., 2008; Voineagu et al., 2011).

In general, omic data can be modelled as a network in which molecules or genes are nodes and their functional relationships with each other are edges. Gene network analysis can be summarized in five basic steps:

1) Node specification:
   a. Seeded (prior-based) – nodes are selected using prior knowledge, e.g. disease-associated genes from genome-wide association studies
   b. Unseeded (genome-wide) – all usable measurements from the genome are included in an unsupervised analysis
2) Edge specification:
   a. experimentally observed pairwise statistical relationships(Butte and Kohane, 2000; Carter et al., 2004; Horvath, 2011) evaluating shared patterns of molecular levels across experiments: e.g. co-expression
   b. experimentally observed or literature-curated physical interactions: e.g. protein interactions from immunohistochemistry and Y2H experiments

c. computationally predicted relationships: e.g. transcription factor binding based on DNA motifs

Notably, edges are susceptible to ascertainment biases (Hakes et al., 2008; Lee and Marcotte, 2009) and confounding factors that can induce spurious relationships (Leek et al., 2010).

3) Modules are identified from an adjacency matrix to simplify biological relationships at a higher-order level. Assessing connectivity can identify hubs and enables comparison of changes between health and disease at the module level.

4) Annotation of modules and gene connectivity by:

a. Relating external measures of gene importance (such as cell-type specificity or GWAS signal) with module membership, intra-modular connectivity, or whole-network connectivity of genes

b. Associating module summary or hub gene measurements, such as eigengenes or average expression levels, to biological traits

c. Assessing preservation or changes in network connectivity for specific genes or modules between health and disease

d. Integrating data at the edge level or the module level across biological levels, such as different cell types or brain regions, or different regulatory levels, such as gene expression and ChIP signal

e. Addressing the crucial issue of reproducibility by validating network observations in independent data or experiments (see Table 1.1 for examples)

Table 1.1 – Different edge types in gene networks: practical and theoretical considerations.

	Gene co-expression	Protein-protein	Motif enrichment

		interaction	**for transcription factors**
Edge relationships (specific example)	Statistical association (correlation, topological overlap)	Physical binding (interacting or not interacting)	Computational inference (motif binding scores)
Main advantages	Indirectly predicts co-regulation, physical interactions, and cell-type specificity, comes from tissue of interest	Based on direct physical interactions, predicts protein complexes	Identifies putative co-regulatory relationships without needing to perform new experiments
Completeness of data across the genome	Genes in most studies are similarly covered genome-wide	Incomplete assessment for most interactions, biased toward most-studied molecules	Predictions restricted to availability and accuracy of motif information
Tissue specificity	Primary data often tissue specific	Primary data rarely tissue specific	Primary data not tissue-specific
Sources of bias	Technical artifacts (RNA quality), postmortem artifacts (cause of death), biological confounders (age, sex)	Literature curated data contains biases toward more studied interactions, which tend to be non-neuronal	Available motif data may not reflect neuronal tissue motifs
Examples of bioinformatic validation	Preservation of co-expression in independent data, enrichment of physical interactions in modules	Enrichment of co-expression from independent data	Enrichment of predicted binding sites from independent ChIP-seq data
Examples of experimental validation	Showing cell-type specificity of hubs by *in situ* hybridization, demonstrating regulatory potential of hubs by hub gene knockdown	Co-immunoprecipitation of proteins, disruption of protein complexes when hubs targeted	Showing changes in transcription of targets on knockdown of regulator(s)

The structure of inter- and intra-modular relationships for genes can be further connected to biological concepts. Depending on the experimental design, network or module hubs may represent key drivers of the underlying biological process being measured in the module, such as transcription factors that are actually driving co-expression. Alternatively, they may represent the biological process itself, such as genes highly expressed in a particular cell type, such as markers for granule cells in the cerebellum or specific interneuron classes in cortex (Oldham et al., 2006; 2008; Winden et al., 2009).

A general advantage of co-expression network analysis in brain is that it has the ability to represent and integrate differing levels of molecular organization within the hierarchy of brain region, cell-type, organelle, and molecular pathways using transcriptional data alone (Geschwind and Konopka, 2009). Networks comparing similar brain regions or sub-regions can identify more specific cell sub-types, sub-cellular compartments, or specific biological processes (Bernard et al., 2012; Hawrylycz et al., 2012). For example, co-expression modules can be compared between studies to assess whether they are preserved (Langfelder et al., 2011). Similarly, how a specific gene's ranking in a module changes in health and disease can be evaluated (Choi et al., 2005; Hudson et al., 2009; Langfelder et al., 2011).

The major question assessed in both situations, which cannot be queried by single gene approaches and is only poorly assessed by standard differential expression methods, is whether there are changes at the level of gene modules between conditions. Alternatively, because modules correspond to elements of biological function, a gene of unknown function can be annotated based on its module membership, so called, "guilt by association" (Langfelder and Horvath, 2008; Oldham et al., 2008). Similarly, if genes within a particular module have been implicated in disease, it suggests that others within that module may also be disease associated at a higher probability than other brain genes. In this context, network analysis has also provided a framework for cross-species comparisons aimed at understanding the drivers of human brain evolution (Konopka et al., 2012b; Oldham et al., 2006), which is critical in light of a neutral model of transcriptome level evolution, which suggests that the majority of gene expression level differences between species are nonfunctional (Khaitovich et al., 2004). In another comparative context, expression data can be used to understand the degree to which model systems recapitulate the human brain.

1.3: Gene networks in neurodevelopmental disorders

Neurodevelopmental disorders are characterized by abnormal behavioural or cognitive phenotypes originating either in utero or during early postnatal life, and can be accompanied by clinical features outside of the nervous system. Various genetic approaches have been successful in identifying the causes of more than a thousand Mendelian, and fewer non-Mendelian, forms of neurodevelopmental disorders: prototypical examples are intellectual disability (ID) (de Ligt et al., 2012; Gilissen et al., 2014; Lubs et al., 2012; Matson and Shoemaker, 2009; Rauch et al., 2012; Ropers, 2008; van Bokhoven, 2011), autism spectrum disorder (ASD) (Abrahams and Geschwind, 2008; De Rubeis et al., 2014; Geschwind, 2011; Iossifov et al., 2014; 2012; Jamain et al., 2003; Neale et al., 2012; O'Roak et al., 2012; Sanders et al., 2012), epilepsy(Epi4K Consortium et al., 2013; Poduri and Lowenstein, 2011), and schizophrenia (SCZ) (Fromer et al., 2014; Purcell et al., 2014; Xu et al., 2011). The overlap and divergence in genetic architecture across these disorders is beyond the scope of this review (Cross-Disorder Group of the Psychiatric Genomics Consortium et al., 2013; Gratten et al., 2014; 2013).

As more risk variants for these disorders are discovered, remarkable pleiotropy has emerged for some even high risk alleles (Zhu et al., 2014). A given rare, highly penetrant mutation can lead to ASD, SCZ, ID, or a learning disability in different individuals (Doherty and Owen, 2014; Hoischen et al., 2014; Zhu et al., 2014). This observation suggests that rather than causing a specific clinically defined disorder, many mutations act by disrupting normal development, which can have several outcomes. This hypothesis hearkens back to the idea of developmental canalization discussed above (Waddington, 1942), whereby normal development buffers against various insults (such as genetic or environmental perturbations) to achieve a typical trajectory. From the canalization perspective, genetic variation or mutations that significantly increase susceptibility to disease do not necessarily have to invariably lead to the cognitive or behavioural phenotypes with which they are causally linked; instead, they disrupt specific developmental processes that can result in several aetiologically related, but functionally distinct phenotypes. Conversely, many distinct mutations in hundreds of different genes might converge on the same set of clinical phenotypes or disorders, as seems to be the case in ASD, SCZ, and ID (Abrahams and Geschwind, 2008; Cross-Disorder Group of the Psychiatric Genomics Consortium et al., 2013; Devlin and Scherer, 2012; Fromer et al., 2014; Iossifov et al., 2014; Purcell et al., 2014).

Thus, for these complex diseases, it is particularly important to evaluate individual genes in the context of genetic background and normal development, a task for which omic and network methods are particularly well suited. In the next sections, I focus on ASD and SCZ, which represent the complexity and molecular systems insights that have been gained in the area of neurodevelopmental disorders more generally. I provide examples of both genome-wide and seed-based approaches for understanding ASD and SCZ, and gene network studies that use co-expression, protein networks, and integrated omic networks.

1.3.a: Dysregulated networks in ASD and SCZ brain

ASD is a phenotypically and aetiologically heterogeneous neurodevelopmental disorder defined by deficits in social communication and mental flexibility, with onset in the first few years of life (Geschwind, 2011). Transcriptional studies of ASD brain have been limited by the paucity of available tissue. Therefore, several transcriptional analyses of ASD postmortem brain have included fewer than ten individuals and were underpowered to identify reproducible pathways with statistical rigor (Chow et al., 2012; Garbett et al., 2008; Ginsberg et al., 2012; Purcell et al., 2001). Nevertheless, some themes emerge across studies, mostly highlighting increased expression of immune-microglial genes and decreased expression of synaptic genes in the cerebral cortex.

The first ASD study to identify reproducible, genome-wide findings evaluated frontal cortex, temporal cortex, and cerebellum in 19 ASD and 17 control individuals using gene expression microarrays (Voineagu et al., 2011). This study identified DGE changes in ASD shared by about 2/3 of cases. Pathway analysis suggested that increased expression of neural immune genes and decreased expression of synaptic and neuronal genes were consistent and global transcriptomic effects in ASD. Weighted correlation network analysis (WGCNA) (Zhang and Horvath, 2005) allowed the authors to identify coherent biological processes represented by 18 co-expression modules, and the module eigengene, or first principal component of each module. The ability to quantitatively relate module eigengene expression to phenotypic or experimental variables, especially potential confounders, reduces the problem of multiple comparisons faced in high-dimensional omics data and highlights the advantages of using networks as an organizing framework. In this study, two modules, one upregulated and one downregulated, were associated with ASD, but not with confounding factors. Transcriptomics alone cannot distinguish whether such changes are causal or reactive, so the investigators assessed whether common variants associated with ASD (represented by genome-wide association study [GWAS] signals) and candidate ASD risk genes harbouring rare mutations were enriched in these two modules. This analysis provided evidence for a causal role of ASD-associated variants in the downregulated neuronal signalling module. Interestingly, the module upregulated in ASD was enriched for markers of microglia and astrocytes, suggesting that processes involving these cell types occur in response to alterations in synaptic function, perhaps to modify synaptic plasticity by neuron-glial interaction (Voineagu et al., 2011). These results overlap with some previous smaller studies (Garbett et al., 2008; Purcell et al., 2001) and the synaptic and microglial modules have been replicated using RNA-seq in larger independent cohorts(Gupta et al., 2014).

SCZ is defined by prolonged or recurrent episodes of psychosis (characterized by hallucinations and delusions) as well as negative symptoms and deficits in cognitive function(van Os and Kapur, 2009). Although diagnosis is usually made in late adolescence or early adulthood, extensive evidence indicates a neurodevelopmental origin (Weinberger, 1987). Transcriptional studies of SCZ have benefitted from considerably larger sample sizes than those of ASD. However, patients with SCZ have greater comorbidity of smoking, alcohol, and substance abuse than those with ASD, which can confound postmortem studies. Overcoming potential confounders requires careful matching of patient and control individuals and accounting for potential covariate effects wherever possible, as has been done in many studies (Mirnics and Pevsner, 2004; Mirnics et al., 2000). Despite wide-ranging results, consistent findings across studies can be identified, including dysregulation of GABAergic signalling (Hashimoto et al., 2007); downregulation of oligodendrocyte- and myelination-related genes (Hakak et al., 2001), mitochondrial function or energy metabolism (Altar et al., 2005), and synaptic genes (Faludi and Mirnics, 2011); and upregulation of immune and inflammatory genes (Arion et al., 2007).

One of the first studies to put SCZ transcriptomics into a genome-wide network framework used a modified WGCNA approach based on the pairwise mutual information between genes(Torkamani et al., 2010). This study showed that, as in ASD, the overall transcriptomic structure in SCZ is intact, but a neural differentiation module associated with SCZ does not follow the normal trajectory of downregulation with age. Another study confirmed that a dysregulated neuronal differentiation module was consistently observed in multiple SCZ postmortem brain studies, and reported preliminary evidence that the same pathways are involved in bipolar disorder(Chen et al., 2012). Moreover, this latter study applied GWAS signal enrichment, as was done by Voineagu and colleagues(Voineagu et al., 2011), to confirm that common variants associated with SCZ and bipolar disorder were enriched in the neuronal differentiation module. This suggests that disorders sharing genetic architecture also share functional transcriptional alterations(Cross-Disorder Group of the Psychiatric Genomics Consortium et al., 2013).

1.3.b: Mapping risk genes onto developmental networks

A shortcoming of transcriptomic studies investigating postmortem brain samples is that tissue is usually obtained long after the disease-causing changes have occurred. Given that the human brain transcriptome has a consistent and reproducible structure(Hawrylycz et al., 2012; Oldham et al., 2008), one useful way to explore how mutations in risk genes perturb typical brain development is to map risk genes onto transcriptional networks that represent normal brain development (Figure 1c). The first study to do this identified co-expression modules using nearly 1,000 adult brain regions and evaluated enrichment for ASD susceptibility genes(Ben-david and Shifman, 2012). The researchers characterized cell-type-specific and region-specific modules and found that modules enriched in ASD GWAS signal and candidate ASD risk genes were enriched in a neuronal module, and that this module is upregulated during fetal brain development. Thus, genes associated with risk for ASD were demonstrated to affect neuronal development.

The identification of genetic risk factors by whole-exome sequencing (WES) (Iossifov et al., 2012; Neale et al., 2012; O'Roak et al., 2012; Sanders et al., 2012) and the availability of transcriptome data spanning multiple brain regions and developmental stages (Colantuoni et al., 2011; Kang et al., 2011) have created new opportunities to map disease risk genes onto developmental transcriptional networks. The study descried in Chapter 2 used WGCNA to construct co-expression networks from 8 weeks post-conception to one year of age across 11 cortical regions and assessed how both ASD and ID risk genes are involved in cortical development, cell types, and circuits (Parikshak et al., 2013). Robust co-expression modules that were reproducible in independent data were identified, and module eigengenes were shown to have trajectories reflecting timing of molecular processes during human cortical development. Five developmentally regulated co-expression modules were enriched for multiple ASD risk gene sets, but not ID or multiple control gene sets, highlighting specific biological processes disrupted by ASD risk genes. Two of the five modules identified were enriched for *de novo* mutations identified through WES, and contained genes involved in transcriptional and chromatin regulation relevant to neural development, such as the BAF complex (Ronan et al., 2013).

The other three modules were upregulated later in cortical development, representing various stages of synaptic development. These three "synaptic" modules were enriched in ASD candidate genes (Basu et al., 2009), genes downregulated in postmortem ASD cortex (Voineagu et al., 2011), and were preferentially affected by inherited genetic risk. Bioinformatic analyses further suggested that the chromatin regulatory modules and synaptic modules were co-regulated at the transcriptional and translated levels. Finally, analysis of cortical laminae and cell types revealed that multiple modules were enriched for genes preferentially expressed in superficial layer glutamatergic neurons during, suggesting this cell type is preferentially affected genetic hits in ASD.

A complementary study searched temporal windows spanning 8 weeks post-conception to 40 years postnatal age in four groupings of brain regions to identify co-expression networks enriched for ASD risk that were seeded around nine high-confidence risk genes identified by WES (Willsey et al., 2013). These investigators asked if, when, and where this subset of nine genes converges during brain development by constructing binary co-expression networks based on the top 20 correlations for each seed gene for various spatial and temporal combinations that resulted in 85 networks. Each network was evaluated for enrichment of 122 additional ASD risk genes supported by WES evidence, which identified three spatiotemporal combinations that passed stringent correction for multiple testing: frontal cortical regions during 10-19 weeks post-conception and 16-24 weeks post-conception as well as thalamic and cerebellar regions from birth to 6 years of postnatal age.

The researchers further identified a potential role for lower layer glutamatergic projection neurons by assessing co-expression to candidate markers of cortical layers and cell types. Interestingly, there was no pathway or PPI enrichment identified in these networks, probably due to the small size of the co-expression modules and the inclusion of both positive and negative correlations when computing co-expression relationships (unsigned networks), which is less sensitive to pathway and protein interaction detection (Ramani et al., 2008; Song et al., 2012).

Importantly, both of these studies found that the greatest convergence for rare *de novo* ASD-associated mutations was during early fetal and midfetal development, with the major enrichment for risk in cortical glutamatergic projection neurons. Thus, despite the fact that the same gene is rarely hit recurrently by rare de novo variants in ASD, this class of variation preferentially disrupts specific cell-types. Notably, the genome-wide study (Parikshak et al., 2013), which assessed both ASD and ID genes, further suggested that disruption of the upper cortical layers (L2-4) results is related to ASD-like phenotypes, while disruption of the lower layers (L5-6) results in more severe consequences leading to ID. Other studies have also found that fetal cortical development and glutamatergic neurons are affected by mutations in ASD, making it likely that this is a robust finding warranting experimental testing (Miller et al., 2014; Stein et al., 2014; Steinberg and Webber, 2013).

A developmental co-expression approach has also been used to identify risk convergence in SCZ. A seeded spatial and temporal search of co-expression between genes identified in a WES study identified fetal development of the prefrontal cortex as a point of convergence for *de novo* variation (Gulsuner et al., 2013). Furthermore, the investigators demonstrated enrichment for PPIs in this gene set and confirmed statistical significance by comparing against rare *de novo* variants found in unaffected individuals. However, this study did not extend the network to genes beyond the seed set, and it did not identify cellular, laminar, or regulatory relationships among these genes. As larger sets of risk genes are becoming available (De Rubeis et al., 2014; Fromer et al., 2014; Iossifov et al., 2012; Purcell et al., 2014), comparisons of how mutations in ASD, SCZ, ID, and other psychiatric disorders affect cells and circuits will become more comprehensive.

1.3.c: The role of transcriptional and translational co-regulation

Another promising approach by which to identify disease-associated networks is to experimentally construct a seed-based network for a candidate regulatory molecule, as has been done using CLIP-seq with the Fragile X Mental Retardation Protein (FMRP) in brain (Darnell et al., 2011). FMRP is an RNA-binding protein critical to neuronal plasticity that is involved in translational repression at the synapse. Recent work has demonstrated that FMRP bound transcripts are enriched for genes dysregulated in ASD (Darnell et al., 2011) and ASD rare *de novo* variants (Iossifov et al., 2012). Co-expression network analysis connected these observations: FMRP targets that are involved in chromatin modification are affected by rare *de novo* variants in ASD and are downregulated in early development, whereas FMRP targets that are ASD risk genes involved in synaptic function increase in expression during early development (Parikshak et al., 2013). Steinberg and colleagues (Steinberg and Webber, 2013) rigorously investigated co-expression networks throughout development by seeding with FMRP targets, and showed FMRP-targeted developmental networks are also affected by ASD-associated copy number variations (CNVs). Furthermore, WES studies of other neurodevelopmental disorders have found enrichment for FMRP targets in rare mutations in SCZ (Purcell et al., 2014), ID(Gilissen et al., 2014) and epilepsy(Epi4K Consortium et al., 2013). Given that many FMRP targets are evolutionarily conserved(Ronemus et al., 2014) and under purifying selection (Iossifov et al., 2012), FMRP-related activity-dependent regulation during fetal brain development might be particularly vulnerable to genetic perturbations, with rare mutations resulting in a disruption of developmental canalization.

Another example of a seed-based approach with a disease-relevant molecular regulator involves the splicing factor RBFOX1 (also known as A2BP1). Postmortem co-expression networks in autism identified *RBFOX1* as a hub in an ASD dysregulated synaptic module

(Voineagu et al., 2011). RNA-seq demonstrated that alterations in *RBFOX1* expression were associated with changes in splicing at RBFOX1 regulated sites (Fogel et al., 2012; Voineagu et al., 2011), and further studies demonstrated that regulation of gene expression levels by 3' UTR stabilization might also contribute to this module (Ray et al., 2013). A recent CLIP-seq study comprehensively mapped RBFOX1/2/3 binding sites and suggested that these splicing regulators share binding targets and are each dysregulated to some extent in ASD (Weyn-Vanhentenryck et al., 2014). Taken together, these findings suggest a pervasive role for altered transcriptional and splicing levels related to perturbations in RBFOX family in ASD.

Finally, one recent study (Sugathan et al., 2014) highlights the use of ChIP-seq to better understand the regulatory targets of the chromatin regulator CHD8, which is to date the most frequently identified gene harbouring rare *de novo* variation in ASD (Iossifov et al., 2012; O'Roak et al., 2012; Talkowski et al., 2012). CHD8 mutations are accompanied by macrocephaly, ID, and gastrointestinal problems, potentially defining a genetically defined subtype of ASD (Bernier et al., 2014). Sugathan and colleagues (Sugathan et al., 2014) evaluated CHD8 binding and DGE on CHD8 knockdown in a genome-wide manner in neural progenitor cells (NPCs) to identify a downstream network of genes related to neuronal differentiation, consistent with CHD8 serving as a master regulator of neurogenesis during brain development. By further integration of the CHD8 network with gene co-expression data, the authors found evidence that CHD8 directly regulates early gene co-expression modules enriched for rare *de novo* mutations and genes found in the proliferating layers of fetal cortex, but only indirectly affects subsequent synaptic development (Parikshak et al., 2013). Given the emerging role of transcriptional and chromatin regulators identified by WES in ASD (Ben-David and Shifman, 2012; De Rubeis et al., 2014; Parikshak et al., 2013), integrative ChIP-seq and DGE studies of other risk genes with putative gene regulatory functions will be important.

1.3.d: PPI networks define new interactions

Genetic investigations in ASD have also used seed-based networks with literature-curated PPIs to identify convergence of ASD risk genes (Neale et al., 2012; O'Roak et al., 2012). The most thorough example of this approach was undertaken by O'Roak and colleagues (O'Roak et al., 2012), who identified a highly interconnected PPI sub-network among *de novo* rare variants hit genes in ASD. The authors performed a follow-up targeted sequencing study (O'Roak et al., 2012) of this sub-network in a larger cohort, and found more new variants affecting these interconnected genes compared to chance. However, the potential lack of tissue specificity in these PPI networks and biases in literature-curated PPIs may have limited identification of novel pathways or circuits that are affected by these risk variants (Table 1.1).

To evaluate more unbiased molecular interactions and define whether candidate disease genes interact at the protein level, Sakai and colleagues (Sakai et al., 2011) performed a Y2H screen of 35 syndromic or candidate ASD genes. This study was the first of its kind in neurodevelopmental disorders, and identified many specific novel interactions. Another Y2H study assessed a larger seed set corresponding to spliced isoforms of candidate genes that were found in whole-brain RNA-seq (Corominas et al., 2014), hypothesizing that isoform-level PPIs would discover tissue-specific PPI networks (Ellis et al., 2012). The resultant PPI network was modestly enriched for products of known ASD genes, proteins that were co-expressed, or proteins that were targeted by common molecular regulators.

These results further demonstrate convergence among known disease-relevant genes at a PPI level and also that evaluating tissue-specific isoforms can reveal more relevant PPI networks. Both of these studies involved state-of-the-art quality control and validation, and identified many novel interactions. However, even with knowledge of isoform-specific interactions, the tissue environment for interaction cannot be recapitulated with current PPI approaches at a large scale (Table 1.1). This is a motivation for beginning with tissue-specific transcriptional networks and then integrating PPI analyses to elucidate protein-level complexes or pathways occurring in the cell type or tissue of interest.

Recently, one study utilized global PPI interactions have been evaluated to identify modules enriched for ASD-associated genes (Li et al., 2014). This study identified a particularly interesting module related to synaptic function and weakly enriched for mutations in whole genomes from individuals with ASD. This module was identified as most highly expressed in oligodendrocytes, and enriched for gene expression in the corpus callosum. Here, gene expression data was essential to arriving at a neurobiological interpretation of the PPI module (Li et al., 2014). This highlights how tissue specific molecular data is essential to interpreting PPI analyses. Given the biases inherent to global PPIs discussed above and in Table 1, these findings warrant replication with new PPI data and investigation of why these relationships are detected at the PPI level, but not the co-expression level.

1.3.e: Integrated networks reveal shared molecular phenotypes

Multiple levels of omics data can contribute unique functional insights and increase power to detect molecular convergence. This has motivated the integration of genetic and functional evidence to support specific genes or pathways. Network-based analysis of genes

(NETBAG) (Gilman et al., 2011) integrates multiple forms of shared molecular phenotype evidence based on a framework demonstrated to be effective for predicting gene essentiality in yeast (Lee et al., 2008).

NETBAG defines a network based on shared genetic phenotypes – genes highly connected in the network likely participate in the same phenotype. Edges in the network are primarily based on a weighting of contributions from direct and indirect PPIs and shared GO or Kyoto Encyclopedia of Genes and Genomes (KEGG) (Ogata et al., 1999) term membership. The contribution of each data type to edges is weighted based on relationships among a known set of disease genes (Feldman et al., 2008). These different weights are combined to construct a phenotypic 'background network' on which gene lists of interest are assessed for non-random clustering using a permutation analysis. Pathway enrichment in resultant modules was performed by assessing GO term clustering in modules, which is expected given that clustering is partly driven by shared GO functions. When seeded with 75 disparate *de novo* CNV deletions spanning 746 genes associated with ASD, NETBAG found a highly interconnected module related to synaptic function, whereas a similarly sized set of genes affected by inherited rare CNVs were sparsely connected and did not form a module (Levy et al., 2011).

Furthermore, genes in CNVs from females contributed more to the module connectivity than those from males, suggesting that females are affected by more severe genetic hits in ASD, an observation that has been replicated in exome sequencing studies (Iossifov et al., 2014; Ronemus et al., 2014). A later study (Noh et al., 2013) that evaluated CNV duplications in addition to deletions also found an interconnected PPI network that was enriched for proteins involved in synaptic transmission, validating the observation that pathogenic CNVs affect similar gene networks (Gilman et al., 2011).

Subsequently, Gilman and colleagues (Gilman et al., 2012) extended this approach (dubbed NETBAG+) to simultaneously evaluate CNVs, SNVs, and common variants, to assess network convergence in schizophrenia. Two modules were identified, one related to axonal guidance, neuronal migration, and synaptic function, and another enriched for chromatin modifiers. The first module was highly connected in the phenotypic background network with candidate genes formerly associated with ASD and ID, and with the previously identified module that was defined by CNVs in ASD (Gilman et al., 2011). This module exhibited a pattern of developmental upregulation during fetal development, while the second module exhibited downregulation during fetal development. Despite the overlap between genes associated with ASD and SCZ, the authors provided evidence suggesting that the mutations in ASD increase neuronal spine or dendrite growth, while those in SCZ decrease it (Gilman et al., 2012).

NETBAG+ has also been applied to SNV and CNV affected genes in ASD, identifying a weakly enriched module that was extensively characterized with gene expression data and phenotypic data from the genetic studies (Chang et al., 2015). This study validated the increase in ASD risk for gene network modules enriched for chromatin regulators and genes susceptible to haploinsufficiency as observed before (Parikshak et al., 2013) and found enrichment of cortical laminae and interneurons, extending the previously observed associations (Parikshak et al., 2013; Voineagu et al., 2011; Willsey et al., 2013) to corticosriatal projection neurons. It also affirmed that early developmental SNVs are more severe and result in intellectual disability (ID) in addition to ASD (Samocha et al., 2014). Thus, integrative network analysis can identify shared pathways among disorders even with networks not based on brain-specific data. However, deriving more refined insights clearly require a direct integration of neurobiological data, such as gene expression normal development, specific cell types, or disease tissue.

An exciting integrative approach is to simultaneously integrate PPIs, co-expression, and mutational burden in neurodevelopmental disorders, as has been done for cancer (Chang et al., 2013; Hoadley et al., 2014). Recent work has constructed networks and identified modules based on seeding with mutation affected genes and known pathways and then constructing network modules around these initial seeds using genome-wide co-expression and PPI data (Hormozdiari et al., 2015). The degree to which modules grow is restricted by an objective function that optimizes the number of genes that are co-expressed and co-regulated in a module against the rate of mutations observed in controls in these genes. This identifies modules containing highly related genes that are enriched for highly pathogenic mutations. Interestingly, the authors found that these mutations were also found in epilepsy, SCZ, and ID, further supporting the idea that very severe mutations disrupt canalization in a manner less specific to a particular disease.

1.4: Conclusions

In this chapter, I covered the utility of transcriptomics and other high-throughput molecular systems approaches in neuroscience. Although there are challenges posed by the spatial heterogeneity and the temporal dynamics of molecular changes in the nervous system, novel insights can still be gained with the appropriate study design and analysis. Gene network analysis is a particularly promising way to analyse these data, as it can integrate diverse types of data into one framework.

In Chapter 2, I use gene expression analysis to organize the molecular changes that occur during cortical development. I then map genes implicated by multiple methodologies as associated with autism onto these networks, and identify co-expression modules that are enriched for autism risk genes. The power of network analysis in revealing biological changes becomes apparent in the functional characterization of these modules.

In Chapter 3, I investigate the value of this developmental transcriptomic network in predicting where new mutations in ASD will be identified. These results suggest that co-expression networks are a powerful approach for identifying the biological processes affected in brain development by mutations in autism.

In Chapter 4, I utilize differential gene expression analysis, differential splicing analysis, and gene co-expression networks to understand the molecular changes that occur in brains from autism spectrum disorder. I compare these changes across brain regions and between idiopathic ASD and a genetic subtype of ASD, and demonstrate that robust, reproducible changes occur in ASD brain.

Although my work utilizes the principles and approaches covered in this chapter, it does not focus very heavily on rigorously optimizing methodology. Much work remains to be done on that frontier, so throughout my work I have attempted to choose the best available methods when possible, but relied on heuristics that yield reproducible results otherwise.

CHAPTER 2:
Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism

"At the same time, it is clear that canalization is not a necessary characteristic of all organic development, since it breaks down in mutants, which may be extremely variable..."
— Conrad H. Waddington, Canalization of Development and the Inheritance of Acquired Characters, *Nature*, 1942

2.1: Abstract

Genetic studies have identified dozens of autism spectrum disorder (ASD) susceptibility genes, raising two critical questions: 1) do these genetic loci converge on specific biological processes, and 2) where does the phenotypic specificity of ASD arise, given its genetic overlap with intellectual disability (ID)? To address this, I mapped ASD and ID risk genes onto co-expression networks representing developmental trajectories and transcriptional profiles representing fetal and adult cortical laminae. ASD genes tightly coalesce in modules that implicate distinct biological functions during human cortical development, including early transcriptional regulation and synaptic development. Bioinformatic analyses suggest translational regulation by FMRP and transcriptional co-regulation by common transcription factors connect these processes. At a circuit level, ASD genes are enriched in superficial cortical layers and glutamatergic projection neurons. Furthermore, I show that the patterns of ASD and ID risk genes are distinct, providing a novel biological framework for investigating the pathophysiology of ASD.

These findings are summarized in Figure 2.1.

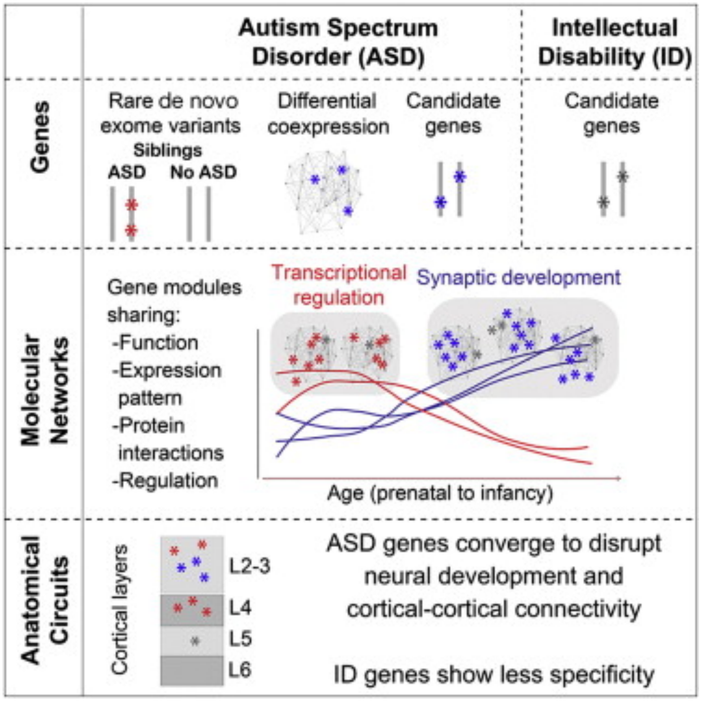

Autism Spectrum Disorder (ASD)
Intellectual Disability (ID)
Genes
Rare de novo exome variants
Siblings
ASD
No ASD
Differential coexpression
Candidate genes
Candidate genes
Molecular Networks
Gene modules sharing:
-Function
-Expression pattern
-Protein interactions
-Regulation
Transcriptional regulation
Synaptic development
Age (prenatal to infancy)
Anatomical Circuits
Cortical layers
L2-3
L4
L5
L6
ASD genes converge to disrupt neural development and cortical-cortical connectivity
ID genes show less specificity

Figure 2.1 Graphical abstract summarizing findings in Chapter 2. Genes implicated in ASD across multiple sources were mapped to molecular networks. These modules identified from the network reflecting cortical development represent shared molecular functions and are co-regulated during cortical development. Two modules related to early fetal brain development and transcriptional regulation are enriched for protein disrupting and missense rare de novo variants linked to ASD, and three modules related to later fetal development and synaptic function are implicated by gene expression changes in ASD brain and inherited variants. These modules involved genes that are highly expressed in layers 2 and 3 of the adult cortex, suggesting that ASD risk genes converge on circuits that are related to inter- and intra- hemispheric connectivity, and that these are cellular circuit-level pathways coherently disrupted in ASD. Finally, genes implicated in intellectual disability (ID) are not enriched in these modules.

2.2: Introduction

Autism Spectrum Disorder (ASD) is a heterogeneous neurodevelopmental disorder, in which hundreds of genes have been implicated (Berg and Geschwind, 2012; Geschwind and Levitt, 2007). Analysis of copy number variation (CNV) and exome sequencing (Iossifov et al., 2012; Neale et al., 2012; O'Roak et al., 2012; Sanders et al., 2012) have identified rare *de novo* variants (RDNVs) that alter dozens of protein coding genes in ASD, none of which account for more than 1% of ASD cases (Devlin and Scherer, 2012). This, and the fact that a significant fraction (40-60%) of ASD is explained by common variation (Klei et al., 2012), points to a heterogeneous genetic architecture.

These findings raise several issues. Based on the background human mutation rate (MacArthur et al., 2012), most genes affected by only one observed RDNV to date are likely false positives that do not increase risk for ASD (Gratten et al., 2013). It is therefore essential to develop approaches that prioritize singleton variants, especially missense mutations. Furthermore, given the heterogeneity of ASD, it would valuable to identify common pathways, cell-types, or circuits disrupted within ASD itself. Recent studies combining gene expression, protein-protein interactions (PPIs), and other systematic gene annotation resources suggest some molecular convergence in subsets of ASD risk genes (Ben-David and Shifman, 2012; Gilman et al., 2011; Sakai et al., 2011; Voineagu et al., 2011). Yet, it remains unclear how the large number of genes implicated through different methods may converge to affect human brain development, which is critical to a mechanistic understanding of ASD (Berg and Geschwind, 2012). Additionally, ASD has considerable overlap with ID at the genetic level, so identifying molecular pathways and circuits that confer the phenotypic specificity of ASD would be of considerable utility (Geschwind, 2011; Matson and Shoemaker, 2009).

Here, I take a stepwise approach to determine if genes implicated in ASD affect convergent pathways during *in vivo* human neural development, and whether they are enriched in specific cells or circuits (Figure 2.2A). First, I constructed transcriptional networks representing genome-wide functional relationships during fetal and early postnatal brain development and mapped genes from multiple ASD and ID resources to these networks. I then assessed shared neurobiological function among these genes, including co-regulatory relationships and enrichment in layer-specific patterns from micro-dissected human fetal and adult primate cortical laminae. I used validation in independent *in vivo* and *in vitro* expression data and additional functional evidence (shared annotated pathways and PPIs) to confirm shared function among genes, and I replicated the enrichment analyses in independent data to ensure robustness. This integration of an unsupervised network analysis with large gene sets from multiple resources permits rigorous interrogation of biological convergence and specificity in ASD that takes its heterogeneity into consideration and enables comparison of ASD with ID.

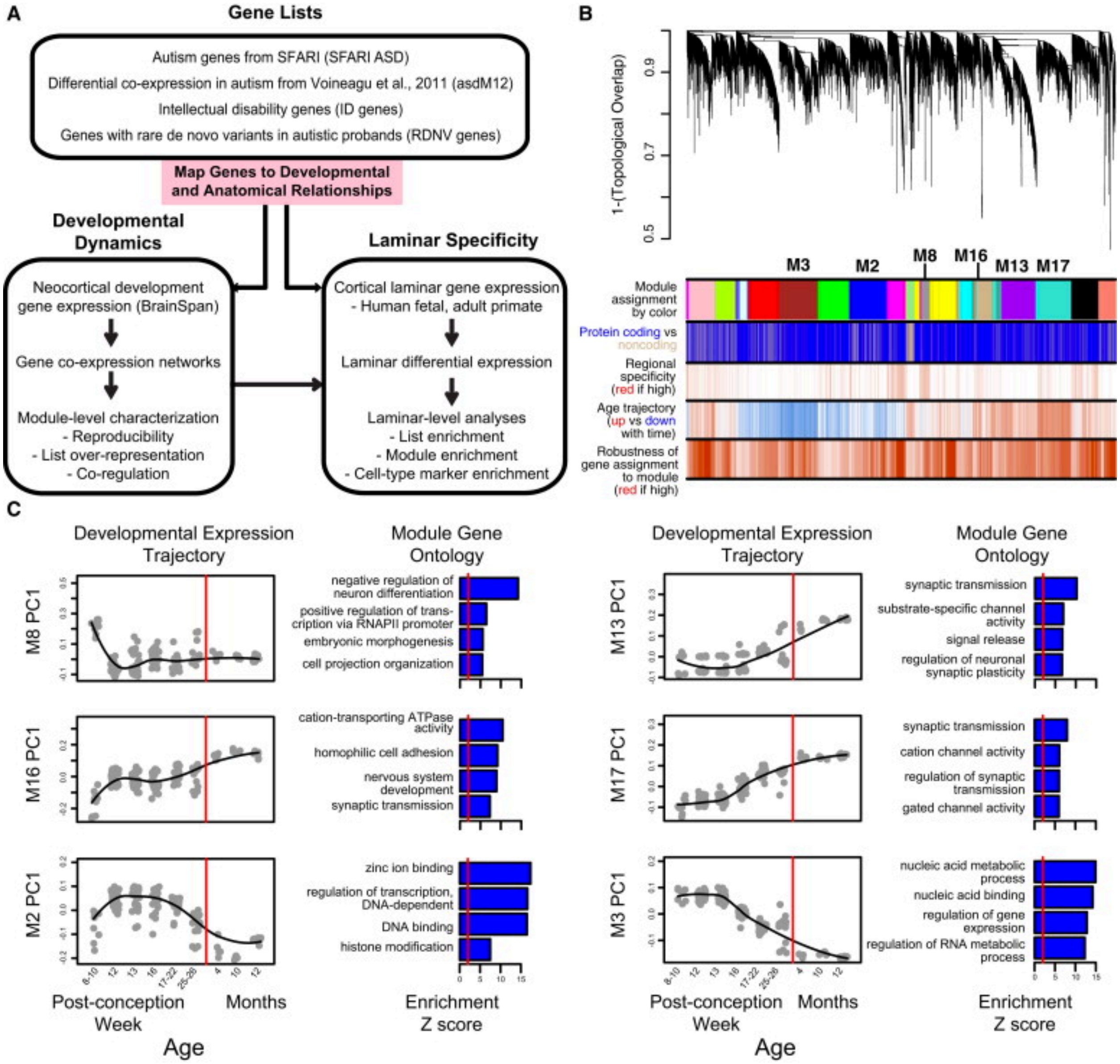


Figure 2.2 Methodological Overview and Coexpression Network Analysis. (A) Flowchart of the overall approach. (B) Network analysis dendrogram showing modules based on the coexpression topological overlap of genes throughout development. Color bars below give information on module membership, gene biotype, cortical region specificity, age trajectory, and robustness of module assignment. (C) Module characterization, including GO enrichment and trajectory throughout development. The fit line represents locally weighted scatterplot smoothing (See Appendix A1 for more details). GO enrichments are adjusted for multiple comparisons (FDR < 0.05), and reported *Z* scores represent relative enrichment in the module compared to all cortex-expressed genes, with the red line at Z = 2. See also Table A1.1 and Figure A1.1.

2.3: Results

2.3.a: Genome-wide co-expression networks reflect biological processes essential to human neocortical development

I reasoned that transcriptomic data from human neocortex would inform understanding of ASD pathophysiology, as the cerebral cortex has been consistently implicated in ASD pathophysiology by multiple modalities (Amaral et al., 2008; Ecker, 2012; Geschwind, 2011; Rubenstein, 2010; Voineagu et al., 2011). I focused on gene expression from cortical development spanning post-conception week (PCW) 8 to 12 months after birth, as this time period reflects many critical molecular processes that orchestrate brain circuit formation that could be disrupted by genetic hits in ASD (Andersen, 2003; Courchesne et al., 2011).

I constructed networks of gene relationships agnostic to ASD candidate genes based on BrainSpan whole-genome transcriptomic data collected by RNA-seq (www.brainspan.org). I applied signed, weighted gene co-expression network analysis (WGCNA, see section 2.5, Materials and Methods; (Zhang and Horvath, 2005) and identified 17 co-expression modules (labeled numerically, e.g. M8, and by color, e.g. magenta, see Table A1.1B for module details). These modules represent genes that share highly similar expression patterns during cortical development (Figure 2.2B), and additional analyses show that these modules identify highly significant shared expression patterns that are replicated in independent data from both *in vivo* and *in vitro* human neural development (Figures A1.1A-C, see section 2.5, Materials and Methods).

First, I investigated each module's developmental trajectory by calculating the module eigengene (ME, the first principal component of the module) and assessed shared function among genes within the module by enrichment for Gene Ontology (GO) annotation terms.

Representative examples for up and down-regulated modules are shown in Figure 2.2C. Module eigengenes for M13, M16, and M17 increase during early cortical development and are each enriched for the GO term synaptic transmission (Figure 2.2C). M16 is upregulated the earliest, starting at PCW 10 and its hubs (most inter-connected genes based on correlation to the ME, kME) include genes coding for the structural synaptic proteins *SV2A* and *NRXN1*. M16 GO terms include cation transporter activity, homophillic cell adhesion, and nervous system development, consistent with early development of synaptic ultrastructure. M17 represents a later phase of synaptic maturation, as it is upregulated after PCW 13 and its hubs include *CAMK2B* and *CACNA1C*, which are important for calcium-dependent regulation of synaptic activity. M13 increases last, after PCW 16, and its hubs include the NMDA and GABA receptor subunits *GRIN2A* and *GABRA1*, while GO terms include substrate-specific channel activity and regulation of neuronal synaptic plasticity. These three modules have closely aligned, yet distinct developmental trajectories that likely reflect sequential phases of synaptic development, maturation, and function, all of which are essential to the development of the cerebral cortex.

In contrast, M2 and M3 have anti-correlated trajectories to M13, M16, and M17 (r = -0.46 to -0.96, Table A1.1B), and are enriched in GO terms associated with DNA binding and transcriptional regulation (Figure 2.2C). Expression in M3 is initially upregulated and then decreases after PCW 12, suggesting its functions may be most important prior to M2, which is upregulated after PCW 10 and peaks later (PCW 12 to PCW 22). Given the GO enrichment and anti-correlation to the synaptic module MEs, genes in these modules may be critical to orchestrating processes such as progenitor proliferation and cell fate specification via initial repression followed by de-repression of neuronal genes (Srinivasan et al., 2012). Furthermore, many of the genes found in M2 and M3 are part of well-studied chromatin remodeling complexes, most notably the BAF complex (*ARID1A* and *SMARCA4* in M2; *ARID1B*, *SMARCB1*, *SMARCC1*, *SMARCC2*, *SMARCD1*, *ARID2*, *DPF2*, *BCL11A*, *BCL11B*, and *ACTL6A* in M3), that have recently been linked to neural differentiation and neurodevelopmental disorders (Ronan et al., 2013; Yoo et al., 2009).

Since, positive correlations among genes also reflect pair-wise interactions between proteins (Ramani et al., 2008), enrichment for protein-protein interactions within modules provides an independent line of validation for shared function in these modules at the protein level. I combined all known PPIs from InWeb (Rossin et al., 2011) and BioGRID (Stark, 2006) into one network, comprising 251,881 interactions between 18,384 proteins, and observed that 12/17 of all co-expression modules, including all the modules in Figure 2.2C are enriched for PPI after stringent multiple testing correction ($p < 0.003$, Table A1.1B). Overall, 10/17 co-expression modules are preserved in independent gene expression data sets, enriched for GO terms, and enriched for PPI, while 2/17 are enriched for two of these three criteria. These results demonstrate the utility of a systems biology approach: instead of analyzing lists of thousands of genes regulated during development, I focused on this set of 12 reproducible and biologically meaningful modules sharing distinct expression patterns and biological functions.

2.3.b: Genes implicated in ASD are highly co-expressed during human cortical development

I next asked whether genes associated with risk for ASD converge on common biological processes. I compiled a set of 155 ASD genetic risk candidates from the Simons Foundation Autism Research Initiative (SFARI) AutDB database (Basu et al., 2009), which I refer to as SFARI ASD. The SFARI ASD list is a manually curated set of candidate genes implicated by common variant association, candidate gene studies, genes within ASD-associated CNV, and, to a lesser extent, syndromic forms of ASD (see section 2.5, Materials and Methods). I mapped this gene set to the protein coding genes in the developmental co-expression network and observed that SFARI ASD genes are most over-represented in M16 ($p = 0.0024$, odds-ratio (OR) = 2.9, 95% confidence interval = [1.4-5.5], false discovery rate (FDR) < 0.05), and less so in M13 and M17 (Figure 2.3A).

I also examined a set of ASD genes previously shown to be dysregulated in postmortem ASD temporal and frontal cortex (asdM12; Voineagu et al., 2011), which represents a shared molecular pathology in ASD brain identified in an unbiased, genome-wide manner. The asdM12 gene set was strongly enriched in the same three modules as SFARI ASD genes, M13, M16 and M17 (asdM12-M13, $p = 3.0\text{x}10^{-15}$, OR 3.6 [2.7-4.8]; asdM12-M16, $p = 3.5\text{x}10^{-15}$, OR 3.9 [2.8-5.3]; asdM12-M17, $p = 1.0\text{x}10^{-7}$, OR 2.5 [1.8-3.4]; each at FDR < 0.05). A remarkable 42% of asdM12 and 25% of the SFARI ASD sets are found in one of these three modules. This analysis, which uses gene sets identified based on different methods (only 15 genes overlap between SFARI ASD and asdM12), converges onto three modules involved in prenatal and early postnatal synaptic development.

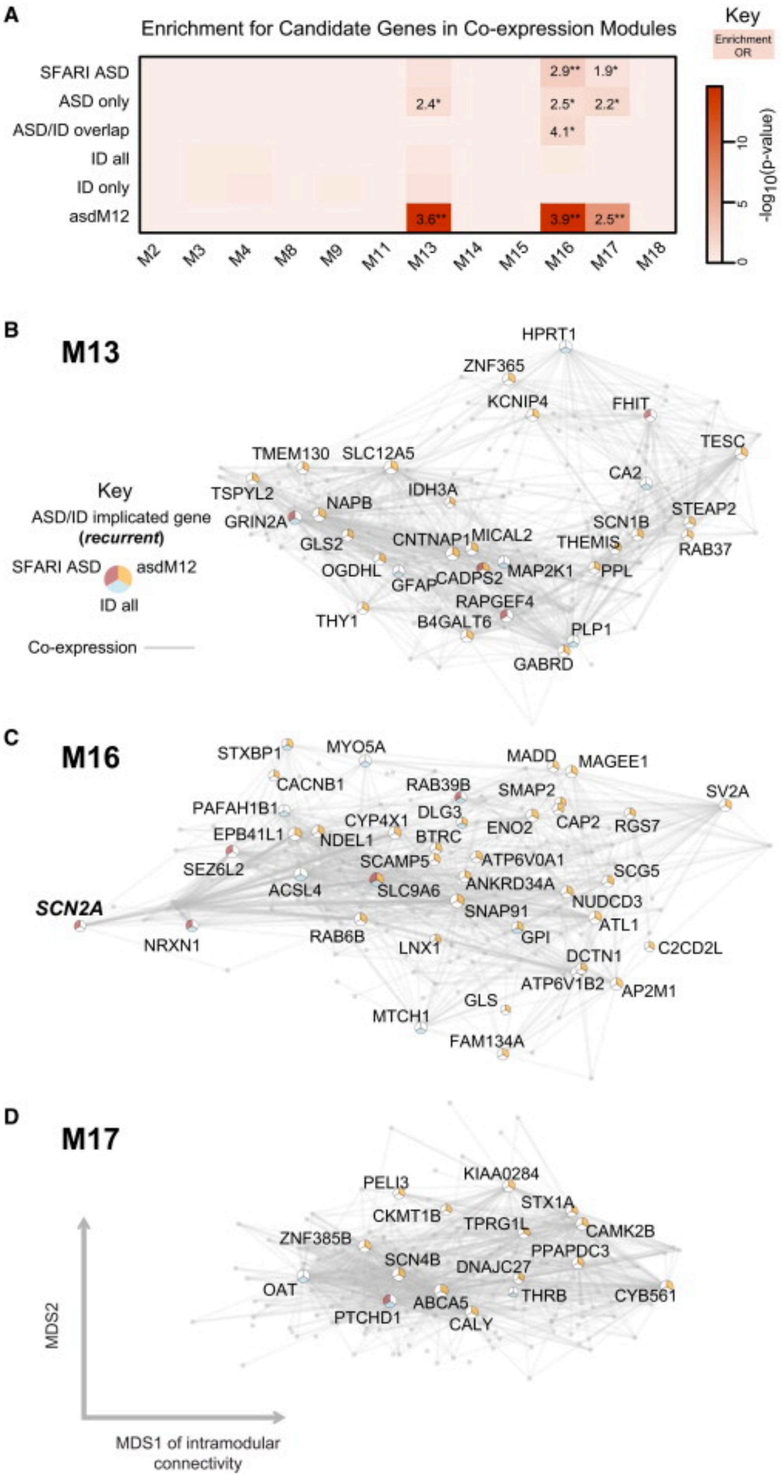


Figure 2.3 Enrichment of SFARI ASD, asdM12, and ID Genes in Developmental Networks. (A) Module-level enrichment for gene sets from a curated set of ASD risk genes (SFARI ASD), a curated set of ID genes ("ID all"), and an unbiased set of ASD risk genes (asdM12). Overlapping (ASD/ID overlap) and nonoverlapping sets ("ASD only" and "ID only") are also shown. All enrichment values for overrepresented lists with $p < 0.05$, $OR > 1$ are

shown to demonstrate enrichment trends ($^{*}p < 0.05$ and $^{**}FDR < 0.05$). Heatmap colors for p values reflect enrichment trends; p values for gene sets with OR < 1 can be seen in Table A1.2B. (B–D) These panels show network plots for M13, M16, and M17, respectively. Most hub genes overlapping with SFARI ASD and asdM12 enrichment are not the same, showing that enrichment of these two sets is not driven by a narrow shared subset of genes. Network plots comprise the top 200 connected genes (based on kME, a measure of intramodular connectivity) and their top 1,000 connections in the subnetwork. By definition, all edges in the network reflect positive correlations. Genes with membership in SFARI ASD, asdM12, or the "ID all" list are labeled and plotted according to multidimensional scaling of gene expression correlations, which graph genes with similar expression patterns closer to each other. See also Table A1.2.

I next hypothesized that mapping ID genes to this network would enable me to assess whether ASD susceptibility genes show any specificity in their developmental expression patterns. I compiled an extensive set of high confidence genes implicated in monogenic forms of ID from multiple publications (Inlow, 2004; Lubs et al., 2012; Ropers, 2008; van Bokhoven, 2011), referred to as "ID all" (see section 2.5, Materials and Methods). Remarkably, this set of 364 genes expressed in human neocortex is not enriched in any of the 12 co-expression modules. Importantly, this lack of enrichment is at a relaxed threshold that reduces the risk of false negatives (uncorrected $p > 0.05$). Removing the small set of 37 genes (<10%) that overlap between ASD and ID to establish exclusive sets ("ASD only", "ID only") further confirms that ASD genes exhibit enrichment, while ID genes do not (Figure 2.3A, Table A1.2B). Thus, it is genes connected with the ASD phenotype that are enriched in 3 specific transcriptional modules related to synaptic function during development, but not those that have been related solely to ID.

2.3.c: ASD-associated protein disrupting rare de novo variants are highly enriched in two co-expression modules in early fetal development

Additional evidence implicating specific genes in ASD comes from whole-exome sequencing in families (Iossifov et al., 2012; Neale et al., 2012; O'Roak et al., 2012; Sanders et al., 2012), which has identified many rare protein disrupting variants (nonsense, splice-site, frameshift) over-represented in individuals with ASD compared to their unaffected siblings (OR > 2). This evidence is largely distinct from the evidence implicating genes in SFARI ASD and asdM12, as it is from purely non-inherited, rare variation discovered in an unbiased, genome-wide manner. I therefore asked whether RDNV-affected genes found in ASD probands shared biological function. I also tested silent RDNVs since they should not exhibit a similar pattern of functional enrichment, providing a key control for gene size, GC content, and other features affecting mutability (Michaelson et al., 2012).

I first tested for enrichment using RDNVs from three studies sharing similar coverage criteria and variant calling methodology (Neale et al., 2012; O'Roak et al., 2012; Sanders et al., 2012), representing 622 ASD probands and 222 unaffected siblings. Strikingly, genes expressed during development and affected by protein disrupting RDNVs in probands (60 genes, Table A1.2A, Discovery Set) are significantly enriched in two modules, M2 and M3, which exhibit highly similar developmental trajectories and functional enrichment, indicative of remarkable biological specificity. Eight genes harboring protein disrupting RDNVs are enriched in M2 ($p = 0.006$, OR = 3.2 [1.3-6.8]; FDR < 0.05) and 10 are enriched in M3 ($p = 0.0011$, OR = 3.6 [1.6-7.2]; FDR < 0.05). A trend for enrichment is observed for M16 as well, but this does not pass the FDR threshold. For comparison, genes affected by RDNVs in unaffected siblings or affected by silent mutations are not enriched in any modules (Table A1.2B, $p > 0.05$). Since missense

RDNVs are only weakly over-represented in ASD (Sanders et al., 2012), I reasoned that overlap with network modules might prioritize specific subsets of this RDNV class. I find that a subset of missense RDNV affected genes is over-represented in the same pathways as the more deleterious protein disrupting RDNVs (M2 and M3, Table A1.2B). Taken together, out of 385 protein disrupting or missense RDNV affected genes expressed in brain, 34 are found in M2 ($p = 2.9 \times 10^{-4}$, OR = 2.1 [1.4-3.0], FDR < 0.05) and 41 in M3 ($p = 2.3 \times 10^{-5}$, OR = 2.2 [1.5-3.1], FDR < 0.05). There is no enrichment for this combined set in any other modules. Furthermore, the combined set of protein disrupting and missense RDNVs from unaffected siblings was not found enriched any modules ($p > 0.05$).

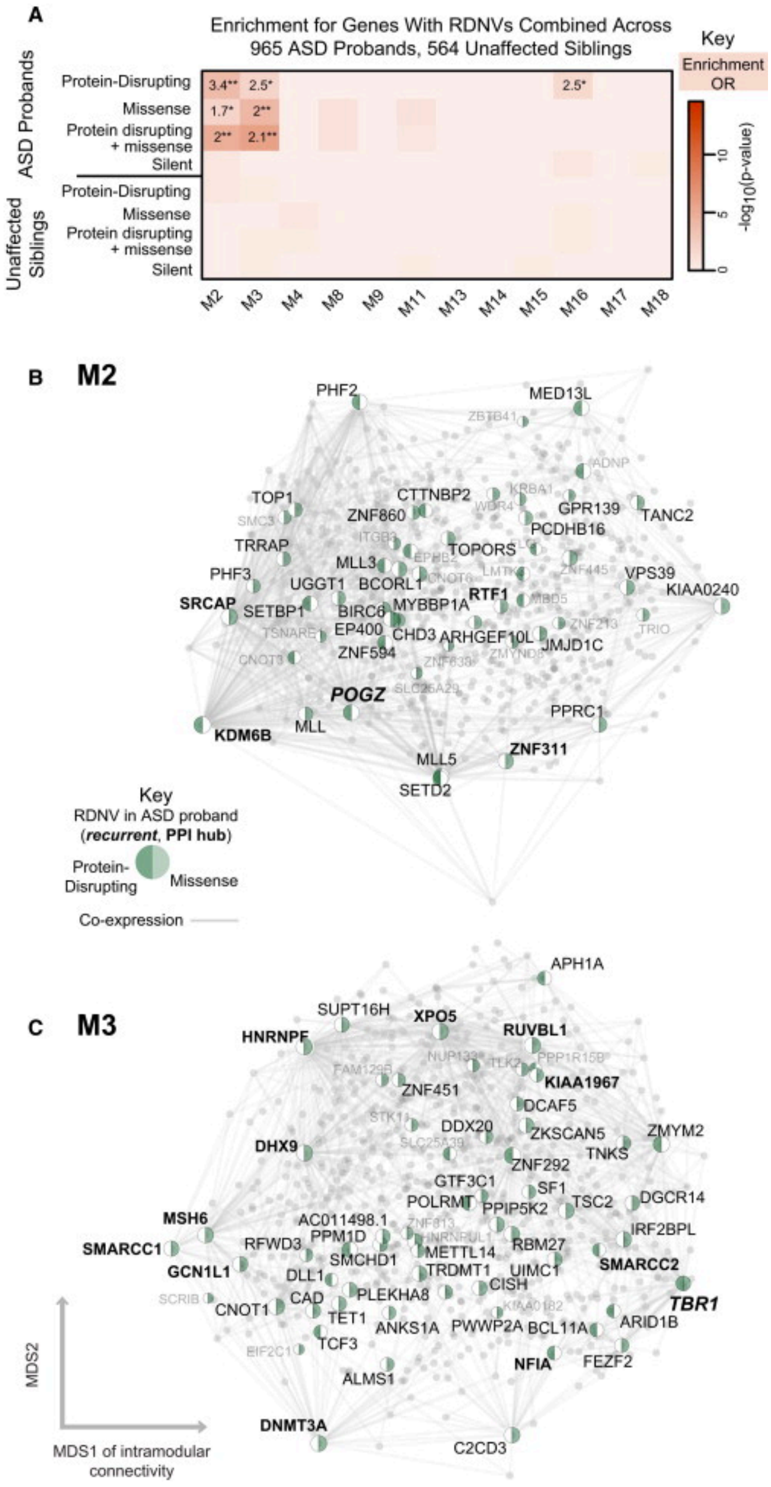


Figure 2.4 Enrichment of Genes Affected by RDNVs in Developmental Networks. (A) Module-level enrichment for multiple categories of RDNV in ASD affected probands and unaffected siblings combined across four studies. M2 and M3 are strongly enriched for protein disrupting and missense RDNV-affected genes in probands. Enrichment for genes affected by silent RDNVs in probands and RDNV gene sets affected in siblings represent

control gene sets and do not show enrichment. All enrichment values for overrepresented lists with $p < 0.05$, OR > 1 are shown to demonstrate enrichment trends ($^{*}p < 0.05$, **validated in replication set). Heatmap colors for p values reflect enrichment trends; p values for gene sets with OR < 1 can be seen in Table A1.2B.
(B and C) (B) and (C) show network plots for M2 and M3, with all genes plotted and all genes carrying RDNVs displayed. Network plots show all genes in the module with protein disrupting or missense RDNV-affected genes highlighted. For visualization, genes with high intramodular connectivity (kME > 0.75) are labeled in black, and the rest are labeled in gray. By definition, all edges in the network reflect positive correlations. The top 1,000 connections are shown, and genes are plotted according to the multidimensional scaling of coexpression as in Figure 2.3. See also Figures A1.2 and A1.3 and Table A1.2.

I further validated the observed RDNV enrichment pattern in M2 and M3 in an independent set of patients from a study with more stringent RDNV calling criteria (Iossifov et al., 2012). In this additional set of 343 ASD probands and unaffected siblings, I found that the patterns of RDNV enrichment replicated, with the set of protein disrupting and missense RDNVs from ASD probands enriched specifically in M2 and M3 ($p < 0.05$), and RDNVs from siblings and silent RDNVS not enriched in any set (Table A1.2B, Replication Set). Combining across all studies, I find that out of 598 protein disrupting or missense RDNV affected genes expressed in brain, 52 are in M2 ($p = 9.6 \times 10^{-6}$, OR = 2.0 [1.5-2.8]) and 61 are in M3 ($p = 8.5 \times 10^{-7}$, OR = 2.1 [1.6-2.8]). Importantly, the enrichment pattern across modules is not only replicated in the independent set, but is stronger in the combined set, is robust to perturbations in module composition (Figure A1.3A), and is not driven by variants from any one study (Table A1.2C-D). I show the enrichment pattern of this combined set across 965 ASD probands and 565 unaffected siblings in Figure 2.4A and use this combined set for the remainder of this analyses.

I next asked whether M2 and M3 prioritized functional subsets of genes with RDNVs. I confirmed that RDNV-affected genes in M2 and M3 are significantly enriched for interactions at a protein level (Figure A1.2A-D), and highlight genes that are both PPI hubs and co-expression hubs in Figure 2.4B-C. Furthermore, M2 and M3 genes harboring RDNVs are also more dosage sensitive, as evidenced by the significant increase in the probability of haploinsufficiency (P(HI), see section 2.5, Materials and Methods) among genes affected by these mutation classes (Huang et al., 2010; Luo et al., 2012). This is consistent with the heterozygous state of variants observed in ASD probands. Overall, a remarkable proportion, 113/598 (19%) of genes affected by known RDNVs are co-expressed in two modules reflecting similar temporal trends of high expression in cortex during the neurodevelopmental period of early neuronal fate determination, migration, and cortical lamination. Of note, as with M13, M16, and M17, which were enriched for asdM12 and SFARI ASD, ID genes showed no enrichment in M2 or M3 ($p > 0.05$).

I also observed that the SFARI ASD genes and asdM12 genes, which are enriched for inherited common variants in ASD (small average effect size), affect the synaptic modules, M13, M16, and M17. In contrast, the non-inherited (larger average effect size) RDNVs preferentially affect the early transcriptional regulation modules (see section 2.5, Materials and Methods). I emphasize that this is not absolute, as M16 includes some genes harboring RDNVs (e.g. in *SCN2A, SHANK2, NRXN1*; Figure 2.3A). To formally assess common variant enrichment using independent data, I compared ASD GWA signals across these modules (see section 2.5, Materials and Methods). Genes in M13 and M16 were more strongly affected by common variation in at least one of two ASD GWA studies (Anney et al., 2012; Wang et al., 2009) than M2 or M3 (Figure A1.3E). This is consistent with susceptibility of distinct biological processes for different mutational classes, and that in general more severe biological consequences would result from early transcriptional dysregulation during neuronal proliferation and differentiation, compared with later disruption of synaptic development and neuronal function.

2.3.d: ASD gene enriched modules are linked by translational and transcriptional regulation

Upregulated and downregulated modules are highly anti-correlated throughout development, so I hypothesized that common molecular regulatory relationships could potentially link genes within these modules. I first used a set of FMRP-RNA interactors from a cross-linking and immunoprecipitation (CLIP) experiment (Darnell et al., 2011), since Iossifov et al. (2012) had previously shown that RDNVs identified in their exome sequencing study were enriched in this class of genes. Remarkably FMRP targets are specifically enriched in modules that also contain ASD-related genes M2, M16, and M17 (FMRP-M2 $p = 1.6 \times 10^{-13}$, OR = 3.0 [2.3-3.9]; FMRP-M16 $p = 2.4 \times 10^{-29}$, OR = 5.7 [4.3-7.6]; FMRP-M17 $p = 9.3 \times 10^{-10}$, OR = 2.4 [1.8-3.1]; all at FDR < 0.05; Figure 2.5A). This provides a strong, independent line of evidence that translational regulation by FMRP not only affects genes harboring RDNVs, but links different molecular pathways that are co-expressed during early fetal cortical development and susceptible to diverse classes of ASD genetic mutation.

I next tested whether ASD associated modules are also linked at the transcriptional level (see section 2.5, Materials and Methods). I found 17 TFs that are predicted to link at least one upregulated and one downregulated module based on binding site enrichment (Figure 2.5B, Table A1.3A-B). Many of these TFs are expressed during fetal development (Table A1.1A), have been previously implicated in relevant neuronal functions, and have DNA binding targets have been experimentally characterized (Table A1. 3B). For example, M*EF2A* and *MEF2C*, both members of a TF family regulating synaptic plasticity and glutamatergic synapse number (Ebert and Greenberg, 2013), are enriched for binding targets in M2 and M17, which are anti-correlated across development (Figure 2.5C-D). *SATB1*, which is required for the development of cortical interneurons (Close et al., 2012), *ELF1*, which is involved in axonal guidance, and *FOXO1* which regulates neuronal polarity (la Torre-Ubieta and Bonni, 2011) also link these two modules (Figure 2.5E-F). To provide further evidence that these are experimentally plausible binding sites, I overlaid these bioinformatic predictions with chromatin immunoprecipitation (ChIP) data where available, supporting many of these predicted interactions, including 39% of MEF2A, 23% of MEF2C and 87% of *ELF1* binding sites (Figure 2.5C, 2.5D, 2.5G; see section 2.5, Materials and Methods). These results implicate existing and novel TFs as putative co-regulators of ASD-associated gene networks during neocortical development.

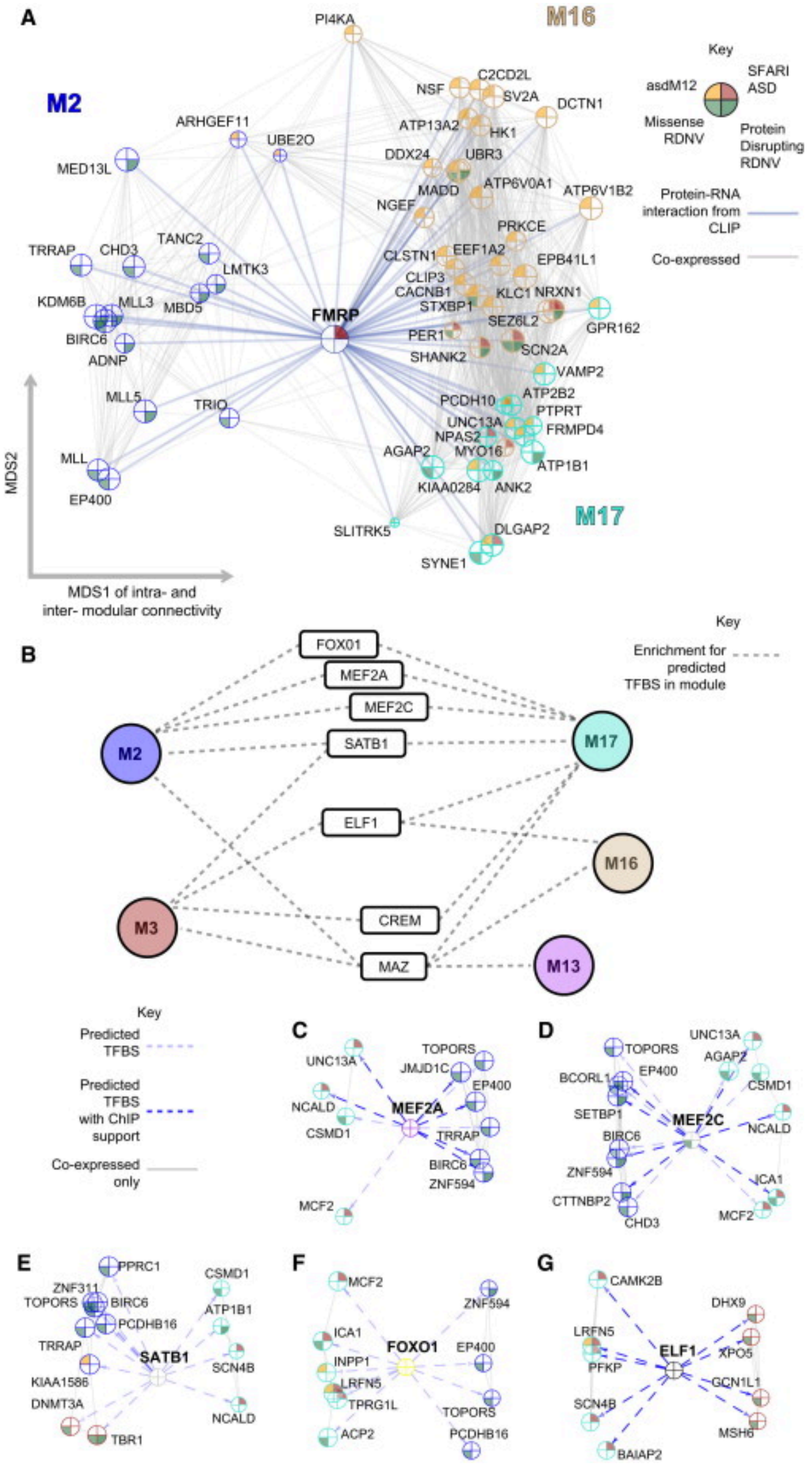


Figure 2.5 Translational and Transcriptional Coregulation Connect Developmentally Distinct ASD-Affected Modules. (A) Coexpression-based network plot of FMRP interactions with genes in M2, M16, and M17 that are either affected by RDNVs or are in an ASD candidate list. Genes are plotted as in Figures 2.3 and 2.4 but now across modules, with FMRP placed at the center. (B) Summary of TF binding site (TFBS) enrichment in modules

for TFs that have evidence for function in a neurodevelopmental context and link anticorrelated modules. Dashed lines indicate enrichment in the module for predicted binding sites. (C–G) MEF2A, MEF2C, SATB1, FOXO1, and ELF1 are all enriched for their binding motifs in the upstream regions of ASD gene-enriched modules following anticorrelated developmental patterns. Network plots highlight genes with a predicted binding site (light dashed arrow) for the TF (placed at the center) contributing to this enrichment that are also affected by RDNVs or in an ASD candidate list. Arrows representing a TFBS found in a ChIP experiment are marked in dark blue.
For network plots, the top 1,000 positive connections between genes are plotted, and node size is proportional to connectivity within the genes' assigned module; therefore, larger nodes are more central hubs. The outer color of each node reflects its module membership, and coexpression edges in the network reflect positive correlations. See also Tables A1.2 and A1.3.

2.3.e: ASD-associated genes exhibit laminar and cellular enrichment

Deficits in cortical patterning and layering have been observed in ASD (Voineagu et al., 2011), I therefore tested whether ASD-affected genes are enriched in the developing laminae of fetal cortex and the terminally differentiated laminae of adult cortex (see section 2.5, Materials and Methods). I compared multiple ASD gene lists with the ID gene sets for enrichment in laminae of the developing and adult cortex, and found a sharp contrast in laminar enrichment between ASD and ID genes (Figure 2.6A-B). Additionally, in adult, asdM12 exhibits strongly significant enrichment in L3 ($Z > 2.7$, $FDR < 0.01$), while other ASD lists follow a similar trend of superficial layer enrichment ($Z > 2$, $p < 0.05$). In contrast, the "ID all" and "ID only" gene sets follow a trend of lower layer enrichment (Figure 2.6B), an across-layer pattern that is significantly different from all of the ASD lists (Figure 2.6C-D, see section 2.5, Materials and Methods).

I also observed a similar trend in superficial layer enrichment for the modules that are enriched in asdM12 genes (M13, M16, and M17; Figure 2.6F). M13 and M16 also exhibit weaker enrichment in L5 and L6. Module-level analysis in fetal brain also highlighted a difference between the RDNV enriched modules, M2 and M3. Although both M2 and M3 are most highly expressed in early human fetal development (prior to PCW 17), M2 reaches its peak later and is enriched in the cortical plate (CPi/CPo), whereas M3 peaks earlier, consistent with its enrichment in the germinal zone (VZ, SZi, SZo; Figure 2.6E). In adult, this distinction is no longer present (Figure 2.6F), with both M2 and M3 showing enrichment in superficial layers (L2, L4). I also asked whether any of these gene sets or modules were enriched for cell-type specific markers paralleling the observed laminar enrichment. I observed enrichment in this set of well-curated upper layer glutamatergic neuron markers among asdM12, M2, and M3 genes (Figure A1.4C-D), which agrees with the L2-4 enrichment of asdM12 and ASD risk gene modules.

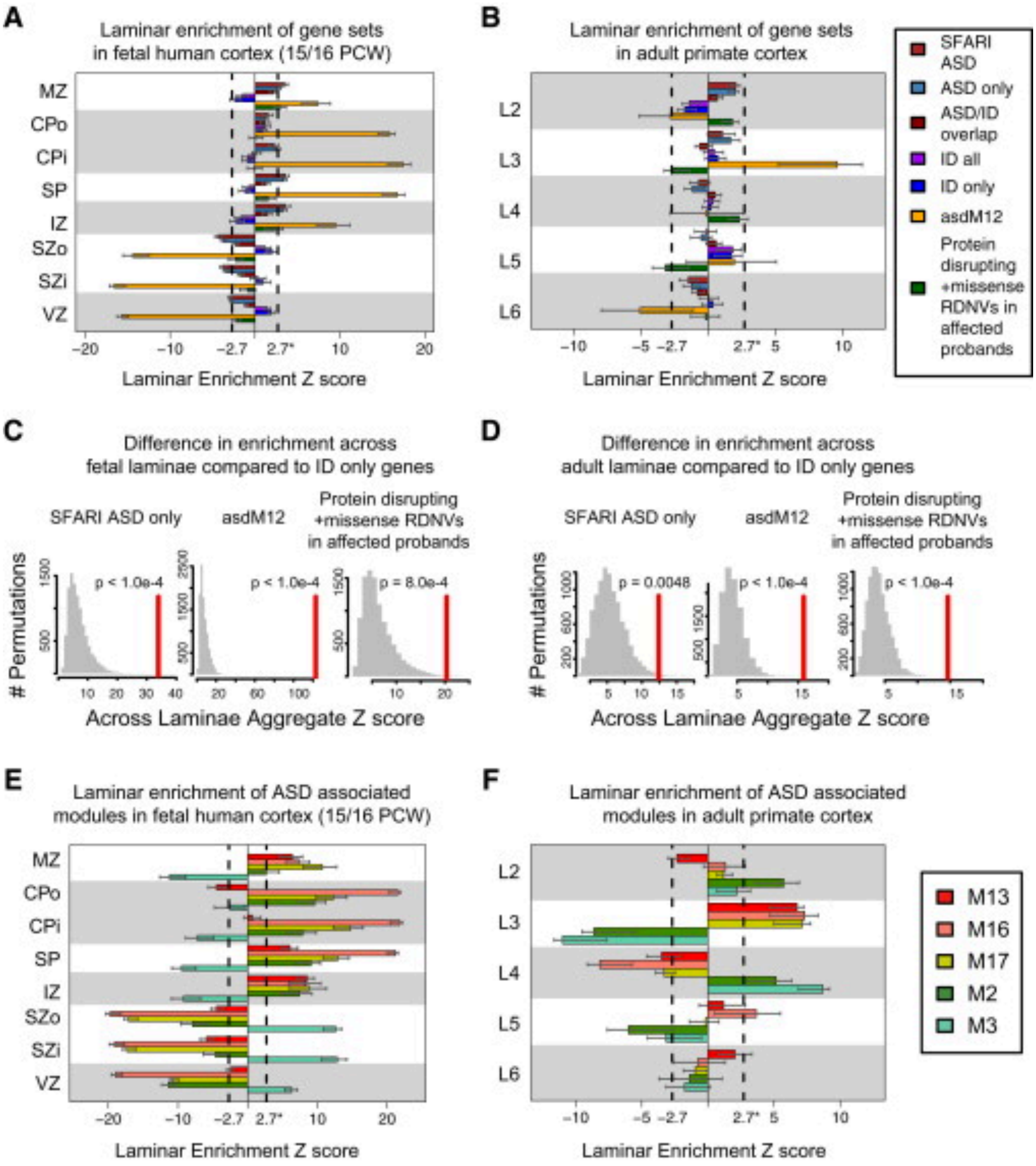


Figure 2.6 Enrichment for Laminar Differential Expression of Gene Sets and Associated Developmental Coexpression Modules in Fetal Human and Adult Primate Cortex. (A) In fetal cortex, ASD sets (SFARI, asdM12, and RDNV affected) are enriched for differential expression in laminae containing postmitotic neurons, whereas genes implicated in ID are weakly enriched in germinal layers. A high *Z* score for a gene set in a layer corresponds to differential expression across the gene set in that layer. (B) In adult cortex, asdM12 sets show strong enrichment in layer 3, whereas ID genes are weakly enriched in layer 5. (C and D) Summing the *Z* score across

layers in (A) and (B) and comparing to randomly permuted sets of genes of similar size demonstrates that, in both fetal and adult cortex, the laminar distribution of multiple ASD implicated gene sets is significantly distinct from that of genes implicated only in ID. (E) SFARI/asdM12-associated developmental coexpression modules M13, M16, and M17 follow enrichment trends similar to the SFARI/asdM12 gene set in fetal brain. However, the modules strongly associated with the RDNV affected genes, M2 and M3, show distinct enrichment patterns.
(F) ASD-associated modules are predominantly enriched in superficial layers 2–4 of adult cortex. Additionally, M16 shows weak enrichment in L5. In contrast to fetal cortex, M2 and M3 are in enriched in the same laminae in adult, suggesting that they serve distinct functions during cortical development that contribute to superficial cortical layers 2–4. Dashed lines in bar plots indicate Z = 2.7 (equivalent to FDR = 0.01); error bars indicate 95% bootstrapped CIs. Laminae: marginal zone (MZ), outer/inner cortical plate (CPo/CPi), subplate (SP), intermediate zone (IZ), outer/inner subventricular zone (SZo/SZi), ventricular zone (VZ), and adult cortical layers 2–6 (L2–6). See also Figure A1.4.

Figure 2.7A highlights adult layer-level expression patterns of several strong ASD candidate genes with enriched expression in superficial layers (e.g. *SHANK2*, *CNTNAP2*) and shows that many genes recurrently affected by protein disrupting RDNVs in the 965 ASD probands and an additional set of patients assessed by targeted sequencing (O'Roak et al., 2012) also show superficial layer enrichment (e.g. *SCN2A*, *POGZ*, Figure 2.7B). I use these mature laminae for cell-marker enrichment analyses because laminar expression patterns are more clearly delineated relative to PCW 15-21 (Figure 2.6A and 2.65E, Figures A1.4A-B). Furthermore, neuronal migration in humans persists into the third trimester, and upper layer neuronal identity is not finalized until after PCW 28 (Bystron et al., 2008). Out of the 6 genes with recurrent RDNVs in probands in which I can detect layer preference, 5 are predominantly expressed in superficial layers in adult. Some of the genes in Figure 2.7 also show expression in a lower layer (*NLGN1, SCN2A, ITPR1, MLL3*), though superficial layer enrichment is stronger (larger differential expression t-value in Table A1.1A).

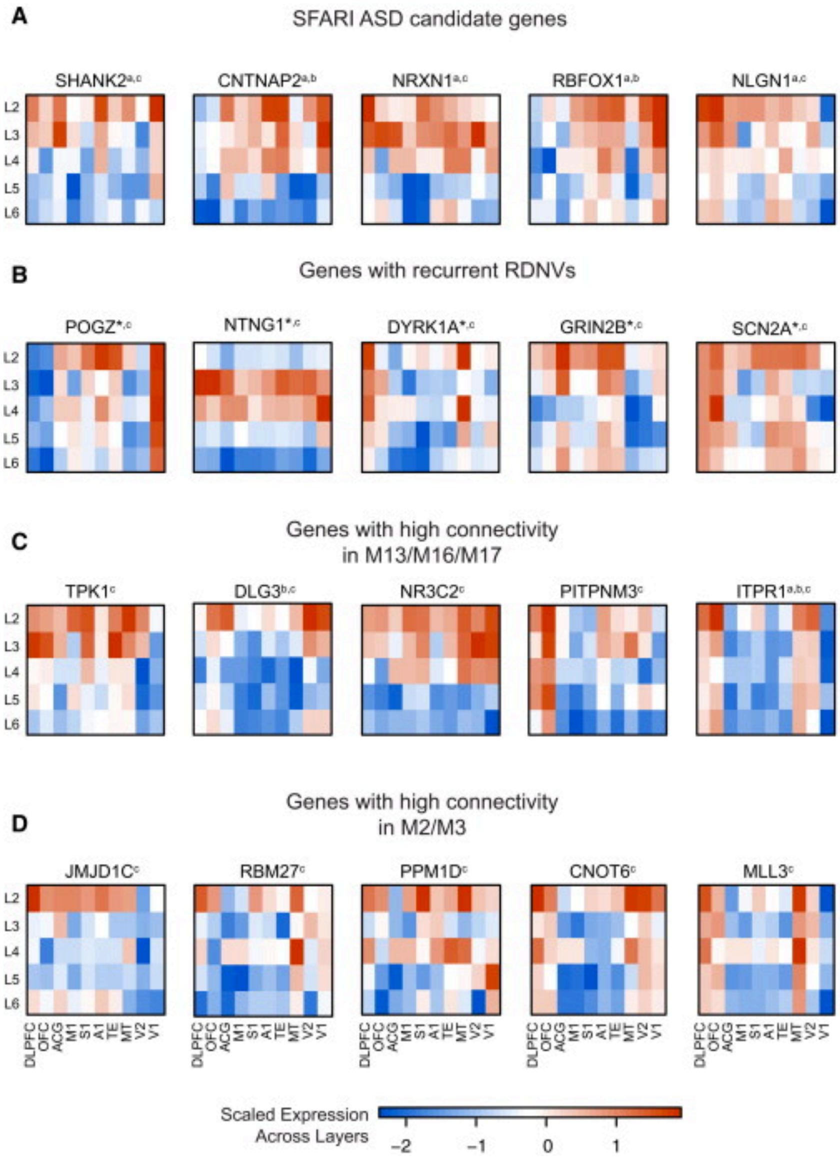


Figure 2.7 Laminar Patterns for Genes Implicated in ASD. (A) SFARI candidate genes for ASD. (B) Genes with recurrent RDNV evidence across studies. Genes not displayed include *TBR1* (lower layer enriched), *CHD8* (no layer enrichment detected), *CUL3* (no layer enrichment detected), and *KATNAL2* (not detected in these data). (C) Genes with high connectivity in M13, M16, and M17. (D) RDNV genes with high connectivity in M2 and M3. ^{a}indicates membership in SFARI ASD, ^{b} indicates membership in asdM12, ^{c} indicates the gene is affected by a RDNV, and the asterisk indicates recurrent RDNVs. Color bar values represent scaled expression (SDs from the mean-centered expression value across layers). All genes shown have t > 2 for enrichment in an upper layer (L2, L3, or L4) over background and t < 2 for lower layers (L5 or L6). Regions: dorsolateral prefrontal (DLPFC), orbitofrontal (OFC), anterior central gyrus (ACG), primary motor (M1), primary somatosensory (S1), primary auditory (A1), higher-order visual area TE (TE), higher-order visual area MT/5 (MT), secondary visual cortex (V2), and primary visual cortex (V1).

2.4: Discussion

These analyses offer a genome-wide neurobiological context to begin to unify the genetics of ASD, providing robust evidence of both molecular pathway and circuit-level convergence (Figure 2.8A-B). Integration of ASD genes with developmental co-expression networks and laminar expression data connects multiple ASD risk enriched modules to glutamatergic neurons in upper cortical layers (L2-L4), tying ASD risk genes to specific brain circuitry (Figure 2.8C). The observation of convergent biology in ASD stands in striking contrast with ID, which does not show the same level of developmental or anatomical specificity. Laminar enrichment in the "ASD/ID overlap" genes show a similar pattern as the "ASD only" genes (in L2, Figure 2.6B). Therefore disruption in ID genes that also cause ASD likely affects superficial layers compared to disruption in genes causing ID only; these analyses lead to the prediction that specific disruption of cortical-cortical connectivity, for example by targeting upper layer glutamatergic neurons which predominantly comprise inter- and intra-hemispheric projections, is more likely to affect core ASD phenotypes such as social behavior, rather than general intellectual ability alone.

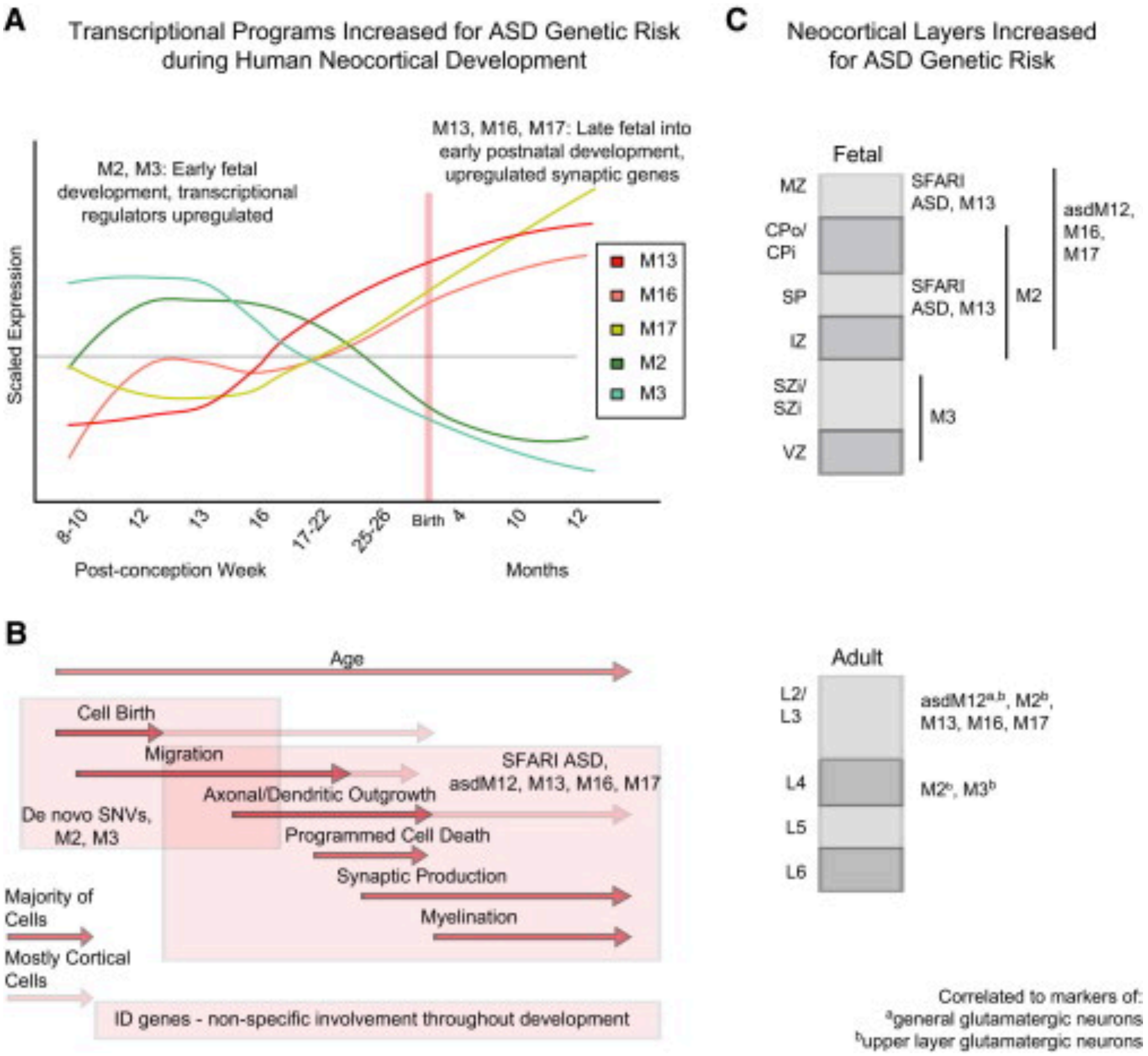


Figure 2.8 Summary of Findings and Model for Effects of ASD Implicated Gene Sets. (A) ASD risk genes from multiple sources were enriched in five coexpression modules throughout development—M2, M3, M13, M16, and M17. (B) Early transcriptional regulators in M2/M3 are enriched for RDNVs, whereas the later expressed synaptic genes are associated with previously studied ASD genes (biological process time periods adopted from Andersen, 2003). (C) ASD genes are most consistently associated with laminae containing postmitotic neurons during early fetal development (broadly in IZ, SP, CPo/CPi, and MZ) and superficial layers in adult (L2–L4). Multiple modules are also strongly associated with markers of upper-layer glutamatergic neurons in adult cortex, suggesting many ASD genes preferentially affect these cell types. (B) and (C) also summarize that ID genes are largely distinct from ASD genes in both developmental trajectory and neocortical layer enrichment. See also Table A1.4. Both (A) and (B) correspond to the same timescale as marked by the axis on the plot in (A). I summarize the strongly enriched findings but note that weaker enrichment for other patterns exists that may be important for subsets of ASD.

Individual genes can be prioritized for biological validation using a combination of network position, bioinformatic scores, and the biological context highlighted here, as shown in Table A1.4.

This analysis further links specific molecules and pathways to the cortical-cortical intra- and inter-hemispheric disconnection that has been hypothesized as a shared circuit-level deficit unifying diverse ASD etiologies (Belmonte, 2004; Geschwind and Levitt, 2007). An illustrative example is the disruption of *ARID1B*, a BAF complex member that harbors a RDNV and is a hub of M3. Severe mutations in *ARID1B* cause corpus callosum abnormalities, ID, and ASD (Halgren et al., 2011; Santen et al., 2012). Another BAF complex member, SMARCC2, implicated by RDNVs in probands, controls cortical thickness by repressing the pool of intermediate progenitors, which preferentially contribute to forming cortical layers 2-4 (Tuoc et al., 2013), providing another molecular link to inter- and intra-hemispheric connectivity. These analyses make the first systematic connection between genes disrupted in ASD and this circuit-level disruption. As additional genes in the early fetal co-expression modules are found to harbor recurrent RDNVs, cortical-cortical connectivity will be a valuable phenotype to assess in both animal models and human patients.

Translational regulation by FMRP during fetal cortical development and transcriptional co-regulation of ASD candidate genes provides another level of convergent biology in ASD, and a rich starting point for further experimental investigation. Notable also are TFs that are predicted to drive the transcriptional co-regulation of molecular and circuit-level processes, including *MEF2A*, *MEF2C*, and *SATB1*, which have binding site enrichment in M2 and M17. This is intriguing in light of decreased *PVALB* expression in ASD brain (Voineagu et al., 2011), the hypothesized convergent mechanism of a shift in the excitation-inhibition balance in ASD (Rubenstein and Merzenich, 2003), and the observation that *SATB1* plays a key role in regulating cortical PVALB+ and SST+ interneuron development (Close et al., 2012; Denaxa et al., 2012). I speculate that M2 and M17 reflect processes involved in the migration and differentiation of inhibitory and excitatory cell populations whose balanced co-regulation may be essential to proper cortical development. These analyses underscore the notion that understanding the structure of the transcriptional and chromatin regulatory networks underlying cortical development and their relationship to translational control will better inform the genetic risk architecture of ASD.

In addition to demonstrating biological convergence, network analysis further allowed me to stratify the full set of 684 RDNV-affected genes to a narrower list of 113 genes (Table A1.1A) that I hypothesize are more likely to confer increased ASD risk based on their enrichment in M2 and M3, and an elevated probability of conferring a phenotype when haploinsufficient. Furthermore, I demonstrate that the observed enrichment is specific by comparison to silent RDNVs and unaffected siblings' RDNVs. As an example of how to prioritize these candidates further based on the functional relationships summarized in Figure 2.8, I constructed a list of candidates using Table S1A, filtering by expression during development, membership in M2 or M3, high predicted haploinsufficiency (P(HI) > 0.5), protein disrupting or missense mutation in probands, and either a layer preference (t > 2 for a particular layer) or a cell-type preference (r > 0.2 for a cell-type) in Table S4. This yields a set of 24 candidates with a hypothesized layer- or cell-type phenotype for investigation. Among these, *TBR1* is known to harbor recurrent mutations, while *CHD3* is a member of the same gene family as *CHD8*, a gene with the strong recurrent *de novo* mutation evidence (O'Roak et al., 2012). Additionally, *SMARCC1* and *SMARCC2* are members of the BAF complex, which is of particular interest since it is statistically associated with ASD: 6/28 BAF complex genes are affected by RDNVs ($p = 1.5x10^{-}$

^{3}, OR = 5.7 [1.9-14.5]). Other RDNV-affected molecular families, including the CCR4-NOT complex members (*CNOT* family) and chromodomain helicase DNA binding proteins (*CHD)*, are also seen in M2 and M3 and have been linked to the regulation of neuronal proliferation and differentiation (Feng et al., 2013; Potts et al., 2011; Ronan et al., 2013; Zheng et al., 2012).

In parallel work, Willsey et al. 2013, find convergence on fetal cortical developmental networks in frontal lobe by seeding with a subset of high confidence ASD genes identified by exome sequencing. Despite the different analytical approaches, there is remarkable overlap between the developmental processes implicated by the gene networks identified in our studies. Although I see the strongest cell-type and layer enrichment in adult L2-4, I also see a signal in CPi during fetal development and a weaker signal in L5-6 of adult, consistent with a subset of genes affecting lower-layer glutamatergic neurons. Together, our studies highlight the importance of understanding the spatial and temporal context of specific genes for future mechanistic investigation.

I also acknowledge several issues that challenged my approach. Many of the genes I identified as putatively involved in ASD do not have complete PPI data, P(HI) scores, TF binding site information, or are not well studied in brain. This is one reason why I rely most heavily on RNA-seq based transcriptome data, as it comprehensively represents relationships present in the developing human brain in an unbiased manner. I did not assess enrichment of genetic hits in other brain regions across development, as sample size and cell-type heterogeneity make it difficult to interpret co-expression across cytoarchitecturally diverse brain regions such as cerebellum and amygdala, which may also be involved in ASD (Amaral et al., 2008). I also focused on single gene disruption in ASD and did not include CNVs affecting multiple genes to improve signal to noise. Additionally, current genetic approaches favor *de novo* mutation detection; as different classes of mutations (e.g. inherited rare coding or non-coding regulatory variants) are identified, I speculate that heritable variants will affect genes in the modules related to synaptic development and function, rather than earlier transcriptional regulation. Likewise, It will also be useful to investigate rare, inherited recessive ASD risk variants (Lim et al., 2013; Yu et al., 2013) when sufficient data are available, so as to compare with other forms of genetic variation.

The conclusions summarized in Figure 2.8 pass a stringent multiple comparisons cut-off; weaker enrichment patterns may become more salient with higher resolution tiling of gene expression during development and increased sample sizes in sequencing studies. To facilitate future studies I have shared the code used in this analysis (http://labs.genetics.ucla.edu/horvath/htdocs/CoexpressionNetwork/developingcortex/) and provided a graphical interface for exploring specific genes within the network context (http://geschwindlab.neurology.ucla.edu/sites/all/files/networkplot/ParikshakDevelopmentalCortexNetwork.html). I have shown how an integrative approach, which is not driven by any small set of samples, candidate genes, or candidate hypotheses, can place heterogeneous genetic etiologies into a unifying structure. These analyses provide a working framework for mechanistic investigation and hypothesis testing, which points to interactions between genes in specific cell types and circuits, as well as the general biological processes in which these genes are implicated.

2.5: Materials and Methods

Developmental expression data: BrainSpan developmental RNA-seq data (obtained from www.brainspan.org) summarized to Gencode v10 (Harrow et al., 2006) gene-level reads per kilobase million mapped reads (RPKM) values were used (see Extended Methods in Appendix A1 for data preprocessing, see Table A1.1D for sample details). Only neocortical regions were used in this analysis and only genes with a normalized RPKM value of 1 in at least one region at one time point for 80% of the available samples were considered expressed.

Weighted gene co-expression network analysis: I used the R package WGCNA (Langfelder et al., 2008) to construct co-expression networks, as previously done (Voineagu et al., 2011) and described in detail in Appendix A1. The modules were characterized using GO Elite to control the network-wide false discovery rate, with all enriched pathways comprising at least 10 genes at $Z > 2$ and $FDR < 0.01$ (Zambon et al., 2012). All network plots were constructed using the igraph package in R (Csárdi and Nepusz, 2006).

Protein-protein interaction enrichment analysis: Protein-protein interactions were compiled from two resources, InWeb (Liu et al., 2011) and BioGRID (Stark, 2006). A union of the two networks was taken, and a degree-matched permutation analysis was applied in order to control for biological and methodological biases in PPI data (see Appendix A1 for details).

Gene Sets: The SFARI ASD set was compiled using the online SFARI gene database, AutDB (https://gene.sfari.org/autdb/, accessed 8/20/2012). I used the Gene Score to restrict the set to those categorized as S (Syndromic) and evidence levels 1-4 (high confidence - minimal evidence). This resulted in 155 total genes. I obtained asdM12 (432 genes) and adsM16 (377 genes) from a gene expression study that profiled expression changes in ASD cortex and applied WGCNA identify modules of dysregulated genes ASD (Voineagu et al., 2011). I curated ID genes from four reviews cataloging all known genes causing ID (Inlow, 2004; Lubs et al., 2012; Ropers, 2008; van Bokhoven, 2011) and supplemental table 6 from Neale et al., 2012, resulting in 471 genes. All candidate gene sets are available in Table A1.9A. I obtained RDNVs from four publications (Iossifov et al., 2012; Neale et al., 2012; O'Roak et al., 2012; Sanders et al., 2012), which in total identify 122 protein-disrupting, 528 missense, and 210 silent RDNV hit genes in affected individuals and 40 protein-disrupting, 307 missense, and 122 silent RDNV affected hit in unaffected siblings.

Gene set over-representation analysis: All enrichments of gene sets were performed using a two-sided Fisher exact test with 95% confidence calculated according to the R function fisher.test. To declare a set as enriched, I required at least 10 genes to overlap and an OR of 1.5. For strong enrichment, I further required the enrichment to pass a Benjamini-Hochberg FDR < 0.01 (Benjamini and Hochberg, 1995).The background set for protein coding genes is defined by the biotype annotation "protein_coding" in GENCODE. Genes were overlapped according to the HUGO symbol, and all conversions among identifiers were performed using the R package biomaRt (see Appendix A1 for details).

Transcription factor binding site enrichment: For each TF in TRANSFAC (Portales-Casamar et al., 2010), I assessed enrichment as follows: 1) putative motifs bound by the TF were obtained from the databases. 2) 1000bp upstream sequences of the top 200 genes by kM were scanned to calculate motif enrichment; and iii) enrichment above background was calculated compared to 3 background datasets to ensure robustness: 1000 bp sequences upstream of all human genes, human CpG islands, and the sequence of human chromosome 20. Only TFs with $p < 0.05$ across all backgrounds are considered enriched (see Appendix A1 for details).

Layer-specific and cell-type marker enrichment: I utilized human fetal neocortical laminar gene expression datasets from BRAINSPAN, two for each of the earlier and later fetal periods and primate neocortical laminar gene expression data from. A Laminar Enrichment Z-score, which is a z statistic quantifying the skew of differential expression t-values of a given gene set in a layer against background was calculated for each gene set in each cortical layer. This normalized distribution of individual gene t-values is expected to follow the same distribution as the background set ($Z = 0$) if the genes in the set exhibit no layer specificity. To quantify cell-marker relationships, I the same method, with the t-value replaced by the correlation of each gene to the first principal component of a set of known cell marker genes (Table A1.6 lists cell-type marker genes, see Extended Experimental Procedures for details). Comparisons between each ASD agene set and ID gene set were performed by 1) computing the difference in enrichment score between the two sets for each layer, 2) summing this difference across all layers, and 3) comparing this to the distribution of summed differences in layers of 10,000 randomly drawn pairs of sets matched in gene set size.

CHAPTER 3:
Comparison of gene network approaches in autism genetics

"Data! Data! Data!" he cried impatiently. "I can't make bricks without clay."

– Sherlock Holmes, Sir Arthur Conan Doyle, *The Adventure of the Copper Beeches*

“All models are wrong, but some are useful.”

– George E. P. Box

3.1. Introduction

In Chapter 2, I described work, which was published in 2013 (Parikshak et al., 2013) that uses gene co-expression networks to understand the neurobiological pathways, cells, and circuits affected by ASD risk genes. This work shows that developmentally co-expressed modules enriched for cell markers of upper cortical layers and excitatory glutamatergic neurons are affected by ASD risk genes from multiple sources, enriched for physical interactions, and enriched for co-regulatory relationships. Since publishing this work, additional genetic data implicating genes in ASD (De Rubeis et al., 2014; Iossifov et al., 2014) as well as higher resolution cellular transcriptomes have become available and been used for identifying more detailed cellular enrichments in disease(Doyle et al., 2008; Molyneaux et al., 2015; Xu et al., 2014; Zhang et al., 2014). I reasoned these data could allow me to assess the robustness of the statements made about the convergent biology of ASD in Chapter 2, as well as enable a formal evaluation of the predictive value of the modules implicated in ASD and the affected cell types and circuits.

Additionally, I used genome-wide co-expression network analysis to identify the modules in Chapter 2 (unseeded, unsupervised, see Chapter 1 [section 1.3] for more on different types of gene networks that have been applied to ASD). Multiple studies have since attempted to identify gene networks related to ASD risk using seeded co-expression networks (Willsey et al., 2013), seeded co-expression and protein interaction networks (Hormozdiari et al., 2015), unseeded protein interaction networks (Li et al., 2014), and seeded integrated molecular phenotype networks (Chang et al., 2015) all of which are discussed in Chapter 1. A common yet somewhat implicit assumption in the literature is that PPI networks are the best way to understand how risk mutations might converge on similar biological processes (Krumm et al., 2014; Neale et al., 2012; O'Roak et al., 2012). However this may not be the case for neuropsychiatric disease, where the developmental processes affected may involve specific cellular populations disrupted at specific times and given that currently available PPIs do not reflect any neurobiological specificity. It is therefore important to understand the relative strengths and weaknesses of these approaches, and compare them against each other to understand their relative predictive power for ASD risk affected genes and pathways and guide future gene network studies in neurodevelopmental disorders such as ASD.

In this chapter, I first describe criteria that will be used for evaluating recently available mutation and cell type data and apply them to the previously identified co-expression modules from Chapter 2. I then compare different gene network methods to evaluate how robustly they identify mutated genes in ASD, and whether they identify a level of biological specificity that is of value beyond the approach used in Chapter 2.

3.2. Background

3.2.a. Overview of gene networks in ASD and criteria for evaluation

Gene networks are constructed by connecting genes through shared functional relationships (see Chapter 1 for details). However, there are a plethora of valid methods by which one can relate genes to each other, and combinations of these methods may be used to identify greater biological specificity (Mitra et al., 2013). Additionally, detecting cliques or modules in these networks is usually guided by heuristics, as it is difficult to truly optimize all possible parameters. Even where this is done using precise objective functions that maximize some desirable properties of genes in a module (e.g. connectivity between genes, pathogenic mutations in the module) and minimize undesirable properties (e.g. connectivity to genes outside the module or in other modules, pathogenic mutations found in control samples) in the module, (Gilman et al., 2011; Hormozdiari et al., 2015; Langfelder et al., 2008; Segal et al., 2003), the inclusion or exclusion of parameters to optimize in the objective function is subjective.

Two guiding principles that can be used in constructing relevant and biologically valid networks are their predictive value, as quantified by their ability to make the same conclusions on new data sources, and their biological usefulness, as quantified by their enrichment in previously identified biological processes.

3.2.b. Overview of analyses in this chapter

Here, I assess the predictive value and biological usefulness of the five ASD associated modules from Chapter 2: M2, M3, M13, M16, and M17 (Parikshak et al., 2013), detected by genome-wide co-expression followed by gene set enrichment analyses. To clearly differentiate these modules from others that are discussed, I refer to them as devM2, devM3, devM13, devM16, devM17.

I first address the predictive value of these modules for identifying biological processes affected by rare *de novo* mutations (RDNVs) in ASD. The mutation enrichment results described in Chapter 2 were initially identified in a discovery cohort, validated in a replication cohort, and shown to pass multiple corrections and be even stronger when combining across 965 ASD exomes (Chapter 2, Figure 2.4). Now, mutations have been reported from over 1,700 new individuals using a uniform *de novo* variant calling pipeline which reduces inconsistencies in the initial studies (Iossifov et al., 2014). If the implicated modules have predictive value, they ought to show a similar signature of risk gene enrichment with these new data. Therefore, I perform enrichment analysis with the previously used gene sets (Figure 2.4) and contrast them to enrichment analysis using only newly implicated genes in each mutational category. This constitutes a *bona fide* replication effort. Additionally, I apply a modified enrichment approach using logistic regression and covariates to control for gene length, which is described and justified in A2 Additional Methods for Chapter 3.

I then use the same enrichment approach to re-assess fetal human cortex (Miller et al., 2014) and adult primate cortex (Bernard et al., 2012) laminar enrichments, and compare these with cell-type enrichments derived from RNA profiling of more homogenous cellular populations in mouse. The first dataset, reported by Zhang and colleagues (Zhang et al., 2014), applied RNA-seq to fluorescent activated cell sorted (FACS) neurons, astrocytes, oligodendrocytes, and microglia from mouse. The second dataset, reported by Doyle and colleagues (Doyle et al., 2008), utilized bacTRAP technology (Gong et al., 2003) to molecularly tag ribosomes from specific cell-types in mouse, purify RNA bound to those ribosomes from that specific cell type, and evaluate these cell-type specific transcriptomes using microarray. This study profiled multiple populations of neurons and glia from the cortex. To evaluate enrichments for lamina and cell types, I utilized a modification of the enrichment analysis used in Chapter 2 combined with the logistic regression framework used for mutational analysis (see A2 Additional methods for Chapter 3). I also use a third dataset, which assess the development of three subpopulations of cortical projection neurons from mid-fetal to early postnatal development in mouse (Molyneaux et al., 2015). Due to a lack of sufficient data from each time point or a whole brain background from the same experiments, I assess module relationships with these transcriptomes using a modified approach (see A2 Additional methods for Chapter 3).

3.2.c. Overview of additional gene network studies utilized for comparison

Finally, I compare modules (also referred to as clusters, cliques, or communities depending on the study) from different methods with each other to evaluate how well each method predicts future ASD genes, and to understand how they might implicate similar or distinct gene sets in ASD. The modules and methods used to define them follow:

- devM2/3/13/16/17 (Parikshak et al., 2013): network constructed by genome-wide signed weighted co-expression network analysis (Zhang and Horvath, 2005) of neocortical regions from BrainSpan (Sunkin et al., 2013), followed by clustering using the topological overlap to identify modules. Modules devM2 and devM3 were implicated as related to ASD using RDNVs if RDNV-affected genes from a discovery set encompassing three WES studies (Neale et al., 2012; O'Roak et al., 2012; Sanders et al., 2012) were enriched and if this enrichment was confirmed by RDNV-affected genes from an independent study (Iossifov et al., 2012). Modules devM13, devM16, and devM17 modules were implicated with ASD gene sets from co-expression in ASD brain (Voineagu et al., 2011), candidate lists (Basu et al., 2009), and GWAS (Anney et al., 2012; Wang et al., 2009a). Finally, circuits and cell types were implicated by assessing biases in the distribution of adult (Bernard et al., 2012) or fetal (Miller et al., 2014) laminar specific differential gene expression or cell-type marker correlations in each module.
- P3-5 PFC-MFC, P4-6: PFC-MFC, P8-10 MD-CBC (Willsey et al., 2013): modules defined by a network constructed from seeded binary unsigned ($|r| >= 0.7$) co-expression modules using different spatial and temporal combinations of brain samples (Kang et al., 2011). P = Period, which reflects developmental periods delineated in previous work (Kang et al., 2011). PFC, MFC, MD, and CBC reflect abbreviations for spatial regions, with PFC being prefrontal cortex, MFC being medial frontal cortex, MD being mediodorsal thalamus, and CBC being cerebellar cortex. The authors of this study

identified 9 high-confidence ASD genes at FDR < 0.5 using four WES studies (Iossifov et al., 2012; Neale et al., 2012; O'Roak et al., 2012; Sanders et al., 2012), their own data (Willsey et al., 2013), and CNVs with breakpoints in protein coding genes (Talkowski et al., 2012). The authors used the high-confidence genes as seeds, and expanded modules using multiple combinations of spatial and temporal co-expression, implicating modules with enrichment for remaining ASD protein disrupting RDNVs as implicated in ASD. This identified three modules, which were then evaluated for neuronal gene enrichment and co-expression in fetal layers. The authors found high co-expression of genes in the inner cortical plate during fetal development, and interpreted this as enrichment for lower layer glutamatergic neurons. Furthermore, modules were found to be enriched for genes implicated by an alternative mutation enrichment approach, but this was not an independent validation.

- MAGI: ASD - ID M1, MAGI ASD + ID M1/2/3 (Hormozdiari et al., 2015): networks constructed by integrating unsigned binary ($|r| >= 0.6$) co-expression from BrainSpan (Sunkin et al., 2013), and PPI from the Human Protein Reference Database, HPRD, (Keshava Prasad et al., 2009) and the Search Tool for the Retrieval of Interacting Genes/Proteins, STRING, (Franceschini et al., 2012). The authors seeded modules with pathways identified using genes mutated in ASD and ID (ASD + ID) or ASD only (ASD - ID), and identified modules that optimized intramodular connectivity and pathogenic mutation burden in cases while minimizing pathogenic mutation burden in controls. They demonstrated these modules were more specific for pathways compared to previous work, though this is somewhat tautological as their modules were seeded on pathways. They also demonstrated that their modules contained a large differential in the burden of pathogenic mutations in cases compared to controls, something not seen in any previous method. However, no reproducibility analyses were conducted.

- ppiM2/13 (Li et al., 2014): a genome-wide PPI network was constructed using BioGRID (Stark, 2006), and modules were defined using the topological overlap using a supposedly parameter-free algorithm (Blondel et al., 2008) for clustering (in reality no clustering approach is really parameter free). They identified two modules, but focused on one that was the most enriched for known ASD genes, ppiM13 (Basu et al., 2009) rather than an interesting one that contained transcriptional regulators such as *CHD8* and *FOXP2*, ppiM2. They identified ppiM13 as enriched for oligodendrocytes markers and gene expression in the corpus callosum, and implicated inter-hemispheric connectivity. They also found weak enrichment for mutations from ASD whole genomes. No robustness or reproducibility criteria were evaluated.
- NETBAG+ ASD (Chang et al., 2015): this approach constructs a background network of molecules participating in shared function by integrating edges from shared function in the KEGG and GO database, direct and indirect PPIs from many databases, and multiple other data sources, and is discussed in more detail in Chapter 1, and described in the original paper (Gilman et al., 2011). Modules are seeded with CNVs or SNVs detected in disease, and the algorithm uses a greedy search in the background network to identify modules of shared function. The authors applied NETBAG+ to rare *de novo* SNVs (referred to here as RDNVs, though technically they also use CNVs which are “rare de novo variants”) from three WES studies (Iossifov et al., 2012; O’Roak et al., 2012; Sanders et al., 2012), and rare *de novo* CNVs from one study (Levy et al., 2011). One module was weakly enriched for ASD mutations, and the remainder of the study analyzed

this module, sub-modules from this module, and various other gene sets to demonstrate biases toward time points and cell-types in gene expression, as well as a separation of low IQ and high IQ ASD cases using further criteria. However, no reproducibility criteria were evaluated, and the primary module identified by NETBAG+ was not used in the majority of the study to inform biologically conclusions.

To compare modules from different studies with each other, I first evaluated each module described above for reproducibility of a genetic signal with RDNV data, then compared the modules with each other to understand differences and similarities between modules, and finally evaluated modules to with laminar and cell type specific gene sets to understand their biological informativeness. Genes in modules were based on what was reported in the respective studies' supplemental materials.

3.3. Results

3.3.a. Genes affected by novel mutations are in previously identified co-expression modules

First, I re-evaluated enrichment for RDNV affected genes in modules using a logistic regression model. This model utilizes binary mutation status in each gene (mutation found in that category or not) as the outcome, and binary module membership (in module or not) and gene length (based on the exome capture size for each gene as reported in Iossifov et al., 2014) as predictors. This analysis is equivalent to controlling for gene length by a stratified permutation analysis, and the effect size associated with the module membership can be transformed into an odds-ratio (see A2 Additional Methods for Chapter 3). Controlling for gene length is important as the rate of any type of *de novo* SNV across genes is highly correlated to gene length (Michaelson et al., 2012; Samocha et al., 2014). Notably, this model yields similar enrichment results as reported from gene-length adjusted permutation analyses when comparing between rare variant implicated gene sets (Iossifov et al., 2014).

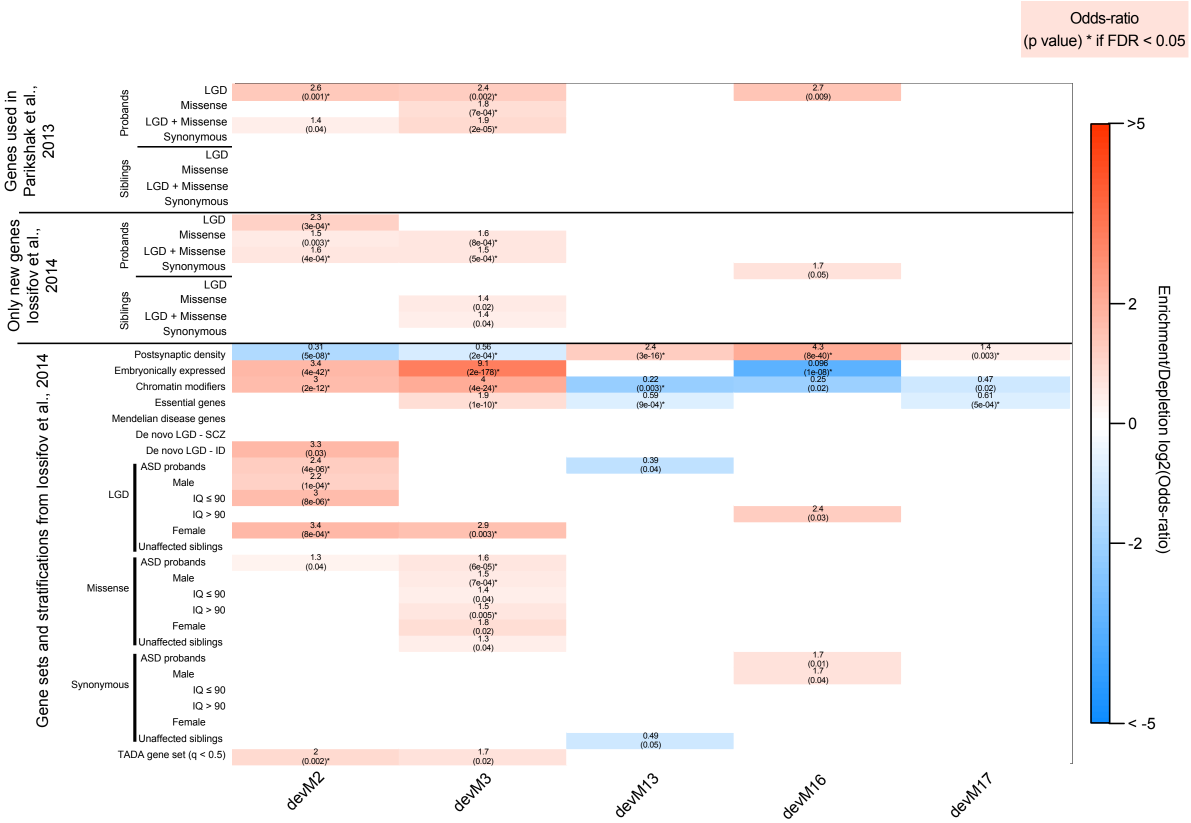


Figure 3.1 Heatmap of gene set enrichment for ASD rare *de novo* mutations comparing enrichment in ASD-associated modules from Parikshak et al., 2013. Tested gene sets include RDNV affected genes used in Parikshak et al., 2013 from four WES studies (Iossifov et al., 2012; Neale et al., 2012; O'Roak et al., 2012; Sanders et al., 2012), new genes implicated by de novo mutation since 2013 (Iossifov et al., 2014), and stratification of mutations. The heatmap dislpays $\log_2$ fold enrichment for genes implicated by RDNV. All odds-ratios with $p < 0.05$ are reported, and those passing an FDR adjusted p value < 0.05 for the comparisons are delineated with an *. Three broad categories have been tested: "Genes used in Parikshak et al., 2013" reflects the same analysis shown in Chapter 2, "Only new genes from Iossifov et al., 2014" reflects genes that were newly implicated from sequencing ~1600 additional ASD containing simplex families, and "Gene sets and stratifications from Iossifov et al., 2014" reflects all genes currently identified from families in the Simons Simplex cohort by WES, stratifying by sex and splitting males by IQ as done in Iossifov et al., 2014. Finally, TADA q < 0.5 contains genes implicated at an FDR < 0.5 by the Transmission and *De novo* Association method (He et al., 2013) as applied to the Simons Simplex cohort and 1000 case and control samples (De Rubeis et al., 2014). The abbreviation LGD (likely gene disrupting) corresponds to the term "protein disrupting" used in Chapter 2.

Figure 3.1 shows that the enrichment signal for RDNV affected genes in ASD generalizes with new data in devM2/3. This suggests that these modules, and conclusions about their relationship to ASD, were not overfit to the initial data, and that devM2/3 are likely to be enriched for RDNVs in any new study. Notably, devM2 is still enriched for protein disrupting (also called likely gene disrupting, LGD) mutations (OR = 2.4, p = $4x10^{-6}$, FDR < 0.05), with weaker enrichment for missense mutations and the combined set. devM3 shows more consistent enrichment for missense mutation affected genes (OR = 1.6, p = $6x10^{-5}$, FDR < 0.05). There is some weaker enrichment for missense mutation affected genes found in siblings, too, suggesting that the enrichment for missense RDNVs is not as clearly associated with ASD as that for protein disrupting RDNVs. This is consistent with the observation in the total Simons cohort that protein disrupting RDNVs are found about twice as frequently in cases compared to controls while the rate of missense mutations is only marginally different between ASD and controls (OR ~ 2, p = $2x10^{-5}$, for protein disrupting RDNVs; OR ~ 1.1, p = 0.01, for missense RDNVs as computed by Iossifov et al., 2014). Interestingly, other studies have not seen an enrichment for missense RDNVs from ASD in gene networks or biological pathways. These analyses clearly identify a missense RDNV affected module, highlighting how partitioning the genome by co-expression modules can identify a stronger genetic signal than genome-wide enrichment analysis alone.

Additionally, it is notable that devM16 shows enrichment only for synonymous mutations after correcting for gene length, likely due to the fact that it contains very long synaptic genes where enrichment is driven by increases in all categories of RDNVs. In Chapter 2, I did not focus on devM16 as it was inconsistently enriched (notably this was without gene length correction). It is important to note that devM16 contains several individual genes implicated in

ID, ASD, SCZ, and epilepsy (e.g. *SCN2A*, *NRXN1*), but these analyses demonstrate that the module as a whole does not exhibit enrichment for RDNVs above chance.

I also assessed the gene sets and stratifications assessed in Iossifov et al., 2014. The authors defined the following functional gene sets: genes whose proteins compose the postsynaptic density proteome (Bayés et al., 2010), genes in the GO term chromatin modification, genes highly expressed and correlated to each other during fetal brain development (embryonically expressed, defined from BrainSpan (Sunkin et al., 2013)), genes resulting in lethal phenotypes in mouse (essential genes), and genes involved in Mendelian disease (Iossifov et al., 2014). Enrichment of these gene sets in modules re-affirms the GO term enrichments seen in Chapter 2 (Figure 2.2), as well as the fact that devM3, which contains more early expressed genes, also reflects earlier biology compared to devM2, as it is more enriched for essential genes and more strongly enriched for embryonically expressed genes.

Combining the mutation data across over 2700 ASD affected individuals, the overall trend for enrichment in protein disrupting RDNVs in devM2 is re-affirmed, as is the enrichment for missense RDNVs in devM3. Looking across all known mutations and stratifying by sex and IQ, it is clear that females and male with IQ < 90 are most strongly affected in devM2 and devM3. It has become apparent from several studies now (Iossifov et al., 2014; Robinson et al., 2014; Ronemus et al., 2014; Samocha et al., 2014) that protein disrupting RDNVs are associated not with ASD alone, but with ASD comorbid with lower IQ. In the Simons cohort, females have lower IQ than males (average IQ in the cohort of about 75 vs 85, respectively) (Ronemus et al., 2014) and are fewer in number due to potential ascertainment biases, so they are not stratified in this comparison.

The main observation from these stratifications is that genes found mutated by protein disrupting RDNVs in males with IQ > 90 are not enriched in devM2 and M3. In contrast, missense mutations in devM3 are enriched across stratifications, but as noted previously, they are also weakly enriched in genes with RDNVs found from unaffected siblings. Taken together, this suggests that weaker protein altering missense mutations in the very early expressed genes from devM3 are likely contributing to ASD in a different manner, perhaps by different mechanisms, than more severe loss-of-function mutations in devM2. Finally, high confidence mutations implicated by TADA, which reflects genes with genome-wide evidence across over 3,500 individuals with ASD including both *de novo* and inherited variants, are clearly enriched in both devM2 and devM3, demonstrating that these modules are robustly enriched for high-confidence genes implicated by both *de novo* and inherited rare variants in ASD.

3.3.b. Extended analysis of cell-type specificity in ASD-associated developmental co-expression modules

I next evaluated enrichment of cell-type specific gene sets using the same logistic regression analysis as used for RNDV-affected genes. I defined cell-type specific gene lists based on differential expression within each individual dataset, using genes differentially expressed at an FDR-adjusted p value < 0.05 for each comparison, and then evaluated how binary module membership predicts cell-type specificity. These results, shown in Figure 3.2, demonstrate that many co-expression modules are highly specific for cortical laminae or cell types – they exhibit strong enrichment for functionally similar categories of laminae or cell types, and strong depletion of genes from other laminae other cell types. However, most modules do not reflect just one cortical lamina or cell type.

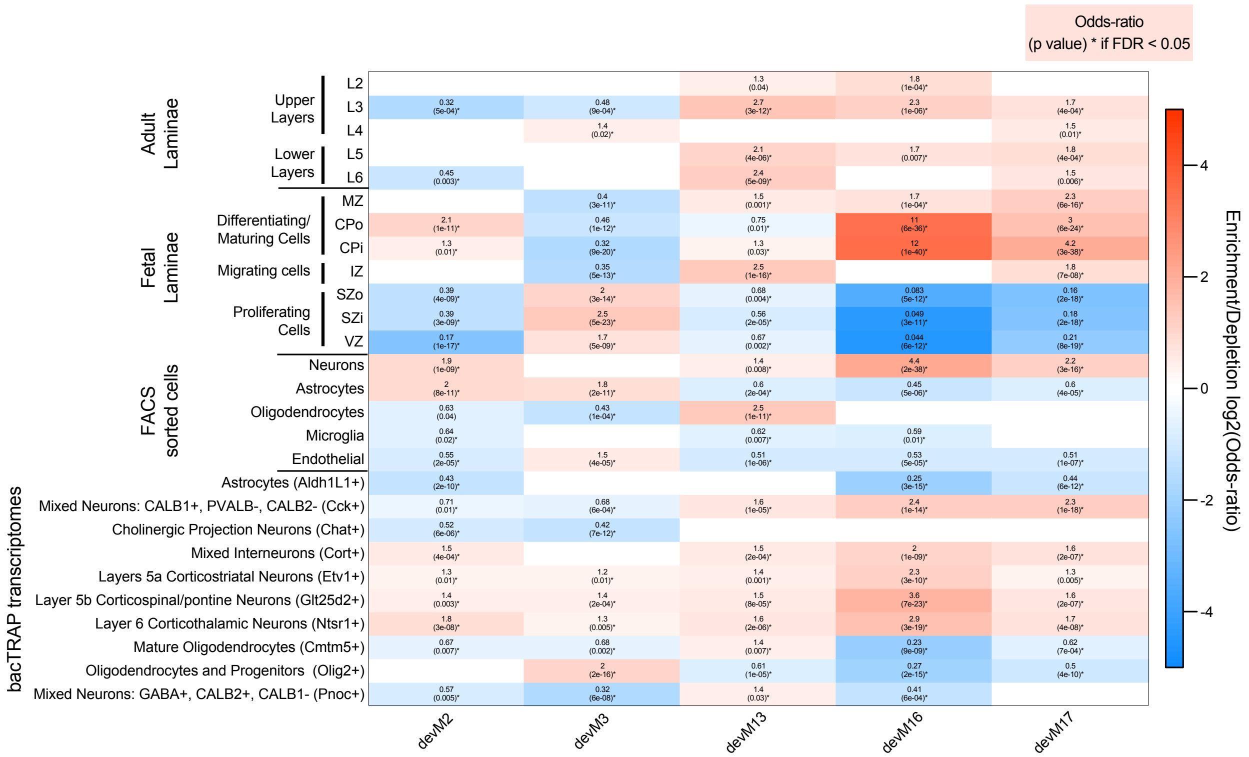


Figure 3.2 Heatmap of gene set enrichment for cell-type specific gene lists with ASD-associated modules from Parikshak et al., 2013. Heatmap displaying $\log_2$ fold enrichment for cell type lists. All odds-ratios with $p < 0.05$ are reported, and those passing an FDR adjusted p value < 0.05 for the comparisons are delineated with an *. Four categories have been tested:

- Adult laminae, from non-human primate brain where each layer was dissected via laser capture microdissection (LCM): RNA was extracted, and microarrays were run (Bernard et al., 2012)
- Fetal laminae, from human prenatal brains evaluated by a similar LCM paradigm followed by microarray as used for adult primates (Miller et al., 2014)
- Transcriptomes from major cell types in mouse cortex, sorted by FACS, followed by RNA extraction and RNA-seq (Zhang et al., 2014)
- Transcriptomes from RNA bound to ribosomes in specific cell types in mouse and then profiled by microarray (Doyle et al., 2008).

Two caveats must be considered when interpreting this heatmap: the bacTRAP lists reflect purer cell populations than the FACS lists, but thes are still mixed cellular populations (marker-level characterization of the cell types are listed where applicable) and some mouse cell types might not accurately reflect of human cellular populations, particularly when considering laminar specificity (Zeng et al., 2012). Laminar abbreviations are as described in Figure 2.6. Laminae: marginal zone (MZ), outer/inner cortical plate (CPo/CPi), subplate (SP), intermediate zone (IZ), outer/inner subventricular zone (SZo/SZi), ventricular zone (VZ), and adult cortical layers 2–6 (L2–6).

The laminar enrichments in Figure 3.2 used the same datasets from Chapter 2, Figure 2.5-6, but with a stricter statistical threshold and different enrichment approach that are easier to interpret from a biological perspective. The Z score for enrichment in Chapter 2 measures the skew in a distribution, and do not offer an easily interpretable effect size (such as a fold change or odds ratio). Such a method, which forces a ranking of all genes, can give equal weight to very low or high expressed genes in adult cortical layers that can result in over-emphasis of weak and biologically non-significant biases. The modified enrichment approach used here agrees with findings from Chapter 2 for devM13, 16, and 17, showing that they are enriched for genes specific to upper layers of adult cortex (L2-3). However, there is also enrichment for these modules in the lower cortical layers, consistent with the general observation that ASD genes affect both upper and lower cortical layers.

Additionally, in fetal cortex, devM2, 16, and 17 are predominantly enriched in both the inner and outer cortical plate (CPo and CPi), while the earlier devM3 shows enrichment in the germinal zone regions (SVo, SVi, VZ), again consistent with the previous method of enrichment (Figure 2.6). Interestingly, one difference in these results is that devM13 is enriched mostly for the intermediate zone (IZ), which contains cells migrating to the cortical plate.

Additionally, in Chapter 2, cell-type enrichments relied on correlations to cell-type marker profiles summarized by their first principal component in the data. Potential shortcomings of this approach include the fact that correlations to markers might not reflect true cell type specificity and that many cellular subpopulations are identifiable only with a combination of cell-type markers. I therefore re-evaluated cell type specificity with expression profiles from cell-type specific transcriptomes in mouse. Figure 3.2 shows that, across major cell types in the mouse cortex, devM2, 16, and 17 (the same modules enriched for CPo and CPi) show strong enrichment for neuron specific gene expression. Both devM2 and 3 also show enrichment for broad astrocyte markers, however, this is likely a reflection of early neural progenitors and radial glial cells, which can also show expression of these markers(Lui et al., 2014; Stein et al., 2014).

Enrichment analysis for the bacTRAP cell transcriptomes, which reflect more homogeneous (but certainly not pure) cellular populations than those profiled via FACS by Zhang et al (Zhang et al., 2014), are only marginally more informative for cell type specificity. First, it is clear from these data that devM2, 16, and 17 are generally enriched for both excitatory and inhibitory neurons. This is consistent with what was found in Figure 2.6, only now the interneuron enrichment is stronger. Overall, this analysis supports putative involvement of both upper and lower layer excitatory neurons, as well as Pvalb+, Calb1+, and Cck+ interneurons which are found across layers in both mouse and human (Zeng et al., 2012).

Another finding from this analysis is that modules are not enriched for genes found exclusively in cholinergic (Chat+) projection neurons, mature oligodendrocytes (Cmtm5+ or Olig2+), oligodendrocytes progenitors (Olig2+), and astrocytes (Aldh1L1+). Interestingly, devM3 is enriched for genes found in Olig2+ cells, which is also a marker for radial glia and further supports the idea that devM3 may reflect a mix of radial glial cells and neural progenitors. Finally, devM13, which was enriched for genes expressed in IZ, is mostly similar to devM16 and 17 in cell type enrichment. However, it shows a distinct patterns of enrichment for Gaba+, Calb2+, Calb1- cells, which may be related to the primate-specific migration of SZo/SZi/VZ derived interneurons (in mouse, interneurons migrate exclusively from the ganglionic eminence (GE), in humans, both from the GE and the GZ), which are known to be CALB2+ (Zeng et al., 2012).

A major shortcoming of these cell-type specific data is that they do not profile a population of purely upper layer neurons in mouse. Recently, a modified cell specific transcriptomic approach has allowed the profiling of specific excitatory pyramidal cell populations in mouse cortex (Molyneaux et al., 2015). The general approach entails the fixation of whole brain tissue, combinatorial labeling of cell types, FACS, and then RNA-seq. Current application of this strategy has profiled cellular transcriptomes from corticothalamic (mostly layers 5-6), subcerebral (mostly layers 5-6), and callosal projection neurons (mostly layers 2-3). I therefore sought to evaluate these data for enrichment analysis, but found the data unsuitable for the type of comparisons shown in Figure 3.2 due to the fact that four time points were considered, only two samples per cellular population were profiled for each time point, and no global background (e.g. whole tissue) was profiled. I therefore asked a simpler question: how do the trajectories of the top hub genes in ASD-associated developmental modules change in these cellular subpopulations over mouse development?

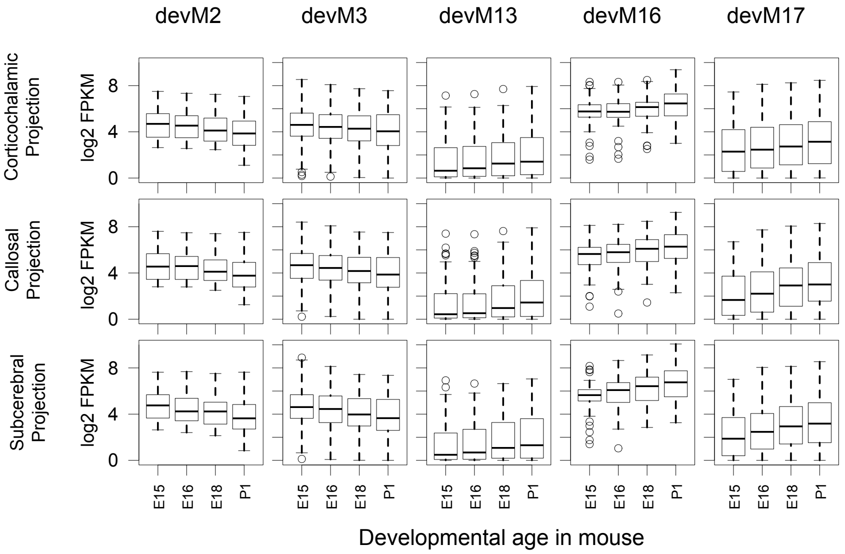


Figure 3.3 Boxplot of different excitatory neuron subtype developmental trajectories using ASD-associated modules from Parikshak et al., 2013. Boxplot displaying gene expression trajectories from embryonic day (E) 15,

E16, E18, and postnatal day (P) 1 in mouse for three categories of excitatory pyramidal cells profiled by Molyneaux et al., 2015 by a modified FACS approach. Genes with kME > 0.9 in the ASD-associated from Chapter 2 are plotted for each time point in each cell type.

Figure 3.3 illustrates temporal trajectories from embryonic (E) days 15, 16, and 18 as well as postnatal (P) day 1 across the three neuronal populations for each module. In general, the trajectories are remarkably similar, suggesting that the hub genes in ASD-associated modules are dynamically changing in each of these cell types during brain development, at least in mouse. Thus it is clear that, as far as excitatory projection neurons are concerned, ASD-associated modules do not identify genes with dramatically different transcriptomic signatures across these cell types.

Taken together, the enrichment analyses in Figure 2.6 and Figure 3.2 suggests that the earliest affected ASD risk enriched module, devM3, reflects genes that are mostly highly expressed in the germinal zone (VZ, SZi, SZo), which contains proliferating neuronal progenitors and radial glia. The next earliest ASD risk enriched module, devM2, reflects maturing neurons in the inner and outer cortical plate (CPi/CPo). Both of these earlier modules show enrichment in L4 when using distribution-based enrichment methods (Figure 2.6), but this is not as well supported by stricter enrichment criteria. Additionally, cell-type specific transcriptomes affirm that these modules contain genes that are important for the maturation of multiple excitatory and inhibitory neurons in the cortex, potentially highlighting some neuronal subpopulations and excluding others, to the extent that mouse cell types can reflect human cell types (Figure 3.3).

The later expressed modules, devM16 and dev17, clearly reflect more mature neurons and are not enriched for genes involved in other cellular processes. They also reflect a mix of interneurons, but predominantly reflect glutamatergic projection neurons as found in Figure 2.6, and affirmed in Figure 3.2 and Figure 3.3. Finally, devM13 remains ambiguous despite these efforts at more detailed characterization. It exhibits enrichment for multiple adult layers, the intermediate zone (IZ) in fetal brain, and potentially CALB2+ interneurons.

These analyses saturate the level of temporal, laminar, and cell type specificity likely to be found with the current developmental co-expression modules and currently available transcriptomic data. It is thought that integrating additional information might enable gene networks to reveal novel biological insights that may be missed by evaluating just one level of biology. Although the most unbiased and genome-wide data is currently available at the transcriptomic level in brain (Chapter 1), multiple gene network studies have been performed claiming to identify distinct, novel, and biologically more specific modules affected in ASD.

3.3.c. Comparison of multiple gene network approaches in the context of ASD

I next sought to evaluate whether different approaches to gene network construction followed by ASD module identification might define more specific or biologically more informative modules compared to the approach used in Chapter 2. I evaluated multiple studies that range in network methods, using different combinations of the genome-wide or the seeded approach; either one or more of co-expression, protein interaction, or other information to guide module construction; and use neuronal and/or non-neuronal data. I first assessed the relative reproducibility of modules implicated by different methods for predicting RDNV-affected modules, then assessed whether different seeded or more integrative methods discover highly distinct modules, and finally asked whether the different methods implicate specific laminae or cell types better than the approach used in Chapter 2.

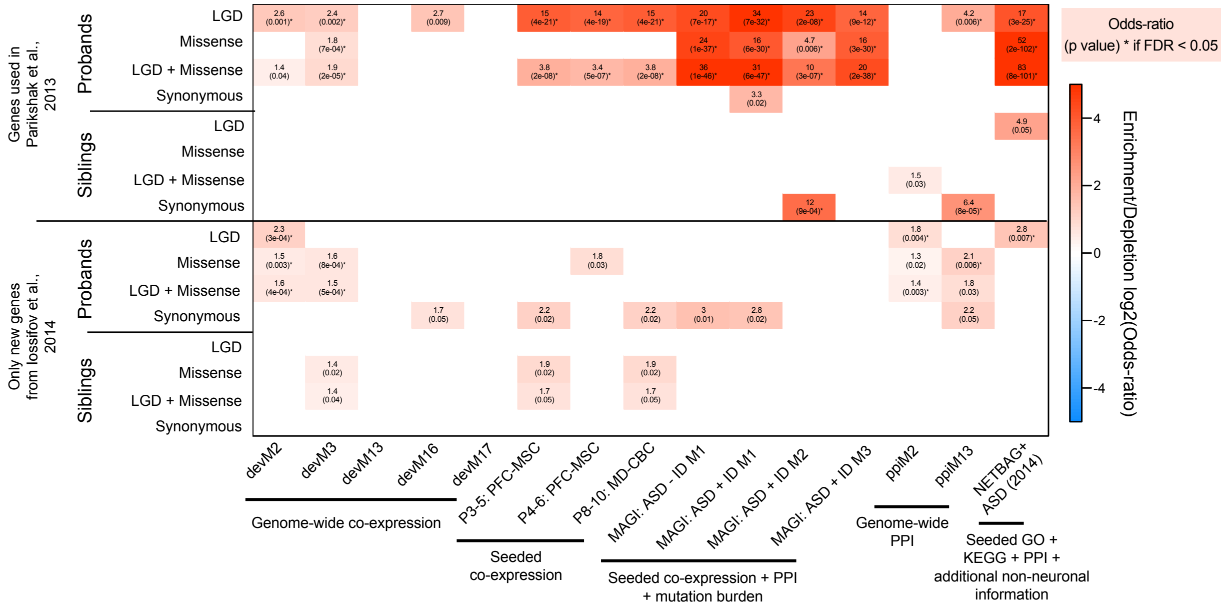


Figure 3.4 Heatmap of gene set enrichment for reproducibility of ASD risk mutations in ASD-associated gene network modules from multiple approaches. Gene sets used for testing reproducibility are the same as those in Figure 3.1. Tested gene sets include RDNV affected genes from Parikshak et al., 2013 and new genes implicated by de novo mutation since 2013 (Iossifov et al., 2012). The heatmap displays $\log_2$ fold enrichment for RDNV implicated genes in modules from genome-wide co-expression networks (Parikshak et al., 2013), seeded co-expression(Willsey et al., 2013), seeded co-expression and protein interaction networks using an objective function for module detection that minimizes the contribution of genes with pathogenic mutations in controls (Hormozdiari et al., 2015), genome-wide protein interaction networks (Li et al., 2014), and a highly integrative method that compiles known pathway annotations (GO, KEGG) and protein interactions (Chang et al., 2015). All odds-ratios with p < 0.05 are reported, and those passing an FDR adjusted p value < 0.05 for the comparisons are delineated with an *. Two RDNV affected gene categories have been tested: “Genes used in Parikshak et al., 2013” reflects the same gene sets used in Chapter 2, “Only new genes from Iossifov et al., 2014” reflects genes that were newly implicated from sequencing ~1600 additional ASD containing simplex families. Importantly, for each mutation category, any genes overlapping from previous findings with have been removed from the “Only new genes” sets, allowing *bona fide* assessment of reproducibility. A well-replicated and generalizable module should exhibit similar enrichment for protein disrupting (LGD), missense, or combined sets in probands, lack of enrichment in synonymous mutations, and weak or no enrichment for mutations in siblings for both the initial and the replication set.

Figure 3.4 shows module enrichments across different methods, including the module enrichments for the genome-wide co-expression approach utilized in Chapter 2 (Parikshak et al., 2013). The methods showing some level of reproducibility are the genome-wide co-expression method (as seen in Figure 3.2), the unseeded PPI method (though this also shows enrichment for synonymous mutations), and the integrative pathway and PPI approach, NETBAG+ (Chang et al., 2015). Notably, all methods experience a dramatic drop in the enrichment odds for protein disrupting (LGD) RDNVs found in ASD probands, except the genome-wide co-expression approach. This suggests that every method, particularly the seeded methods that used initial RDNV affected gene sets for module definition, overfit to initial findings to the extent that their modules fail to generalize to genes affected by new mutations. Additionally, it is notable that the robustness of devM2 and devM3, which show excellent reproducibility, was assessed carefully by module preservation, bootstrapping, and replication in the original work, demonstrating the value of using good statistical and data analysis practices when defining gene networks.

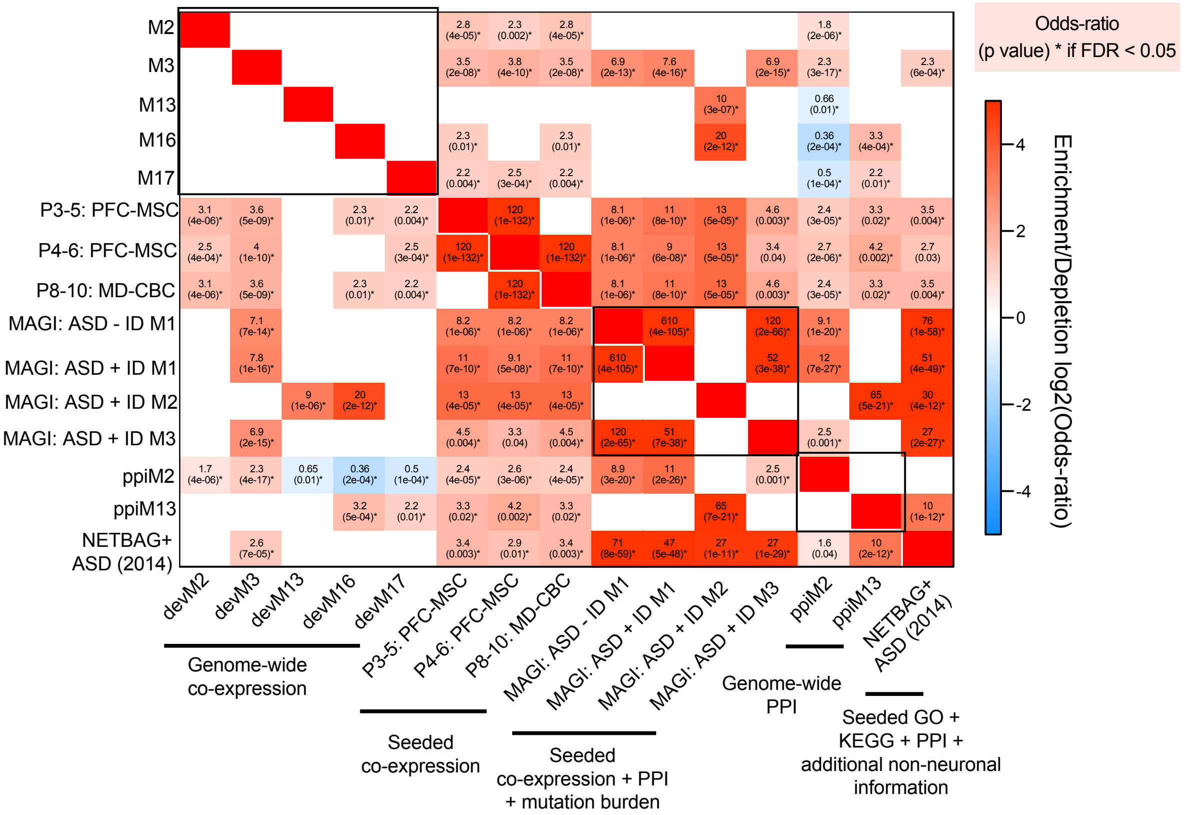


Figure 3.5 Heatmap of gene set enrichment between ASD-associated modules from multiple gene network methods. Modules used in Figure 3.4 are evaluated for gene set enrichment with each other here to highlight similarities or differences between methods. All odds-ratios with p < 0.05 are reported, and those passing an FDR adjusted p value < 0.05 for the comparisons are delineated with an *. Module-module overlaps between modules from the same method are encompassed in boxes, and the diagonal is reflects infinite overlap since modules are compared against themselves. The values on the off-diagonals are not symmetric due to slight differences in logistic regression that arise from switching variables from outcomes to predictors and vice versa.

Although some of the network methods do not generalize well, it is possible they hold value in discovering novel biological processes related to the specific study in which they are applied. To evaluate whether these different network approaches identify distinct modules specific to a particular study, I evaluated overlaps among their resultant modules. Importantly, although this is not an ideal systematic comparison of gene network methods, the modules evaluated here have all been implicated in ASD and are constructed by methods that cover a diverse range of criteria for network nodes (seeded vs genome-wide), different edge information (co-expression, protein interaction, annotated pathways, or combinations of these), and different module definition approaches (hierarchical clustering, greedy search, pre-set module sizes, or combinations of these). Figure 3.5 shows module-module overlaps. Several themes emerge:

- modules defined by genome-wide, unsupervised approaches are more distinct from each other, as these methods define modules by partitioning all genes from the genome into non-overlapping sets
- modules identified by seeded approaches tend to be highly overlapping, likely due to the use of overlapping seed gene sets, suggesting they identify more redundant pathways.
- Modules identified by seeded approaches are generally smaller than those identified by genome-wide methods, and it is claimed that they identify unique or more specific biological processes due to the prior information provided by the seeds (Gilman et al., 2011; Hormozdiari et al., 2015; Willsey et al., 2013). However, every module identified by a seeded approach also overlaps a module from the unseeded approach, suggesting they are not distinct or specific to the extent that unseeded methods fail to capture the same genes without seeding on known biology or known genes.
- The NETBAG+ approach has extremely high overlap with modules from the MAGI approach (OR = 27 – 71). This is likely because they are seeded on similar genes and use similar protein interaction databases. This demonstrates that, despite employing rather different methods and databases, the utilization of seeded gene networks with known protein interactions as edges results in very similar modules. Part of this is that the seeds are similar in the two approaches, but the other factor is the likely bias across databases

for PPI edges between frequently studied or more easily studied proteins in PPI databases (Hakes et al., 2008).

- The unseeded PPI approach (ppiM2, ppiM13) shows very high overlap with MAGI ("MAGI: ASD + ID M1" – ppiM2, OR = 12, other overlaps very high) and NETBAG+ ("NETBAG + ASD" - ppiM13, OR = 10), further suggesting that using protein interactions results in highly similar modules.
- The developmental modules from genome-wide co-expression from Chapter 2 overlap most highly with MAGI modules, then with supervised co-expression modules (Willsey et al., 2013), and finally show weaker overlap with modules from the methods that do not use data with neurobiological context (genome-wide PPI and NETBAG+).

Taken together, these results demonstrate that all methods show some overlap, though this ought to be interpreted cautiously for the seeded methods as the overlap is inflated by the use of common seed genes. Notably, methods using PPIs tend to overlap very highly with each other, and this is likely due to the fact that the same studies go into the literature curated PPIs compiled by BioGRID (Stark, 2006), STRING (Franceschini et al., 2012), and HPRD (Prasad et al., 2009) due to biases in curating the literature (Hakes et al., 2008; Hart et al., 2006). Finally, taking the results from Figure 3.4 into consideration, the genome-wide co-expression method seems ideal since it is generalizable to new mutations and overlaps considerably with all modules defined by other methods. It is, however desirable to get to more specific modules, and I discuss this issue later (see section 3.4 Conclusions).

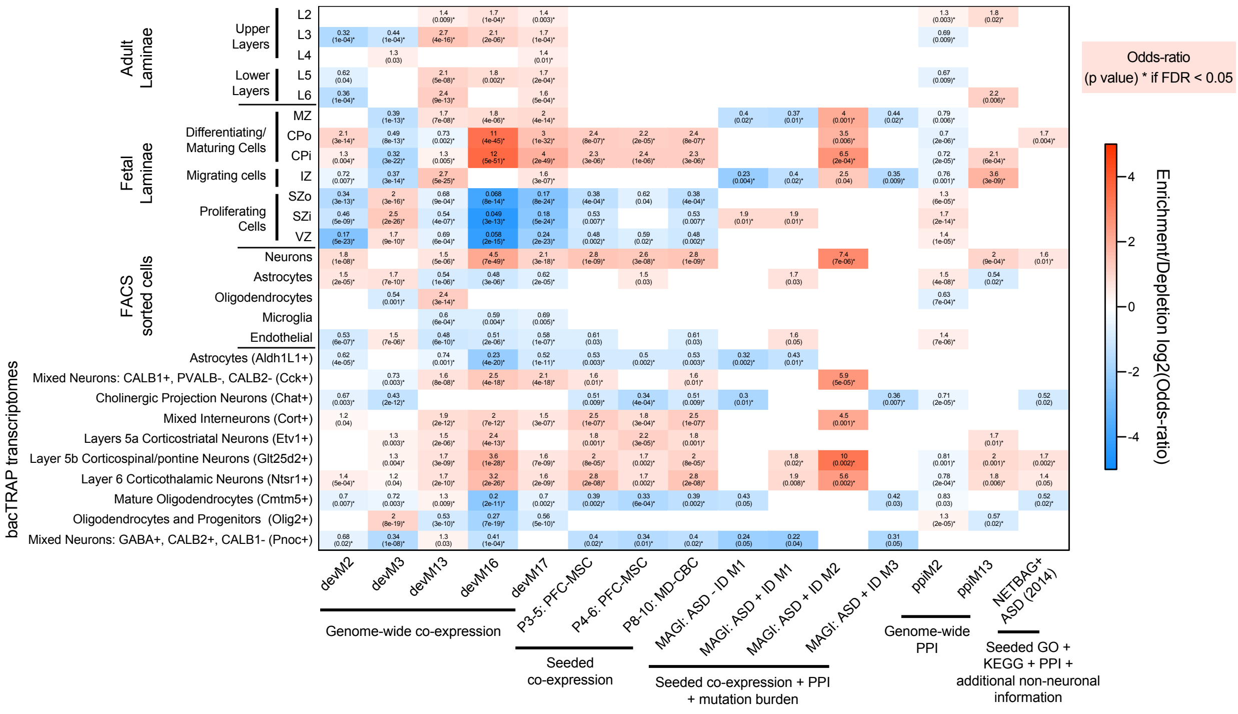


Figure 3.6 Heatmap of gene set enrichment between ASD-associated modules from multiple methods and laminar and cell-type specific gene sets. Modules used in Figure 3.4 are evaluated against laminar and cell-type specific gene sets described in Figure 3.2. All odds-ratios with p < 0.05 are reported, and those passing an FDR adjusted p value < 0.05 for the comparisons are delineated with an *. Laminae: marginal zone (MZ), outer/inner cortical plate (CPo/CPi), subplate (SP), intermediate zone (IZ), outer/inner subventricular zone (SZo/SZi), ventricular zone (VZ), and adult cortical layers 2–6 (L2–L6).

As mentioned previously, methods found by the seeded approaches find small modules (10- 200 genes) while those from genome-wide methods find modules of variable sizes (10-3000 genes), with the most interesting modules being the somewhat larger modules (modules in Parikshak et al., 2013 are generally > 400 protein coding genes in size, while those found by Li et al., 2014 were over 1400 genes (ppiM2) and 120 genes (ppiM13), though they focused on the latter). It is often claimed that these smaller modules define more specific biological processes, and are more biologically relevant in some manner or another, but this claim has not been tested in a systematic manner with neurobiologically relevant gene sets. It is already clear that many of these smaller modules lack predictive value (Figure 3.4) but it is possible they identify enrichment in specific laminae or cell types. To assess whether this is the case, I evaluated the laminar and cellular enrichments shown in Figure 3.2 for modules from different methods. Figure 3.6 demonstrates that the genome-wide co-expression method shows the most enrichment and depletion across the board. Although this is partly due to the larger module sizes, it demonstrates that each module reflects different levels of laminar and cell-type enrichment despite using non-overlapping modules (Figure 3.5).

The other network methods that also use co-expression (seeded co-expression from Willsey et al., 2013, and co-expression + PPI in MAGI) also demonstrate strong fetal laminar and cell specific gene enrichment, but none identifies a pattern of enrichment that isn't already seen by a module from genome-wide co-expression. This suggests that seeding co-expression modules does not yield greater neurobiological specificity for identifying ASD-associated biological processes.

Methods that do not include any neurobiological information (the PPI method from Li et al., 2014, and NETBAG+) identify mostly weaker enrichment patterns, and these patterns are more difficult to interpret from a neurobiological perspective. For example ppiM2 is enriched for L2, SZ, VZ, astrocytes, endothelia, and oligodendrocytes supporting weak enrichment for very different cellular populations that are not associated with each other in a clearly connected neuroanatomical or developmental manner. ppiM13 shows a more clear enrichment trend in L2, L6, CPi, IZ, and mixed neuron types. This is more consistent than ppiM2, however, it is difficult to know what this might mean without further dividing this module using neurobiological information, for example gene expression. Interestingly Li et al., 2014 functionally implicated this module based on where it was most highly expressed in brain, and arrived at the conclusion that it is related to cells in the corpus callosum, a neuroanatomical region that is not assessed here.

The same issue of ambiguous interpretability applies to the NETBAG+ module, where the authors identify many different functionally disparate cellular populations as being weakly associated with the modules' function (Chang et al., 2015). The NETBAG+ study relies heavily on neurobiological data to sub-stratify and characterize modules, suggesting that methods using non-neuronal information alone do not identify neurobiologically interpretable modules. Clearly, neurobiological context is necessary to interpret these potentially counterintuitive enrichments, and it therefore future work should include neurobiological data in module construction, rather than only in module evaluation.

3.4. Discussion

In this chapter, further analyses of the modules from Chapter 2 (Parikshak et al., 2013) and alternative gene network approaches (Chang et al., 2015; Hormozdiari et al., 2015; Li et al., 2014; Willsey et al., 2013) reveal important technical and biological insights that suggest a better way forward with gene network approaches to understand the disruption of normal neurodevelopmental pathways.

First, it is clear that the modules identified in chapter 2 with a genome-wide co-expression gene network approach have predictive value for identifying where future mutations in ASD will be found. This evaluation focused on whether there was predictive value for genes affected by RDNVs, and ignored common variation, which will likely play a greater role in explaining ASD risk (Gaugler et al., 2014; Klei et al., 2012; Stein et al., 2013). This was largely due to the lack of an ideal genome-wide association study (GWAS) dataset for testing and replication, as even the total set of GWAS data available in ASD is currently underpowered (Anney et al., 2012; Cross-Disorder Group of the Psychiatric Genomics Consortium et al., 2013; Wang et al., 2009a). As results from larger GWAS of ASD become available, it will be important to assess whether M13, 16, or 17 are predictive for the biological processes affected by common variation, as suggested by initial analyses in Figure A1.3.

Next, these analyses suggest that the current modules, which are relatively large gene sets, cannot offer many new neurobiological insights based on enrichments with existing laminar and cell-type specific gene sets. One possible reason for this is that the set of biological pathways affected by RDNV in ASD is actually very large, and these networks capture potentially affected genes and genes that may never have observed mutations, as discussed below. However, this is unlikely, and further specificity likely exists and it will be valuable to identify more specific modules.

Toward this end, I evaluated modules constructed by different gene network methods, covering a wide range of approaches described in Chapter 1. Enrichment analysis revealed that divergent approaches yield similar results, but signed genome-wide co-expression is the least biased based on the fact that it replicates mutation enrichment at approximately the same enrichment strength using new data. Additionally, although modules from genome-wide co-expression are quite large, they yield as much neurobiological information as methods identifying smaller modules based on similar (if not superior) enrichment results for laminae and cell types. This approach could be improved considerably by including increasing the temporal resolution (more time points), spatial resolution (cell-type specific data), or transcriptome resolution (isoform-level measurements).

Taking the results from this cross-method analysis as a whole highlights a major biological insight. It is clear that RDNV affected genes in ASD do not coalesce into ASD-specific modules. If they did, seeded approaches would potentially have high predictive value. Instead, RDNVs disrupt endogenous biological processes that occur during brain development, reflecting a disruption of canalization (Suliman et al., 2014; Waddington, 1942). To identify more specific modules, it will be important to evaluate genome-wide networks derived from neurobiological data. Not every gene in these early biological processes, which include transcriptional regulation and chromatin modification, will be observed as mutated with an observable high-impact effect. Many mutations might by lethal and cause spontaneous abortion while others may not yield a phenotype because the gene is compensated for in some manner, either by another gene serving a similar biological function, or by compensation for haploinsufficiency by the opposite allele.

The seeded approach, followed by extending the modules to include additional nodes that may not be affected by mutation but increase intramodular connectivity, may be effective once a large fraction of ASD mutations are identified with truly high confidence. However, taking all WES data with both *de novo* and inherited rare variant contributions into consideration, just over 100 genes are implicated in ASD at an FDR < 0.5, suggesting only 50 out of nearly 1000 genes are currently identified (De Rubeis et al., 2014; He et al., 2013). This lack of appropriate prior evidence is likely another reason the seeded approaches fail to generalize.

Finally, based on the distinct biological processes implicated by devM2 and devM3, I predict that disruption of genes that peak at different points of development yields different outcomes. To evaluate this properly, a very high temporal resolution transcriptomic atlas will be necessary – likely 20-30 individuals per every few days of development, allowing whole-genome co-expression networks to be constructed for each time point. Current datasets lack this temporal resolution as only 1-3 individuals are available for each time point, and most time points are weeks apart from each other. The best temporal sampling in currently available data is at mid-fetal time points (Willsey et al., 2013), so it is possible that the discovery of mid-fetal development as a first time point of convergence is a product of the available data, and interpreting these findings as evidence of "specificity" is likely incorrect. Additionally, the appropriate data may never be available to evaluate the effect of mutations during the third trimester of fetal development, so it may be necessary to evaluate these in nonhuman primates (Bernard et al., 2012; Sunkin et al., 2013) or *in vitro*, once viable comparisons between time points can be established (Stein et al., 2014).

Finally, to gain additional biological insights about the affected cell-types, it will be necessary to have temporal trajectories in individual cell types, or a lineage tree of cortical cell types. Molyneaux et al., 2015 demonstrate how this could be done using mouse, and claim a similar approach could work in human (Molyneaux et al., 2015). Given that currently available PPIs do not contain any cell-type specific information, a promising avenue is to use cell-type specific transcriptomes and epigenomes to elucidate the important regulatory networks during cell-fate determination in the cortex, as has been done for the development of blood cells (Lara-Astiaso et al., 2014). Such a method, which would likely track cells from neural progenitor status to differentiated neuronal subtypes and profile transcriptomes, histone marks, and open chromatin for homogeneous populations defined by combinations of cellular markers, could identify the important regulatory changes at each lineage branch point and identify which steps of cortical development might specifically be affected in ASD by different mutational processes.

3.5 Materials and methods

Please see A2 Additional Methods and Figures for Chapter 3 for all methodological information.

CHAPTER 4:
Dysregulation of the transcriptome in autism spectrum disorder

"All happy families are alike; each unhappy family is unhappy in its own way."

— Leo Tolstoy, *Anna Karenina*

"A set is a Many that allows itself to be thought of as One."

— Georg Cantor as quoted by Rudy Rucker, *Infinity and the Mind*

4.1. Abstract

Autism spectrum disorder is a genetically complex neuropsychiatric disorder with immense locus heterogeneity. Despite this, a shared gene expression signature has been previously identified in the frontal and temporal cortex of postmortem brains from autistic individuals compared to neurotypical individuals (Voineagu et al., 2011). Here, I replicate this gene expression signature using RNA sequencing and demonstrate that it generalizes to independent brain samples. Using differential gene expression, differential splicing, and co-expression network analyses, I also find a shared transcriptomic signature in the noncoding transcriptome and in transcript splicing. Furthermore, I show that transcriptomic changes in duplication 15q syndrome, a genetically defined cause of ASD, strongly recapitulate those observed in idiopathic ASD. Finally, I utilize co-expression network analysis to explore the role of transcriptional regulators and chromatin modifiers in ASD.

4.2. Introduction

Autism spectrum disorder (ASD) is a heterogeneous collection of neurodevelopmental disorders that share deficits in social communication and mental flexibility (Geschwind, 2011). A key objective in understanding ASD, and indeed the brain at large, is to link genetic and environmental etiologies to phenotypes. Genome-scale association studies have provided insight into the genetic architecture of ASD by associating diagnoses with whole-genome genotyping (Anney et al., 2012; Gaugler et al., 2014; Wang et al., 2009a; Weiss et al., 2009), assessment of copy number variation (CNV) (Levy et al., 2011; Pinto et al., 2014; 2010; Sanders et al., 2011), whole exome sequencing (WES) (De Rubeis et al., 2014; Iossifov et al., 2012; 2014; Lim et al., 2013; Neale et al., 2012; O'Roak et al., 2012; Sanders et al., 2012; Yu et al., 2013), and, more recently, whole genome sequencing (WGS) (Jiang et al., 2013; Yuen et al., 2015). These studies have demonstrated that no single category of genetic variation – be it single nucleotide or copy number, common or rare, inherited or *de novo* – can fully explain the genetic underpinnings of autism, and that hundreds of genetic loci will be involved in ASD risk (Gaugler et al., 2014; He et al., 2013; Stein et al., 2013).

Despite the genetic heterogeneity, significant progress has been made in understanding the functional genomic architecture of ASD. Our lab previously identified a shared molecular signature in postmortem frontal cortex (FC, Brodmann area [BA] 9) and temporal cortex (TC, BA41/42/22) from individuals with ASD (19 ASD and 17 controls [CTL]) by differential gene expression (DGE) analysis and co-expression network analysis with gene expression microarrays (Voineagu et al., 2011). This study identified a downregulated module of synaptic genes (adsM12) enriched for neuronal markers, and an upregulated group of genes involved in inflammation enriched for astrocyte and microglial markers (asdM16) (Voineagu et al., 2011).

Moreover, asdM12 was enriched in genome-wide association (GWA) signals from an ASD GWA study, suggesting that common variants associated with ASD may causally contribute to the changes seen in asdM12. Additionally, analyses of gene expression during cortical development *in* vivo and *in vitro* implicated asdM12 and inherited variants in late prenatal and postnatal synaptic development, while protein disrupting rare *de novo* variants (RDNVs) were enriched in co-expression modules related to early transcriptional and chromatin regulation (Parikshak et al., 2013; Stein et al., 2014). Together, these studies show that there exists a shared gene expression signature in ASD, and suggest that at least some the changes observed in adult brain are the product of genetic variation disrupting cortical development.

However, all studies in ASD brain to date are small, and replication of the shared signature in ASD brain is warranted. Furthermore, gene expression in postmortem ASD brain has relied on microarrays or poly(A) tail selection (polyA+) followed by RNA sequencing (RNA-seq) on degraded RNA, restricting transcriptomic analysis to mostly protein coding genes, and biasing gene expression to the 3' end of transcripts. This has resulted in limited coverage of the transcriptome and prohibits accurate detection of long noncoding RNAs (lncRNAs) and transcript splicing.

Additionally, one promising avenue for understanding biological mechanisms and evaluating therapies is to focus on genetically defined subtypes of autism, which may show distinct and more homogeneous phenotypes as has been demonstrated by clinical phenotyping of patients harboring *de novo* mutations in the SWI/SNF (BAF) complex (Helsmoortel et al., 2014) and *CHD8* (Bernier et al., 2014). However, it is unclear whether genetic subtypes of autism will have very distinct gene expression signatures, or whether they will share a transcriptomic signature with idiopathic ASD.

Here I evaluate transcriptomes from a large sample set of ASD and control brain samples from FC, TC, and cerebellum in 81 individuals (46 ASD, 35 CTL, 205 samples) and replicate and refine the shared gene expression signature in ASD. I use ribosomal RNA depleted library preparation followed by RNA-seq (Figure A3.1A), which reduces sequencing coverage bias across transcripts, even in postmortem brain samples (Figure A3.1B-E). This allows more accurate evaluation of long noncoding RNA (lncRNA) expression and transcript splicing. Additionally, by analyzing samples from 8 individuals with duplication 15q syndrome (dup15q) who were diagnosed with ASD, I show that a genetic subtype of ASD largely shares the transcriptomic abnormalities seen idiopathic ASD. Finally, I construct a co-expression network that reveals how diverse ASD-associated perturbations in the transcriptome can result in a shared downstream signature in ASD cortex.

4.3. Results

4.3.a. Replication of differential gene expression in autism cortex

Given the extreme heterogeneity of ASD, I first aimed to assess whether DGE between ASD and CTL cortex (FC and TC, referred to broadly as CTX) could be reproduced with samples from new ASD and CTL individuals, independent of those previously evaluated by microarray (Voineagu et al., 2011) (see Appendix A3.1 for more details). I analyzed 58 previously published microarray gene expression profiles from FC and TC for DGE between 16 ASD and 17 CTL individuals using a mixed linear regression framework (see Extended Experimental procedures), and compared these results with a similar analysis in 56 independent gene expression profiles from RNA-seq in 15 ASD and 17 CTL new, covariate-matched individuals.

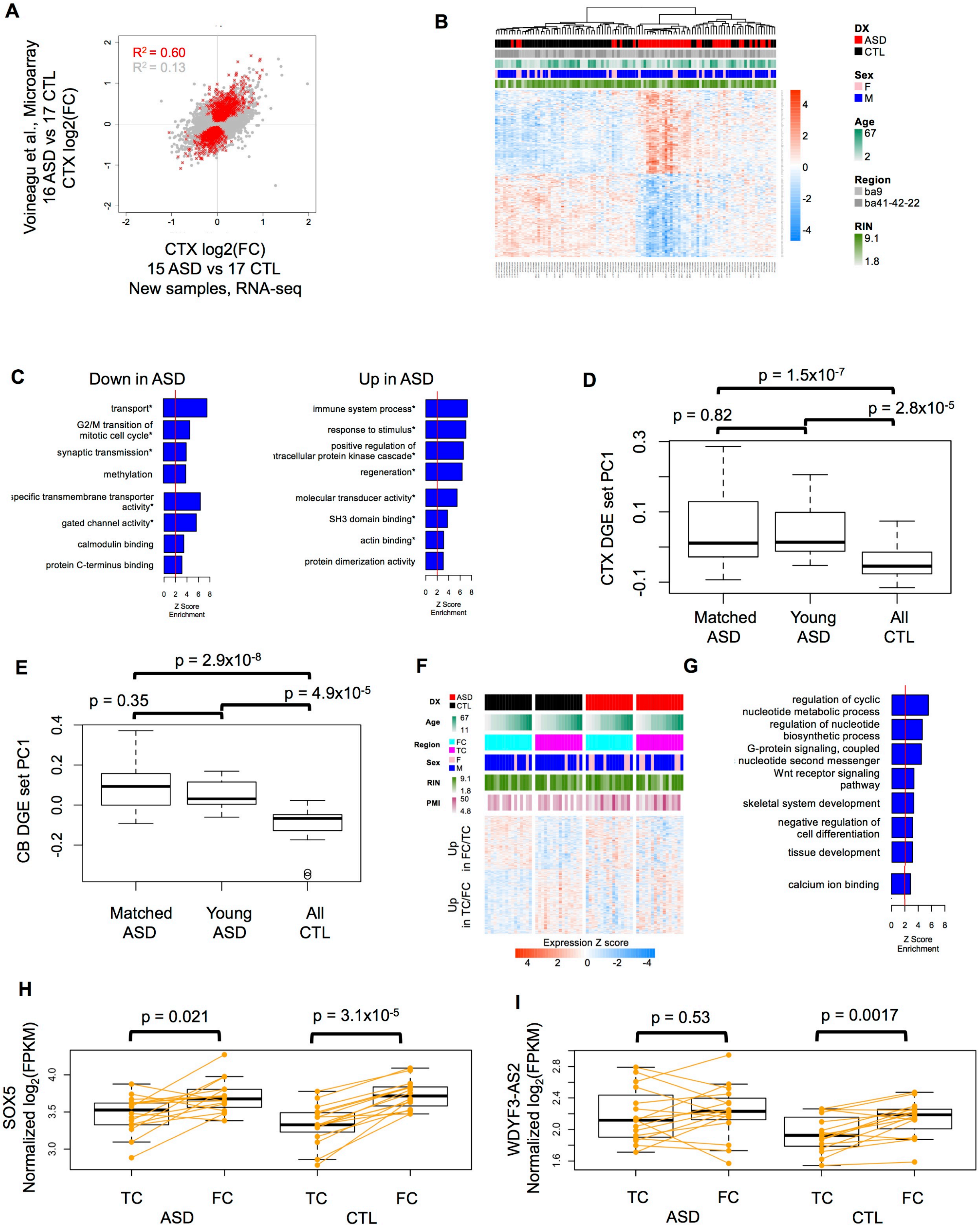

A
Voineagu et al., Microarray
16 ASD vs 17 CTL
CTX log2(FC)
$R^2 = 0.60$
$R^2 = 0.13$
CTX log2(FC)
15 ASD vs 17 CTL
New samples, RNA-seq
B
DX
ASD
CTL
Sex
F
M
Age
67
2
Region
ba9
ba41-42-22
RIN
9.1
1.8
C
Down in ASD
transport*
G2/M transition of mitotic cell cycle*
synaptic transmission*
methylation
specific transmembrane transporter activity*
gated channel activity*
calmodulin binding
protein C-terminus binding
Z Score Enrichment
Up in ASD
immune system process*
response to stimulus*
positive regulation of intracellular protein kinase cascade*
regeneration*
molecular transducer activity*
SH3 domain binding*
actin binding*
protein dimerization activity
Z Score Enrichment
D
p = 1.5x10-7
p = 0.82
p = 2.8x10-5
CTX DGE set PC1
Matched ASD
Young ASD
All CTL
E
p = 2.9x10-8
p = 0.35
p = 4.9x10-5
CB DGE set PC1
Matched ASD
Young ASD
All CTL
F
DX ASD CTL
Age 67 11
Region FC TC
Sex F M
RIN 9.1 1.8
PMI 50 4.8
Up in FC/TC
Up in TC/FC
Expression Z score
G
regulation of cyclic nucleotide metabolic process
regulation of nucleotide biosynthetic process
G-protein signaling, coupled nucleotide second messenger
Wnt receptor signaling pathway
skeletal system development
negative regulation of cell differentiation
tissue development
calcium ion binding
Z Score Enrichment
H
p = 0.021
p = 3.1x10-5
SOX5
Normalized log2(FPKM)
TC FC TC FC
ASD CTL
I
p = 0.53
p = 0.0017
WDYF3-AS2
Normalized log2(FPKM)
TC FC TC FC
ASD CTL

Figure 4.1 Differential gene expression in ASD and attenuation of cortical patterning between cortical regions in ASD. A) Replication of effect sizes from ASD vs CTL differences in CTX using new brain samples in this study, highlighting genes found to be at $p < 0.05$ with the microarray profiles. The R^2 values reflect the concordance of effect sizes for the highlighted set (red) and the remaining genes in the background (grey). B) Average linkage hierarchical clustering of samples using correlation-based distance from the top 100 up- and down- regulated genes in ASD CTX from the DGE set (FDR < 0.05). C) Top GO term enrichment for biological processes and molecular function (* if FDR adjusted $p < 0.05$). D) Differences between the "Matched ASD" set and "Young ASD" set compared to controls (N = 123 samples total) using PC1 of the DGE set, pairwise Wilcoxon rank-sum test p values are given above the boxplots. Young age (< 10 years) samples were held out from the DGE analysis to match covariates. E) Similar to D), but evaluating weaker DGE changes in CB at $p < 0.01$ and comparing across similar sample categorizations RNA-seq. F) Heatmap of genes with attenuated cortical patterning (FDR < 0.05 in CTL, FDR > 0.05 in ASD). G) GO term enrichment for the set of genes with attenuated cortical patterning in ASD. H-I) The transcriptional regulator SOX5 exhibits attenuated cortical patterning, as does a noncoding transcript antisense to the gene *WDFY3*. Pairs of expression values from the brains of the same individuals are connected to illustrate expression patterns between FC and TC. See also Figure A3.1 and Figure A3.2.

I find strong agreement in the DGE signal in cortex (Figure 1A) suggesting a highly reproducible signature of DGE exists in ASD CTX. Genes with $p < 0.05$ for ASD vs CTL in the microarray analysis had highly concordant ASD vs CTL effect sizes compared with the independent RNA-seq set (Pearson's $R^2 = 0.60$), and agreement was much greater than that between the changes above this significance threshold ($p >= 0.05$, $R^2 = 0.13$). Overall, out of 7339 genes overlapping in the two analyses, 522 genes overlap in DGE between the two sets (OR = 2.4, $p = 4x10^{-39}$), with all but 6 genes changing in the same direction. As expected, the odds-ratio and statistical significance of this overlap increases with more stringent thresholds, and similar concordance in DGE signature is seen when comparing RNA-seq in samples from the microarray study to these independent samples ($p < 0.05$: $R^2 = 0.58$, $p >= 0.05$: $R^2 = 0.22$, 411/16399 overlap: OR = 1.8, $p = 2.6x10^{-20}$, Figure A3.1F).

I performed a similar analysis of reproducibility of the DGE signature in cerebellum (CB). In the microarray study (Voineagu et al., 2011), no significant DGE signature for the ASD vs CTL comparison in CB was found. Comparing the previously evaluated and new samples as above, I analyzed 10 ASD and 11 CTL CB samples from the microarray study and 15 ASD and 16 CTL new CB samples in this study. I find no clearly reproducible signature ($p < 0.05$: $R^2 = 0.033$; $p >= 0.05$: $R^2 = 0.001$) and the overlap is not statistically significant (14/7320 overlap: OR = 0.58, p = 0.04). However, utilizing RNA-seq, I found greater concordance in the DGE signal ($p < 0.05$: $R^2 = 0.29$, $p >= 0.05$: $R^2 = 0.13$, 82/15970 overlap: OR = 1.5, p = 0.002, Figure A3.1G). This supports the observation that, although there may be a DGE signature in CB, it is considerably weaker than what is seen in CTX. Taken together, these results demonstrate that there exists a highly reproducible DGE signal in ASD vs CTL cortex, and a weaker signal in CB.

4.3.b. Dysregulated synaptic function and inflammation in the ASD cortex

Next, I combined the samples above to analyze the full covariate matched set (“Matched ASD”) in CTX (26 ASD and 33 CTL individuals, N = 106 samples, Figure A3.1H), leaving younger ASD samples (age < 10, “Young ASD”) and all dup15q samples for other analyses. After filtering out genes where expression might have been due to pre-mRNA signal from other genes (e.g. for genes contained largely in introns of other genes, Extended Experimental Procedures), 16403 genes remained for the cortical analysis (13688 protein coding, 2715 lncRNAs). Using the linear mixed regression framework on log2(FPKM) gene expression quantifications, I identified 1156 genes as differentially expressed in ASD brain, with 582 increased and 574 decreased in ASD compared to CTL at an FDR adjusted $p < 0.05$.

Importantly, DGE analyses changing assumptions about modeling the effect of sequencing covariates and the statistical distribution of RNA-seq data (Figure A3.2A-C, Table A3.2A). A heatmap and clustering dendrogram of the top 100 increased and top 100 decreased genes reveals that the majority of ASD samples cluster together, and that factors such as age, sex, and RNA quality are not responsible for this signature (Figure 4.1B).

Of the genes downregulated at FDR < 0.05, the most downregulated gene was *PVALB* (0.53 fold decrease, FDR < 0.05). *PVALB* is a marker for an interneuron subpopulation and codes for a protein that binds to calcium. Interestingly, *SST*, whose protein product also binds calcium but serves as a marker for a different population of interneurons, is also among the top downregulated genes (0.61 fold, FDR < 0.05). Additional genes of interest at FDR < 0.05 among the top downregulated include *NEUROD6*, which is involved in neuronal differentiation (0.60 fold), several ion channels (*SLC38A5*, 0.53-fold decrease; *SLC5A11*, 0.64-fold decrease), and *KDM5D*, a lysine demethylase (0.66-fold decrease). This highlights a diverse set of biological processes downregulated in ASD cortex. The top upregulated gene, *HSPA6* (2.6 fold increase, FDR < 0.05) is involved in the cellular stress response, as are other top upregulated genes such as *HSPB1* (2.1 fold, FDR < 0.05) and *GADD45G* (1.78 fold, FDR < 0.05). Additionally, the microglial marker *CD93* (1.88 fold, FDR < 0.05) and multiple members of the complement cascade implicated in microglial-neuronal interactions (*C4A*, 1.94-fold; *C1QB*, 1.65-fold, FDR < 0.05) are upregulated in ASD.

In order to gain a systematic understanding of biological pathways and cell types underlying the DGE signature in CTX, I evaluated gene ontology (GO) term enrichment of the gene sets increased and decreased in expression in ASD vs CTL (Appendix A3.1). Top enriched biological function and molecular function pathways are shown in Figure 4.1C. Genes decreased in ASD are involved in synaptic and neuronal function (synapse, postsynaptic membrane, synaptic transmission, gated channel activity, regulation of ion transport, ion channel complex) and mitochondrial function, while upregulated genes are enriched for inter- and intra- cellular signaling (response to stimulus, positive regulation of intracellular protein kinase cascade, activation of MAPK activity) and inflammatory pathways (immune system process, regulation of cytokine production).

Additionally, upregulated genes were enriched for the GO terms “regeneration” and “cellular developmental process.” In order to better understand the role of genes in these pathways, I evaluated cell-type specific expression changes more systematically. I assessed enrichment for genes expressed with high specificity in neurons, astrocytes, myelinating oligodendrocytes, and microglia (Appendix A3.1) and found significant enrichment in the upregulated DGE set for astrocytes and microglia genes, and significant enrichment of neuron and oligodendrocytes specific genes in the downregulated DGE set (Figure A3.2D). This suggests that, in ASD CTX, there is a downregulation of neuronal signaling and upregulation of astrocyte and microglia signaling. However, similar changes could also be seen due to alterations in cell type proportions. However, major markers of neurons are not significantly altered in ASD vs CTL (neurons: *RBFOX3*, p = 0.078; *RELN*, p = 0.30; *MAP2*, p = 0.1; glia: *GFAP*, p = 0.19; *S100B*, p = 0.71), demonstrating that global shifts in cellular populations are unlikely to drive this DGE signature.

I next sought to evaluate whether the DGE signature identified in the “Matched ASD” set generalizes to the “Young ASD” set. I utilized the DGE set to cluster the younger ASD samples (age < 10, which were held out to match the initial DGE analysis), and found that these samples, which were not included in defining the DGE set, cluster similarly suggesting this DGE set is generalizable to new samples (Figure A3.2E). To more formally evaluate the robustness of the DGE signal, I evaluated whether the young ASD samples shared the same signature identified in the matched cohort. I used principal components analysis (PCA) on the DGE set across 123 total samples and found that the 1st principal component (PC) explained 39% of the variance, is significantly different between the matched ASD samples and controls ($p=1.5x10^{-7}$) and between the young ASD samples and controls ($p = 2.8x10^{-5}$), but is not significantly different between the matched ASD samples and the younger samples (p=0.82, Figure 4.1D).

4.3.c. Gene expression changes in ASD cerebellum are weaker than those seen in CTX

Next, I used a similar DGE analysis as above for CB samples (22 ASD, 26 CTL). Although this analysis doubled the sample size from previous investigations (Ginsberg et al., 2013; Voineagu et al., 2011), no gene expression changes passed multiple comparisons at FDR < 0.05 (the most significant were at FDR ~ 0.30) (Table A.3.2B). Given that the CTX and CB dramatically differ in the cells composting the whole tissue, it is possible that similar underlying biological processes change in CB, but only in a weak manner or only in a minority of cells. This would manifest as a weaker, but similar DGE signature as that seen in CTX.

I therefore asked whether some of the most downregulated transcripts from CTX were also changed in CB, and found that *PVALB* was downregulated, but at a lower fold change (0.58 fold, p = 0.007) and *SST* did not pass criteria to be called as sufficiently expressed in CB. Other top downregulated genes from CTX were downregulated in CB, but at a lower magnitude and at with greater variability (*NEUROD6*, 0.67 fold, p = 0.09; *SLC38A5*, 0.73 fold, p = 0.02; *SLC5A11*, 0.64 fold, p = 0.08; *KDM5D*, 0.66 fold, p = 0.04). Comparing top upregulated gene in CTX, changes in CB were much weaker and suggested reduced inflammation in CB (*HSPA6*, 1.83 fold, p = 0.23; *HSP1B* was not detected; *GADD45G*, 1.22 fold, p = 0.21; *CD93*, 1.10 fold, p

= 0.66; *C4A*, 1.20 fold, p = 0.46; *C1QB*, 1.22 fold, p = 0.32). These findings reveal that upregulated ASD genes from CTX show a very weak upregulation signature in CB. In conclusion, for both down- and up- regulated genes in ASD, nearly all changes in CB are of lower magnitude and greater variability.

Given these weaker changes in CB, I asked whether the CTX DGE signature could be detected in CB more systematically by using the same analysis used for assessing replication in Figure 4.1A. This revealed that many of the genes changed in ASD CTX were, on average, also altered in ASD CB (genes DGE in CTX at $p < 0.05$: $R^2 = 0.47$ with CB; genes DGE in CTX at $p >= 0.05$: $R^2 = 0.10$ in CB; 673/15239 overlap, OR = 2.3, $p = 7.4x10^{-48}$). Moreover, the slope of the best-fit line among the points in the CTX DGE set between CTX and CB is 1.46. This demonstrates that, on average, DGE in the CTX is of about ~1.5x greater magnitude than that in CB (Figure A3.2G). This supports the idea that some of the same pathways are affected in both regions, but to a different extent.

I next evaluated whether the weaker changes in ASD CB were enriched for particular biological processes or were generalizable to new samples. On a DGE set at an unadjusted $p < 0.01$ comprising 357 genes, GO enrichment identified some evidence of pathways agreeing with those found in CTX but none were significantly enriched (data not shown). PCA on this DGE set across all CB samples revealed that the first PC explains 34% of the variance and distinguishes ASD from CTL in both the initially evaluated set and younger samples which were held out (Figure 4.1E). Furthermore, as with CTX, PC1 was not strongly related to measured covariates (Figure A3.2H). This shows that weak changes in CB can distinguish ASD from CTL, and this is generalizable to independent samples.

Taken together, comparison of CTX to CB suggests that similar underlying molecular changes may affect these cytoarchitecturally distinct brain regions, but the CTX exhibits selective vulnerability in ASD. However, given the relative cellular homogeneity of the CB (which is mostly comprised of granule cells) compared to CTX, the weaker gene expression changes in CB suggest that this region is less susceptible to changes in ASD. This could be due to the unique cell types in CB relative to CTX, or due to some other aspect of the molecular milieu of the CB that renders it resilient to the changes seen in CTX.

4.3.d. Attenuation of cortical patterning in ASD

The human brain exhibits regional specialization for behavioral and cognitive tasks, which is driven by patterning of gene expression that is related to developmental differentiation, neuronal signaling, and cortical cytoarchitecture (Hoch et al., 2009; Johnson et al., 2009; Kang et al., 2011; Khaitovich, 2004). Previous work with microarrays demonstrated that patterning between the frontal and temporal cortex was attenuated in ASD cortex (Voineagu et al., 2011). I evaluated DGE between FC and TC in this study with a paired Wilcoxon rank-sum test using the technical variable corrected (Appendix A3.1) expression profiles of 16 ASD and 16 CTL individuals who were matched for age and sex.

I find a similar loss of patterning as previously observed, with 551 genes at FDR < 0.05 between FC and TC in controls, but only 51 in ASD (Figure 4.1F). I refer to the set of 523 genes with this patterning in CTL but not ASD as the “Attenuated Cortical Patterning” set. This attenuation of patterning is also evident from the global distribution of differences between FC and TC in ASD and CTL (Figures A3.2I-J). GO term enrichment analysis on the genes with attenuated cortical patterning revealed enrichment for metabolic processes, G protein coupled signaling, Wnt receptor signaling, calcium binding, and additional developmental and differentiation related functions (Figure 4.1G). Additionally, the attenuated cortical patterning set was enriched for genes specific to neurons (OR = 1.6, p = 5.4x10^{-4}) and astrocytes (OR = 1.4, p = 6.9x10^{-3}), but not markers of oligodendrocytes or microglia (Zhang et al., 2014). This suggests that genes in neurons and astrocytes are primarily losing cortical region differences in ASD, and microglia and oligodendrocyte genes are affected similarly across the cortical regions.

Genes in the "Attenuated Cortical Patterning" set includes multiple molecules known to be involved in cell-cell communication and cortical patterning *PCDH10*, *PCDH17*, and *CDH12. MET*, which is among the most cortically patterned genes (Hawrylycz et al., 2012), is also seen as having diminished regional differences. Interestingly, *PDGFD*, which was recently shown to be necessary for human but not mouse cortical development, is also in this set (Lui et al., 2014).

I next evaluated whether the attenuation of patterning between cortical regions was due to increased variability in gene expression or a severe, global loss of cortical patterning. I used Bartlett's test for differences in variance between gene expression levels in ASD vs CTL, and found that there is a difference in variance for thousands of genes in ASD compared to CTL (Figure A3.2K, Table A3.2A). Alterations in variance between conditions can be due to many technical or biological factors, and this study is not optimally designed to understand differences in variance between ASD and CTL. However, it is clear that the genes exhibiting attenuated cortical patterning are not more likely to exhibit greater variance in ASD vs CTL than other genes (Kolmogorov-Smirnov test, two-tailed p = 0.11; 139/523 genes with p < 0.05 on Bartlett's test).

Next, to assess the extent of the loss of DGE in ASD compared to CTL FC and TC, and ensure the attenuation of patterning is not due to poor dissection quality or tissue degradation in ASD postmortem brains, I used an independent set of gene expression data to classify cortical regions. I trained a cross-validated Lasso regression model (Tibshirani et al., 2004) to differentiate frontal cortex and temporal cortex using BrainSpan gene expression data (Sunkin et al., 2013). This identified 14 genes that robustly differentiate FC and TC in BrainSpan and consistently differentiate FC and TC in both the CTL (AUC = 0.97, Figures A3.2L-M) and ASD samples (AUC = 0.96, Figures A3.2N-O). This suggests that the loss of cortical patterning in ASD is not so severe that cortical regions are indistinguishable, and demonstrates that dissections of brain regions and the brain samples themselves are of sufficiently high quality to separate cortical regionalization. Together, these results confirm an attenuation of cortical patterning between FC and TC in ASD, and identify that this alteration is not due to global differences in ASD and CTL samples.

Given that there is not a loss of global cortical patterning, I next sought to evaluate whether molecular pathways regulated by specific transcriptional regulators might be altered between regions. I used transcription factor binding site (TFBS) enrichment analysis (see Appendix A1.1) to evaluate whether common transcriptional regulators may bind upstream of the 523 genes in the attenuated cortical patterning set, and found that SP1, SP2, EGR2, KLF5, SOX5, and ARID3A may potentially bind to mediate cortical patterning. To prioritize a TF for future experiments, I evaluated whether any of the genes for these TFs were in the attenuated cortical patterning gene set. Out of these factors, only *SOX5*, which has been implicated in coritcofugal projection neuron development in mouse (Kwan et al., 2008; Lai et al., 2008), shows a difference between ASD and CTL between cortical regions (Figure 4.1H). This suggests its attenuation in patterning may be upstream of the loss of patterning in its targets. Further validation and analyses will be necessary to understand the role of SOX5 in cortical patterning in ASD.

4.3.e. Dysregulated lncRNAs in ASD

Multiple lncRNAs were found dysregulated between ASD and CTL (33 lincRNA, 19 antisense transcripts, and 10 processed transcripts at FDR < 0.05) and found to have reduced patterning in FC compared to TC in ASD (20 lincRNA, 6 antisense transcripts, 8 processed transcripts in the Attenuated Cortical Patterning set). Most of these lncRNAs are developmentally regulated (Jaffe et al., 2015) and contain chromatin states indicative of transcription start sites (TSSs) at their 5' end in brain (http://www.roadmapepigenomics.org/). For example, *SNHG11* (0.88 fold, FDR < 0.05) and *PART1* (0.75 fold, FDR < 0.05) are a processed transcript and lincRNA, respectively, that are downregulated in ASD cortex. *SNHG11* is highly specific to neurons (Zhang et al., 2014), most upregulated during infancy and childhood (Jaffe et al., 2015), and chromatin marks are indicative of a TSS at its 5' end. *PART1* shows no pairwise alignments in mouse, though is detected in primates suggesting its sequence is primate specific. It is highly developmentally regulated (Jaffe et al., 2015), with consistent increase in expression from fetal to teenage development, followed by plateauing in expression throughout, and its 5' end contains a TSS chromatin state, and it shares this bidirectional promoter with *PDE4D*. Finally, several lncRNAs show a loss of cortical patterning. For example, *WDFY3-AS2*, a transcript antisense to an ASD-implicated gene involved in cortical neurogenesis (Iossifov et al., 2014), exhibits attenuated patterning in ASD cortex (Figure 4.1I).

I plan to evaluate these lncRNA changes more systematically, particularly with a focus on the primate-specific sequences which could be interesting from the perspective of brain development (Geschwind and Rakic, 2013). It is important to note that sequence evolution and lack of alignment in a species is not sufficient to declare a lncRNA as species-specific as the sequence can change substantially but transcription and function may be unaffected. The current best practice is to show that syntenic regions are intact in mouse, but there is a lack of expression where the lincRNA ought to be expressed (Chodroff et al., 2010).

4.3.f. Alteration of alternative splicing in ASD

Previous work has shown that dysregulated splicing plays a role in ASD (Irimia et al., 2014; Voineagu et al., 2011; Weyn-Vanhentenryck et al., 2014). However, these studies have largely focused on subsets of samples showing extreme gene expression changes in *RBFOX1* (Voineagu et al., 2011; Weyn-Vanhentenryck et al., 2014) or in other selected subsets of patients with ASD (Irimia et al., 2014). Furthermore, previous splicing analyses in ASD pooled samples together to obtain sufficient depth for splicing event detection, which averages out inter-individual variation. I therefore performed a differential splicing (DS) analysis to assess whether a shared splicing pattern exists in ASD using the same mixed multiple regression framework and experimental design used in DGE, with percent spliced in (PSI) values at events of sufficient depth.

I evaluated 34025 splicing events in CTX and 32954 in CB (Table A3.3), encompassing skipped exons (SE), alternative 5' splice sites (A5SS), alternative 3' splice site (A3SS), and mutually exclusive exons (MXE) using the MATS pipeline for PSI calculation (Shen et al., 2012) (see Appendix A3.1 for event criteria). I found no events passing correction for multiple comparisons at FDR < 0.05 (or FDR < 0.1 or 0.2, which are more common thresholds for DS analysis) in CTX or CBL, and therefore explored more relaxed thresholds for significance. At p < 0.01 in CTX, which corresponds to a FDR < 0.45, 639 events are different between ASD and CTL. At p < 0.01 in CBL, I identified 411 events, but the most significant of these are at FDR > 0.50 suggesting a much weaker splicing signature in CBL and that CTX has a generally stronger signature of DS. Highly concordant results were found with an alternative splice junction mapping and quantification approach (Figure A3.3A). Additionally, there was no detectable global overlap between CBL and CTX above chance at $p < 0.01$ (OR = 1.1, p = 0.59) or $p < 0.05$ (OR = 1.1, p = 0.21). I therefore focused the DS analysis on cortical samples.

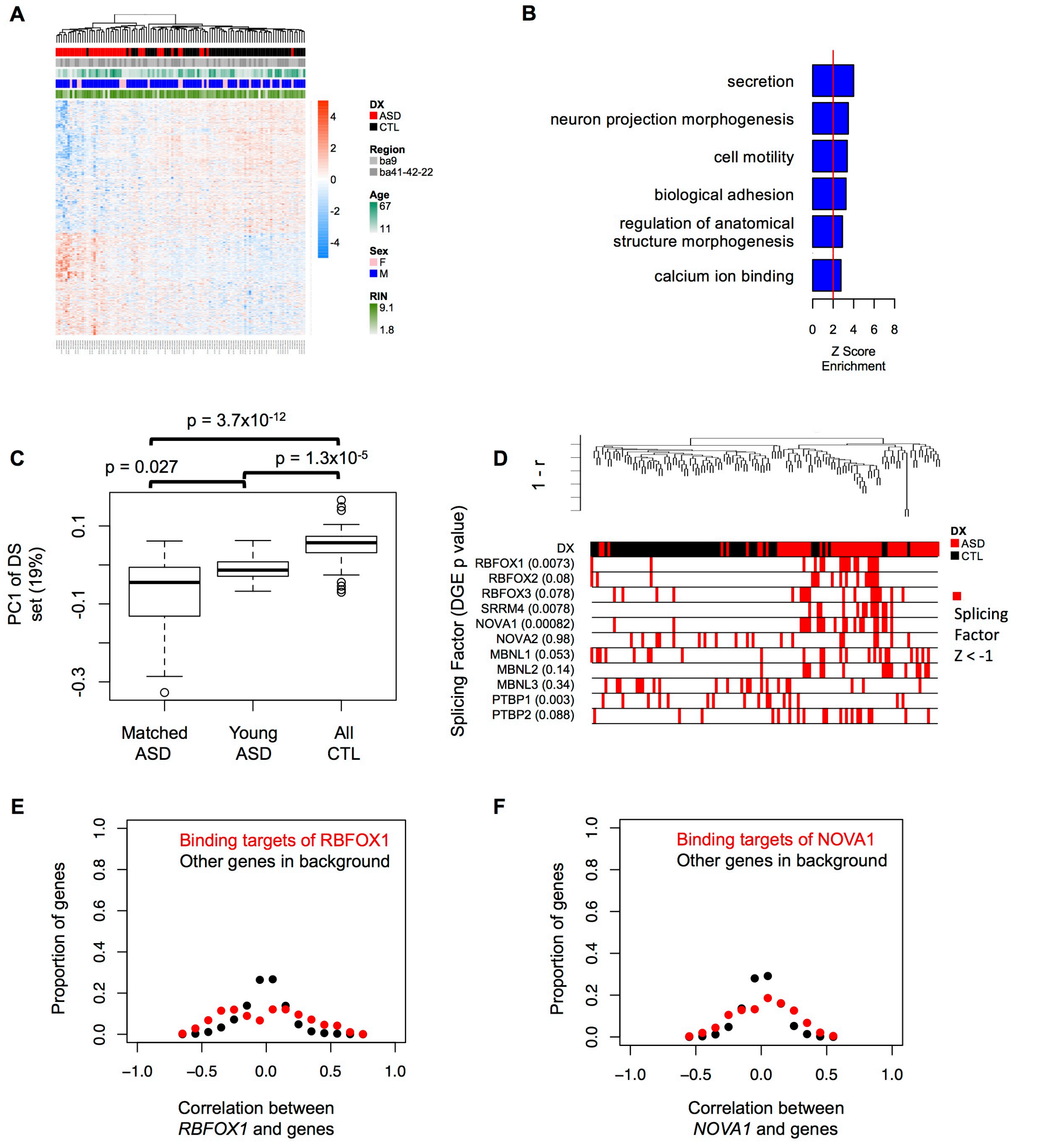


Figure 4.2 Differential splicing in ASD and the role of splicing regulators. A) Average linkage hierarchical clustering of samples using correlation-based distance with all DS events in ASD CTX at $p < 0.01$. B) GO term enrichment analysis of DS genes at $p < 0.01$. C) PC1 of the DS set shows clear differences between the "Matched ASD" set CTL, and generalizes to differences between "Young ASD" and CTL. D) Hierarchical clustering of samples with the DS set (excluding nominally DGE genes containing splicing events) with splicing factors plotted below. Instances where the splicing factor is 1 standard deviation below the mean across samples ($Z < -1$) are highlighted. E-F) Correlations between *RBFOX1* and *NOVA1* with genes split by targets of the respective splicing factors compared to background across all samples.

The most significantly altered splicing event was the inclusion of an exon in *ASTN2* (ΔPSI = 0.058 [5.8%], $p = 7.8 \times 10^{-6}$), a gene implicated by CNVs in ASD and other developmental disorders(Lionel et al., 2014). Several genes involved in synaptic function harbored evidence for multiple exons being alternatively spliced in ASD (*ANK2*, *NRXN1*, *NRCAM*, multiple events at $p < 0.01$). These patterns of events may be indicative of the presence of alternative isoforms in ASD, though the effect sizes are generally small ($|\Delta PSI|<0.10$) suggesting this occurs only to a mild degree or in a small fraction of the cells. Notably, at a genome-wide level, SE events contributed the most to the DGE signature (Figure A3.3B).

Splicing changes may be observed as an artifact when substantial DGE changes occur between conditions. I confirmed that the DS signature was not driven by DGE of the genes in which splicing events occurred (Figure A3.3C) by removing 228 events in genes with even nominal ($p < 0.05$) DGE in ASD vs CTL. The remaining events still identified a strong difference between ASD and CTLs for PC1 of the DS events ($p = 1.3 \times 10^{-12}$, Wilcoxon rank-sum test, Figure A3.3D). These findings establish that there exists a strong DS signature in ASD CTX that is independent of DGE, but that detection of individual high-confidence events at stricter statistical thresholds will likely require larger sample sizes in ASD. Clustering the 639 events at

$p < 0.01$ in CTX demonstrates that ASD samples cluster together (Figure 4.2A), and PC1 of the DS set was associated with ASD status but no other covariates or measured variables (Figure A3.3E).

Next, to evaluate the functional implications of these DS changes in ASD, I utilized GO term enrichment analysis using the genes harboring DS events. This identified biological and molecular processes related to neuronal function (secretion, neuron projection morphogenesis, calcium ion binding) as well as the synapse and related cellular compartments (Figure 4.2B). Additionally, the DS set shows enrichment for genes found in neurons (OR = 2.0, $p = 0.0062$), but no other cell types (Figure A3.3F). It is possible that longer genes, which contain more exons, also contain more detected splicing events. This could bias pathway and cell type enrichment to more neuronal and synaptic genes, which are, on average, longer than other genes in the genome. However, the correlation between the number of detected events in genes and gene length is minimal ($R^2 = 0.004$), and the correlation is even smaller for events at $p < 0.01$ ($R^2 = 0.00012$) demonstrating that longer genes are not more likely to contain DS events.

Finally, I evaluated whether these DS events generalize to younger ASD samples. Clustering analysis suggested this is the case (Figure A3.3G), and comparison of PC1 from the DS set confirmed this (Figure 4.2C). Taken together, there is a robust splicing signature in CTX that differentiates ASD from CTL, this signature is distinct from the DGE signature, and it is mostly related to changes in synaptic and neuronal function. There is a notable absence of splicing changes in genes found specifically in inflammatory pathways or astrocyte, oligodendrocytes, or microglia, suggesting that changes in transcript structure is exclusive to neurons.

4.3.g. Multiple splicing factors contribute to the shared splicing signature in ASD

Previous work has shown that splicing factors such as those from the RBFOX family (Fogel et al., 2012; Irimia et al., 2014; Voineagu et al., 2011; Weyn-Vanhentenryck et al., 2014) and SRRM4 may play a role in ASD. I hypothesized that the shared splicing signature in ASD might be a product of perturbations in specific splicing factors. I evaluated whether specific neuronal splicing factors are consistently perturbed in individuals with ASD. Figure 4.2D demonstrates that there is not one simple pattern of splicing dysregulation in ASD, and that different combinations are altered in different individuals.

Notably, *RBFOX1-3* all show one standard deviation ($Z < -1$) drop from the mean more frequently in ASD than CTL, and not necessarily in the same individuals. *SRRM4*, *NOVA1*, MBNL1, *MBNL2*, and *PTBP2* also exhibit this pattern, and several show evidence for DGE at $FDR < 0.05$ across ASD individuals (Table A3.2A). Notably, most splicing factors are decreased in ASD with the exception of *PTBP1*, which is increased. *PTBP1* is most commonly studied for its effect on splicing in the context of neuronal development, but is predominantly expressed in in microglia in adult brain (Zhang et al., 2014). Moreover, none of the splicing factors found in the DGE set from CTX show evidence of DGE in CB (those with $FDR < 0.05$ in CTX have $p > 0.5$ in CB). This suggests that splicing changes in ASD may be a product of regionally specific changes in these factors.

To evaluate whether some of these splicing factors might regulate the DS changes observed in CTX, I performed DS analysis as above with the 10 samples showing the greatest downregulation of *RBFOX1* and *NOVA1*, and compared their splicing event profiles to all CTL samples (Figures A3H, J). This identified many changes at $FDR < 0.1$ and $|\Delta PSI| > 0.10$ which were highly correlated to weaker changes in the full sample set (Figures A3I, K), suggesting that subsets of individuals with perturbations in specific splicing factors can identify more homogenous splicing alterations in ASD that are representative of the global alterations seen in ASD. Moreover, the events found at FDR < 0.1 and |ΔPSI| > 0.10 in the *RBFOX1* analysis overlap highly with known Rbfox1 targets from cross-linking immunoprecipitation (CLIP) experiments in mouse(Weyn-Vanhentenryck et al., 2014) (OR = 6.0, p = 3.9x10-40; out of 755 events, 106 overlap 535 orthologous events). A high overlap is also seen comparing events from the *NOVA1* analysis to Nova1 targets from CLIP data in mouse (Zhang et al., 2010) (OR = 8.7, p = 4.6x10-38; out of 1002 events, 77 overlap 228 orthologous events). Moreover, events predicted to be regulated by RBFOX1 and NOVA1 show a greater correlation to *RBFOX1* and *NOVA1* gene expression levels across all samples (Figures 2E-F). A similar concordance between splicing factor and putative regulatory sites has been observed for *SRRM4*, which regulates microexon events (Irimia et al., 2014).

Together, these analyses show that specific splicing factors in CTX are likely to underlie the changes seen in CTX. The shared splicing signature in ASD CTX may at least be partly mediated by primary alterations in different splicing factors driving overlapping splicing alterations. Several claims about this putative splicing level convergence need to be evaluated more rigorously. First, additional data is available for splicing factors other than *NOVA1* and *RBFOX1* that can be utilized to buttress the claims made. Second, a similar analysis as shown in Figures 2D-F can be performed in CB to evaluate the regional specificity and the CTX-specificity claims made here. Finally, DS events from stratified analyses using *RBFOX1*, *NOVA1*, and other factors need to be formally overlapped to identify the core set of splicing events that are regulated by these factors.

4.3.h. Duplication 15q syndrome exhibits widespread and stronger gene expression changes that recapitulate those in idiopathic ASD

Next, I sought to examine transcriptomic alterations in 8 individuals (6 FC, 8 TC, 3 CB, see Table A3.1) with duplication 15q syndrome (dup15q), a genetically defined cause of autism defined by a maternally inherited duplication of chromosome 15q11-13. Duplications along the 5 known breakpoints (BPs) were re-evaluated in 7/8 individuals via genotyping (Appendix A3, Table A3.4) and obtained for the remaining individual from a previous report {Scoles:2011jw}. For most individuals there were 4 copies of the region from breakpoints 1-4, and 3 between breakpoints 4-5. Thus, as expected, most genes in the 15q11.1-13.2 region have higher expression in dup15q CTX compared to CTL (Figure 3.3A). Changes in the dup15q region in CB are similar, though potentially weaker (Figure A3.4A, but this should be interpreted with caution with N = 3). Notably, although there is general overexpression in dup15q in this region, *SRNPN* and *SNURF* were downregulated as were additional genes flanking region near BP5 *SCG5* and *FMN1*. Additionally, no changes in idiopathic ASD were significant and in the same direction as the changes in dup15q in CTX or CB.

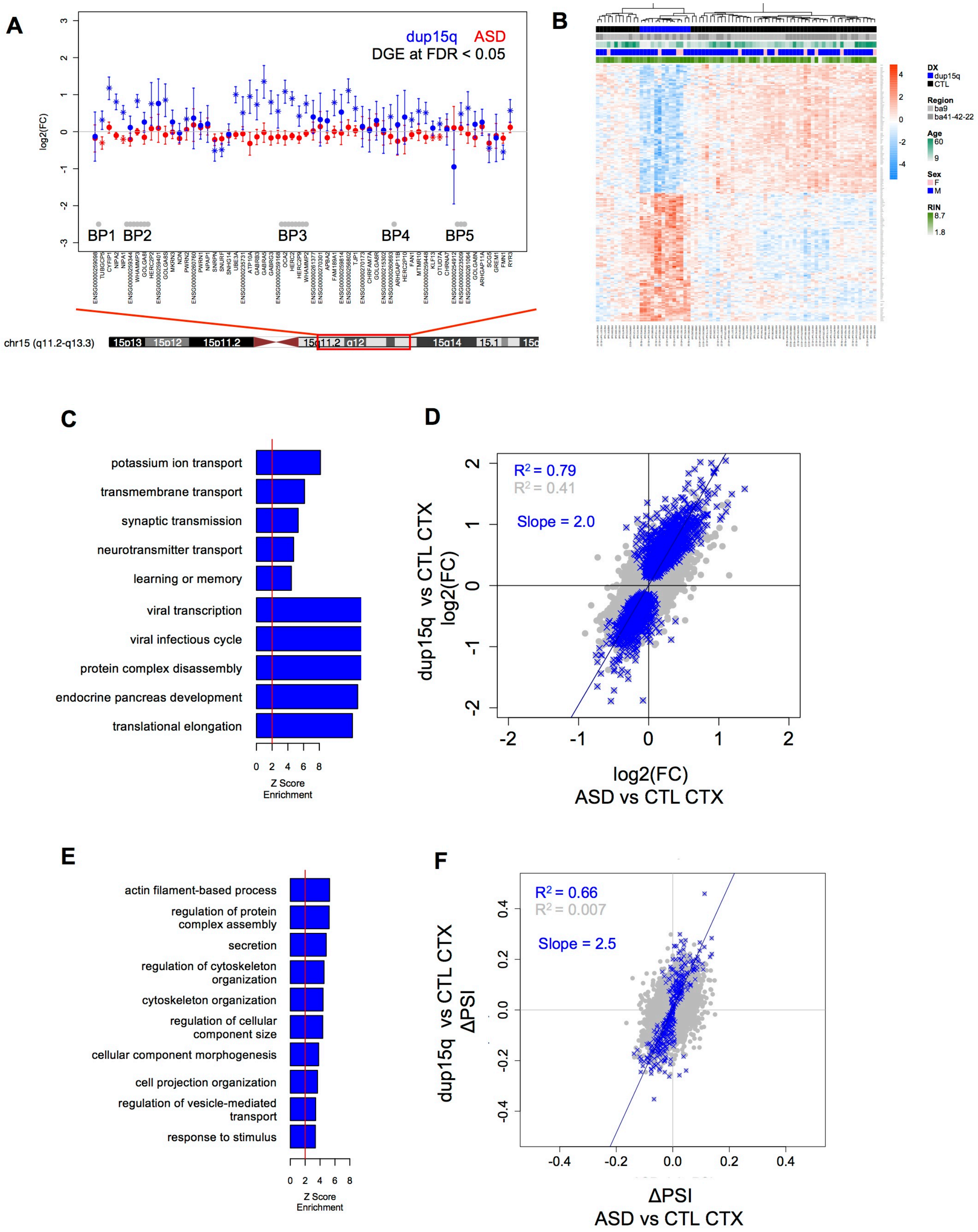


Figure 4.3 Differential gene expression and differential splicing in duplication 15q syndrome. A) DGE changes across the 15q11-13.2 region for ASD and dup15q compared to CTL, error bars are +/- 95% confidence intervals for

the effect sizes. B) Average linkage hierarchical clustering and heatmap of DGE top 100 up- and down- regulated changes in dup15q. C) GO term enrichment for 2875 genes DGE in dup15q at FDR < 0.05. D) Comparison of effect sizes in dup15q vs CTL and ASD vs CTL, with changes in dup15q at FDR < 0.05 highlighted. E) GO term enrichment with 330 genes with DS events in dup15q vs CTL. F) Comparison of DS changes in dup15q vs CTL and ASD vs CTL, highlighting 402 events at FDR < 0.2 in dup15q.

Moreover, with only these 8 dup15q individuals evaluated, there is a clear dup15q vs CTL signal in CTX with 2875 genes differential at FDR < 0.05 (1506 upregulated, 1369 downregulated). Clustering using the top 100 upregulated and bottom 100 downregulated transcripts, this DGE signal separates all dup15q samples from CTL (Figure 4.3A) demonstrating that dup15q changes are far more homogeneous than those seen in idiopathic ASD. Comparison with CB (Figures A3.4B-C) did not identify as strong agreement of changes as found in idiopathic ASD, though this may be due to the low sample size in CB for dup15q. GO term enrichment with the dup15q vs CTL in CTX DGE set implicated pathways and cell types similar to the idiopathic ASD analysis, but with greater enrichment (Figure 4.3C, Figure A3.4D). Notably, major neuronal cell type markers change considerably more in dup15q, suggesting there may be consistent cell loss (neurons: *RBFOX3*, p = 2.2×10^{-4}; *RELN*, p = 4.1×10^{-3}; *MAP2*, p = 0.012; glia: *GFAP*, p = 0.42; *S100B*, p = 0.17), potentially related to previously reported microcephaly in some of these individuals {Wegiel:2012vm}. Additionally, many dup15q individuals shared sudden unexpected death in epilepsy or had seizures as a reported cause of death, so we evaluated the relationship between all measured covariates (including whether the individual had seizures) but found only a weak association between PC1 of the DGE set and any seizure status, and minimal associations to other factors other than diagnosis (Figure A3.4E).

Next, I asked whether dup15q, which appears to reflect a more homogeneous DGE signature compared to CTL, shares gene expression patterns with idiopathic ASD. Remarkably, using the dup15q vs CTL CTX DGE set (at FDR < 0.05), there is substantial overlap (FDR < 0.05: $R^2 = 0.79$, FDR > 0.05 $R^2 = 0.41$, suggesting substantial sharing of the signature above this threshold). Moreover, the slope of the best-fit line through these changes is 2.0, demonstrating that, on average, the changes in dup15q CTX are twice the magnitude of those in ASD CTX. Overlapping the 1156 genes DGE in ASD vs CTL CTX at FDR < 0.05 with the 2875 genes DGE in dup15q vs CTL CTX at FDR < 0.05 confirms this remarkable overlap, with 700 genes in common (OR = 9.2, $p = 8.8 \times 10^{-259}$; 325 downregulated, 375 upregulated). Taken together, these findings demonstrate that dup15q syndrome exhibits DGE in CTX that is remarkably similar to ASD, demonstrating that a genetically defined subtype of dup15q has a convergent DGE signature with idiopathic ASD.

Next, I sought to evaluate DS changes in dup15q vs CTL in CTX. There is only one DS change at $p < 0.01$ in the dup15q region (Figure A3.4F), consistent with the idea that duplication in this region simply duplicates all isoforms of the genes resulting in no alteration of transcript structure. Global DS analysis in dup15q compared to CTL revealed a stronger signature that that seen in CTX, with 402 events at FDR < 0.2 that clearly discriminate dup15q samples from CTL (Figure A3.4G). Given the widespread changes in gene expression in dup15q, this signature could be driven by gene expression alterations, but eliminating all genes DGE at $p < 0.05$ from the DS set retains a strong signature that separates dup15q from CTL (Figure A3.4H-I). Additionally, as with the DGE signature, PC1 of this DS set shows weak association with seizure status, but is otherwise largely associated with dup15q status (Figure A3.4J).

GO term enrichment analysis clearly implicates cytoskeletal components and genes involves in changes in cellular morphogenesis. Cell-type enrichment of genes harboring the DS events shows enrichment for genes found in neurons (OR = 2.6, p = 2.6×10^{-4}), but not other cell types suggesting these pathways are acting largely in neurons. Comparing this DS set with the changes seen in idiopathic ASD vs CTL, there is clear overlap (FDR < 0.2: $R^2 = 0.66$, FDR > 0.2: $R^2 = 0.007$) suggesting that DS changes in dup15q syndrome recapitulate those of idiopathic ASD. Finally, the slope of the best fit line through the DS events in dup15q CTX compared to those in ASD CTX is 2.5, suggesting that the splicing changes in dup15q are greater than those in ASD, on average. Taking the dup15q and ASD DS analyses together, it is clear that our study had sufficient power and sequencing depth to detect DS, but the heterogeneity of DS in idiopathic ASD is a major obstacle to identifying large and consistent changes at the PSI level. It will be important to assess the status of neuronal splicing factors in dup15q brain and assess the splicing factor – target relationship as assessed above.

These results strongly support the idea that the weaker DGE and DS signature seen in idiopathic ASD is due to biological changes (as opposed to technical factors such as brain quality), and support the idea that a causal genetic alteration can lead to the transcriptomic changes in idiopathic ASD. Moreover, they demonstrate that dup15q may serve as a viable model for better understanding additional molecular alterations in ASD, as the changes in dup15q brain are far more homogeneous and of greater magnitude than those seen in idiopathic ASD.

4.3.i. Co-expression network analysis reveals novel ASD modules

DGE analyses rely on prior knowledge of the ASD and CTL status of individuals to identify significant gene sets, and ignore the molecular context in which each gene functions. I therefore utilized weighted gene co-expression network analysis (WGCNA), which identifies modules of shared biological function in an unsupervised manner and then allows assignment of module-level relationships to diagnosis and other measured factors. I utilized a modified version of signed weighted whole-genome WGCNA, which ensures robustness given heterogeneous samples (see Appendix A3.1). This was applied to 137 cortical samples (combining the "Matched ASD", the "Young ASD", dup15q, and all control samples). This resulted in 16 robustly identified modules (Figure A3.5A).

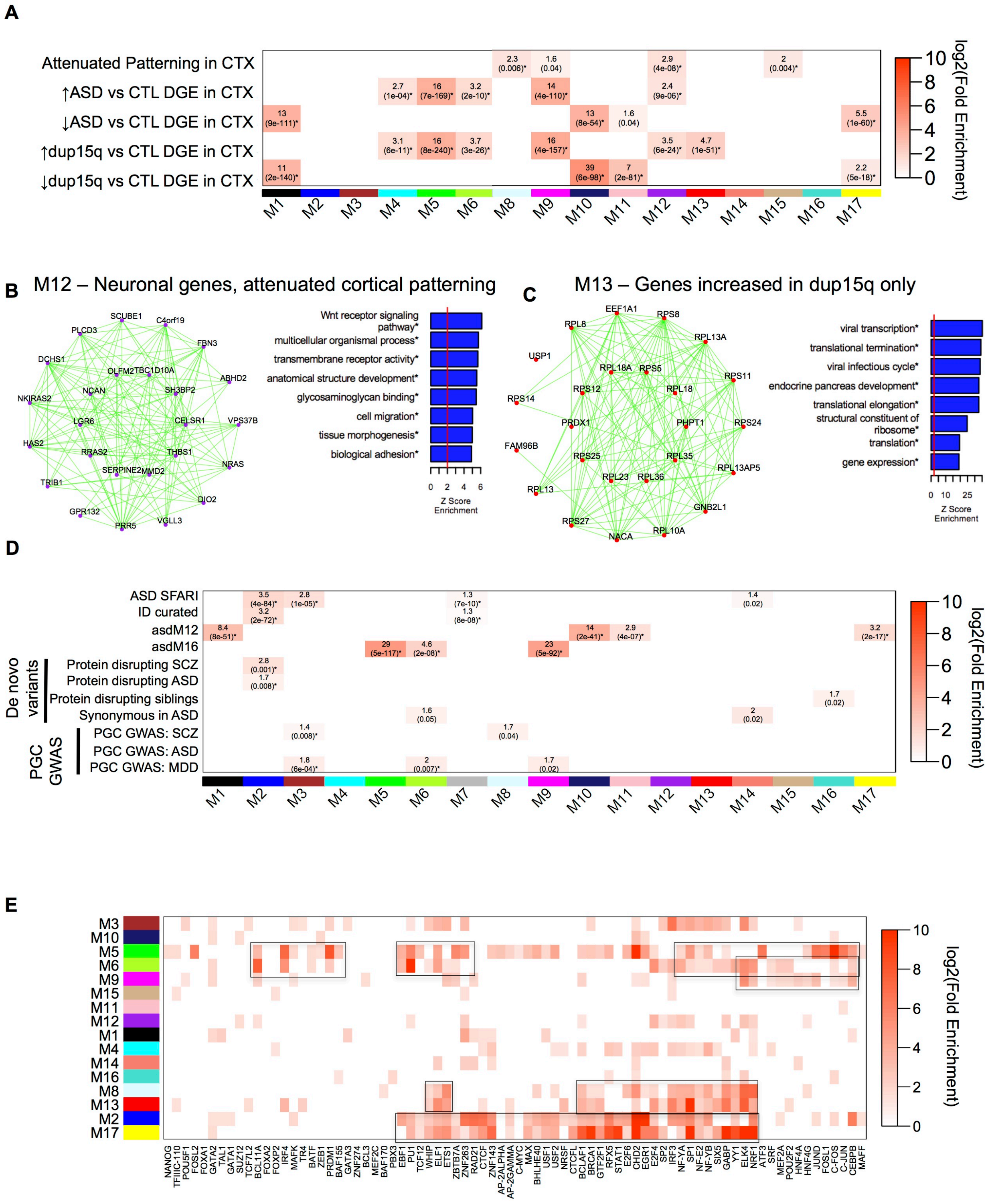


Figure 4.4 Co-expression network analysis across all ASD and CTL samples in CTX. A) Enrichment for gene sets related to DGE between FC and TC, between ASD and CTL in CTX, and between dup15q and CTL in CTX.

*FDR < 0.05 for this Heatmap. See also the module-trait enrichments in Figure A3.5B. B-C) Module and GO term enrichment plots for two modules of interest. The top 25 hub genes are displayed. *FDR < 0.05 for GO term analysis across all 16 modules. D) Module enrichment for gene sets previously implicated in ASD, including curated genes (ASD SFARI, ID curated), ASD modules from Voineagu et al., 2011 (asdM12 and asdM16), genes implicated by protein disrupting rare variants in SCZ, ASD, and control sets (genes affected by protein disrupting in siblings and synonymous mutations in ASD), and genes implicated by GWAS p < 0.001 in the PGC cohort. All enrichments shown here are adjusted for gene length. E) TF binding site enrichment assessing predicted TF binding (Arbiza et al., 2013) in open chromatin regions in frontal cortex (from the Roadmap Epigenetics Mapping Consortium, http://www.roadmapepigenomics.org/). TFs are assigned to genes using DNaseI hypersensitivity based open chromatin correlations to the promoter of genes (Thurman et al., 2012). The heatmap is clustered on both the x- and y-axis to group similar TF-module pairs, and boxes identify pairings corresponding to potential clusters.

In order to identify modules associated with ASD or dup15q, I first systematically tested enrichment for the “Attenuated Cortical Patterning” set, the up- and down-regulated DGE in ASD vs CTL CTX sets, and the up- and down-regulated DGE in dup15q vs CTL CTX sets. At a FDR < 0.05, this identified 3 modules associated with ASD (M1/10/17) and one with dup15q (M11, which is more weakly enriched for the ASD set). Many more modules were upregulated in ASD and dup15q, with 5 in both ASD and dup15q (M4/5/6/9/12) and one specific to dup15q (M13). Notably, the modules with highest enrichment with DGE sets are concordant with the module eigengene (ME)-trait associations associated with ASD (the model accounts for biological and technical covariates) (Figure A3.5B). Weaker enrichments are observed that are missed by ME-trait enrichment, likely because the enrichment is driven by a smaller subset of genes that are not captured by the ME. Cell-type (Figure A3.5C) and GO term enrichments in modules identified that most modules downregulated in ASD were enriched for neuronal genes, while modules upregulated in ASD were either enriched for markers of microglia or astrocytes.

For example, M1 (Figure A3.4D) is downregulated in ASD and dup15q, is enriched for genes specific to neurons, and contains GO terms related to calcium signaling, synaptic transmission, and ion channel activity. Hubs of M1 include *CNTNAP1*, *SV2A*, and *SYT1*, which are markers of neurons and synapses. Intriguingly, M12 is a module that is enriched for the attenuated cortical patterning set and upregulated genes in ASD and dup15q, but does not show a change in ASD in the ME-trait comparison. M12 is enriched for neuron-specific genes, and pathways related to cortical patterning and brain development (Wnt signaling, anatomical structure development, cell migration). This suggests that the attenuation of cortical patterning might occur due to alterations in the normal functioning of this module.

Upregulated modules in ASD show less concordance with both the enrichment analysis and the ME-trait analysis. The module most strongly enriched for upregulated genes in ASD and dup15q, M5, is also upregulated in both conditions and highly enriched for microglial markers. This module contains GO terms related to inflammation and cytokine signaling (Figure A3.4E). Interestingly, both gene set enrichment and ME-trait association for upregulated genes support M13 in dup15q, but not ASD supports one module. This module contains many genes related to translation and transcription, warranting further investigation into why it is altered to a greater extent in dup15q compared to idiopathic ASD.

Gene set enrichment analysis (Figure 4.4D) with known ASD genes confirmed that the downregulated ASD modules were all enriched for asdM12, the downregulated co-expression module from Voineagu et al., 2011. Additionally, M4/5/9 were enriched for asdM16, the upregulated co-expression module. The larger sample set in this study not only stratifies these modules into more specific biological processes but also further identifies novel models related to cortical patterning and dup15q syndrome.

Interestingly, M2 is highly enriched for curated ASD and ID genes and protein disrupting RDNVs in SCZ and ASD. M2 is not associated with any particular cell type, but is enriched for the GO terms polymerase II promoter, chromatin binding, and establishment of RNA localization. Remarkably, it also contains an unusually large fraction of lncRNAs (15% of the genes in M2 are classified as lncRNAs, while other modules are 1-5% lncRNA). However, this module is not altered in ASD and is not enriched for any ASD-associated set from at the transcriptomic level. M2 contains genes more highly expressed during brain development, so perturbations in it may not be evident in adult.

Finally, no module was strongly enriched for genes near common variants associated with ASD using GWA statistics (genes with $p < 0.001$ for the best SNP within 20kB). However, several steps are necessary to optimize this analysis, including assessment of additional GWA datasets and evaluation of better SNP assignment methods to genes. Taken together, co-expression network analysis re-affirms many observations from previous analyses, including the pathways involved in attenuated cortical patterning and the shared effects between dup15q and ASD compared to CTL. However, the module specific to dup15q warrants further investigation, as do general developmental trajectories that are associated with these co-expressed modules. Investigating earlier expression patterns might enable further stratification of modules to more developmentally relevant pathways. It will also be of value to attempt to assign splicing changes into these modules by correlating PSI levels with the eigengenes, and evaluate splicing factor co-expression with putatively regulated events.

4.3.j. Transcriptional regulators and ASD-associated changes

Co-expression among genes may indicate that they are co-regulated. To assess whether this is the case, transcription factor (TF) and chromatin regulator (CR) binding site enrichment can be evaluated in putatively co-regulated sets. I obtained regions of the genome with high-confidence binding sites from previously published work (Arbiza et al., 2013), and assigned them to genes by combining open chromatin states in FC (from the Roadmap Epigenomics Mapping Consortium, http://www.roadmapepigenomics.org/) with cross-tissue DNaseI hypersensitivity sites (Thurman et al., 2012). The latter correlates DNaseI signal across over 100 cell types, allowing distal regions of the genome to be correlated to genes by high correlation ($r > 0.7$) to DNaseI signal in the promoter of genes. In this manner, each TF was assigned genes that it was likely to regulate in brain.

I then utilized Fisher's exact test to assess enrichment across all modules for all TFs (Figure 4.4E). For each TF, this identified a putative enrichment signature across all modules, allowing the grouping of similarly enriched modules (Figure 4.4E, y-axis) and TFs that share similar patterns of enrichment across modules (Figure 4.4E, x-axis). Several clear "blocks" of enrichment stand out, related to modules sharing similar TF binding patterns. M5, M6, and M9 all share common TFs and are all upregulated in ASD. These TFs include ones known to be involved in inflammation or inflammatory signaling cascades (IRF1, C-JUN), consistent with these modules' enrichment for these pathways and upregulation in ASD.

Interestingly, M2, which was enriched for ASD related mutations but not altered in ASD, and M17, which is not enriched for genetic signatures but is enriched for neuronal markers and downregulated in ASD, share a large set of TFs that might co-regulate them. These include ELF1, which was previously found enriched between developmental co-expression modules (Parikshak et al., 2013), CHD2 which is implicated in ASD (Iossifov et al., 2014), and SP1/2 which are involved in activity dependent signaling.

These analyses of transcriptional regulation convergence are preliminary, and require additional investigation. First, whether these TFs are differentially expressed or the hubs of particular modules will help inform their biological relevance to ASD. Second, understand their developmental trajectories will help identify if they share common functions during brain development. Third, utilizing additional binding data or experimental knockdowns will help ensure these are likely to be real regulatory relationships. Despite these shortcomings, this analysis begins to identify an approach toward identifying co-regulation in a novel way that accounts for cortex-specific TF binding by integrating the appropriate epigenetic data.

4.4. Discussion

Despite challenges posed by the heterogeneity of ASD, these results confirm clearly shared alterations at the DGE, DS, and co-expression level in ASD brain. Our lab had previously demonstrated that there is a shared DGE signature in ASD CTX at the DGE and co-expression level, and this study confirms that this is the case in independent samples. Moreover, it extends this finding by adding transcript splicing and long noncoding RNA to the picture. Finally, it confirms that this signature is unlikely to be due to confounding factors by rigorous evaluation of covariates at every level, and, more importantly, by comparison to a genetically defined subtype of ASD, dup15q syndrome.

The results presented here are more along the lines of an extensive analysis of this large amount of data, and many efforts were made to choose the best analytical methods and statistical models to account for potential technical biases and confounders. Despite this, one potential limitation is that these data are from unstranded RNA-seq, and therefore miss many antisense and lincRNA transcripts. Furthermore, I have not attempted to perform novel splice junction detection or transcriptome assembly, so putative events or transcripts may be missed. Finally, our average read depth (~20M fragments per sample, Figure A3.1A) is on the lower end for lncRNA and DS event detection, which often require deep sequencing of protein coding regions (100M fragments). However, the depth we use is clearly sufficient to detect ASD associated changes across analyses, including DS analyses. This is supported by the fact that we obtain concordant results with more homogeneous biological changes (dup15q shows stronger effects, for DS and lncRNA changes). Sequencing deeper in future work, or pooling some of the existing samples together, could help identify novel events that might be missed otherwise. However, washing out biological variation and exacerbating potential technical biases is a potential shortcoming of the latter approach.

Finally, the following analyses are still in progress for this study at the time of writing this work:

- evaluation of primate-specific lincRNAs using RNA-seq in mouse and macaque
- deeper evaluation of splicing factors with differential splicing data from mouse experiments knocking down splicing factors such as SRRM4, RBFOX1, NOVA1, etc.
- better evaluation of transcription factor enrichment, including evaluating changes combining binding information and TF knockdown experiments, and TF developmental trajectories
- More rigorous evaluation of the relationship between co-expression modules and GWA findings by properly assigning genes to transcripts after linkage disequilibrium pruning
- integration of these transcriptomic changes with epigenetic changes (this may not go into the final version of this work, but be published in collaboration with others)

4.5. Materials and methods

Please see A3 Additional Methods and Figures for Chapter 4 for all methodological information.

CHAPTER 5:
Conclusions and future directions

“But I don’t want to go among mad people," Alice remarked.

"Oh, you can’t help that," said the Cat: "we’re all mad here. I’m mad. You’re mad."

"How do you know I’m mad?" said Alice.

"You must be," said the Cat, "or you wouldn’t have come here.”

— Lewis Carroll, *Alice in Wonderland*

5.1: Toward high-resolution gene regulatory networks in brain development and autism

The work I present here identifies convergent biological pathways affected by genetic risk factors in ASD during brain development as well as pathways altered in brains of individuals who have an ASD diagnosis. However, this work is limited in that it cannot identify causal pathways or precise mechanisms. Here I discuss the major factors underlying these difficulties, and highlight potential solutions. Notably, these are all factors are related to the broader difficulties of omics methods in brain, as discussed in Chapter 1.

5.1.a. Heterogeneity in ASD and lack of brain tissue

Despite my finding concordance in ASD-associated signals at the transcriptomic level, understanding deeper molecular mechanisms is still hindered by the heterogeneity of ASD. For example, it would be valuable to understand if individuals might be sub-stratified based on combinations of splicing factor dysregulation and the consequence splicing changes. Toward this end, it would be valuable to understand whether these individuals share common mutations in particular splicing factors, or in a splicing factor network. However, this would require sample sizes at least a magnitude greater (but likely more – it is difficult to know without more data).

One possible way around this issue is to collect more genetically defined subtypes of ASD. Brain samples are available for individuals with Fragile X syndrome and other rare monogenic forms of ASD, which could add considerably to the current transcriptomic analyses. Another approach to obtain genetically similar patients might be to take a large set of genotyped control brain samples (e.g. 1000 brain samples from various brain banks) and stratify them based on polygenic risk scores for ASD using a combination of common variants (Cross-Disorder Group of the Psychiatric Genomics Consortium, 2013; Euesden et al., 2014). Once large and well-powered ASD GWAS are available, this could be used to identify a top and bottom 10% of the cohort (e.g. comprising 100 individuals each) that are highly discordant in their risk for ASD (e.g. the top 10% might be 10x as likely to get ASD). Contrasting these groups could identify transcriptomic signatures related to ASD risk, but not confounded by downstream pathology such as microglial inflammation.

Finally, given the large number of informative molecular assays, there is not enough biological material from a postmortem brain to perform all desirable analyses. Toward this end, biological assays requiring smaller amounts of input material without sacrificing assay resolution are necessary. For example, currently available low-input RNA-seq kits rely on poly(A) tail selection and are very sensitive to RNA degradation, and are therefore not ideal for postmortem brain transcriptomics.

5.1.b. Whole tissue profiling of postmortem disease tissue remains a challenge

Another major challenge in this work is the biological interpretation of transcriptomic results from whole brain tissue. It is difficult to truly know whether the genome-wide transcriptomic signature is influenced by changes in neuronal subpopulations or by changes in signaling in those neuronal subpopulations without the appropriate detailed experiments. And even then, these experiments would be limited to looking at one or a few marker genes or proteins at a time, and this fundamentally loses the value of the genome-wide approach.

To address this issue, some groups have used neuronal nuclei selection (Cheung et al., 2010), and additional groups are pursuing this as an approach to sequencing more homogeneous transcriptomes and epigenomes. However, there is considerable heterogeneity within the NeuN+ population itself, and this will not resolve the issue of understanding accurate splicing alterations as these changes occur largely outside the nucleus. Additionally, transcriptomes from nuclei lose cytosolic and synaptic transcripts that contain isoform-level information, which clearly contribute a unique transcriptomic signature. Single cell transcriptomics is promising (Pollen et al., 2014), but this method is in its infancy and is mostly applied to fresh tissue. Developing this method further for frozen tissue and moving it beyond shallow depth sequencing and marker detection will be necessary to fully utilize postmortem brain samples from disease.

5.1.c. Current approaches cannot infer causality

Ultimately, the major limitation to many areas of postmortem and disease brain profiling come down to the lack of causality in profiling experiments. Several approaches can be leveraged to obtain causal insights. First, high-resolution temporal profiling of representative samples may allow the inference of causal trajectories, but shifts in cellular populations may hinder this (Jaffe et al., 2015), so this ought to be done after purifying cell populations in some manner.

Constructing appropriate *in vitro* samples from humans is also a promising approach (Dolmetsch et al., 2011), but the relevance of induced pluripotent stem cell models to human brain development and disease is still poor (Stein et al., 2014). Although these *in vitro* models may be valuable for therapeutic testing or directed experimental assays, they are unlikely to appropriately recapitulate disease due to the lack of *in vivo* cellular circuitry. Organoid models may change this with time (Lancaster et al., 2013), but currently no convincing proof of principle exists for a complex neurodevelopmental disease to be reliably modeled at a circuit level *in vitro*.

Finally, a systems genetics approach (Civelek and Lusis, 2013), which leverages genetic changes in the population (and is at some level very similar to the idea of profiling individuals with high vs low polygenic risk scores discussed above) and relates them to molecular alterations might be of value. However, these methods, which are discussed below, should be interpreted with caution in brain, as they are also susceptible to changes in cell type proportions.

5.2: Integrating the genome and transcriptome for causality

As discussed above, even after accounting for cell loss and postmortem effects, a causal role cannot be assumed for transcriptional changes identified in postmortem tissue without evidence from causal perturbations. The most ambitious and exciting goal in systems biology is to elucidate the functional genetic architecture of diseases by systematically identifying causal effects using genome-wide variation to disambiguate primary and secondary changes that occur in disease. Two recent studies show that this goal is possible(Rhinn et al., 2013; Zhang et al., 2013) in Alzheimer's disease (AD) by using genetic variation as a filter or causal anchor to define genetically driven network-level changes in AD and provide experimental validation for the network predictions.

Rhinn and colleagues(Rhinn et al., 2013) leveraged the differential susceptibility for AD conferred by alleles of the *APOE* gene; the *APOE4* allele is a major genetic risk factor for AD, accounting for almost 30% of the genetic variance(Guerreiro et al., 2012). The researchers used existing transcriptomic and genetic variation data in AD derived from an eQTL study(Webster et al., 2009)to compare DGE genes between individuals homozygous for *APOE3* (low-risk allele) with and without AD, and DGE between unaffected individuals with and without one *APOE4* allele (high-risk allele). The expression signature in *APOE4* carriers was similar to that of patients with AD(Rhinn et al., 2013), suggesting that unaffected *APOE4* carriers show a prodromal molecular AD-like transcriptomic state before major pathological alterations manifest. The authors applied DCA to compare the transcriptome-wide correlation of each gene in the unaffected *APOE4* group versus the affected *APOE3* group(Rhinn et al., 2013) and filtered the changes to a list of candidate genes predicted to cause transcriptomic dysregulation in AD. This approach is similar to assessing co-expression connectivity within a module, but at a genome-wide network scale. Experimental validation in a neuroblastoma cell line and in human neurons from induced pluripotent stem cells confirmed that selective pharmacologic inhibition of one of the top candidates, *SV2A*, decreased pathological beta amyloid accumulation. The authors further related their findings back to patients by demonstrating that an interaction between two single nucleotide polymorphisms in top candidate genes (SNPs; in *FYN* and *RNF219*) decrease the age of onset in AD in *APOE4* non-carriers using GWAS data. Furthermore, they showed that the *RNF219* SNP modulates beta amyloid levels through assessing positron emission tomography scans from patients using publicly available data from the Alzheimer's Disease Neuroimaging Initiative(Mueller et al., 2005). This approach relates genotype to phenotype to discover novel mechanisms and demonstrates how genotype, transcriptomic and phenotypic data may be integrated across studies using publicly available data.

An even more generalizable strategy using an eQTL network approach was undertaken by Zhang and colleagues(Zhang et al., 2013), who applied WGCNA on hundreds of postmortem brain samples from individuals with AD, other neurodegenerative disease, and controls. The team used transcriptomic networks to show that multiple transcriptional modules were remodelled in AD. Gain of connectivity was observed in immune and neurogenesis pathways, whereas loss of connectivity was predominant in pathways related to GABA signalling and myelination. eQTL analysis identified over 11,000 SNP-gene associations that were enriched in several modules. These eQTL associations provide a causal anchor since the gene expression changes are caused by genetic variation, and not vice versa. The researchers then applied Bayesian network analysis to evaluate putative causality in network structure on an AD-related microglial module, implicating *TYROBP* as a regulatory hub. The regulatory role of *TYROBP* was validated by overexpressing two forms of the gene in mouse and subsequently performing

RNA-seq. Remarkably, the magnitude of DGE of genes in the *Tyrobp*-overexpressing mice was related to the distance of the gene from *TYROBP* in the human causal network, with closer genes showing the greatest changes in expression(Zhang et al., 2013). This suggests that the network structure is predictive of effects of hub genes on other genes in the network, similar to previous studies using WGCNA(Winden et al., 2009)**.**

These two studies demonstrate the power of integrating genomics with transcriptomics and correlational networks using large sample sizes (>100 cases and controls in each), and establishing causality by evaluating genotype-phenotype and eQTL associations. Future studies that combine the mechanistic rigor and resourceful study design of Rhinn and colleagues(Rhinn et al., 2013) with the eQTL methodology and causal network approach of Zhang and colleagues(Zhang et al., 2010) will advance the capacity of network studies to leverage causal genetic effects in CNS disorders.

In the context of neurodevelopmental diseases, a systems genetics approach over early developmental time periods in individuals with disease might be extremely valuable, though this would require hundreds to thousands of fetal brain samples evaluated for eQTL. Alternatively, using a polygenic risk score, one might stratify individuals as done by Rhinn et al. to see whether the prodromal effects of incipient disease might be detected. This could be particularly informative at prenatal time points, as it would identify the aggregate effect of ASD-associated common variant risk on the prenatal transcriptome.

5.3: Conclusions and future directions

In this work, I have discussed how transcriptomic and other omic approaches can discover new biology in ASD in an unbiased, powerful, and reproducible manner. Combining resources across individual studies permits novel and aggregate analyses that identify genes, pathways, or other relationships that cannot be detected at the level of individual studies.

Currently, much of neuroscience research is still focused on candidate genes and candidate hypotheses, so sceptics may question the value of measuring entire systems as done throughout this work. However, complexity cannot be ignored; measuring genome-wide changes in conjunction with studying individual genes and pathways will be essential to address the true underlying mechanisms of neurodevelopmental disorders. In the future, I hope to perform well-designed and reproducible transcriptomic (and other omic) studies in brain that simultaneously evaluate hypotheses in an unbiased manner and generate new hypotheses. The results of such high-quality genome-wide studies will be essential to develop and test hypotheses that look beyond where current knowledge ends.

APPENDIX

“We conclude that the human appendix contains a robust and varied microbiota distinct from the microbiotas in other niches within the human microbiome. The microbial composition of the human appendix is subject to extreme variability and comprises a diversity of biota that may play an important, as-yet-unknown role in human health.”

― Guinane et al., 2013 supporting the idea that the appendix may not be useless and justifying that its contents are expected to be variable across the three following appendices (Guinane et al., 2013).

A1. Additional Methods and Figures for Chapter 2

A1.1. Extended Methods

Developmental expression data: BrainSpan developmental RNA-seq data (publicly available via www.brainspan.org) summarized to Gencode v10 (Harrow et al., 2006) gene-level reads per kilobase million mapped reads (RPKM) values were used (Table A1.1C for sample details, BrainSpan website for data collection methods). I used the RNA-seq level data instead of microarray data as it better reflects the dynamic range of transcripts across development, particularly low-expressed transcription factors. Furthermore, networks based on RNA-seq have the advantage of including relevant intermediary non-coding transcripts in the co-expression relationships and will allow for future studies to investigate non-coding regions for putative function by mapping relevant mutations to these networks.

Count-level RNA-seq data from 52,376 transcripts across 528 samples was normalized for GC content (Hansen et al., 2012) followed by batch effect (Johnson et al., 2006) and outlier removal. Only the neocortical regions were used in this analysis – dorsolateral prefrontal cortex (DFC), ventrolateral prefrontal cortex (VFC), medial prefrontal cortex (MFC), orbitofrontal cortex (OFC), primary motor cortex (M1C), primary somatosensory cortex (S1C), primary association cortex (A1C), inferior parietal cortex (IPC), superior temporal cortex (STC), inferior temporal cortex (ITC), and primary visual cortex (V1C).

Genes were defined as expressed if they were present at an RPKM of 1 in 80% of the samples from at least one neocortical region at one major temporal epoch (based on the BrainSpan periods), resulting in 22,084 coding and non-coding transcripts. Of these, 15,591 (representing 15,585 unique gene symbols) were protein coding as annotated by Gencode v10 (which corresponds to Ensembl 65), a similar number to those observed in a microarray analysis of a subset of these brain samples (Kang et al., 2011).

The samples were split into development (PCW 8 to 3 years of age) and later maturation (after 3 years of age to 40 years of age). RNA integrity number (RIN, a surrogate marker for RNA quality) was highly correlated ($r = -0.33$, $p = 1.3 \times 10^{-6}$) to age during early development, so I filtered for RIN >=9, this left 146 samples for the developmental time points and reduced the RIN effect ($r = -0.10$, $p = 0.24$). This resulted in the 146 high-quality samples ranging from PCW 8 to 1 year of age that were used to construct the developmental network. Time points prior to PCW 10 and between 1 year of age and 8 years of age were not used as the anatomy of earlier regions is less well defined for the former, and samples did not pass the RIN threshold for the latter. Finally, expression values were log-transformed ($\log_2$[RPKM+1]). The processed data used for network analysis are available with the supplemental code.

Weighted Gene Co-expression Network Analysis: All analyses were carried out in R (version 2.15.1) on a 64-bit Linux system equipped with a 2xIntel Xeon X5690 with Westmere 3.47Ghz processors and 96GB RAM. All network plots were constructed using the igraph package in R (Csárdi and Nepusz, 2006).

Briefly, correlations were estimated in a robust manner using the biweight midcorrelation (Langfelder and Horvath, 2012). Next a signed weighted correlation network was used to identify co-expression modules comprised of positively correlated genes with high topological overlap (Zhang and Horvath, 2005). Modules were defined as branches of a hierarchical cluster tree using the hybrid dynamic tree cut method (Langfelder et al., 2008). For each module, the expression patterns were summarized by the module eigengene (ME), defined as the right singular vector of the standardized expression patterns. Pairs of modules with high module eigengene correlations ($r > 0.85$) were merged. This maintains a level of decorrelation among MEs (Table A1.1B). MEs for modules are plotted in Figures 2.2 and A1.1, with trajectories visualized using the scatter.smooth function in R with a second order polynomial fit to the points (otherwise default parameters were used) after grouping by age as shown on the axes.

In more detail, the biweight midcorrelation, which is more robust to outliers compared to Pearson correlation, was implemented as defined in the default settings of the bicor function in the WGCNA package. A weighted signed network was computed based on a fit to scale-free topology, and a thresholding power of 26 was chosen (as it was the smallest threshold that resulted in a scale-free R^2 fit of 0.8), and the pair-wise topological overlap (TO) between genes was calculated (Zhang and Horvath, 2005). These transformations effectively monotonically transform pair-wise correlation values from [-1,1] to TO co-expression values from [0,1], where values close to 1 represent highly shared neighborhoods of co-expression. The TO captures relationships among neighborhoods of genes, and is therefore more robust than pairwise correlation alone for clustering genes by similarity. In fact, the TO approach has been shown to be as effective as mutual information in defining modules for large-scale gene networks (Allen et al., 2012).

This TO dendrogram was used to define modules using the cutreeHybrid function in WGCNA (Langfelder et al., 2008), with a minimum module size set to 200 genes and the deepSplit parameter set to 2. The connectivity of every gene in every module (assessed by correlation to the ME, kME) is available in Table A1.1A. I tested additional parameters and found that the modules I focus on were robustly identified under variations of these parameters.

Further network characterization – permutation, resampling, and preservation analyses: Further characterization of the co-expression network and modules was carried out by asking 1) which modules represented co-expression above chance, 2) whether modules were robustly defined in the current set of samples, 3) whether modules reproduced in independent data. I first compared the summed correlation of genes in each module with 10,000 randomly drawn gene sets representing modules of the same size, and found that every module exhibited co-expression above chance (all modules, $p < 1x10^{-4}$).

Next, I asked whether module structure was highly sensitive to removing samples involved in the initial calculation of the network. I reconstructed networks 100 times with the same parameters but by randomly resampling from the initial sample set. Modules were found to be reproducible with perturbations to the initial individual subject, regional, and temporal structure, and the fraction of times each gene was assigned to the same module is reported in Table A1.1A. To validate co-expression in independent data, I asked whether modules represented co-expressed sets of genes that could be found in other datasets. Module preservation analysis was used to calculate the $Z_{summary}$ statistic for each module. This measure combines module separability, module density, and intramodular connectivity metrics to give a composite statistic where $Z > 2$ suggests moderate preservation and $Z > 10$ suggests high preservation (Langfelder et al., 2011). The preservation analysis was performed in three epochs from an independent dataset of prefrontal cortex microarray spanning development (Colantuoni et al., 2011). M13, M16, and M17 are well-preserved at all time points, while M2 and M3 are highly preserved only at the earliest time point (Figure A1.1B). The preferential preservation of M2 and M3 during early development was also observed in the BrainSpan adult data (Table A1.1B). This analysis demonstrated that these findings are highly reproducible. Modules that were enriched for ASD risk genes are set off in bolded italics, and M2, M3, M13, M16, and M17 are all –log10(p-values) < -40 in the independent data from early development, indicating that these modules are highly reproducible in human fetal cortex.

I also assessed preservation in normal human neural progenitor (NHNP) development (Konopka et al., 2012). Plots of MEs and the average normalized expression in modules enriched for ASD risk genes show that NHNPs show a similar temporal trend as early *in vivo* development (PCW 8-20). Comparing Figure 2.2C to Figure A1.1C suggests that at least part of the transcriptional trajectory captured by these modules reflects the differentiation of neurons, and conversely, that the differentiation of neurons can model these transcriptional trajectories *in vitro*.

Gene Ontology Analysis: Genes in network modules were characterized using GO Elite (version 1.2.5 updated on 7/7/12) to control the network-wide false discovery rate using the cortex-expressed genes as background (Zambon et al., 2012). GO Elite uses a Z-score approximation of the hypergeometric distribution to assess term enrichment, and removes redundant GO terms to give a concise output. I used 10,000 permutations and required at least 10 genes to be enriched in a given pathway at a Z-score of at least 2. Gene Ontology enrichment results fulfilling these criteria are reported for all modules in Figure 2.2C, Figure A1.1D, and Table A1.1B.

Protein-Protein Interaction Analyses: Protein-protein interactions were compiled from two resources, InWeb (Rossin et al., 2011) and BioGRID (Stark, 2006). Data were downloaded for InWeb via the DAPPLE web resource at http://web.mit.edu/~erossin/Public/ on 3/13/2013. BIOGRID 3.2.98 data were downloaded on 3/21/2013 and restricted to physical interactions observed in *Homo sapiens*. I used only non-redundant interactions, and defined all interactions as undirected edges in a binary network. A union of the two networks was taken, and a degree-matched permutation analysis was applied in order to control for biological and methodological biases in PPI data. For every module, the subset of compiled PPIs between genes in that module was extracted and all edges were counted. The entire PPI dataset was split into percentiles based on the degree of connectivity of every gene to other genes, and equally sized null modules matching the degree percentiles in the observed module were generated, and their interactions were counted over 10,000 iterations. A p-value was calculated based on the rank of the observed module count among the null module counts (Table A1.1B).

I also assessed whether the enriched subsets of RDNVs are interconnected by PPIs above chance using DAPPLE (Rossin et al., 2011), which uses a within-degree within-node permutation method that allows one to rank PPI hubs by p-value. RDNV-affected genes in both M2 and M3 show increased PPI connectivity (Figure A1.2). Thus, independent data supports the coherent nature of these co-expressed RDNV affected genes, as they are highly connected at the protein-protein interaction level.

Criteria for Shared Function by Multiple Systems Biology Resources: For downstream characterization of modules, I kept modules that fulfilled two of the following three criteria: 1) significant preservation in independent developmental expression data after Bonferonni correction; 2) enrichment for protein-protein interaction after Bonferonni correction; 3) enrichment for GO terms at an FDR < 0.01. From the initial set of 17 modules, 12 passed these criteria of reproducibility and independent functional validation. These 12 modules were used for downstream analyses.

Gene set over-representation: All enrichments of gene sets were performed using a two-sided Fisher exact test with 95% confidence calculated according to the R function fisher.test. I preferred p-values from this two-sided approach to the one-sided test (which is equivalent to the hypergeometric p-value) as I do not *a priori* assume there will be enrichment (Rivals et al., 2007). To reduce the likelihood of false positives, I focus on FDR adjusted p-values (Benjamini and Hochberg, 1995). I computed the false discovery rate for all gene set enrichments relevant to the primary analysis (candidate genes, RDNV discovery set, and FMRP interactors) based on 204 tests (Table A1.3), and focused on enrichments with OR > 1 passing FDR < 0.05. For the RDNV replication set, given the smaller sample size and differing methodology of the replication study, I required an OR > 1, p < 0.05 or validation. The stricter FDR threshold may result in false negatives when the claim is made that a gene set is not enriched in a given module, e.g. when I claim ID genes are not enriched in M2, M3, M13, M16, or M17. Therefore, to reduce the risk of false negatives for such claims I require p > 0.05 for non-enrichment. I make note of the enrichment trends that do not reach significance where applicable, as these processes or pathways may be significant once additional data is available. Thus, in favoring stronger enrichment when claiming enrichment and in favoring weaker enrichment when claiming lack of enrichment, I ensure my claims are more accurate.

It is critical that the background set in an over-representation analysis reflect the claim being made. I use a cortex-expressed protein coding gene set for enrichment analyses unless otherwise specified in Table A1.2B. Protein coding is defined by the biotype annotation “protein_coding” in Gencode. Ensembl Gene IDs in Gencode v10 were overlapped according to the HUGO symbol, and all conversions among identifiers were performed using the R package biomaRt. This set of 20,007 genes was intersected with the 22,084 cortex-expressed transcripts, resulting in a "cortex-expressed background" set of 15,585 unique gene symbols. Gene membership in each set is delineated in Table A1.1A. In the case of asdM12 and asdM16, I used the set of 8,108 protein coding genes that had probes called as expressed in Voineagu et al., 2012. In the case of FMRP interactors, I use one-to-one human-mouse protein coding orthologs as the background set.

ASD and ID implicated gene sets: The ASD SFARI list was compiled using the online SFARI gene database, AutDB (https://gene.sfari.org/autdb/, accessed 8/20/2012). The database contains ASD candidates based on varying levels of evidence from the published literature (Basu et al., 2009). I restricted this list to genes with strong genetic evidence by filtering by the category S (syndromic) and evidence levels 1-4 (high confidence - minimal evidence). The minimal evidence category encompasses any gene in an ASD-associated multigenic CNV, genes near GWAS variants, convincing but not replicated association study results, and genes with multiply identified mutations that were not identified in a genome-wide statistical context. Importantly, this prioritization excludes genes with equivocal evidence (one of multiple genes found under a linkage peak, for example) and genes that functionally interact with higher-confidence genes (PPI, co-expression, or other network-based categorizations). This resulted in 155 total candidate ASD risk genes that have observed genetic evidence.

I obtained asdM12 (444 genes) and adsM16 (386 genes) from an independent gene expression study that identified reproducible gene expression changes in ASD post-mortem cortex and applied WGCNA to identify modules of dysregulated genes ASD (Voineagu et al., 2011). An important rationale behind using asdM12 is that it is possible that the SFARI ASD gene set is biased by the likely over-representation of studies investigating candidate synaptic genes in ASD, but asdM12 is agnostic to such biases.

I also considered a set of 72 CNV-affected genes that were highly interconnected by a published functional network analysis called NETBAG, which I refer to as NETBAG CNV genes (Gilman et al., 2011). Although this gene set exhibited elevated ORs (>1.5) in M2, M3, M16, and M17, I did not observe significant enrichment for the NETBAG genes in any module due to the small size of the gene set. Since the NETBAG network constructed without considering brain tissue specificity or molecular relationships from brain development among genes (shared phenotype, shared PPI, shared annotated terms from many resources), I asked if it was coalescing genes from different neurobiological pathways into one network. Post-hoc analysis testing enrichment of NETBAG genes in ASD modules (M2, M3, M13, M16, or M17 as one set) confirmed enrichment for NETBAG genes in this larger non-specific set ($p = 5.3x10^{-3}$, OR = 2.1 [1.2-3.5]), suggesting that incorporating brain gene expression added specificity above previous methods in these analysis.

The ID set was compiled based on four reviews (Inlow, 2004; Lubs et al., 2012; Ropers, 2008; van Bokhoven, 2011) of genes that have been associated with ID. After removing redundant gene symbols, this resulted in 401 genes. Notably, GO enrichment of the ID set using DAVID implicates the terms disease mutation, mental retardation, and epilepsy as the top three enriched terms, but many other terms associated with syndromic disorders and brain disorders are also enriched, suggesting this gene set agrees well with systematic annotation.

I also analyzed enrichment for overlapping ASD and ID genes (ASD/ID overlap, 38 genes) and ASD genes with ID-implicated genes removed (ASD only, 117 genes) and ID genes with ASD implicated genes removed (ID only, 363 genes). These were also used in Figure 2.6

and Figure A1.4 to ensure the overlapping genes did not confound the layer enrichment analyses. The intersection of candidate gene sets with genes expressed in cortex is available in Table A1.1A, while the full lists are in Table A1.2A.

I obtained RDNVs from four publications (Iossifov et al., 2012; Neale et al., 2012; O'Roak et al., 2012; Sanders et al., 2012). Sanders et al., O'Roak et al., and Neale et al. were used as a discovery set as they shared similar criteria for calling variants, while Iossifov et al. was used as a replication set as more stringent filters were used to avoid false positives which may have increased the false negative rate. I note that the two studies predominantly sequenced trios (Neale et al., 2012; O'Roak et al., 2012), thus due to this and the reduced hit rate of protein-disrupting and missense RDNVs in siblings, I have fewer total genes available from unaffected siblings which decreases power to assess gene set enrichment in siblings (Table A1.2B). I compiled all rare *de novo* mutation affected genes reported in these studies and categorized them as protein disrupting (there is a protein-coding change that induces a nonsense, splice-site, or frameshift mutation), missense (an amino acid change), or silent (no amino acid change). Contributions of each variant type from each study are delineated in Tables A1.2C-D. In total, these studies identify 125 protein disrupting, 559 missense, and 236 silent RDNV hit genes in 965 affected individuals and 36 protein disrupting, 307 missense, and 126 silent RDNV hit genes in 565 unaffected siblings, though the variant counts are higher as some genes are affected by recurrent RDNVs. I mapped the position of mutation for these variants to Ensembl gene models using biomaRt to ensure that all symbols complied with Gencode v10.

Robustness of RDNV enrichments: One concern when taking variants across multiple studies is the difference in exome capture, DNA sequencing, and bioinformatic analyses that could lead to results being driven by only one study (Gratten et al., 2013; Leek et al., 2010). Table A1.2D shows that all studies contribute to the observed results. The ratio of protein disrupting and missense variants to silent variants is similar across studies (2.4-4.7, combined across studies 3.0), while the ratio of protein disrupting, missense, and silent variants in probands compared to siblings is similar in the two studies with matching probands-siblings pairs. Finally, M2 and M3 were enriched for protein disrupting and missense variation in M2 and M3 in the vast majority of resampled networks (Figure A1.2A), demonstrating that enrichment for RDNVs in M2 and M3 is extremely robust.

Comparative Enrichment of Common Variants from Genome-Wide Association Studies: In order to test whether common variants differentially affect ASD implicated modules M2, M3, M13, M16, and M17, I compared the distribution of p-values from two genome-wide association (GWA) studies, one from the Autism Genome Project (AGP) and another from the Autism Genome Resource Exchange/Children's Hospital Philadelphia (AGRE/CHOP). Both have been published previously, though no single finding was replicated between the two at a genome-wide significant level (Anney et al., 2012; Wang et al., 2009). Of note, the published AGRE cohort overlaps with the AGP cohort. I used a modified set of AGRE subjects that is independent of the AGP cohort for this analysis, and obtained GWA p-values by re-running the association using permutation-based tests in PLINK (Purcell et al., 2007). It had previously been shown that common variants from the AGRE/CHOP GWA were enriched in asdM12 (Voineagu et al., 2011) using a modified Kolmogorov-Smirnov (K-S) test (Wang et al., 2010). For each gene model and the 30kB upstream from that gene, the SNP with a minimum p-value from the GWA is taken to tag the gene, and an enrichment statistic is calculated. Genes were tagged by SNPs using the Ensembl gene model and dbSNP137 SNPs and their locations in on hg19. The distribution of SNP p-values near genes is then calculated for a given pathway (or in this case, a module).

I utilized a similar approach, but applied a permutation-based procedure to assess pair-wise differential enrichment for low p-value SNPs from GWA in order to control or differences in gene size, since longer genes are more likely to contain a lower p-value SNP by the above definition of tagging. First, I compared the distribution of the SNPs in a pair of modules, and calculated the Kolmogorov-Smirnov test statistic. Next, I drew 10,000 pairs of distributions sharing the same number of tagging SNPs as the initial pair (controlling for gene size and haplotype structure), and re-computed the K-S statistic each time. Finally, I calculated a p-value based on the rank of the observed K-S statistic in this distribution. I also conducted this test with p-values from a GWAS in psoriasis (Nair et al., 2009), and found that none of the pairwise comparisons passed FDR < 0.05.

Transcription Factor Binding Site (TFBS) Enrichment: TFBS enrichment analysis was performed by scanning the canonical promoter region (1000bp upstream of the transcription start site) for the genes in each co-expression module. For each TF, I assessed the top 200 connected genes (ranked by kME) in each module using the following steps: 1) putative motifs bound by the TF were obtained from TRANSFAC (Matys, 2003). 2) upstream sequences of these 200 genes were scanned with the Clover algorithm (Frith, 2004) to calculate motif enrichment; and iii) enrichment above background was calculated using the MEME algorithm (Bailey and Elkan, 1994). I compared enrichment for each TF motif in three different background datasets to ensure robustness: 1000 bp sequences upstream of all human genes, human CpG islands, and the sequence of human chromosome 20. P values are computed by drawing 1000 sequences of the same length and testing them for enrichment using MEME, and computed the p-value based on the observed motif enrichment rank versus the randomized sets. I report TFs with $p < 0.05$ across all 3 backgrounds (Table A1.3A-B).

I also asked whether existing ChIP data for TFs from either ENCODE (The ENCODE Project Consortium, 2011) or other compiled genome-wide ChIP data (Lachmann et al., 2010) (ENCODE ChIP data website: http://genome.ucsc.edu/ENCODE/dataMatrix/encodeChipMatrixHuman.html, accessed 2/6/2013; ChEA data website: http://amp.pharm.mssm.edu/lib/chea.jsp, accessed 5/8/2013) supported computationally predicted binding sites. For TFs where a dataset was readily available, I report the fraction of sites that overlap in the existing data. Most of these TF binding sites come from non-neuronal cell lines, and many come from proliferating cells (most ENCODE lines are cancer cell lines). I therefore cautiously interpret this experimental support for computationally predicted binding sites, but find it encouraging that a moderate to large fraction of predicted binding sites have been observed in experiments, suggesting that it is at least possible for the TF to bind near the predicted target gene.

Layer-specific and Cell-type Marker Enrichment: To quantify layer-specific gene expression during development, I utilized micro-dissected human fetal neocortical laminar gene expression datasets from BrainSpan, two for each of the earlier and later fetal periods. The 15 PCW and 16 PCW data together comprises 351 samples in total, including 6 regions and 8 layers, while the two 21 PCW brains' data comprises 337 samples. Entrez gene IDs corresponding to array probes were mapped to Gencode v10 gene symbols using biomaRt. Since multiple probes can cover each gene, I picked the probe with maximum mean expression level to represent each gene. For adult layers, I used primate neocortical laminar gene expression data from macaque data comprising 10 cortical regions and 5 layers within each region (Bernard et al., 2012). I used adult cortical dissections for cell-type analyses, as I found that laminar differential expression exhibits gradients in the fetal data. In contrast, differential expression of genes in adult primate cortical data at $t > 2$ reflected well ISH patterns of laminar specificity seen in human (compare specific genes where overlaps occur with Zeng et al., 2012).

For laminar enrichment, the limmar package in R was used to calculate the t-statistics of differential expression for all genes in each layer against all the other layers. Then, for each gene set, the difference in the distribution of t-values in each layer for that set versus background was computed using a Z statistic. This quantifies the skew of differential expression t-values of each gene set in each layer. If there is no layer specificity, the distribution of t-values from a gene set is expected to follow the same distribution as the background set (with $Z = 0$), while a significant skew toward differential expression in a layer results in a positive Z score. I calculated an FDR cut-off across all enrichments in all layers ($Z = 2.7$, $FDR = 0.01$) and computed bootstrapped confidence intervals for each enrichment. To quantify cell-marker relationships, I used the same method, with the t-value replaced by the correlation of each gene to the first principal component of a set of known cell marker genes in the adult layer data (Table A1.1A lists cell-type marker genes and r values). I reported both strong, FDR-corrected enrichment, as well as nominal enrichment to emphasize trends. Statistical comparison of enrichment trends across layers between ASD and ID gene sets set was performed by 1) computing the difference in the Z score between the two sets for each layer, 2) summing this difference across all layers, and 3) comparing this to the distribution of summed differences in layers of 10,000 permuted pairs of sets matched for gene set size (see Extended Experimental Methods for details).

Network analysis resources and R code parameters for network analysis: Table A1.4 shows an example of prioritizing RDNV affected genes from a module using information from these analyses (see Discussion for details). Future work using the same methods here can increase the temporal tiling (which could result in more specific modules) and expand the pool of mutations implicated in ASD (this would improve the signal-to-noise in the enrichments). I have provided the code used in this analysis that will allow future work to easily update and incorporate this level of analysis. I provide a template for other to build upon by providing the R code and processed expression data at:
http://labs.genetics.ucla.edu/horvath/htdocs/CoexpressionNetwork/developingcortex
to allow reconstruction of networks with additional data and mapping of new genes as they are discovered.

A1.2. Extended Figures

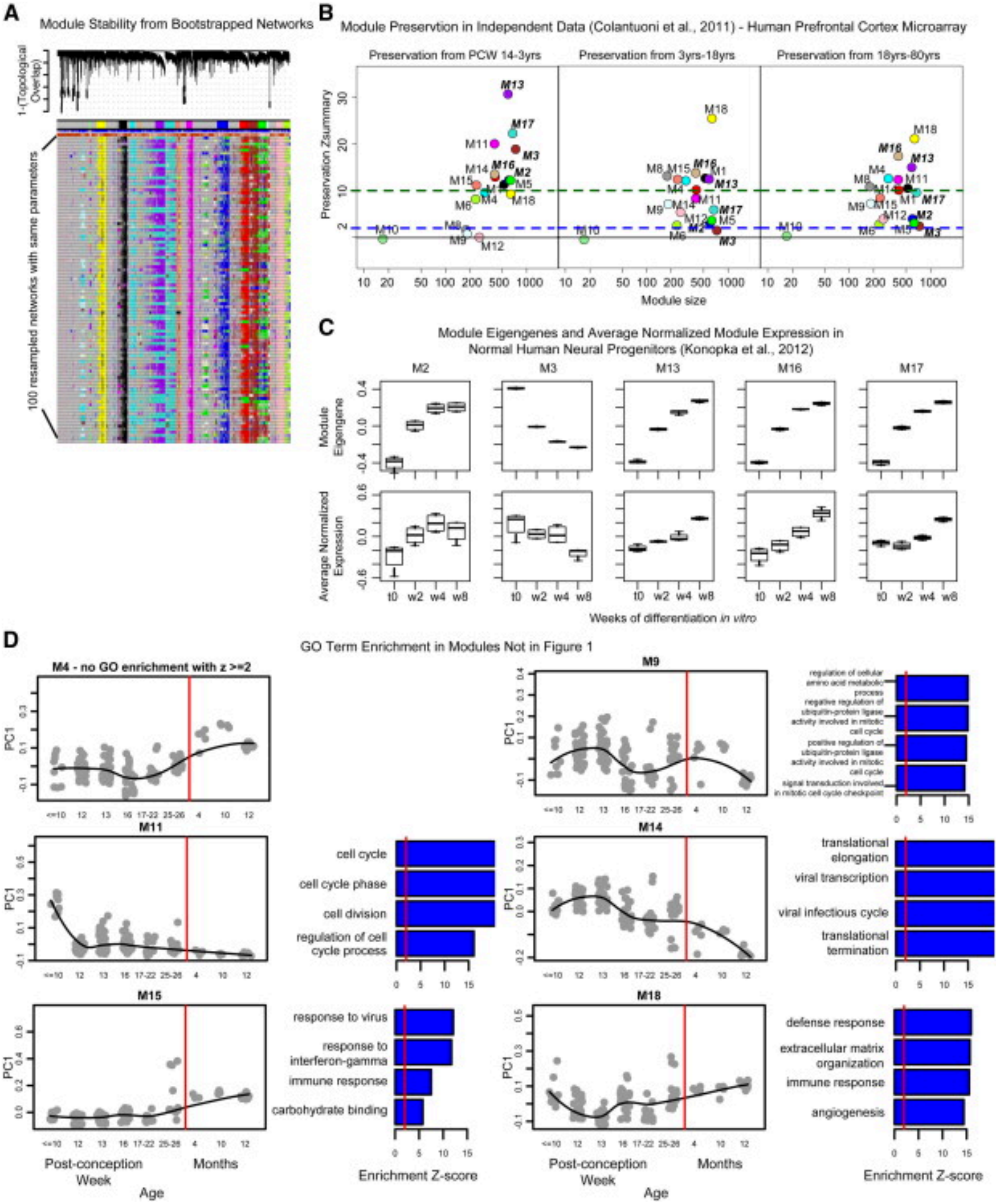


Figure A1.1, related to Figure 2.2. Further Characterization and Validation in Independent Data of Coexpression Network and Modules (A) Module robustness analysis was carried out by reconstructing networks with the same parameters and randomly resampling from the initial set of samples (as described in Langfelder and Horvath, 2012). Modules were found to be reproducible with perturbations to the initial individual subject, regional, and temporal structure, and the fraction of times each gene was assigned to the same module is reported in Table A1.1A. (B) Module preservation analysis was used to calculate the $Z_{summary}$ statistic for each module (Langfelder et al., 2011). This permutation test assesses whether module density and intramodular connectivity are preserved. An advantage of the $Z_{summary}$ statistic is that it allows one to rigorously argue that a module is not preserved ($Z_{summary}<$ 2), if it is moderately preserved ($Z_{summary} > 2$), or if it is highly preserved ($Z > 10$). I applied the module preservation analysis to assess whether modules are preserved in three epochs from an independent dataset of prefrontal cortex microarray spanning development (Colantuoni et al., 2011). At the first time window (Panel B, left), many modules were preserved, while at later time points some modules were not preserved. Of note, M13 and M16 are highly preserved at all time points, while M2 and M3 are highly preserved at the earliest time point, and moderately or weakly preserved thereafter. This non-preservation of M2 and M3 in later time points is also seen in the adult BrainSpan data (Table A1.1B). Bonferonni-corrected p-values for these Z-scores are reported in Table A1.1B. Modules that were enriched for ASD risk genes are shown in bolded italics. (C) Module preservation analysis in Table A1.1B indicates that modules are preserved during normal human neural progenitor (NHNP) development (Konopka et al., 2012a), despite differences in *in vivo* and *in vitro* neural development. Plots of expression trajectories based on the eigengene and the average normalized expression in modules enriched for ASD risk genes show that NHNPs show a similar temporal trend as early *in vivo* development (PCW 8-20). Compare modules from Figure A1.1D to Figure 2.2C. This suggests that modules, at least in part, reflect the differentiation of neuronal progenitors to neurons. (D) Module eigengenes and gene ontology for remaining 6 modules that pass 2/3 replication criteria described in A1 Extended Methods. For each plot, the eigengene trajectory and gene ontology are plotted, with top 4 GO terms passing $Z > 2$ (for FDR values see Table A1.1B). See Table A1.1B for eigengenes and additional details for GO term enrichment.

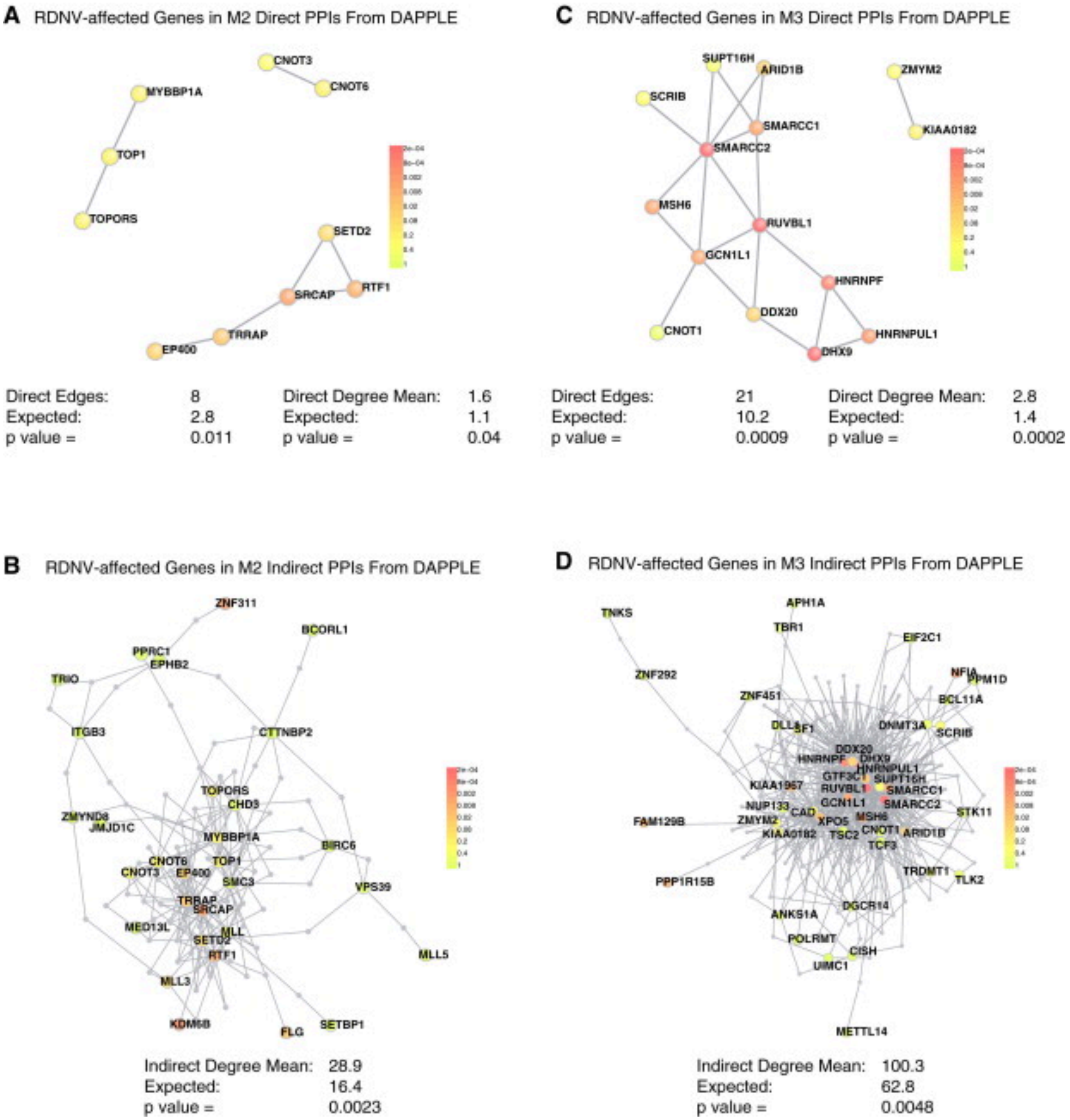


Figure A1.2, related to Figure 2.4. Direct and indirect protein-protein interaction networks for RDNV-affected genes in M2 and M3. (As obtained from InWeb PPIs via DAPPLE (Rossin et al., 2011)). (A) M2 direct interaction network obtained by inputting genes in M2 with a variant identified in the combined set of protein disrupting and missense RDNVs. (B) M2 indirect interaction network, which allows for one node to be skipped when calculating enrichment and network relationships. (C-D) Same as A-B, for M3. Expected values and p-values are calculated via the DAPPLE's within-degree within-node permutation methodology that allows ranking of PPI hubs by p-value. PPI hubs in the RDNV sub-network from M2 with p < 0.01 include: *KDM6B, SRCAP, ZNF311,* *and RTF1*. PPI hubs for the RDNV sub-network from M3 with p < 0.01 include: *DHX9, RUVBL1, SMARCC2, HNRNPF, HNRNPUL1, MSH6, SMARCC1, GCN1L1, NFIA, KIAA1967, PPP1R15B, FAM129B, XPO5*.

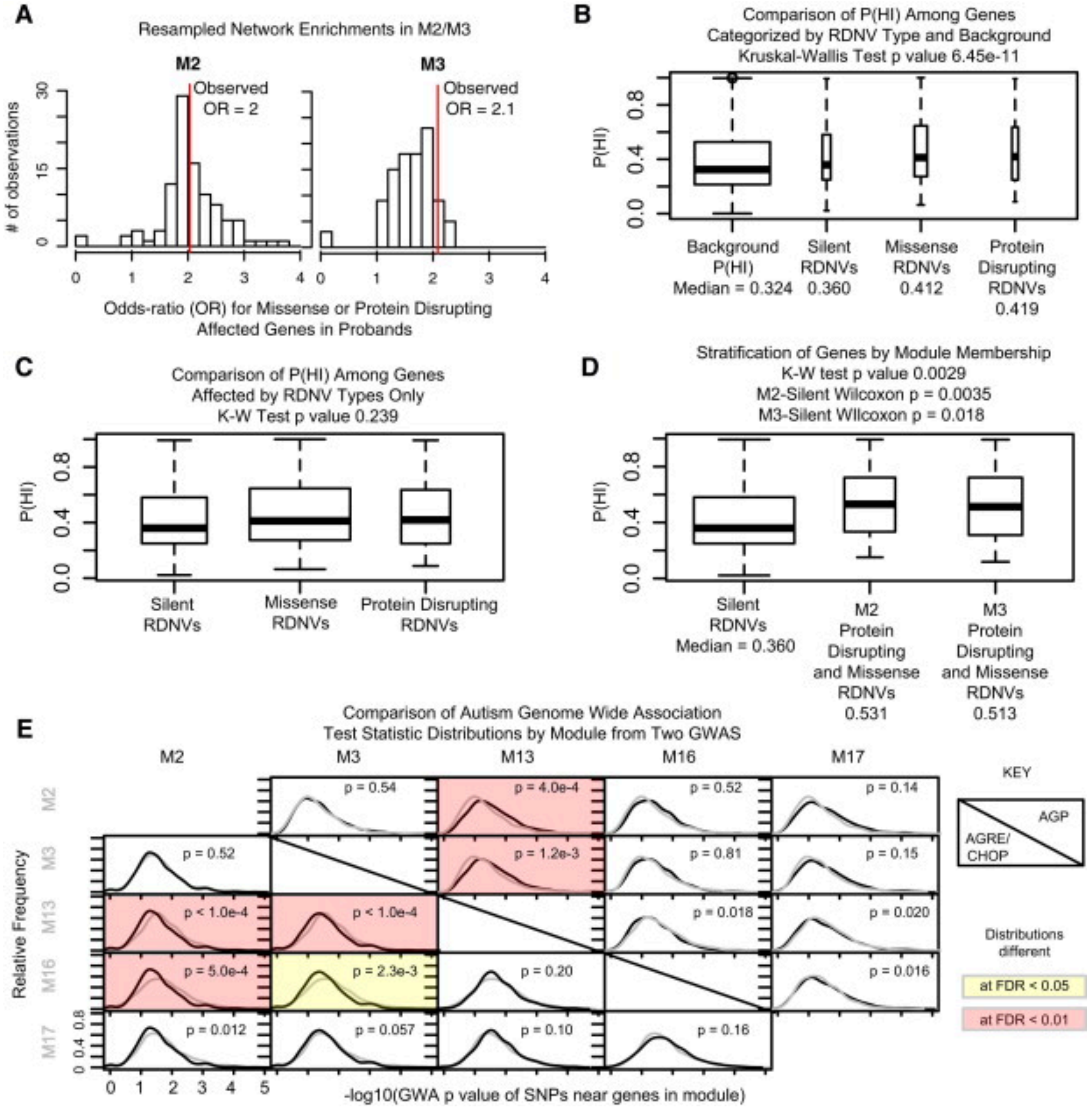


Figure A1.3, related to Figure 2.4. Further characterization of ASD risk gene enrichment in modules and P(HI) score comparisons. (A) Enrichment for Protein Disrupting and Missense RDNV affected genes from probands in M2 and M3 from the 100 resampled networks shows that the enrichment is robust to perturbations in network structure. (B-C) In B), comparison of P(HI) scores among background and three mutation-affected gene categories reveals a significant difference, but in C) a comparison of scores excluding background reveals no significant difference, suggesting that all RDNV-affected gene sets, including those affected by silent variants, are predicted to be more deleterious than background. Thus, stratification of protein disrupting and missense RDNV affected genes distinguish them from the silent set requires additional information. (D) Stratifying the RDNV affected gene sets by co-expression relationships, based on membership in M2 or M3, yields a significantly elevated P(HI) score both sets compared to the silent set. (E) Differential enrichment for common variants suggests that M13 is enriched for common variant genome-wide association (GWA) signal compared to M2 and M3 (see A1 Extended Methods for details). M16 is also preferentially affected by common variation, but only in one GWA. AGP refers to the GWA performed in by Anney et al. (Anney et al., 2012), while AGRE/CHOP refers to a GWA by Wang et al. (Wang et al., 2009a). Of note, the latter set was re-analyzed with samples overlapping the AGP cohort removed to establish independent sets.

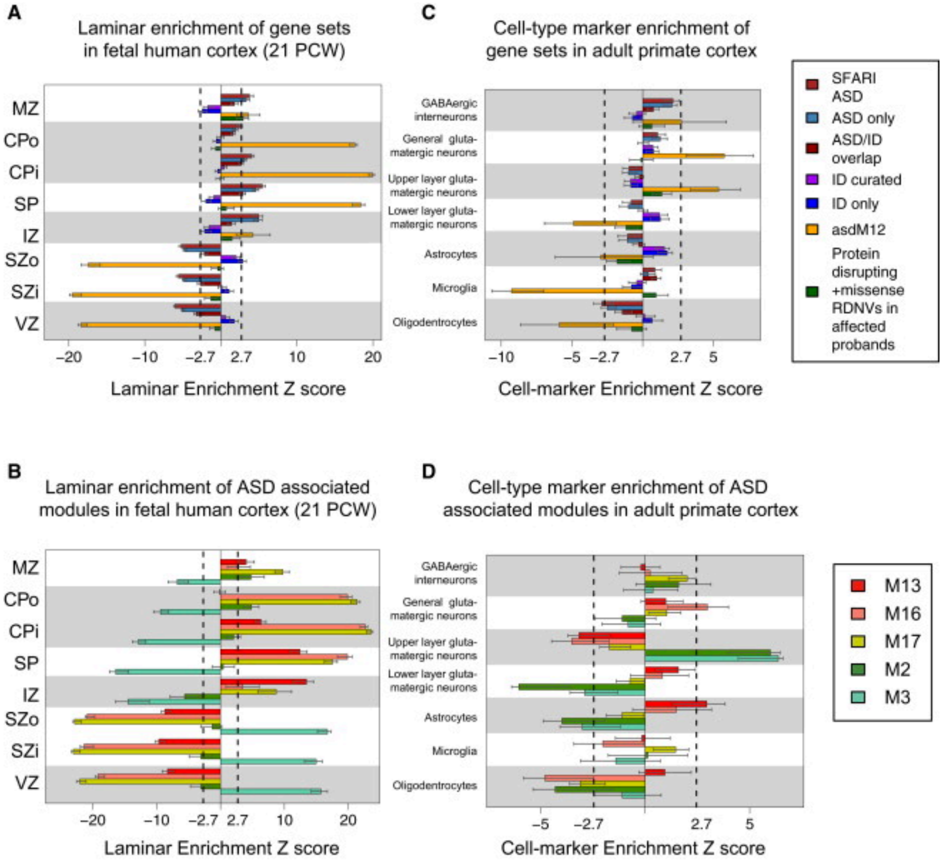


Figure A1.4, related to Figure 2.6. Laminar enrichment for PCW 21 and cell-type marker enrichment in adult primate cortex. (A-B) Laminar enrichment at PCW 21 shows an identical pattern of enrichment to PCW 15/16 for candidate genes and B) modules. (C-D) Cell-type marker enrichment in the adult primate cortex for candidate genes and D) modules. This enrichment was calculated using the distribution of Pearson correlation values to the first principal components of the set of cell markers for each cell-type as delineated in Table A1.1. Z > 2.7 is equivalent to an FDR < 0.01, and 95% CIs were derived from bootstrapping the underlying expression data 10,000 times, as in Figure 2.6. For detailed t-values for each gene as well as cell markers used for Figures A1.4C and D, see Table A1.1.

A1.3. Extended Tables

Please see the electronic tables associated with this document for extended tables (Tables A1.1-4). Descriptions for these tables follow:

Table A1.1, related to Figures 2.2. Network analysis details and analysis results for genes and modules compiled across analyses. (A) Network statistics, gene set membership, laminar specificity, and cell-type enrichment information for all 22,084 protein-coding and noncoding transcripts. For each gene, this includes (in order from left to right) gene ID and associated information from Gencode v10, module assignment (by color and label) and robustness (P(Assigned to same module in resampled network)), average expression level at multiple temporal epochs, square root of the adjusted R^2 when using neocortical region as the factor in a linear model with expression values as the outcome, correlation of the transcript to RIN, connectivity to the module eigengene (kME, correlation to the ME, a measure of centrality in the module) across all modules (all related to Figure 2.2), membership in candidate and RDNV gene sets (related to Figures 2.3 and 2.4), t-values for layer-specificity from limma differential expression analysis (Figure 2.6, Figure A1.3A-B), and r values from correlation to marker genes from cell-type marker analysis (Figure A1.3C-D). In the Module Label column, genes assigned to M7, the grey module, are marked with a "-" as the grey module represents the set of genes not strongly co-expressed with other genes. (B) Module-level statistics, including GO term enrichment, enrichment of PPIs, module preservation, and module eigengenes across all samples. This includes (in order from left to right), a summary of the independent levels of validation supporting each module, the proportion of variance explained by the ME, the top 5 GO terms from GO Elite, PPI enrichment statistics (p-value, observed interactions, and percentiles of randomized distributions), module preservation in 4 independent datasets (including the $Z_{summary}$ and the associated Bonferonni-corrected p-values), pairwise correlation between MEs, and the ME value across all samples for each module. For details, see A1 Extended Methods.

C) List of BrainSpan neocortical samples used to construct developmental networks.

Table A1.2, related to Figure 2.3, 2.4, and 2.5. Gene sets and enrichment analysis for curated lists and RDNV lists. (A) Gene lists used for enrichment analysis in this study. (B) Network-wide enrichment for candidate gene sets, RDNV-associated gene sets, and FMRP interactors. Contains number of genes in each module overlapping with each set. The background set was all 15,585 cortex-expressed protein coding genes, except in the case of asdM12/asdM16 where I restricted the background to genes Illumina probes used in Voineagu et al., 2012, and in the case of FMRP interactors where I used the background set of all protein coding one-one human-mouse orthologs. Enrichments with OR > 1, FDR < 0.05 are bolded and italicized while enrichments with OR > 1, p < 0.05 are bolded. FDR values are reported as corrected across enrichments performed for the candidate ASD and ID gene sets, the RDNV discovery set, and the FMRP target set. (C) Compilation of RDNV data from four studies – Iossofiv et al. 2012, Sanders et al. 2012, O'Roak et al. 2012, and Neale at al. 2012. (D) Assessment of the contribution of RDNVs from four studies to the enrichments in the network analysis.

Table A1.3, related to Figure 2.5. TF binding site analysis results for enrichments connecting two or more modules. A) Summary of enrichment by module. A "Y" is marked for enrichment if the motif for the TF from TRANSFAC is enriched ($p < 0.05$) above the three background sets. If the TF is associated with neuronal function or neuronal development, a reference is provided. If ChIP data exists for the TF, I report the fraction of overlapping sites. B) Individual TF-motif enrichments driving module-level enrichment above background.

Table A1.4, related to Figure 2.8. Example prioritization of M2 and M3 RDNV affected genes using information from Table A1.1. As described in section 2.5 Discussion.

A2. Additional Methods and Figures for Chapter 3

A2.1. Extended Methods

Gene sets used for enrichment: The sources for all gene sets used in this chapter are described in the main text. For laminar and cell-type specific gene sets, transcriptomes from each lamina or cell were compared against all other lamina or cells from that study, including global whole-tissue in the comparison for the cellular transcriptome data. Comparisons were performed with a two-tailed t-test and all genes with Benjamini-Hochberg FDR adjusted $p < 0.05$ in a given lamina or cell type were used.

Logistic regression for enrichment analysis: Methodologically, most gene set enrichment analysis studies utilize the hypergeometric test to evaluate whether a gene set is enriched over background, providing a p value and enrichment value for gene set enrichment. This is equivalent to a one-sided Fisher's exact test, which is arguably preferable as it can offer an upper and lower confidence interval to the odds-ratio for enrichment, and does not assume *a priori* that the gene set of interest is enriched, and allows for it to be depleted (Rivals et al., 2007). An assumption with either of these tests is that the background set reflects the gene set considered for enrichment for all factors other than pathway membership.

In the specific case of identifying genes affected by *de novo* variants in genes by WES, gene length alone is highly correlated to the mutation rate on that gene (Samocha et al., 2014). The goal of my investigations is to identify whether specific gene sets are predictive of mutations in ASD, and in the initial study in Chapter 2, I utilized enrichment for synonymous mutations as a control set to ensure that factors such as gene length, GC content, or other unmeasured variables affecting detection of mutated genes did not drive enrichment. I had also checked results by stratified permutation analysis using gene-wise mutation rates (Michaelson et al., 2012) to ensure mutation rate alone was not driving observed enrichments. However, upon assessing the same sets with new data, there was considerable enrichment in synonymous gene sets.

To correct for this gene length bias, Iossifov et al., 2014 utilized a stratified permutation analysis. This is equivalent to using gene length as a covariate, and similar to what I had done for PPI enrichment in Chapter 2, where it was necessary to control the global PPI degree for genes in order to identify true PPI enrichment in modules. I therefore adopted a modified enrichment analysis using logistic regression. Logistic regression involves specifying predictors and a binary outcome. Beta values can be interpreted as odds-ratios, and odds-ratios and p values computed by logistic regression for binary gene set membership as a predictor and binary gene set membership as an outcome, with the intersected set of background genes as the full data results in highly equivalent enrichment results. The major benefit of this approach is that I can add gene length as a predictor, and remove the effect of gene length on the enrichment result.

A3. Additional Methods and Figures for Chapter 4

A3.1. Extended Methods

Sample description: Tissue samples for this study were acquired from the Autism Tissue Program (ATP) brain bank at the Harvard Brain and Tissue Bank and the National Institute for Child Health and Human Development (NICHD) Eunice Kennedy Shriver Brain and Tissue Bank for Developmental Disorders. Sample acquisition and Material Transfer Agreement protocols were followed between UCLA and both brain banks, and subjects were de-identified prior to acquisition.

Up to three brain regions from each individual were assessed in this study: dorsolateral or medial prefrontal cortex (frontal cortex, FC, from BA9), superior temporal gyrus (temporal cortex, TC, from BA41, BA42, or BA22 unless otherwise noted), and cerebellar vermis (CB). Dissections were batched for maximal balance of age, sex, brain region, and diagnostic status. Brain samples were dissected on dry ice in a dehydrated dissection chamber to reduce degradation effects from sample thawing or humidity. Approximately 100mg of tissue across the cortical region of interest was isolated from each sample for up to two RNA extractions using the miRNeasy kit (Qiagen).

Library preparation and RNA-seq: For each RNA sample, RNA quality was quantified using the RNA Integrity Number (RIN) (Schroeder et al., 2006). I compared both poly(A) selection (referred to as polyA+) and depletion of cytoplasmic and mitochondrial rRNAs for RNA-seq using existing data and a pilot experiment, and found that the quality of polyA+ RNA-seq drops off with RIN < 9, and considerably after RIN < 8, largely due to transcript degradation, resulting in a strong 3′ bias. Given that RIN values were rarely above 8 in the samples for this study, I opted for rRNA depletion with the RiboZero Gold kit. This library preparation approach captures a large fraction of transcripts which are not polyadenylated (Cheng, 2005; 2012), and about 40% of reads are expected to align to intronic regions (Ameur et al., 2011). Due to the start date of library preparation and the need to keep a consistent protocol, I did not use strand-specific library preparation protocols. I took several precautions to ensure intronic and antisense transcription did not confound findings, as discussed below. Additionally, I assessed 50bp and 100bp single-end and paired-end (PE) RNA-seq through artificially trimming the same 2x100bp (PE) data. I found highly similar mapping and transcriptome quantification results using 2x50bp or 2x100bp RNA-seq, and opted to use 2x50bp RNA-seq to maximize the number of fragments given a constant sequencing depth. Taken together, these assessments led to the use of libraries prepared with RiboZero Gold rRNA depletion for 50bp PE RNA-seq and aim for an average sequencing depth of 50 million base pairs per sample.

Specifically, ribosomal RNA was depleted from 2 μg total RNA with the Ribo-Zero Gold kit (Epicentre). Remaining RNA was then size selected with AMPure XP beads (Beckman Coulter) and resuspended in 8.5 μL of Illumina resuspension buffer and an additional 8.5 μL of 2x EPF buffer. Subsequent steps followed the Illumina TruSeq protocol (starting at page 84 of the sample prep v2 guide, no changes). After this protocol was followed, libraries were quantified with the Quant-iT PicoGreen assay (Life Technologies) and validated on an Agilent 2200 TapeStation system. Libraries were pooled to multiplex 24 samples per lane using Illumina TruSeq barcodes, and each pool was sequenced six times on a HiSeq2000/2500 instrument using high output mode with standard chemistry and protocols for 50bp paired end reads. Most libraries were sequenced across two sequencing cores at UCLA. After confirming that read quality was similar across both core facilities and lanes, reads from six lanes for each sample were pooled after demultiplexing (with Casava v1.8) for downstream analysis.

RNA-seq read alignment: The paired-end raw reads were mapped to the human reference genome assembly GRCh37.73 (Harrow et al., 2006) using Tophat2 (Trapnell et al., 2012a) as follows:

```
tophat -o outputfolder -g 10 -p 8 -r 99 --no-novel-juncs -G
GRCh37.73.gtf GRCh37bowtieindex pairedread1.fastq pairedread2.fastq
```

Aligned reads were sorted and alignments mapped to different chromosomes were removed from the BAM file using samtools (Li, 2011).

Genotyping-based quality control: Genotypes were called from RNA-seq data using a modification of an existing pipeline shown to detect SNPs optimally from RNA-seq data (Quinn et al., 2013). Genotypes reflecting sites that are heterozygous or homozygous for the minor allele relative to the reference genome were called using by removing duplications (*rmdup*) from .bam files and constructing .vcf files from piled up (Li, 2011) reads:

```
samtools mpileup -I -S -gu -f GRCh37reference sorted_reads_rmdup.bam |
bcftools view -bvcg - > var.raw.bcf
bcftools view var.raw.bcf | vcfutils.pl varFilter -D100 > var.flt.vcf
```

Genotypes were then coded as NA (homozygous for the major allele or not enough depth to detect), 1 (detected heterozygous), or 2 (homozygous for the minor allele). Pairwise spearman correlations were assessed between samples, and any sample from an individual that did not match the genotype of another sample from the same individual was assessed for contamination or sample mix-up.

RNA-seq quality control: I performed quality control (QC) analysis after read alignment using PicardTools v1.100 (commands *ReorderSam*, *CollectAlignmnetSummaryMetrics*, *CollectRnaSeqMetrics*, *CollectGcBiasMetrics*) and samtools (duplication metrics other statistics from the *flagstat* command). Sequencing metrics were used to remove samples with poor sequence quality based on the following sequencing metrics: %Total Reads, %High-quality Aligned Reads, %mRNA Bases, %Intergenic Bases, Median 5 to 3' Bias, GC Dropout, and AT dropout. To detect outliers, a quality z-score was calculated for each metric, and samples with low quality (Z > 2 for %Intergenic Bases, GC Dropout, or AT Dropout and Z < -2 for %Total Reads, %High-quality Aligned Reads, %mRNA Bases, or Median 5 to 3' Bias) in this matrix were identified as outlier values, and any sample with greater than one outlier value was removed due to sequencing quality concerns. This QC was performed with 263 initial samples: 28 samples were not for this study, and QC removed 30 samples (13%). Of these one-third had very low RIN (< 4), one-fourth had a high 5' to 3' bias (>0.7), and one-third had a high proportion of reads aligned to intergenic regions (intergenic read proportion > 20%).10%). The remaining 205 samples from 79 individuals along with assessed QC metrics are shown in Table A3.1.

Brain sample metadata: Brain samples were obtained from 33 neurotypical controls and 46 ASD individuals (38 idiopathic with no identified cause of autism and 8 with confirmed duplications in the 15q region). Individuals defined as autistic for this study had either a confirmed ADI-R diagnosis (30/46), duplication 15q syndrome with confirmed ASD (8/46), or a diagnosis of autism supported by other factors such as clinical history (8/46).

Available metadata from brain banks included age, sex, medical history, and sample quality information. Variable levels of detail were available regarding medical case history for the individual and previous sample quality information. Medical history information was used to identify history of psychiatric medications, history of seizures, co-morbidities, post-mortem interval, neuropsychiatric test results, and cause of death where possible. Notably, medication status, seizures, ADI-R, and IQ test results were available only for individuals with ASD, with 22/45, 23/45, 25/45, and 6/45 ASD individuals having available measurements for these, respectively. History of medication and seizures were categorized as having a reported history as supported by medical information (Yes) or not having mention of such medication or co-morbidities (No) despite other medical records being available (one individual did not have enough information to evaluate these criteria).

Additional post-mortem tissue metrics included previously recorded RNA integrity number (RIN), pH, and brain weight, but these were available only for the NICHD brain bank. Notably, where I could assess it, 86% of cortical and 93% of cerebellar RNA extractions were within 2 RIN values of previously documented RIN values, and 46% of cortical and 52% of cerebellar RNA extractions had better RIN values than what was previously documented. A comparison of pH between 12 case and 9 control samples from NICHD revealed weak correlation between diagnosis and pH (r^2=0.12, p = 0.12), and there was no correlation between diagnosis and brain weight (r^2=0.01, p=0.4). Together, these data demonstrate that 1) elapsed time at the brain bank had minimal effect of RNA quality despite some brains being stored for many years, 2) the pH in ASD and control brains is similar, and 3) post-mortem brain size is not different between ASD and controls.

Cause of death was categorized based on agonal state as previously described in a study assessing relative effects of multiple factors on post-mortem gene expression (Monoranu et al., 2009). The main difference between cases and controls in agonal state was that 6 individuals with dup15q died of sudden death from epilepsy (SUDEP), while no controls did. Otherwise, the proportion of individuals with agonal state classifications was similar between groups. Complete and quantitative data types were used as described below in covariate analyses while incomplete variables were assessed to evaluate their effect on observed relationships where appropriate. All phenotypic information used in the analysis is provided in Table A3.1.

Quantification of gene expression: Gene expression levels were quantified for samples passing QC using multiple methods:

HTSeq (v.0.6.1) with a union exon model:

```
python -m HTSeq.scripts.count --stranded=no --mode=union --type=exon --
quiet inputfile.sam --GTF GRCh37.73.gtf >> outputfolder
```

HTSeq with a whole gene (union exon + introns) model:

```
python -m HTSeq.scripts.count --stranded=no --mode=union --type=gene --
quiet inputfile.sam --GTF GRCh37.73.gtf >> outputfolder
```

Cufflinks v2.1.1:

```
cufflinks -o outputfolder --num-threads 8 --GTF GRCh37.73.gtf --frag-
bias-correct GRCh37reference --multi-read-correct --compatible-hits-
norm inputfile.bam
```

Each approach quantifies gene-level expression in a slightly different manner, but the overall expression values are highly correlated, as are the principal components in the data (**Figure S1**). Notably, HTSeq counts simply counts paired read fragments on the given gene models and does not use reads with multiple alignments or that overlap gene models, while Cufflinks uses these multi-mapped reads, quantifies transcripts first, and then provides a gene level estimate. For primary analyses, I use the HTSeq union exon quantifications, and apply the other methods for ensuring data quality. Genes were kept if they pass the following criteria within all cortical and cerebellar samples separately:

- Expressed in 80% of samples with HTSeq union exon quantification of 10 counts or more (to remove genes supported by only a few reads)
- Expressed in 80% of samples with HTSeq whole gene quantification 10 counts or more (removes genes supported solely by intronic reads of other genes)
- Expressed in 80% of samples with Cufflinks (lower bound FPKM estimate > 0)

The resulting read counts, reflecting fragments mapped to each union exon model, were converted to log2 transformed and GC content, gene length, and library size normalized fragments per kilobase million mapped reads (FPKM) values using the cqn package in R (default options, but setting cqn=FALSE which turns of quantile normalization) (Hansen et al., 2012). Where used, these values are referred to as log2(Normalized FPKM).

Exploratory data analysis and adjustment of covariates: Normalized FPKM data were assessed for effects from biological covariates (condition, age, sex, brain region), technical variables related to sample processing (RIN, Brain Bank, Sequencing Batch), and technical variables related to sequencing quality metrics (sequencing metrics plus the proportion of exonic reads, as defined by total quantified reads in the union exon model divided by total quantified reads in the whole gene model for each sample).

For biological variation, two major observations were made. There was an age imbalance as there were few controls younger than 10 years old, and several ASD samples younger than 10 years. Additionally, dup15q samples were more likely to be outliers. Based on these observations, I opted to analyze all idiopathic aged 10 or older in the primary analysis, and use the younger ASD samples for independent validation and extension of findings. Dup15q samples were analyzed separately against the same controls used in the idiopathic set.

For technical variation, there was no strong relationship between any factor and diagnosis, largely due to the fact that I randomized all samples at multiple points (dissection, RNA extraction, and multiplexing) over all biological covariates. However, there was a strong relationship between the first principal component of gene expression and RNA and sequencing quality metrics, including the RIN, sequencing 5'-3' bias, and the proportion of exonic reads. Given the large number of sequencing quality features, I performed PCA on these data and found that the first two PCs explain nearly 99% of the variance. Consequently, I opted to use these 2 sequencing surrogate variables (seqSVs) as covariates (Figure A3.1). The use of biological and technical covariates is discussed below as appropriate for differential gene expression (DGE), differential splicing (DS), and weighted gene co-expression network analysis (WGCNA).

Differential gene expression analysis: Differential gene expression (DGE) analysis: Differential expression analysis was performed with the normalized gene expression levels (not normalized for technical variation). Cortical samples (ba9 and ba41-42-22) were analyzed separately from cerebellar vermis samples. A linear mixed effects model framework was used to assess differential expression in log2(Normalized FPKM) values for each gene for cortical regions (as multiple brain regions were available from the same individuals) and a linear model was used for cerebellum (where one brain region was available in each individual, with a handful of technical replicates removed). Individual brain ID was treated as a random effect, while age, sex, brain region (except in the case of cerebellum, where there is only one region), and diagnoses were treated as fixed effects. I also coded the technical variables discussed above (RIN, proportion of reads mapping to exons, sequencing batch, and brain bank batch) as fixed effects into this model. Effect sizes and p-values for diagnosis were extracted from the model across all genes for both HT-seq Counts and Cufflinks FPKMs, and q-values were computed to assess the false discovery rate.

Quantification of transcript splicing and analysis: I utilized multiple approaches for quantifying transcript splicing. The main analyses are performed using Multivariate Analysis of Transcript Splicing (MATS, v3.08), utilizing the PSI values in the linear mixed regression framework described above for DGE (Shen et al., 2012). The results from ssplicing analyses are sensitive to the use of different splice junction databases and different aligners, so I also computed PSI values with OLego and Quantas (Wu et al., 2013). Using PSI values from these two different methods, the linear mixed regression framework gives similar results in DS analysis (Figure A3.3A). MATS identifies five type of splicing events: spliced exons (SE), alternative 5' splice sites (A5SS), alternative 3' splice sites (A3SS), mutually exclusive exons (MXE), and retained introns (RI). The last category was excluded from these analyses since many false events are called due to retention of pre-mRNA by the ribosomal RNA depletion library preparation used in this study.

For each event, MATS reports counts supporting the inclusion (I) or exclusion (E) of a splicing event. To reduce spurious events due to low counts, I set a filter requiring at least 80%

of samples to have I + S >= 10. For these events, the percent spliced in, PSI = I / (I + S) was calculated. PSIs for events were used in all statistical analyses, for example in the linear mixed effects modeling. This approach is advantageous over existing methods as it allows me to model covariates and take into consideration the variance across samples when assessing event significance with ASD.

Genotyping and CNV calling for dup15q samples: Previously the type of duplication and the copy number in the breakpoint 2-3 region were available for these brains (Scoles et al., 2011). To expand this to the regions between each of the recurrent breakpoint in these samples, 7/8 dup15q brains were genotyped. The number of copies between each of the breakpoints is reported in Table A3.3, with discrepancies with previous studies noted.

Weighted gene co-expression network analysis: The R package WGCNA was used to construct co-expression networks using the technical variation normalized data (Langfelder and Horvath, 2008; Zhang and Horvath, 2005). I used the biweight midcorrelation to assess correlations between log2(Normalized FPKM) values for both Cufflinks and HT-seq Counts data (Langfelder and Horvath, 2012). Parameters for network analysis were the same as used previously (section A1.1). I utilized a modified version of WGCNA which involves bootstrapping the underlying dataset 100 times and constructing 100 networks. The consensus of these networks (50th percentile across all edges) was then used as the final network (Langfelder and Horvath, 2007). The first principal component of each module (eigengene) was related to ASD diagnosis, age, sex, and brain region in a linear mixed effects framework as above, only replacing the expression values of each gene with the eigengene.

Enrichment analyses: All enrichment analyses were performed either with Fisher's exact test (see A1.1 Extended Methods) or logistic regression (see A2.1 Extended Methods). GO term enrichment analysis was performed using GO Elite (Zambon et al., 2012) as in Chapter 2. I focused on molecular function and biological process terms for display, but discuss cellular compartment terms where relevant. Cell type specificity analysis was performed using the gene sets described in Chapter 3 (see section A2.1).

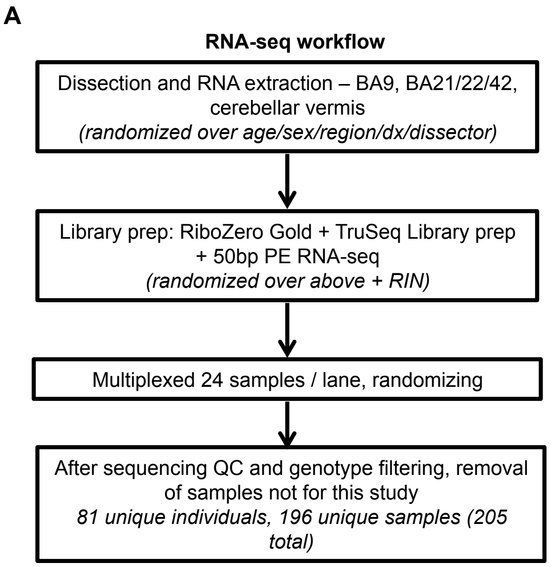
Transcription Factor Binding Site Enrichment:

Transcriptional factor and chromatin regulator binding sites were obtained from a published study (Arbiza et al., 2013), and intersected with brain-specific data using bedtools (Quinlan, 2002) and discussed in the main text.

GWAS enrichment: PGC cross-disorder data

GWAS data were obtained from the Psychiatric Genetics Consortium (PGC, http://www.med.unc.edu/pgc/downloads) and correspond to the 2013 release of the cross-disorder data. GWAS SNPs were associated to genes using the methodology described in section A1.1.

A3.2. Extended Figures


A
RNA-seq workflow
Dissection and RNA extraction – BA9, BA21/22/42, cerebellar vermis
(randomized over age/sex/region/dx/dissector)
Library prep: RiboZero Gold + TruSeq Library prep + 50bp PE RNA-seq
(randomized over above + RIN)
Multiplexed 24 samples / lane, randomizing
After sequencing QC and genotype filtering, removal of samples not for this study
81 unique individuals, 196 unique samples (205 total)

B **Mapping Statistics (this study, RIN = 7.6 [3.0-8.6])**

PicardTools Metric	Median [2.5%-97.5%]
Reads Aligned	43 million [16-76]
% of mRNA bases	53% [34-71]
% of intronic bases	40% [23-58]
% intergenic bases	6.5% [4.9-16]
Median 5'-3' bias	0.60 [0.52-0.66]

C **Mapping Statistics (polyA+ on other ASD and CTL samples, RIN = 4.8 [2.1-6.9])**

PicardTools Metric	Median [2.5%-97.5%]
Reads Aligned	14 million [2.2-66]
% of mRNA bases	75% [48-86]
% of intronic bases	6% [2.7-18]
% intergenic bases	18% [10-40]
Median 5'-3' bias	0.16 [0.0027-1.0]

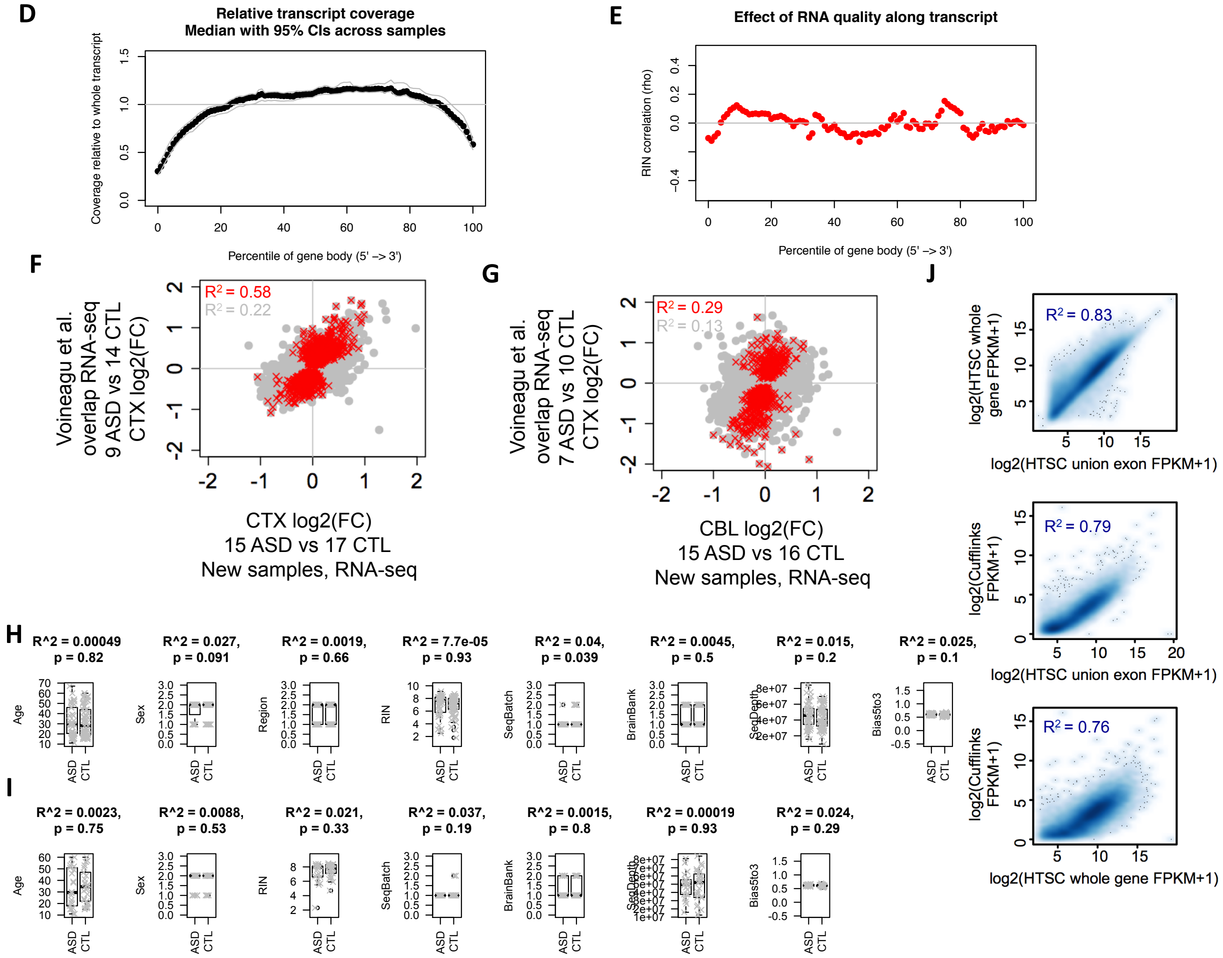


Figure A.3.1 RNA-seq methodology, sequencing quality metrics, reproducibility analyses, covariate effects, and alternate methods of gene quantification. A) RNA-seq workflow. B) Mapping statistics from this study, , with mean and [95% confidence interval]. C) Mapping statistics from another study(Gupta et al., 2014) utilizing polyA+ RNA-seq on similar samples in ASD. D) RNA-seq read coverage relative to normalized gene length across transcripts from the 5' to the 3' end. E) Dependence between coverage and RNA integrity across the normalized gene models from D), where a high magnitude r value on the y-axis would suggest strong dependence on RNA quality. F) Correlation of ASD vs CTL effect sizes between previously evaluated and new ASD samples in cortex by RNA-seq, with red highlighting genes that were at $p < 0.05$ the old samples. G) Correlation between effect sizes as in F), but for cerebellar samples. H-I) Correlation between covariates and ASD vs CTL status in cortex and cerebellum, respectively. J) Correlation between gene-level quantifications in cortex when utilizing different methods and models for gene expression quantification, results are similar in cerebellar samples.

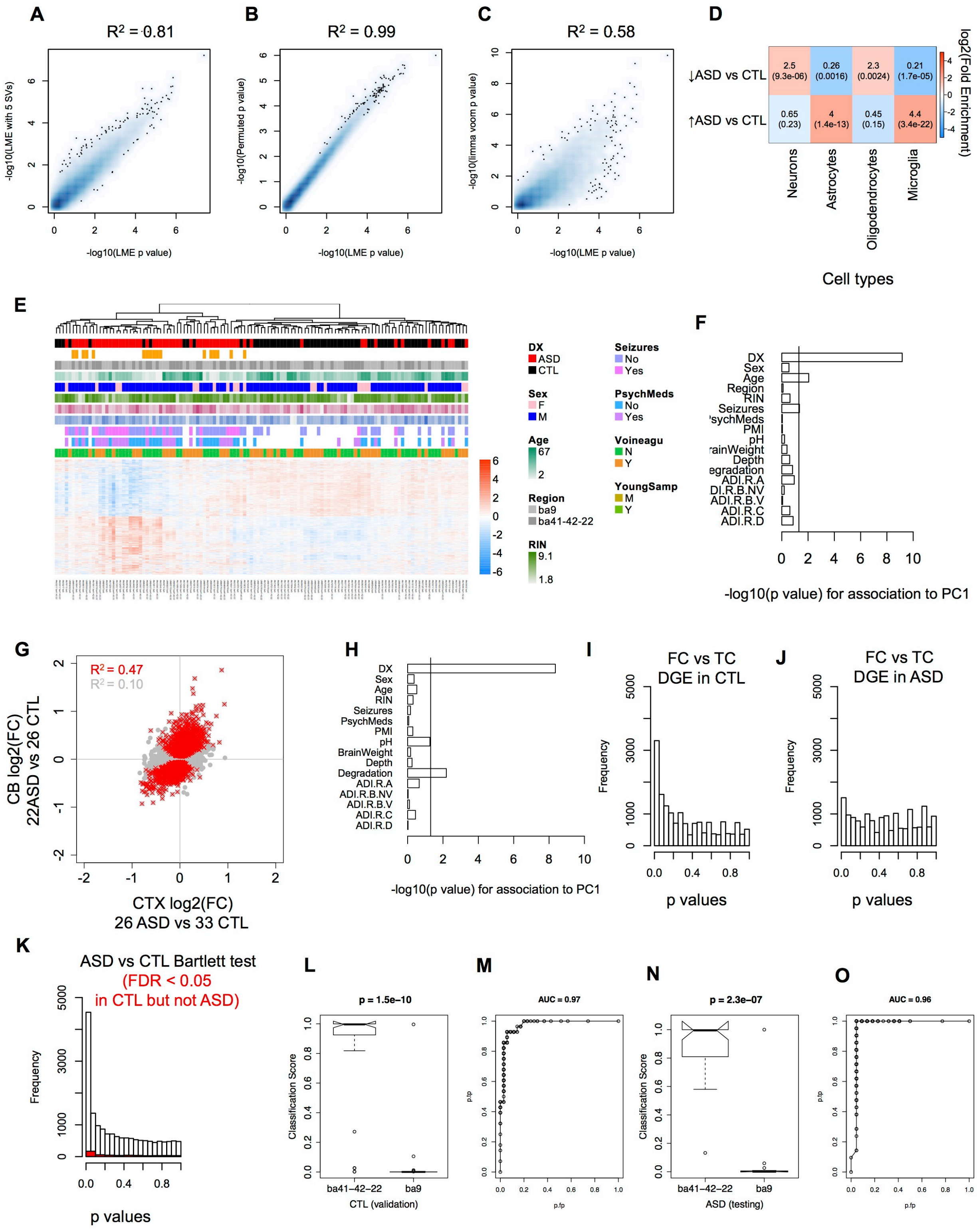


Figure A.3.2 Additional analyses for differential gene expression and cortical patterning analyses. A-C) Agreement of main DGE analysis p values with DGE analyses using additional sequencing quality associated

surrogate variables (SVs), permutation analysis, and limma voom, respectively. R^2 values are from Spearman's rank correlations. D) Cell-type enrichment analysis of four major cell types in genes significantly increased and decreased in ASD. E) Heatmap of DGE clustering using all significant genes over all cortical samples from ASD (N = 123). F) Association of the first PC from the DGE set with all measured covariates. G) Comparison of changes in CB compared to CTX, highlighting those with $p < 0.05$ in CTX. H) Evaluation of ASD signature in CB using the first PC of genes at $p < 0.01$. I-J) Histograms of p values from paired Wilcoxon rank-sum test DGE between FC and TC in CTL and ASD, respectively. K) Histogram of Bartlett's test p values for differences in gene expression variance between ASD and CTL, with difference in variance for genes with attenuated cortical patterning highlighted. L-M) Results of learned cortical region classifications using LASSO regression from BrainSpan on CTL samples from this study, where 1 = TC, 0 = FC. N-O) Results of learned cortical region classifications in ASD. In receiver-operator characteristic plots, the area under the curve (AUC) is given and the x-axis is the rate of true positives, while the y-axis is the rate of false positives.

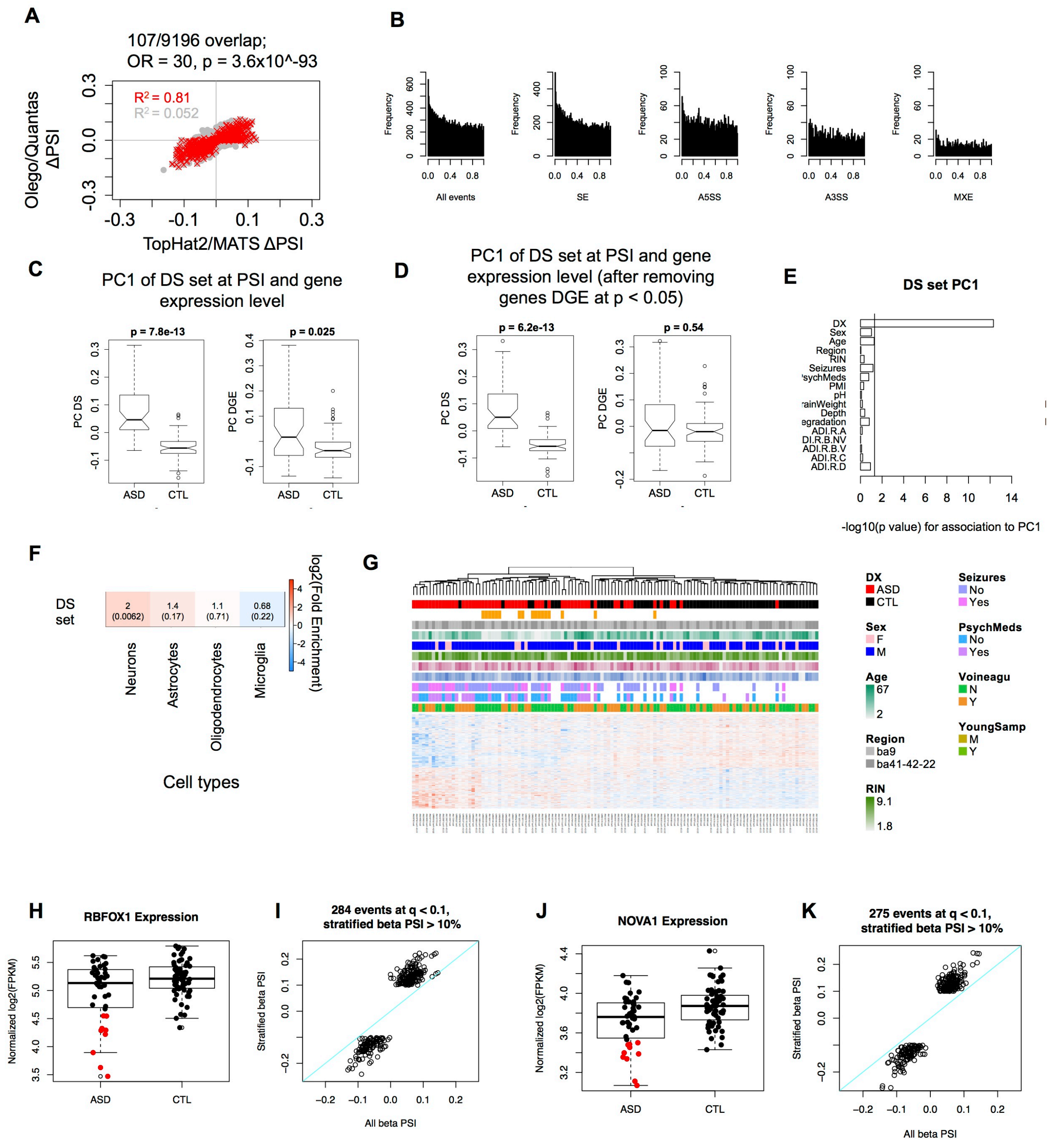


Figure A.3.3 Additional analyses for differential splicing and splicing factor analysis. A) Comparison of the CTX splicing analyses in when using PSI values obtained via read alignment by TopHat2 (Trapnell et al., 2012b) followed by the MATS (Shen et al., 2012) pipeline (used throughout this study) against read alignment by OLego followed by Quantas (Wu et al., 2013). B) Distribution of p values for changes in the PSI between ASD and CTL for all events (left) and event subtypes (SE, spiced exon; A5SS, alternative 5’ splice site; A3SS, alternative 3’ splice site; MXE, mutually exclusive exons). C) Difference between ASD and CTL in the DS set based on PC1 of the DS

set at the PSI level, and PC1 of the gene expression levels of genes in the DS set. D) Similar to C), but with nominally DGE genes ($p < 0.05$) removed. E) Evaluation of association between PC1 of the DS set and all measured covariates. F) Enrichment for cell-type specific genes in the genes harboring DS events at $p < 0.01$. G) Hierarchical clustering and heatmap of all samples ("Matched ASD" + "Young ASD") with the DS set. H) The 10 samples with the lowest expression of *RBFOX1* in ASD are compared against all CTL samples and run in the same DS analysis used for all samples, with effect sizes in this stratified set compared with the effect sizes across all samples in I). J-K) Similar to H-I) but for *NOVA1*.

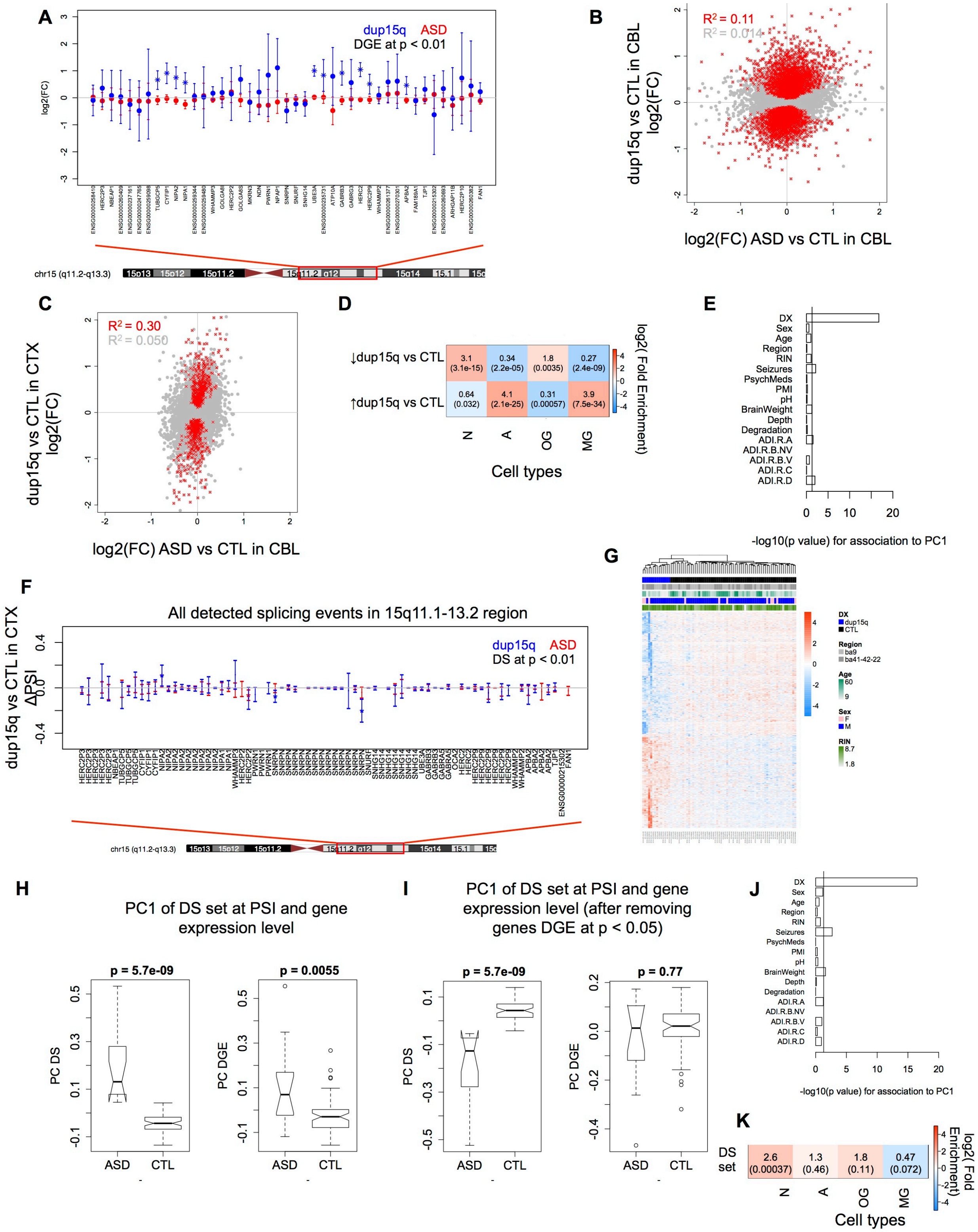


Figure A.3.4 Additional analyses for differential gene expression and splicing analysis in dup15q. A) DGE across the 15q region of interest in dup15q vs CTL and ASD vs CTL CB. B) Comparisons of effect sizes in dup15q

vs CTL and ASD vs CTL in CB. C) Comparisons of effect sizes in dup15q vs CTL CB and dup15q vs CTL CTX. A-C) Should be interpreted with caution, as only 3 samples are available for dup15q CB. D) Cell type enrichments for genes DGE in dup15q vs ASD CTX at FDR < 0.05. E) Association of PC1 of the dup15q vs CTX DGE set with all measured covariates. F) Splicing changes in dup15q and ASD compared to CTL in the 15q region of interest, showing all detected splicing events. G) Average linkage hierarchical clustering heatmap of correlations between samples at the PSI level using all events at FDR < 0.2 in dup15q vs CTL CTX. H) PC1 of the DS set using event-level PSIs and gene-level expression values for the genes on which events were identified. I) Removing all nominally DGE events (p < 0.05) from H) and re-evaluating PC1 differences. J) PC1 of the DS set associated with all measured covariates. K) Cell type enrichments for the genes harboring events in the DS set. For cell type enrichment heatmaps, N = neurons, A = astrocytes, OG = oligodendrocytes, MG = microglia).

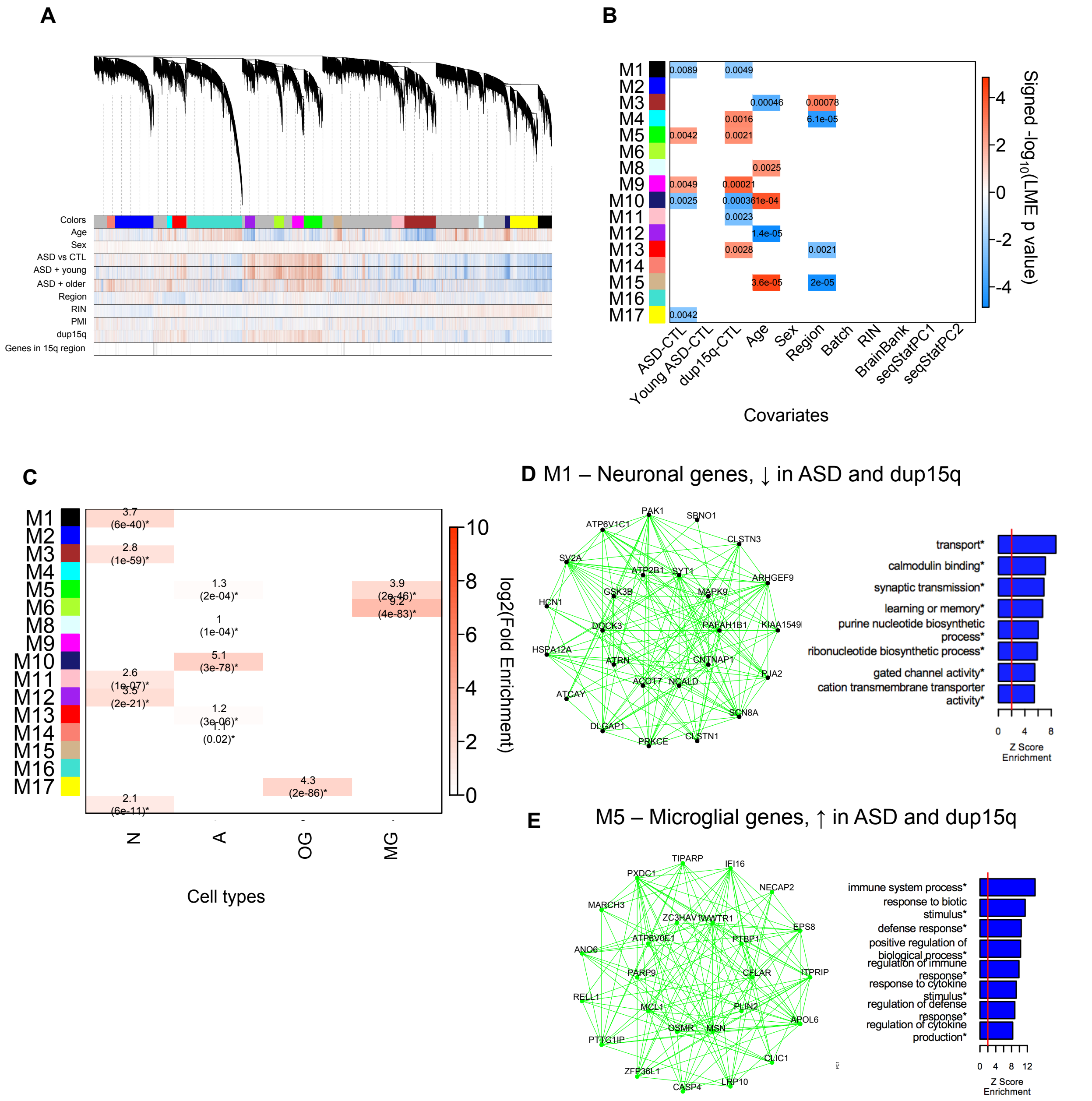


Figure A.3.5 Additional analyses related to the co-expression network analysis. A) Modules identified from a dendrogram constructed from a consensus of 100 bootstrapped datasets using the 137 CTX samples. Correlations for each gene to each measured factor are delineated below the dendrogram (blue = negative, red = positive correlation). B) Module-trait associations as computed by a linear mixed effects model with all factors on the x-axis used as covariates. All p values are displayed where the coefficient passed p < 0.01. C) Module enrichments for cell type specific markers, N = neurons, A = astrocytes, OG = oligodendrocytes, MG = microglia (see Materials and Methods for details). D-E) Two modules related to ASD and dup15q, with top 25 genes and top 8 GO terms shown.

A3.3. Extended Tables

Please see the electronic tables associated with this document for extended tables (Tables A3.1-3). Descriptions for these tables follow:

Table A3.1 Biological and technical metadata for samples used in this study. l

Table A3.2 Differential gene expression changes in CTX and CB, cortical patterning results, and co-expression network module assignments for CTX, related to Figure 1, 3, and 4. A) Results for DGE between ASD and CTL in ASD and dup15q, as well as cortical patterning analysis results and co-expression network results for CTX. B) DGE results for cerebellum in CB.

Table A3.3 Differential splicing changes in CTX and CB, related to Figures 2 and 3. A) DS results for cortex. B) DS results for CB.

Table A3.4 Copy number between dup15q breakpoints, related to Figure 3.

Breakpoints	1-3	2-3	3-4	4-5
AN09402	4	4[b]	2	2
AN14829	4	4	4	3
AN17138	4[c]	4	2	2
AN03935	4	4	4	3
AN05983	4	4	4	3
AN06365	4	4	4	3
AN11931	4	4	4	3
AN14762	-	4[a]	-	-

[a] Obtained from Scoles et al., 2011 who evaluated duplication in this region by RT-PCR of *SNRPN*/*GABRB3*/*UBE3A* vs *B2M*

[b] Discrepancy with Scoles et al., who report 5 here (Scoles et al., 2011)

[c] Discrepancy with Wintle et al., who report 2 here (Wintle et al., 2011)

REFERENCES

"*Dicebat Bernardus Carnotensis nos esse quasi nanos, gigantium humeris insidentes, ut possimus plura eis et remotiora videre, non utique proprii visus acumine, aut eminentia corporis, sed quia in altum subvenimur et extollimur magnitudine gigantea.*"

Translation: “Bernard of Chartres used to compare us to [puny] dwarfs perched on the shoulders of giants. He pointed out that we see more and farther than our predecessors, not because we have keener vision or greater height, but because we are lifted up and borne aloft on their gigantic stature.”

― Bernard of Chartres, as attributed by John of Salisbury, quote and translation taken from Wikipedia by Neelroop of Los Angeles

Allen, J.D., Xie, Y., Chen, M., Girard, L., and Xiao, G. (2012). Comparing Statistical Methods for Constructing Large Scale Gene Networks. PLoS ONE *7*, e29348.

Altar, C.A., Jurata, L.W., Charles, V., Lemire, A., Liu, P., Bukhman, Y., Young, T.A., Bullard, J., Yokoe, H., Webster, M.J., et al. (2005). Deficient Hippocampal Neuron Expression of Proteasome, Ubiquitin, and Mitochondrial Genes in Multiple Schizophrenia Cohorts. Biological Psychiatry *58*, 85–96.

Ameur, A., Zaghlool, A., Halvardson, J., Wetterbom, A., Gyllensten, U., Cavelier, L., and Feuk, L. (2011). Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. Nat Struct Mol Biol *18*, 1435–1440.

Andersen, S.L. (2003). Trajectories of brain development: point of vulnerability or window of opportunity? Neuroscience & Biobehavioral Reviews *27*, 3–18.

Angevine, J.B., Bodian, D., Coulombre, A.J., Edds, M.V., Hamburger, V., Jacobson, M., Lyser, K.M., Prestige, M.C., Sidman, R.L., Varon, S., et al. (1970). Embryonic vertebrate central nervous system: Revised terminology. The Anatomical Record *166*, 257–261.

Anney, R., Klei, L., Pinto, D., Almeida, J., Bacchelli, E., Baird, G., Bolshakova, N., Bolte, S., Bolton, P.F., Bourgeron, T., et al. (2012). Individual common variants exert weak effects on the risk for autism spectrum disorders. Human Molecular Genetics *21*, 4781–4792.

Arbiza, L., Gronau, I., Aksoy, B.A., Hubisz, M.J., Gulko, B., Keinan, A., and Siepel, A. (2013). Genome-wide inference of natural selection on human transcription factor binding sites. Nat Genet *45*, 723–729.

Arion, D., Unger, T., Lewis, D.A., Levitt, P., and Mirnics, K. (2007). Molecular Evidence for Increased Expression of Genes Related to Immune and Chaperone Function in the Prefrontal Cortex in Schizophrenia. Biological Psychiatry *62*, 711–721.

Association, A.P. (2013). DSM 5.

Attanasio, C., Nord, A.S., Zhu, Y., Blow, M.J., Li, Z., Liberton, D.K., Morrison, H., Plajzer-Frick, I., Holt, A., Hosseini, R., et al. (2013). Fine Tuning of Craniofacial Morphology by Distant-Acting Enhancers. Science *342*, 1241006–1241006.

Azevedo, F.A.C., Carvalho, L.R.B., Grinberg, L.T., Farfel, J.M., Ferretti, R.E.L., Leite, R.E.P., Jacob Filho, W., Lent, R., and Herculano-Houzel, S. (2009). Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. J. Comp. Neurol. *513*, 532–541.

Barabasi, A.L. (2009). Scale-Free Networks: A Decade and Beyond. Science *325*, 412–413.

Barabási, A.-L., and Oltvai, Z.N. (2004). Network biology: understanding the cell's functional organization. Nat Rev Genet *5*, 101–113.

Barnstable, C.J., and Dräger, U.C. (1984). Thy-1 antigen: A ganglion cell specific marker in rodent retina. Neuroscience *11*, 847–855.

Basu, S.N., Kollu, R., and Banerjee-Basu, S. (2009). AutDB: a gene reference resource for autism research. Nucleic Acids Res *37*, D832–D836.

Bayés, À., Bayés, A., van de Lagemaat, L.N., van de Lagemaat, L.N., Collins, M.O., Collins, M.O., Croning, M.D.R., Croning, M.D.R., Whittle, I.R., Whittle, I.R., et al. (2010). Characterization of the proteome, diseases and evolution of the human postsynaptic density. Nat. Neurosci. *14*, 19–21.

Ben-David, E., and Shifman, S. (2012). Combined analysis of exome sequencing points toward a major role for transcription regulation during brain development in autism. Molecular Psychiatry *18*, 1054–1056.

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

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. 289–300.

Berg, J.M., and Geschwind, D.H. (2012). Autism genetics: searching for specificity and convergence. Genome Biol *13*, 247.

Bernard, A., Lubbers, L.S., Tanis, K.Q., Luo, R., Podtelezhnikov, A.A., Finney, E.M., McWhorter, M.M.E., Serikawa, K., Lemon, T., Morgan, R., et al. (2012). Transcriptional Architecture of the Primate Neocortex. Neuron *73*, 1083–1099.

Bernier, R., Golzio, C., Xiong, B., Stessman, H.A., Coe, B.P., Penn, O., Witherspoon, K., Gerdts, J., Baker, C., Vulto-van Silfhout, A.T., et al. (2014). Disruptive CHD8 Mutations Define a Subtype of Autism Early in Development. Cell *158*, 263–276.

Blondel, V.D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. J. Stat. Mech. *2008*, P10008.

Butte, A.J., and Kohane, I.S. (2000). Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. Pac Symp Biocomput.

Bystron, I., Blakemore, C., and Rakic, P. (2008). Development of the human cerebral cortex: Boulder Committee revisited. Nat Rev Neurosci *9*, 110–122.

Capra, J.A., Erwin, G.D., McKinsey, G., Rubenstein, J.L.R., and Pollard, K.S. (2013). Many human accelerated regions are developmental enhancers. Philosophical Transactions of the Royal Society B: Biological Sciences *368*, 20130025–20130025.

Carter, H., Hofree, M., and Ideker, T. (2013). Genotype to phenotype via network analysis. Current Opinion in Genetics & Development *23*, 611–621.

Carter, S.L., Brechbühler, C.M., Griffin, M., and Bond, A.T. (2004). Gene co-expression network topology provides a framework for molecular characterization of cellular state. Bioinformatics *20*, 2242–2250.

Casey, B.J., Craddock, N., Cuthbert, B.N., Hyman, S.E., Lee, F.S., and Ressler, K.J. (2013). DSM-5 and RDoC: progress in psychiatry research? Nat Rev Neurosci *14*, 810–814.

Chang, J., Gilman, S.R., Chiang, A.H., Sanders, S.J., and Vitkup, D. (2015). Genotype to phenotype relationships in autism spectrum disorders. Nature Publishing Group *18*, 191–198.

Chang, K., Creighton, C.J., Davis, C., Donehower, L., Drummond, J., Wheeler, D., Ally, A., Balasundaram, M., Birol, I., Butterfield, Y.S.N., et al. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet *45*, 1113–1120.

Changeux, J.-P., and Danchin, A. (1976). Selective stabilisation of developing synapses as a mechanism for the specification of neuronal networks. Nature *264*, 705–712.

Chen, C., Cheng, L., Grennan, K., Pibiri, F., Zhang, C., Badner, J.A., Gershon, E.S., and Liu, C. (2012). Two gene co-expression modules differentiate psychotics and controls. Molecular Psychiatry *18*, 1308–1314.

Cheng, J. (2005). Transcriptional Maps of 10 Human Chromosomes at 5-Nucleotide Resolution. Science *308*, 1149–

1154.

Cheung, I., Shulha, H.P., Jiang, Y., Matevossian, A., Wang, J., Weng, Z., and Akbarian, S. (2010). Developmental regulation and individual differences of neuronal H3K4me3 epigenomes in the prefrontal cortex. Proc. Natl. Acad. Sci. U.S.a. *107*, 8824–8829.

Chodroff, R.A., Goodstadt, L., Sirey, T.M., Oliver, P.L., Davies, K.E., Green, E.D., Molnár, Z., and Ponting, C.P. (2010). Long noncoding RNA genes: conservation of sequence and brain expression among diverse amniotes. Genome Biol *11*, R72.

Choi, J.K., Yu, U., Yoo, O.J., and Kim, S. (2005). Differential coexpression analysis using microarray data and its application to human cancer. Bioinformatics *21*, 4348–4355.

Chow, M.L., Pramparo, T., Winn, M.E., Barnes, C.C., Li, H.-R., Weiss, L., Fan, J.-B., Murray, S., April, C., Belinson, H., et al. (2012). Age-Dependent Brain Gene Expression and Copy Number Anomalies in Autism Suggest Distinct Pathological Processes at Young Versus Mature Ages. PLoS Genet. *8*, e1002592.

Civelek, M., and Lusis, A.J. (2013). Systems genetics approaches to understand complex traits. Nat Rev Genet *15*, 34–48.

Colantuoni, C., Lipska, B.K., Ye, T., Hyde, T.M., Tao, R., Leek, J.T., Colantuoni, E.A., Elkahloun, A.G., Herman, M.M., Weinberger, D.R., et al. (2011). Temporal dynamics and genetic control of transcription in the human prefrontal cortex. Nature *478*, 519–523.

Coppola, G., and Geschwind, D.H. (2006). Technology Insight: querying the genome with microarrays—progress and hope for neurological disease. Nat Clin Pract Neurol *2*, 147–158.

Corominas, R., Yang, X., Lin, G.N., Kang, S., Shen, Y., Ghamsari, L., Broly, M., Rodriguez, M., Tam, S., Trigg, S.A., et al. (2014). Protein interaction network of alternatively spliced isoforms from brain links genetic risk factors for autism. Nat Comms *5*.

Crino, P.B., and Eberwine, J. (1996). Molecular Characterization of the Dendritic Growth Cone: Regulated mRNA Transport and Local Protein Synthesis. Neuron *17*, 1173–1187.

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

Cross-Disorder Group of the Psychiatric Genomics Consortium, Lee, S.H., Ripke, S., Neale, B.M., Faraone, S.V., Purcell, S.M., Perlis, R.H., Mowry, B.J., Thapar, A., Goddard, M.E., et al. (2013). Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. Nat Genet *45*, 984–994.

Csárdi, G., and Nepusz, T. (2006). The igraph software package for complex network research. *1695*.

Dalal, J., Roh, J.H., Maloney, S.E., Akuffo, A., Shah, S., Yuan, H., Wamsley, B., Jones, W.B., Strong, C.D.G., Gray, P.A., et al. (2013). Translational profiling of hypocretin neurons identifies candidate molecules for sleep regulation. Genes & Development *27*, 565–578.

Darnell, J.C., Van Driesche, S.J., Zhang, C., Hung, K.Y.S., Mele, A., Fraser, C.E., Stone, E.F., Chen, C., Fak, J.J., Chi, S.W., et al. (2011). FMRP Stalls Ribosomal Translocation on mRNAs Linked to Synaptic Function and Autism. Cell *146*, 247–261.

de Ligt, J., Willemsen, M.H., van Bon, B.W.M., Kleefstra, T., Yntema, H.G., Kroes, T., Vulto-van Silfhout, A.T., Koolen, D.A., de Vries, P., Gilissen, C., et al. (2012). Diagnostic Exome Sequencing in Persons with Severe Intellectual Disability. N. Engl. J. Med *367*, 1921–1929.

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

De, S. (2011). Somatic mosaicism in healthy human tissues. Trends in Genetics *27*, 217–223.

DeFelipe, J., López-Cruz, P.L., Benavides-Piccione, R., Bielza, C., Larrañaga, P., Anderson, S., Burkhalter, A., Cauli, B., Fairén, A., Feldmeyer, D., et al. (2013). New insights into the classification and nomenclature of cortical GABAergic interneurons. Nat Rev Neurosci *14*, 202–216.

Deneris, E.S., and Hobert, O. (2014). Maintenance of postmitotic neuronal cell identity. Nat. Neurosci. *17*, 899–907.

Devlin, B., and Scherer, S.W. (2012). Genetic architecture in autism spectrum disorder. Current Opinion in Genetics & Development *22*, 229–237.

Doherty, J.L., and Owen, M.J. (2014). Genomic insights into the overlap between psychiatric disorders: implications for research and clinical practice. Genome Med *6*, 29.

Dolmetsch, R., Geschwind, D.H., and Geschwind, D.H. (2011). The Human Brain in a Dish: The Promise of iPSC-Derived Neurons. Cell *145*, 831–834.

Dougherty, J.D., Maloney, S.E., Wozniak, D.F., Rieger, M.A., Sonnenblick, L., Coppola, G., Mahieu, N.G., Zhang, J., Cai, J., Patti, G.J., et al. (2013). The Disruption of Celf6, a Gene Identified by Translational Profiling of Serotonergic Neurons, Results in Autism-Related Behaviors. J. Neurosci. *33*, 2732–2753.

Doyle, J.P., Dougherty, J.D., Heiman, M., Schmidt, E.F., Stevens, T.R., Ma, G., Bupp, S., Shrestha, P., Shah, R.D., Doughty, M.L., et al. (2008). Application of a Translational Profiling Approach for the Comparative Analysis of CNS Cell Types. Cell *135*, 749–762.

Ebert, D.H., and Greenberg, M.E. (2013). Activity-dependent neuronal signalling and autism spectrum disorder. Nature *493*, 327–337.

Ellis, J.D., Barrios-Rodiles, M., Çolak, R., Irimia, M., Kim, T., Calarco, J.A., Wang, X., Pan, Q., O'Hanlon, D., Kim, P.M., et al. (2012). Tissue-Specific Alternative Splicing Remodels Protein-Protein Interaction Networks. Molecular Cell *46*, 884–892.

Epi4K Consortium, Epilepsy Phenome/Genome Project, Allen, A.S., Berkovic, S.F., Cossette, P., Delanty, N., Dlugos, D., Eichler, E.E., Epstein, M.P., Glauser, T., et al. (2013). De novo mutations in epileptic encephalopathies. Nature *501*, 217–221.

Euesden, J., Lewis, C.M., and O'Reilly, P.F. (2014). PRSice: Polygenic Risk Score software. Bioinformatics btu848.

Faludi, G., and Mirnics, K. (2011). Synaptic changes in the brain of subjects with schizophrenia. International Journal of Developmental Neuroscience *29*, 305–309.

Feldman, I., Rzhetsky, A., and Vitkup, D. (2008). Network properties of genes harboring inherited disease mutations. Proc. Natl. Acad. Sci. U.S.a. *105*, 4323–4328.

Fogel, B.L., Wexler, E., Wahnich, A., Friedrich, T., Vijayendran, C., Gao, F., Parikshak, N., Konopka, G., and Geschwind, D.H. (2012). RBFOX1 regulates both splicing and transcriptional networks in human neuronal development. Human Molecular Genetics *21*, 4171–4186.

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., Mering, von, C., et al. (2012). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res *41*, D808–D815.

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

Garbett, K., Ebert, P.J., Mitchell, A., Lintas, C., Manzi, B., Mirnics, K., and Persico, A.M. (2008). Immune transcriptome alterations in the temporal cortex of subjects with autism. Neurobiology of Disease *30*, 303–311.

Gaugler, T., Klei, L., Sanders, S.J., Bodea, C.A., Goldberg, A.P., Lee, A.B., Mahajan, M., Manaa, D., Pawitan, Y., Reichert, J., et al. (2014). Most genetic risk for autism resides with common variation. Nat Genet *46*, 881–885.

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

Geschwind, D.H. (2011). Genetics of autism spectrum disorders. Trends Cogn. Sci. (Regul. Ed.) *15*, 409–416.

Geschwind, D.H., and Gregg, J.P. (2002). Microarrays for the Neurosciences (MIT Press).

Geschwind, D.H., and Konopka, G. (2009). Neuroscience in the era of functional genomics and systems biology. Nature *461*, 908–915.

Geschwind, D.H., and Levitt, P. (2007). Autism spectrum disorders: developmental disconnection syndromes. Current Opinion in Neurobiology *17*, 103–111.

Geschwind, D.H., and Rakic, P. (2013). Cortical Evolution: Judge the Brain by Its Cover. Neuron *80*, 633–647.

Gilissen, C., Hehir-Kwa, J.Y., Thung, D.T., van de Vorst, M., van Bon, B.W.M., Willemsen, M.H., Kwint, M., Janssen, I.M., Hoischen, A., Schenck, A., et al. (2014). Genome sequencing identifies major causes of severe intellectual disability. Nature *511*, 344–347.

Gilman, S.R., Chang, J., Xu, B., Bawa, T.S., Gogos, J.A., Karayiorgou, M., and Vitkup, D. (2012). Diverse types of genetic variation converge on functional gene networks involved in schizophrenia. Nat. Neurosci. *15*, 1723–1728.

Gilman, S.R., Iossifov, I., Levy, D., Ronemus, M., Wigler, M., and Vitkup, D. (2011). Rare De Novo Variants Associated with Autism Implicate a Large Functional Network of Genes Involved in Formation and Function of Synapses. Neuron *70*, 898–907.

Ginsberg, M.R., Rubin, R.A., and Natowicz, M.R. (2013). Patterning of Regional Gene Expression in Autism: New Complexity. Scientific Reports *3*.

Ginsberg, M.R., Rubin, R.A., Falcone, T., Ting, A.H., and Natowicz, M.R. (2012). Brain Transcriptional and Epigenetic Associations with Autism. PLoS ONE *7*, e44736.

Gong, S., Zheng, C., Doughty, M.L., Losos, K., Didkovsky, N., Schambra, U.B., Nowak, N.J., Joyner, A., Leblanc, G., Hatten, M.E., et al. (2003). A gene expression atlas of the central nervous system based on bacterial artificial chromosomes. Nature *425*, 917–925.

Grant, S. (2003). Systems biology in neuroscience: bridging genes to cognition. Current Opinion in Neurobiology *13*, 577–582.

Gratten, J., Visscher, P.M., Mowry, B.J., and Wray, N.R. (2013). Interpreting the role of de novo protein-coding mutations in neuropsychiatric disease. Nat Genet *45*, 234–238.

Gratten, J., Wray, N.R., Keller, M.C., and Visscher, P.M. (2014). Large-scale genomics unveils the genetic architecture of psychiatric disorders. Nat. Neurosci. *17*, 782–790.

Greig, L.C., Woodworth, M.B., Galazo, M.J., Padmanabhan, H., and Macklis, J.D. (2013). Molecular logic of neocortical projection neuron specification, development and diversity. Nat Rev Neurosci *14*, 755–769.

Guerreiro, R.J., Gustafson, D.R., and Hardy, J. (2012). The genetic architecture of Alzheimer's disease: beyond APP, PSENs and APOE. Neurobiology of Aging *33*, 437–456.

Guinane, C.M., Tadrous, A., Fouhy, F., Ryan, C.A., Dempsey, E.M., Murphy, B., Andrews, E., Cotter, P.D., Stanton, C., and Ross, R.P. (2013). Microbial Composition of Human Appendices from Patients following Appendectomy. mBio *4*, e00366–12–e00366–12.

Gulsuner, S., Walsh, T., Watts, A.C., Lee, M.K., Thornton, A.M., Casadei, S., Rippey, C., Shahin, H., Nimgaonkar, V.L., Go, R.C.P., et al. (2013). Spatial and Temporal Mapping of De Novo Mutations in Schizophrenia to a Fetal Prefrontal Cortical Network. Cell *154*, 518–529.

Gupta, S., Ellis, S.E., Ashar, F.N., Moes, A., Bader, J.S., Zhan, J., West, A.B., and Arking, D.E. (2014). Transcriptome analysis reveals dysregulation of innate immune response genes and neuronal activity-dependent genes in autism. Nat Comms *5*, 5748.

Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P., et al. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature *458*, 223–227.

Hakak, Y., Walker, J.R., Li, C., Wong, W.H., Davis, K.L., Buxbaum, J.D., Haroutunian, V., and Fienberg, A.A. (2001). Genome-wide expression analysis reveals dysregulation of myelination-related genes in chronic schizophrenia. Proc. Natl. Acad. Sci. U.S.a. *98*, 4746–4751.

Hakes, L., Pinney, J.W., Robertson, D.L., and Lovell, S.C. (2008). Protein-protein interaction networks and biology—what's the connection? Nat Biotechnol *26*, 69–72.

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

Harrow, J., Denoeud, F., Frankish, A., Reymond, A., Chen, C.-K., Chrast, J., Lagarde, J., Gilbert, J.G., Storey, R., Swarbreck, D., et al. (2006). GENCODE: producing a reference annotation for ENCODE. Genome Biol *7*, S4–S9.

Hart, G.T., Ramani, A.K., and Marcotte, E.M. (2006). How complete are current yeast and human protein-interaction networks? Genome Biol *7*, 120.

Hashimoto, T., Arion, D., Unger, T., Maldonado-Avilés, J.G., Morris, H.M., Volk, D.W., Mirnics, K., and Lewis, D.A. (2007). Alterations in GABA-related transcriptome in the dorsolateral prefrontal cortex of subjects with schizophrenia. Molecular Psychiatry *13*, 147–161.

Hawrylycz, M.J., Lein, E.S., Guillozet-Bongaarts, A.L., Shen, E.H., Ng, L., Miller, J.A., van de Lagemaat, L.N., Smith, K.A., Ebbert, A., Riley, Z.L., et al. (2012). An anatomically comprehensive atlas of the adult human brain transcriptome. Nature *489*, 391–399.

He, X., Sanders, S.J., Liu, L., De Rubeis, S., Lim, E.T., Sutcliffe, J.S., Schellenberg, G.D., Gibbs, R.A., Daly, M.J., Buxbaum, J.D., et al. (2013). Integrated Model of De Novo and Inherited Genetic Variants Yields Greater Power to Identify Risk Genes. PLoS Genet. *9*, e1003671.

Heiman, M., Heilbut, A., Francardo, V., Kulicke, R., Fenster, R.J., Kolaczyk, E.D., Mesirov, J.P., Surmeier, D.J., Cenci, M.A., and Greengard, P. (2014). Molecular adaptations of striatal spiny projection neurons during levodopa-induced dyskinesia. Proc. Natl. Acad. Sci. U.S.a. *111*, 4578–4583.

Helsmoortel, C., Vulto-van Silfhout, A.T., Coe, B.P., Vandeweyer, G., Rooms, L., van den Ende, J., Schuurs-

Hoeijmakers, J.H.M., Marcelis, C.L., Willemsen, M.H., Vissers, L.E.L.M., et al. (2014). A SWI/SNF-related autism syndrome caused by de novo mutations in ADNP. Nat Genet *46*, 380–384.

Hoadley, K.A., Yau, C., Wolf, D.M., Cherniack, A.D., Tamborero, D., Ng, S., Leiserson, M.D.M., Niu, B., McLellan, M.D., Uzunangelov, V., et al. (2014). Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin. Cell *158*, 929–944.

Hoch, R.V., Rubenstein, J.L.R., and Pleasure, S. (2009). Genes and signaling events that establish regional patterning of the mammalian forebrain. Semin. Cell Dev. Biol. *20*, 378–386.

Hockfield, S., and McKay, R.D. (1985). Identification of major cell classes in the developing mammalian nervous system. *5*, 3310–3328.

Hoischen, A., Krumm, N., and Eichler, E.E. (2014). Prioritization of neurodevelopmental disease genes by discovery of new mutations. Nat. Neurosci. *17*, 764–772.

Hormozdiari, F., Penn, O., Borenstein, E., and Eichler, E.E. (2015). The discovery of integrated gene networks for autism and related disorders. Genome Res *25*, 142–154.

Horvath, S., Zhang, B., Carlson, M., Lu, K.V., Zhu, S., Felciano, R.M., Laurance, M.F., Zhao, W., Qi, S., Chen, Z., et al. (2006). Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target. Proc. Natl. Acad. Sci. U.S.a. *103*, 17402–17407.

Horvath, S. (2011). Weighted Network Analysis: Applications in Genomics and Systems Biology (New York: Springer).

Hudson, N.J., Reverter, A., and Dalrymple, B.P. (2009). A Differential Wiring Analysis of Expression Data Correctly Identifies the Gene Containing the Causal Mutation. PLoS Comput Biol *5*, e1000382.

Inlow, J.K. (2004). Molecular and Comparative Genetics of Mental Retardation. Genetics *166*, 835–881.

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D.S., Quinn, K., Sanislow, C., and Wang, P. (2010). Research Domain Criteria (RDoC): Toward a New Classification Framework for Research on Mental Disorders. American Journal of Psychiatry *167*, 748–751.

Iossifov, I., O'Roak, B.J., Sanders, S.J., Ronemus, M., Krumm, N., Levy, D., Stessman, H.A., Witherspoon, K.T., Vives, L., Patterson, K.E., et al. (2014). The contribution of de novo coding mutations to autism spectrum disorder. Nature *515*, 216–221.

Iossifov, I., Ronemus, M., Levy, D., Wang, Z., Hakker, I., Rosenbaum, J., Yamrom, B., Lee, Y.-H., Narzisi, G., Leotta, A., et al. (2012). De Novo Gene Disruptions in Children on the Autistic Spectrum. Neuron *74*, 285–299.

Irimia, M., Weatheritt, R.J., Ellis, J.D., Parikshak, N.N., Gonatopoulos-Pournatzis, T., Babor, M., Quesnel-Vallières, M., Tapial, J., Raj, B., O'Hanlon, D., et al. (2014). A highly conserved program of neuronal microexons is misregulated in autistic brains. Cell *159*, 1511–1523.

Jaffe, A.E., Shin, J., Collado-Torres, L., Leek, J.T., Tao, R., Li, C., Gao, Y., Jia, Y., Maher, B.J., Hyde, T.M., et al. (2015). Developmental regulation of human cortex transcription and its clinical relevance at single base resolution. Nature Publishing Group *18*, 154–161.

Jamain, S., Quach, H., Betancur, C., Råstam, M., Colineaux, C., Gillberg, I.C., Soderstrom, H., Giros, B., Leboyer, M., Gillberg, C., et al. (2003). Mutations of the X-linked genes encoding neuroligins NLGN3 and NLGN4 are associated with autism. Nat Genet *34*, 27–29.

Jiang, Y.-H., Yuen, R.K.C., Jin, X., Wang, M., Chen, N., Wu, X., Ju, J., Mei, J., Shi, Y., He, M., et al. (2013).

Detection of Clinically Relevant Genetic Variants in Autism Spectrum Disorder by Whole-Genome Sequencing. The American Journal of Human Genetics *93*, 249–263.

Johnson, M.B., Kawasawa, Y.I., Mason, C.E., Krsnik, Ž., Coppola, G., Bogdanović, D., Geschwind, D.H., Mane, S.M., State, M.W., and Šestan, N. (2009). Functional and Evolutionary Insights into Human Brain Development through Global Transcriptome Analysis. *62*, 494–509.

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

Kang, H.J., Kawasawa, Y.I., Cheng, F., Zhu, Y., Xu, X., Li, M., Sousa, A.M.M., Pletikos, M., Meyer, K.A., Sedmak, G., et al. (2011). Spatio-temporal transcriptome of the human brain. Nature *478*, 483–489.

Keshava Prasad, T.S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., et al. (2009). Human Protein Reference Database--2009 update. Nucleic Acids Res *37*, D767–D772.

Khaitovich, P. (2004). Regional Patterns of Gene Expression in Human and Chimpanzee Brains. Genome Res *14*, 1462–1473.

Khaitovich, P., Weiss, G., Lachmann, M., Hellmann, I., Enard, W., Muetzel, B., Wirkner, U., Ansorge, W., and Pääbo, S. (2004). A Neutral Model of Transcriptome Evolution. PLoS Biol *2*, e132.

Khalil, A.M., Guttman, M., Huarte, M., Garber, M., Raj, A., Rivea Morales, D., Thomas, K., Presser, A., Bernstein, B.E., van Oudenaarden, A., et al. (2009). Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc. Natl. Acad. Sci. U.S.a. *106*, 11667–11672.

Kim, T.K., Hemberg, M., Gray, J.M., Costa, A.M., Bear, D.M., Wu, J., Harmin, D.A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., et al. (2010). Widespread transcription at neuronal activity-regulated enhancers. Nature *465*, 182–187.

Klei, L., Sanders, S.J., Murtha, M.T., Hus, V., Lowe, J.K., Willsey, A., Moreno-De-Luca, D., Yu, T.W., Fombonne, E., Geschwind, D., et al. (2012). Common genetic variants, acting additively, are a major source of risk for autism. Mol Autism *3*, 9.

Konopka, G., Wexler, E., Rosen, E., Mukamel, Z., Osborn, G.E., Chen, L., Lu, D., Gao, F., Gao, K., Lowe, J.K., et al. (2012a). Modeling the functional genomics of autism using human neurons. Mol Psychiatry *17*, 202–214.

Konopka, G., Bomar, J.M., Winden, K., Coppola, G., Jonsson, Z.O., Gao, F., Peng, S., Preuss, T.M., Wohlschlegel, J.A., and Geschwind, D.H. (2009). Human-specific transcriptional regulation of CNS development genes by FOXP2. Nature *462*, 213–217.

Konopka, G., Friedrich, T., Davis-Turak, J., Winden, K., Oldham, M.C., Gao, F., Chen, L., Wang, G.-Z., Luo, R., Preuss, T.M., et al. (2012b). Human-Specific Transcriptional Networks in the Brain. Neuron *75*, 601–617.

Krumm, N., O'Roak, B.J., Shendure, J., and Eichler, E.E. (2014). A de novo convergence of autism genetics and molecular neuroscience. Trends Neurosci. *37*, 95–105.

Kwan, K.Y., Lam, M.M.S., Krsnik, Z., Kawasawa, Y.I., Lefebvre, V., and Sestan, N. (2008). SOX5 postmitotically regulates migration, postmigratory differentiation, and projections of subplate and deep-layer neocortical neurons. Proc. Natl. Acad. Sci. U.S.a. *105*, 16021–16026.

Lai, T., Jabaudon, D., Molyneaux, B.J., Azim, E., Arlotta, P., Menezes, J.R.L., and Macklis, J.D. (2008). SOX5 Controls the Sequential Generation of Distinct Corticofugal Neuron Subtypes. Neuron *57*, 232–247.

Lancaster, M.A., Renner, M., Martin, C.-A., Wenzel, D., Bicknell, L.S., Hurles, M.E., Homfray, T., Penninger, J.M., Jackson, A.P., and Knoblich, J.A. (2013). Cerebral organoids model human brain development and microcephaly. Nature *501*, 373–379.

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

Langfelder, P., and Horvath, S. (2007). Eigengene networks for studying the relationships between co-expression modules. BMC Syst Biol *1*, 54.

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

Langfelder, P., and Horvath, S. (2012). Fast R Functions for Robust Correlations and Hierarchical Clustering. Journal of Statistical Software *46*.

Langfelder, P., Luo, R., Oldham, M.C., and Horvath, S. (2011). Is My Network Module Preserved and Reproducible? PLoS Comput Biol *7*, e1001057.

Lara-Astiaso, D., Weiner, A., Lorenzo-Vivas, E., Zaretsky, I., Jaitin, D.A., David, E., Keren-Shaul, H., Mildner, A., Winter, D., Jung, S., et al. (2014). Chromatin state dynamics during blood formation. Science *345*, 943–949.

Lee, I., and Marcotte, E.M. (2009). Effects of Functional Bias on Supervised Learning of a Gene Network Model. In Computational Systems Biology, (Totowa, NJ: Humana Press), pp. 463–475.

Lee, I., Lee, I., Lehner, B., Crombie, C., Crombie, C., Wong, W., Wong, W., Fraser, A.G., Fraser, A.G., Marcotte, E.M., et al. (2008). A single gene network accurately predicts phenotypic effects of gene perturbation in Caenorhabditis elegans. Nat Genet *40*, 181–188.

Leek, J.T., Scharpf, R.B., Bravo, H.C., Simcha, D., Langmead, B., Johnson, W.E., Geman, D., Baggerly, K., and Irizarry, R.A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet *11*, 733–739.

Lein, E.S., Hawrylycz, M.J., Ao, N., Ayres, M., Bensinger, A., Bernard, A., Boe, A.F., Boguski, M.S., Brockway, K.S., Byrnes, E.J., et al. (2006). Genome-wide atlas of gene expression in the adult mouse brain. Nature *445*, 168–176.

Lessard, J., Wu, J.I., Ranish, J.A., Wan, M., Winslow, M.M., Staahl, B.T., Wu, H., Aebersold, R., Graef, I.A., and Crabtree, G.R. (2007). An Essential Switch in Subunit Composition of a Chromatin Remodeling Complex during Neural Development. Neuron *55*, 201–215.

Levy, D., Ronemus, M., Yamrom, B., Lee, Y.-H., Leotta, A., Kendall, J., Marks, S., Lakshmi, B., Pai, D., Ye, K., et al. (2011). Rare De Novo and Transmitted Copy-Number Variation in Autistic Spectrum Disorders. Neuron *70*, 886–897.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics *27*, 2987–2993.

Li, J., Shi, M., Ma, Z., Zhao, S., Euskirchen, G., Ziskin, J., Urban, A., Hallmayer, J., and Snyder, M. (2014). Integrated systems analysis reveals a molecular network underlying autism spectrum disorders. Mol Syst Biol *10*, 774–774.

Li, Q., Seo, J.-H., Stranger, B., McKenna, A., Pe'er, I., LaFramboise, T., Brown, M., Tyekucheva, S., and Freedman, M.L. (2013). Integrative eQTL-Based Analyses Reveal the Biology of Breast Cancer Risk Loci. Cell *152*, 633–641.

Lim, E.T., Raychaudhuri, S., Sanders, S.J., Stevens, C., Sabo, A., MacArthur, D.G., Neale, B.M., Kirby, A., Ruderfer, D.M., Fromer, M., et al. (2013). Rare Complete Knockouts in Humans: Population Distribution and Significant Role in Autism Spectrum Disorders. Neuron *77*, 235–242.

Lionel, A.C., Tammimies, K., Vaags, A.K., Rosenfeld, J.A., Ahn, J.W., Merico, D., Noor, A., Runke, C.K., Pillalamarri, V.K., Carter, M.T., et al. (2014). Disruption of the ASTN2/TRIM32 locus at 9q33.1 is a risk factor in males for autism spectrum disorders, ADHD and other neurodevelopmental phenotypes. Human Molecular Genetics *23*, 2752–2768.

Lister, R., Mukamel, E.A., Nery, J.R., Urich, M., Puddifoot, C.A., Johnson, N.D., Lucero, J., Huang, Y., Dwork, A.J., Schultz, M.D., et al. (2013). Global Epigenomic Reconfiguration During Mammalian Brain Development. Science *341*, 1237905–1237905.

Liu, C., Ackerman, H.H., and Carulli, J.P. (2011). A genome-wide screen of gene–gene interactions for rheumatoid arthritis susceptibility. Hum Genet *129*, 473–485.

Lubs, H.A., Stevenson, R.E., and Schwartz, C.E. (2012). Fragile X and X-Linked Intellectual Disability: Four Decades of Discovery. The American Journal of Human Genetics *90*, 579–590.

Lui, J.H., Nowakowski, T.J., Pollen, A.A., Javaherian, A., Kriegstein, A.R., and Oldham, M.C. (2014). Radial glia require PDGFD–PDGFRβ signalling in human but not mouse neocortex. Nature *515*, 264–268.

MacArthur, D.G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., Habegger, L., Pickrell, J.K., Montgomery, S.B., et al. (2012). A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes. Science *335*, 823–828.

Malik, A.N., Vierbuchen, T., Hemberg, M., Rubin, A.A., Ling, E., Couch, C.H., Stroud, H., Spiegel, I., Farh, K.K.-H., Harmin, D.A., et al. (2014). Genome-wide identification and characterization of functional neuronal activity–dependent enhancers. Nat. Neurosci. *17*, 1330–1339.

Margolin, A.A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R., and Califano, A. (2006). ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. BMC Bioinformatics *7*, S7.

Masland, R.H. (2004). Neuronal cell types. Current Biology *14*, R497–R500.

Matson, J.L., and Shoemaker, M. (2009). Intellectual disability and its relationship to autism spectrum disorders. Research in Developmental Disabilities *30*, 1107–1114.

Meredith, M.A., Kryklywy, J., McMillan, A.J., Malhotra, S., Lum-Tai, R., and Lomber, S.G. (2011). Crossmodal reorganization in the early deaf switches sensory, but not behavioral roles of auditory cortex. Proc. Natl. Acad. Sci. U.S.a. *108*, 8856–8861.

Michaelson, J.J., Shi, Y., Gujral, M., Zheng, H., Malhotra, D., Jin, X., Jian, M., Liu, G., Greer, D., Bhandari, A., et al. (2012). Whole-Genome Sequencing in Autism Identifies Hot Spots for De Novo Germline Mutation. Cell *151*, 1431–1442.

Miller, J.A., Ding, S.-L., Sunkin, S.M., Smith, K.A., Ng, L., Szafer, A., Ebbert, A., Riley, Z.L., Royall, J.J., Aiona, K., et al. (2014). Transcriptional landscape of the prenatal human brain. Nature *508*, 199–206.

Mirnics, K., and Pevsner, J. (2004). Progress in the use of microarray technology to study the neurobiology of disease. Nat. Neurosci. *7*, 434–439.

Mirnics, K., Middleton, F.A., Marquez, A., Lewis, D.A., and Levitt, P. (2000). Molecular Characterization of Schizophrenia Viewed by Microarray Analysis of Gene Expression in Prefrontal Cortex. Neuron *28*, 53–67.

Mitra, K., Carvunis, A.-R., Ramesh, S.K., and Ideker, T. (2013). Integrative approaches for finding modular structure in biological networks. Nat Rev Genet *14*, 719–732.

Molyneaux, B.J., Arlotta, P., Menezes, J.R.L., and Macklis, J.D. (2007). Neuronal subtype specification in the cerebral cortex. Nat Rev Neurosci *8*, 427–437.

Molyneaux, B.J., Goff, L.A., Brettler, A.C., Chen, H.-H., Brown, J.R., Hrvatin, S., Rinn, J.L., and Arlotta, P. (2015). DeCoN: Genome-wide Analysis of In Vivo Transcriptional Dynamics during Pyramidal Neuron Fate Selection in Neocortex. Neuron *85*, 275–288.

Monoranu, C.M., Apfelbacher, M., Grünblatt, E., Puppe, B., Alafuzoff, I., Ferrer, I., Al-Saraj, S., Keyvani, K., Schmitt, A., Falkai, P., et al. (2009). pH measurement as quality control on human post mortembrain tissue: a study of the BrainNet Europe consortium. Neuropathol Appl Neurobiol *35*, 329–337.

Mueller, S.G., Weiner, M.W., Thal, L.J., Petersen, R.C., Jack, C., Jagust, W., Trojanowski, J.Q., Toga, A.W., and Beckett, L. (2005). The Alzheimer's disease neuroimaging initiative. Neuroimaging Clinics of North America *15*, 869–77–xi–xii.

Neale, B.M., Kou, Y., Liu, L., Ma'ayan, A., Samocha, K.E., Sabo, A., Lin, C.-F., Stevens, C., Wang, L.-S., Makarov, V., et al. (2012). Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature *485*, 242–245.

Nelson, S.B., Hempel, C., and Sugino, K. (2006). Probing the transcriptome of neuronal cell types. Current Opinion in Neurobiology *16*, 571–576.

Noh, H.J., Ponting, C.P., Boulding, H.C., Meader, S., Betancur, C., Buxbaum, J.D., Pinto, D., Marshall, C.R., Lionel, A.C., Scherer, S.W., et al. (2013). Network Topologies and Convergent Aetiologies Arising from Deletions and Duplications Observed in Individuals with Autism. PLoS Genet. *9*, e1003523.

O'Roak, B.J., Vives, L., Fu, W., Egertson, J.D., Stanaway, I.B., Phelps, I.G., Carvill, G., Kumar, A., Lee, C., Ankenman, K., et al. (2012). Multiplex Targeted Sequencing Identifies Recurrently Mutated Genes in Autism Spectrum Disorders. Science *338*, 1619–1622.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res *27*, 29–34.

Oldham, M.C., Horvath, S., and Geschwind, D.H. (2006). Conservation and evolution of gene coexpression networks in human and chimpanzee brains. Proc. Natl. Acad. Sci. U.S.a. *103*, 17973–17978.

Oldham, M.C., Konopka, G., Iwamoto, K., Langfelder, P., Kato, T., Horvath, S., and Geschwind, D.H. (2008). Functional organization of the transcriptome in human brain. Nat. Neurosci. *11*, 1271–1282.

O'Roak, B.J., Vives, L., Girirajan, S., Karakoc, E., Krumm, N., Coe, B.P., Levy, R., Ko, A., Lee, C., Smith, J.D., et al. (2012). Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature *485*, 246–250.

Parikshak, N.N., Luo, R., Zhang, A., Won, H., Lowe, J.K., Chandran, V., Horvath, S., and Geschwind, D.H. (2013). Integrative Functional Genomic Analyses Implicate Specific Molecular Pathways and Circuits in Autism. Cell *155*, 1008–1021.

Pattabiraman, K., Golonzhka, O., Lindtner, S., Nord, A.S., Taher, L., Hoch, R., Silberberg, S.N., Zhang, D., Chen, B., Zeng, H., et al. (2014). Transcriptional Regulation of Enhancers Active in Protodomains of the Developing Cerebral Cortex. Neuron *82*, 989–1003.

Pinto, D., Delaby, E., Merico, D., Barbosa, M., Merikangas, A., Klei, L., Thiruvahindrapuram, B., Xu, X., Ziman, R., Wang, Z., et al. (2014). Convergence of Genes and Cellular Pathways Dysregulated in Autism Spectrum Disorders. The American Journal of Human Genetics *94*, 677–694.

Pinto, D., Pagnamenta, A.T., Klei, L., Anney, R., Merico, D., Regan, R., Conroy, J., Magalhaes, T.R., Correia, C., Abrahams, B.S., et al. (2010). Functional impact of global rare copy number variation in autism spectrum disorders. Nature *466*, 368–372.

Poduri, A., Evrony, G.D., Cai, X., and Walsh, C.A. (2013). Somatic Mutation, Genomic Variation, and Neurological Disease. Science *341*, 1237758–1237758.

Poduri, A., and Lowenstein, D. (2011). Epilepsy genetics—past, present, and future. Current Opinion in Genetics & Development *21*, 325–332.

Pollard, K.S., Salama, S.R., Lambert, N., Lambot, M.-A., Coppens, S., Pedersen, J.S., Katzman, S., King, B., Onodera, C., Siepel, A., et al. (2006). An RNA gene expressed during cortical development evolved rapidly in humans. Nature *443*, 167–172.

Pollen, A.A., Nowakowski, T.J., Shuga, J., Wang, X., Leyrat, A.A., Lui, J.H., Li, N., Szpankowski, L., Fowler, B., Chen, P., et al. (2014). Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. Nat Biotechnol *32*, 1053–1058.

Prabhakar, S., Noonan, J.P., Paabo, S., and Rubin, E.M. (2006). Accelerated Evolution of Conserved Noncoding Sequences in Humans. Science *314*, 786–786.

Purcell, A.E., Jeon, O.H., Zimmerman, A.W., Blue, M.E., and Pevsner, J. (2001). Postmortem brain abnormalities of the glutamate neurotransmitter system in autism. Neurology *57*, 1618–1628.

Purcell, S.M., Moran, J.L., Fromer, M., Ruderfer, D., Solovieff, N., Roussos, P., O'Dushlaine, C., Chambert, K., Bergen, S.E., Kähler, A., et al. (2014). A polygenic burden of rare disruptive mutations in schizophrenia. Nature *506*, 185–190.

Quinlan, A.R. (2002). BEDTools: The Swiss-Army Tool for Genome Feature Analysis (Hoboken, NJ, USA: John Wiley & Sons, Inc.).

Quinn, E.M., Cormican, P., Kenny, E.M., Hill, M., Anney, R., Gill, M., Corvin, A.P., and Morris, D.W. (2013). Development of Strategies for SNP Detection in RNA-Seq Data: Application to Lymphoblastoid Cell Lines and Evaluation Using 1000 Genomes Data. PLoS ONE *8*, e58815.

Ramani, A.K., Li, Z., Hart, G.T., Carlson, M.W., Boutz, D.R., and Marcotte, E.M. (2008). A map of human protein interactions derived from co-expression of human mRNAs and their orthologs. Mol Syst Biol *4*, 180.

Ramasamy, A., Trabzuni, D., Guelfi, S., Varghese, V., Smith, C., Walker, R., De, T., Hardy, J., Ryten, M., Weale, M.E., et al. (2014). Genetic variability in the regulation of gene expression in ten regions of the human brain. Nat. Neurosci. *17*, 1418–1428.

Rauch, A., Wieczorek, D., Graf, E., Wieland, T., Endele, S., Schwarzmayr, T., Albrecht, B., Bartholdi, D., Beygo, J., Di Donato, N., et al. (2012). Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. The Lancet *380*, 1674–1682.

Ray, D., Kazan, H., Cook, K.B., Weirauch, M.T., Najafabadi, H.S., Li, X., Gueroussov, S., Albu, M., Zheng, H., Yang, A., et al. (2013). A compendium of RNA-binding motifs for decoding gene regulation. Nature *499*, 172–177.

Rhinn, H., Fujita, R., Qiang, L., Cheng, R., Lee, J.H., and Abeliovich, A. (2013). Integrative genomics identifies APOE ε4 effectors in Alzheimer's disease. *500*, 45–50.

Rhodes, D.R., and Chinnaiyan, A.M. (2005). Integrative analysis of the cancer transcriptome. Nat Genet *37*, S31–S37.

Rivals, I., Personnaz, L., Taing, L., and Potier, M.C. (2007). Enrichment or depletion of a GO category within a class of genes: which test? Bioinformatics *23*, 401–407.

Robinson, E.B., Samocha, K.E., Kosmicki, J.A., McGrath, L., Neale, B.M., Perlis, R.H., and Daly, M.J. (2014). Autism spectrum disorder severity reflects the average contribution of de novo and familial influences. Proc. Natl. Acad. Sci. U.S.a. *111*, 15161–15165.

Ronan, J.L., Wu, W., and Crabtree, G.R. (2013). From neural development to cognition: unexpected roles for chromatin. Nat Rev Genet *14*, 347–359.

Ronemus, M., Iossifov, I., Levy, D., and Wigler, M. (2014). The role of de novo mutations in the genetics of autism spectrum disorders. Nat Rev Genet *15*, 133–141.

Ropers, H.H. (2008). Genetics of intellectual disability. Current Opinion in Genetics & Development *18*, 241–250.

Rosen, E.Y., Wexler, E.M., Versano, R., Coppola, G., Gao, F., Winden, K.D., Oldham, M.C., Martens, L.H., Zhou, P., Farese, R.V., Jr., et al. (2011). Functional Genomic Analyses Identify Pathways Dysregulated by Progranulin Deficiency, Implicating Wnt Signaling. Neuron *71*, 1030–1042.

Rossin, E.J., Lage, K., Raychaudhuri, S., Xavier, R.J., Tatar, D., Benita, Y., International Inflammatory Bowel Disease Genetics Constortium, Cotsapas, C., and Daly, M.J. (2011). Proteins Encoded in Genomic Regions Associated with Immune-Mediated Disease Physically Interact and Suggest Underlying Biology. PLoS Genet. *7*, e1001273.

Saenz, M., Lewis, L.B., Huth, A.G., Fine, I., and Koch, C. (2008). Visual Motion Area MT+/V5 Responds to Auditory Motion in Human Sight-Recovery Subjects. J. Neurosci. *28*, 5141–5148.

Sakai, Y., Shaw, C.A., Dawson, B.C., Dugas, D.V., Al-Mohtaseb, Z., Hill, D.E., and Zoghbi, H.Y. (2011). Protein interactome reveals converging molecular pathways among autism disorders. Science Translational Medicine *3*, 86ra49.

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

Sanders, S.J., Ercan-Sencicek, A.G., Hus, V., Luo, R., Murtha, M.T., Moreno-De-Luca, D., Chu, S.H., Moreau, M.P., Gupta, A.R., Thomson, S.A., et al. (2011). Multiple Recurrent De Novo CNVs, Including Duplications of the 7q11.23 Williams Syndrome Region, Are Strongly Associated with Autism. Neuron *70*, 863–885.

Sanders, S.J., Murtha, M.T., Gupta, A.R., Murdoch, J.D., Raubeson, M.J., Willsey, A.J., Ercan-Sencicek, A.G., DiLullo, N.M., Parikshak, N.N., Stein, J.L., et al. (2012). De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature *485*, 237–241.

Scoles, H.A., Urraca, N., Chadwick, S.W., Reiter, L.T., and LaSalle, J.M. (2011). Increased copy number for methylated maternal 15q duplications leads to changes in gene and protein expression in human cortical samples. Mol Autism *2*, 19.

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., and Friedman, N. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet *34*, 166–176.

Shen, S., Park, J.W., Huang, J., Dittmar, K.A., Lu, Z.X., Zhou, Q., Carstens, R.P., and Xing, Y. (2012). MATS: a

Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data. *40*, e61–e61.

Shirasaki, D.I., Greiner, E.R., Al-Ramahi, I., Gray, M., Boontheung, P., Geschwind, D.H., Botas, J., Coppola, G., Horvath, S., Loo, J.A., et al. (2012). Network Organization of the Huntingtin Proteomic Interactome in Mammalian Brain. Neuron *75*, 41–57.

Siegert, S., Scherf, B.G., Del Punta, K., Didkovsky, N., Heintz, N., and Roska, B. (2009). Genetic address book for retinal cell types. Nat. Neurosci. *12*, 1197–1204.

Song, L., Langfelder, P., and Horvath, S. (2012). Comparison of co-expression measures: mutual information, correlation, and model based indices. BMC Bioinformatics *13*, 328.

Srinivasan, K., Leone, D.P., Bateson, R.K., Dobreva, G., Kohwi, Y., Kohwi-Shigematsu, T., Grosschedl, R., and McConnell, S.K. (2012). A network of genetic repression and derepression specifies projection fates in the developing neocortex. Proc. Natl. Acad. Sci. U.S.a. *109*, 19071–19078.

Stark, C. (2006). BioGRID: a general repository for interaction datasets. Nucleic Acids Res *34*, D535–D539.

Stein, J.L., la Torre-Ubieta, de, L., Tian, Y., Parikshak, N.N., Hernández, I.A., Marchetto, M.C., Baker, D.K., Lu, D., Hinman, C.R., Lowe, J.K., et al. (2014). A Quantitative Framework to Evaluate Modeling of Cortical Development by Neural Stem Cells. Neuron *83*, 69–86.

Stein, J.L., Parikshak, N.N., and Geschwind, D.H. (2013). Rare Inherited Variation in Autism: Beginning to See the Forest and a Few Trees. Neuron *77*, 209–211.

Steinberg, J., and Webber, C. (2013). The Roles of FMRP-Regulated Genes in Autism Spectrum Disorder: Single- and Multiple-Hit Genetic Etiologies. The American Journal of Human Genetics *93*, 825–839.

Stevens, C.F. (1998). Neuronal diversity: Too many cell types for comfort? Current Biology *8*, R708–R710.

Stuart, J.M. (2003). A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. Science *302*, 249–255.

Sugathan, A., Biagioli, M., Golzio, C., Erdin, S., Blumenthal, I., Manavalan, P., Ragavendran, A., Brand, H., Lucente, D., Miles, J., et al. (2014). CHD8 regulates neurodevelopmental pathways associated with autism spectrum disorder in neural progenitors. Proceedings of the National Academy of Sciences *111*, E4468–E4477.

Suliman, R., Ben-david, E., and Shifman, S. (2014). Chromatin regulators, phenotypic robustness, and autism risk. Front. Gene. *5*.

Sunkin, S.M., Ng, L., Lau, C., Dolbeare, T., Gilbert, T.L., Thompson, C.L., Hawrylycz, M., and Dang, C. (2013). Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system. Nucleic Acids Res *41*, D996–D1008.

Talkowski, M.E., Rosenfeld, J.A., Blumenthal, I., Pillalamarri, V., Chiang, C., Heilbut, A., Ernst, C., Hanscom, C., Rossin, E., Lindgren, A.M., et al. (2012). Sequencing Chromosomal Abnormalities Reveals Neurodevelopmental Loci that Confer Risk across Diagnostic Boundaries. Cell *149*, 525–537.

Thompson, C.L., Ng, L., Menon, V., Lee, C.-K., Sunkin, S.M., Lau, C., Dang, C., Rubenstein, J.L.R., Hohmann, J., Dee, N., et al. (2014). A High-Resolution Spatiotemporal Atlas of Gene Expression of the Developing Mouse Brain. Neuron *83*, 309–323.

Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., et al. (2012). The accessible chromatin landscape of the human genome. *489*, 75–82.

Tibshirani, R., Johnstone, I., Hastie, T., and Efron, B. (2004). Least angle regression. The Annals of Statistics *32*, 407–499.

Torkamani, A., Dean, B., Schork, N.J., and Thomas, E.A. (2010). Coexpression network analysis of neural tissue reveals perturbations in developmental processes in schizophrenia. Genome Res *20*, 403–412.

Trapnell, C., Hendrickson, D.G., Sauvageau, M., Goff, L., Rinn, J.L., and Pachter, L. (2012a). Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol *31*, 46–53.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012b). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc *7*, 562–578.

Tuoc, T.C., Boretius, S., Sansom, S.N., Pitulescu, M.-E., Frahm, J., Livesey, F.J., and Stoykova, A. (2013). Chromatin Regulation by BAF170 Controls Cerebral Cortical Size and Thickness. Dev. Cell *25*, 256–269.

van Bokhoven, H. (2011). Genetic and Epigenetic Networks in Intellectual Disabilities. Annu. Rev. Genet. *45*, 81–104.

van Os, J., and Kapur, S. (2009). Schizophrenia. The Lancet *374*, 635–645.

Vaske, C.J., Benz, S.C., Sanborn, J.Z., Earl, D., Szeto, C., Zhu, J., Haussler, D., and Stuart, J.M. (2010). Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics *26*, i237–i245.

Visel, A., Taher, L., Girgis, H., May, D., Golonzhka, O., Hoch, R.V., McKinsey, G.L., Pattabiraman, K., Silberberg, S.N., Blow, M.J., et al. (2013). A High-Resolution Enhancer Atlas of the Developing Telencephalon. Cell *152*, 895–908.

Voineagu, I., Wang, X., Johnston, P., Lowe, J.K., Tian, Y., Horvath, S., Mill, J., Cantor, R.M., Blencowe, B.J., and Geschwind, D.H. (2011). Transcriptomic analysis of autistic brain reveals convergent molecular pathology. Nature *474*, 380–384.

Waddington, C.H. (1942). Canalization of Development and the Inheritance of Acquired Characters. Nature *150*, 563–565.

Wang, D.O., Martin, K.C., and Zukin, R.S. (2010). Spatially restricting gene expression by local translation at synapses. Trends Neurosci. *33*, 173–182.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature *456*, 470–476.

Wang, K., Zhang, H., Ma, D., Bućan, M., Glessner, J.T., Abrahams, B.S., Salyakina, D., Imielinski, M., Bradfield, J.P., Sleiman, P.M.A., et al. (2009a). Common genetic variants on 5p14.1 associate with autism spectrum disorders. *459*, 528–533.

Wang, Z., Gerstein, M., and Snyder, M. (2009b). RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet *10*, 57–63.

Webster, J.A., Gibbs, J.R., Clarke, J., Ray, M., Zhang, W., Holmans, P., Rohrer, K., Zhao, A., Marlowe, L., Kaleem, M., et al. (2009). Genetic Control of Human Brain Transcript Expression in Alzheimer Disease. The American Journal of Human Genetics *84*, 445–458.

Weinberger, D.R. (1987). Implications of normal brain development for the pathogenesis of schizophrenia. Archives of General Psychiatry *44*, 660–669.

Weiss, L.A., Arking, D.E., Daly, M.J., Chakravarti, A., Brune, C.W., West, K., O'Connor, A., Hilton, G., Tomlinson, R.L., West, A.B., et al. (2009). A genome-wide linkage and association scan reveals novel loci for autism. Nature *461*, 802–808.

Weyn-Vanhentenryck, S.M., Mele, A., Yan, Q., Sun, S., Farny, N., Zhang, Z., Xue, C., Herre, M., Silver, P.A., Zhang, M.Q., et al. (2014). HITS-CLIP and Integrative Modeling Define the Rbfox Splicing-Regulatory Network Linked to Brain Development and Autism. Cell Reports *6*, 1139–1152.

Willsey, A.J., Sanders, S.J., Li, M., Dong, S., Tebbenkamp, A.T., Muhle, R.A., Reilly, S.K., Lin, L., Fertuzinhos, S., Miller, J.A., et al. (2013). Coexpression Networks Implicate Human Midfetal Deep Cortical Projection Neurons in the Pathogenesis of Autism. Cell *155*, 997–1007.

Winden, K.D., Oldham, M.C., Mirnics, K., Ebert, P.J., Swan, C.H., Levitt, P., Rubenstein, J.L., Horvath, S., and Geschwind, D.H. (2009). The organization of the transcriptional network in specific neuronal classes. Mol Syst Biol *5*, 291.

Wintle, R.F., Lionel, A.C., Hu, P., Ginsberg, S.D., Pinto, D., Thiruvahindrapduram, B., Wei, J., Marshall, C.R., Pickett, J., Cook, E.H., et al. (2011). A genotype resource for postmortem brain samples from the Autism Tissue Program. Autism Res *4*, 89–97.

Wu, J., Anczuków, O., Krainer, A.R., Zhang, M.Q., and Zhang, C. (2013). OLego: fast and sensitive mapping of spliced mRNA-Seq reads using small seeds. Nucleic Acids Res *41*, 5149–5163.

Xu, B., Roos, J.L., Dexheimer, P., Boone, B., Plummer, B., Levy, S., Gogos, J.A., and Karayiorgou, M. (2011). Exome sequencing supports a de novo mutational paradigm for schizophrenia. Nat Genet *43*, 864–868.

Xu, X., Wells, A.B., O'Brien, D.R., Nehorai, A., and Dougherty, J.D. (2014). Cell Type-Specific Expression Analysis to Identify Putative Cellular Mechanisms for Neurogenetic Disorders. J. Neurosci. *34*, 1420–1431.

Yamada, T., Yang, Y., Hemberg, M., Yoshida, T., Cho, H.Y., Murphy, J.P., Fioravante, D., Regehr, W.G., Gygi, S.P., Georgopoulos, K., et al. (2014). Promoter Decommissioning by the NuRD Chromatin Remodeling Complex Triggers Synaptic Connectivity in the Mammalian Brain. Neuron *83*, 122–134.

Yu, T.W., Chahrour, M.H., Coulter, M.E., Jiralerspong, S., Okamura-Ikeda, K., Ataman, B., Schmitz-Abe, K., Harmin, D.A., Adli, M., Malik, A.N., et al. (2013). Using Whole-Exome Sequencing to Identify Inherited Causes of Autism. Neuron *77*, 259–273.

Yuen, R.K.C., Thiruvahindrapuram, B., Merico, D., Walker, S., Tammimies, K., Hoang, N., Chrysler, C., Nalpathamkalam, T., Pellecchia, G., Liu, Y., et al. (2015). Whole-genome sequencing of quartet families with autism spectrum disorder. Nat. Med. *21*, 185–191.

Zambon, A.C., Gaj, S., Ho, I., Hanspers, K., Vranizan, K., Evelo, C.T., Conklin, B.R., Pico, A.R., and Salomonis, N. (2012). GO-Elite: a flexible solution for pathway and ontology over-representation. Bioinformatics *28*, 2209–2210.

Zeng, H., Shen, E.H., Hohmann, J.G., Oh, S.W., Bernard, A., Royall, J.J., Glattfelder, K.J., Sunkin, S.M., Morris, J.A., Guillozet-Bongaarts, A.L., et al. (2012). Large-Scale Cellular-Resolution Gene Profiling in Human Neocortex Reveals Species-Specific Molecular Signatures. Cell *149*, 483–496.

Zhang, B., and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. Statistical Applications in Genetics and Molecular Biology *4*, Article17.

Zhang, B., Gaiteri, C., Bodea, L.-G., Wang, Z., McElwee, J., Podtelezhnikov, A.A., Zhang, C., Xie, T., Tran, L., Dobrin, R., et al. (2013). Integrated Systems Approach Identifies Genetic Nodes and Networks in Late-Onset Alzheimer's Disease. Cell *153*, 707–720.

Zhang, C., Frias, M.A., Mele, A., Ruggiu, M., and Eom, T. (2010). Integrative modeling defines the Nova splicing-regulatory network and its combinatorial controls. Science.

Zhang, Y., Chen, K., Sloan, S.A., Bennett, M.L., Scholze, A.R., O'Keeffe, S., Phatnani, H.P., Guarnieri, P., Caneda, C., Ruderisch, N., et al. (2014). An RNA-Sequencing Transcriptome and Splicing Database of Glia, Neurons, and Vascular Cells of the Cerebral Cortex. J. Neurosci. *34*, 11929–11947.

Zhu, X., Need, A.C., Petrovski, S., and Goldstein, D.B. (2014). One gene, many neuropsychiatric disorders: lessons from Mendelian diseases. Nat. Neurosci. *17*, 773–781.